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

Prashansa Panda and Shalabh Bhatnagar

Indian Institute of Science, Bengaluru, India.
E-mail: 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 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 which is 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 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 A 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 A).

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

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

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	$\widetilde{\mathcal{O}}(\tau_{mix}^2 \epsilon^{-2})$	MLMC
Our work	Two-timescale CA	Markovian	Shown	$\tilde{\mathcal{O}}(\epsilon^{-(2+\delta)})$	TD(0)

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 Gharesifard 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 A 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 $|r(s, a)| \le U_r$, $\forall s \in S, \forall a \in A$ 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 $s_{t+1} \sim P(\cdot|s_t, a_t)$):

$$V^\theta(s) = E\left[\sum_{t=0}^{\infty}(r(s_t, a_t) - L(\theta))|s_0 = s\right].$$

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

$$Q_\theta(s, a) = \mathbb{E}_\theta[\sum_{t=0}^{\infty}(r(s_t, a_t) - L(\theta))|s_0 = s, a_0 = a]$$

$$\overset{(i)}{=} r(s, a) - L(\theta) + \mathbb{E}[V^\theta(s')],$$

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.

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

Function Approximation

In order to save on the computational effort needed to find exact solutions, one often uses value function approximation 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. Specifically, the details of the non-asymptotic analysis are provided in Appendix A. We further show the asymptotic convergence analysis in Appendix A.

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: $\mathbf{A} := \mathbb{E}_{s,a,s'} \left[\phi(s)\big(\phi(s') - \phi(s)\big)^\top \right]$, 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 $-\lambda \overset{\triangle}{=} \sup_\theta \lambda_\theta < 0$.

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

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

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

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 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,

$$\tau_t := \min\left\{m \geq 0 \mid bk^{m-1} \leq \min\{\alpha_t, \beta_t, \gamma_t\}\right\}, \quad (2)$$

where b, k are defined as in Assumption 4.3.

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 A 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

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

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,*

$$\sum_{k=\tau_t}^{t} \mathbb{E}[(L_k - L(\theta_k))^2] \leq \mathcal{O}(\log^2 t \cdot t^{1-\nu}) + \mathcal{O}(t^\nu)$$

$$+ 2\frac{(G + U_w)^2}{(1 - \frac{c_\alpha}{c_\gamma} U_w B)^2} \frac{c_\alpha^2}{c_\gamma^2} \sum_{k=\tau_t}^{t} \mathbb{E}\|M(\theta_k, v_k)\|^2,$$

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 $M(\theta_t, v_t) = \mathbb{E}_{s_t \sim \mu_{\theta_t}, a_t \sim \pi_{\theta_t}, s_{t+1} \sim P(.|s_t,a_t)}[(r(s_t, a_t) - L(\theta_t) + \phi(s_{t+1})^\top v_t - \phi(s_t)^\top v_t)\nabla \log \pi_{\theta_t}(a_t|s_t)]$.

Proof. See Appendix A. □

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 A. □

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,*

$$\frac{1}{1 + t - \tau_t} \sum_{k=\tau_t}^{t} E\|v_k - v^*(\theta_k)\|^2$$

$$= \mathcal{O}(\log^2 t \cdot t^{\sigma-2\nu}) + \mathcal{O}(t^{2\sigma-\nu-1}) + \mathcal{O}(\log^2 t \cdot t^{-3\nu+2\sigma}),$$

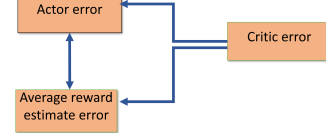*where $v^*(\theta_k)$ is as defined in Equation (1).*

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

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:

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

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

$$\lambda \sum_{k=\tau_t}^{t} E\|z_k\|^2 \leq \underbrace{\sum_{k=\tau_t}^{t} \frac{1}{2\beta_k} E[\|z_k\|^2 - \|z_{k+1}\|^2]}_{I_1}$$

$$+ \underbrace{\sum_{k=\tau_t}^{t} E[\langle z_k, \delta_t\phi(s_k) - E_{\theta_k}[\delta_k\phi(s_k)]\rangle]}_{I_2}$$

$$+ \underbrace{\sum_{k=\tau_t}^{t} \frac{1}{\beta_k} E\langle z_k, v^*(\theta_k) - v^*(\theta_{k+1}) + (\nabla v_k^*)^T(\theta_{k+1} - \theta_k)\rangle}_{I_3}$$

$$+ \underbrace{\sum_{k=\tau_t}^{t} \frac{1}{\beta_k} E\langle z_k, (\nabla v_k^*)^T(\theta_k - \theta_{k+1})\rangle}_{I_4}$$

$$+ \underbrace{\sum_{k=\tau_t}^{t} \beta_k E[\delta_k^2\|\phi(s_k)\|^2]}_{I_5} + \underbrace{\sum_{k=\tau_t}^{t} \frac{1}{\beta_k} E\|v^*(\theta_k) - v^*(\theta_{k+1})\|^2}_{I_6},$$

where $-\lambda = \sup_\theta \lambda_\theta$, see Assumption 4.2. After analysing the terms $I_1, \ldots, I_6$, we get the desired result. Please refer to Appendix A for the 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 over 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,

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

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 $\tilde{\mathcal{O}}(\epsilon^{-2.08})$. We refer the reader to Appendix A for the detailed analysis.

Remark 4.10. The analysis of single time-scale AC cannot be easily applied here. While carrying out the finite time analysis, we are focusing on finding the upper bound on various terms that should approach zero as time tends to infinity. If we apply the analysis of single time-scale AC to that of CA, some terms will not have such upper bounds due to the difference in timescales.

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 as well as Assumptions 5.1 and 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 analyse in the appendix, 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 refer the reader to Appendix A for the detailed analysis.

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. In the context of AC or CA algorithms, ours is the first analysis that incorporates this level of sophistication and generality.

The critic update then takes the following form (see Appendix A for details of the derivation):

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

where $y_t = \sum_s \mu_{\theta_t}(s) \sum_a \pi^{\theta_t}(s,a)(R(s,a) - L^{\theta_t} - v_t^T\phi(s)$

$+v_t^T \sum_{s'} p(s,a,s')\phi(s'))\phi(s),$

$$Y_t = -E[\delta_t\phi(s_t)|\mathcal{F}_2(t)] + \delta_t\phi(s_t),$$

$\kappa_t = E[(R(s_t,a_t) - L^{\theta_t} + v_t^T \sum_{s_{t+1}} p(s_t,a_t,s_{t+1})\phi(s_{t+1})$

$-v_t^T\phi(s_t))\phi(s_t)|\mathcal{F}_2(t)] - y_t.$

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

Proof. See Appendix A. □

Consider the following ODE associated with the slower recursion:

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

where $\hat{\Gamma}_2(v(y)) = \lim_{0 < \eta \to \infty}\left(\frac{\Gamma_2(y + \eta v(y)) - y}{\eta}\right)$ and e^{π^θ} is an error term, see the appendix. Consider also the DI associated with the faster recursion:

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

where $h(v)$ denotes the following set-valued function of v:

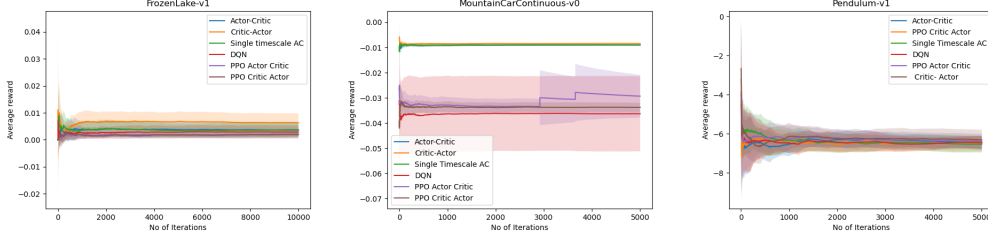$$h(v) = \{\sum_s \mu_\theta(s) \sum_a \pi^\theta(s,a)(R(s,a) - L^\theta$$

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

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

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

$$+v^T \sum_{s'} p(s,a,s')\phi(s') - v^T\phi(s))\phi(s)|\theta \in \bar{\theta}^*(v)\}.$$

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.

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). 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

---

[1]The entire code used for the experiments has been uploaded as supplementary material.

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

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, wherein 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 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), 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.

Acknowledgments

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.

Bertsekas, D. 2017. *Dynamic Programming and optimal control Vol. 2*. Athena Scientific: Belmont, MA, USA.

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. 2012. *Probability theory: an advanced course*. Springer Science & Business Media.

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.

Kushner, H.; and Clark, D. 2012. *Stochastic approximation methods for constrained and unconstrained systems*, volume 26. Springer Science & Business Media.

Kushner, H. J.; and Yin, G. G. 2003. *Stochastic Approximation Algorithms and Applications, 2'nd Ed*. New York: Springer Verlag.

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.

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

Ramaswamy, A.; and Bhatnagar, S. 2018. Stability of stochastic approximations with "controlled markov" noise and temporal difference learning. *IEEE Transactions on Automatic Control*, 64(6): 2614–2620.

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

Schweitzer, P. 1968. Perturbation theory and finite Markov chains. *Journal of Applied Probability*, 5(2): 401–413.

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.

A Appendix

The appendix comprises of three parts. First, we present the complete finite-time (non-asymptotic) analysis of the CA algorithm where we show that it achieves a sample complexity of $\tilde{O}(\epsilon^{-(2+\delta)})$ for the mean-squared error of the critic to be upper bounded by ϵ.

Subsequently, we present the complete analysis of asymptotic convergence of the two-timescale CA algorithm. In particular, we show that the iterates of the algorithm remain uniformly bounded almost surely, i.e., the iterates are stable, and in addition the algorithm is almost surely convergent. It is important to note that asymptotic convergence guarantees for many algorithms in the literature such as the single-timescale AC algorithms of (Olshevsky and Gharesifard 2023; Chen and Zhao 2023) are not available. In fact, it may not be possible to show such guarantees in the case of single-timescale AC algorithms because of the potential violation of the inherent nested loop structure required in policy iteration algorithms that gets mimicked via a difference in timescales.

Finally, we provide details of the hyper-parameter settings used for the various algorithms as well as the performance comparisons of all the algorithms in terms of the training time required in seconds.

Finite-Time Analysis

We present here the details of the finite time analysis of our two-timescale CA algorithm. Recall that our algorithm comprises of three recursions, viz., the average reward recursion, the actor update and the critic update, respectively. The actor update in our algorithm proceeds faster than the critic update. Further, the average reward estimate is used in the temporal difference term δ_t, that in turn is used in both the actor and the critic updates. Moreover, there is only a one-way coupling between the average reward estimate and the actor/critic estimates in the sense that the actor and critic estimates depend on the average reward but not vice versa. Hence, we use the actor's timescale to update the average reward recursion as well. Thus, 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 (which is slower). In Sections A–A below, we present the analysis for these three recursions and obtain the sample complexity estimate for the algorithm. Notice the difference of our scheme with standard 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 recursion while the actor recursion proceeds slower.

We start off with a basic result that provides some sufficient conditions that imply Assumption 4.5.

Theorem A.1. *Under Assumptions 4.4 and 5.1, the stationary distribution μ_θ of the Markov chain $\{s_t\}$ is continuously differentiable in $\theta \in \mathbb{R}^d$, where θ is the policy parameter. Further, if $\nabla^2 \mu_\theta(s)$ exists for each $\theta \in \mathbb{R}^d$ and $s \in S$, and further, $\sup_{\theta,s} \|\nabla^2 \mu_\theta(s)\| \le L_\mu$ for some $L_\mu > 0$, then μ_θ is L_μ-smooth.*

Proof. Let $P(\theta)$ denote the transition probability matrix with policy parameter θ. Also, let

$$Z(\theta) = [I - P(\theta) + P^\infty(\theta)]^{-1},$$

where $P^\infty(\theta) = \dfrac{1}{m}\sum_{k=1}^{m} P^k(\theta)$ is the time averaged transition probability matrix, where $P^k(\theta)$ is the k-step transition probability matrix. Since the state-valued process is ergodic Markov for any θ, it follows that $P_{ij}^\infty(\theta) = \mu_j(\theta), \forall i,j = 1,\ldots,n$. From Assumption 4.4, $\nabla \pi_\theta$ exists and is in fact uniformly bounded over all $\theta \in \mathbb{R}^d$. Thus, $\nabla P(\theta)$ exists as well (and is also uniformly bounded). It now follows from Theorem 2 of (Schweitzer 1968), that μ_θ is continuously differentiable and in fact,

$$\nabla \mu_\theta = \mu_\theta \nabla P(\theta) Z(\theta).$$

Now observe that from the mean-value theorem, for any $s \in S$,

$$\|\nabla \mu_{\theta_1}(s) - \nabla \mu_{\theta_2}(s)\| \le \|\nabla^2 \mu_\xi(s)\|\|\theta_1 - \theta_2\|,$$

where $\xi = \alpha\theta_1 + (1-\alpha)\theta_2$ for some $\alpha \in [0,1]$. The claim now follows from the fact that $\sup_{\xi,s} \|\nabla^2 \mu_\xi(s)\| \le L_\mu$. $\qquad\square$

Convergence of the Average Reward Estimate   Notations:-

$$
\begin{aligned}
O_t &:= (s_t, a_t, s_{t+1}) \\
y_t &:= (L_t - L(\theta_t)) \\
M(\theta_t, v_t) &:= E_{s_t \sim \mu_{\theta_t}, a_t \sim \pi_{\theta_t}, s_{t+1} \sim p}[(r(s_t,a_t) - L(\theta_t) + \phi(s_{t+1})^\top v_t \\
&\qquad - \phi(s_t)^\top v_t)\nabla \log \pi_{\theta_t}(a_t|s_t)] \\
W(v,\theta) &:= E_{s \sim \mu_\theta, a \sim \pi_\theta, s' \sim P}[(V^\theta(s') - v^T\phi(s') - V^\theta(s) + v^T\phi(s))\nabla \log \pi_\theta(a|s)] \qquad (7) \\
N(O_t, \theta_t, v_t, L_t) &:= (r(s_t,a_t) - L_t + \phi(s_{t+1})^\top v_t - \phi(s_t)^\top v_t)\nabla \log \pi_{\theta_t}(a_t|s_t) \\
\Omega(O_t, \theta_t, v_t, L_t) &:= y_t\langle W(v_t, \theta_t), -N(O_t, \theta_t, v_t, L_t) + E_{\theta_t}[N(O_t, \theta_t, v_t, L_t)]\rangle \\
U_w &:= 2B(U_v + \bar{U}_v) \\
G &:= 2B(U_r + U_v)
\end{aligned}
$$

We have , $|V^\theta(s)| \leq \bar{U}_v, \forall \theta \in \mathbb{R}^d, \forall s \in S$.

The following lemmas will be useful in proving the convergence of the average reward estimate.

Lemma 1. *For the performance function $L(\theta)$, there exists a constant $L_{J'} > 0$ such that for all $\theta_1, \theta_2 \in \mathbb{R}^d$, it holds that*

$$\|\nabla L(\theta_1) - \nabla L(\theta_2)\| \leq L_{J'}\|\theta_1 - \theta_2\|, \tag{8}$$

which further implies

$$L(\theta_2) \geq L(\theta_1) + \langle \nabla L(\theta_1), \theta_2 - \theta_1 \rangle - \frac{L_{J'}}{2}\|\theta_1 - \theta_2\|^2, \tag{9}$$

$$L(\theta_2) \leq L(\theta_1) + \langle \nabla L(\theta_1), \theta_2 - \theta_1 \rangle + \frac{L_{J'}}{2}\|\theta_1 - \theta_2\|^2. \tag{10}$$

Proof. Please refer proof of Lemma C.1 in (Wu et al. 2022). $\qquad\square$

Lemma 2. *Under assumptions 4.1, 4.3, 4.4, 4.6, for any $t \geq \tau > 0$, we have*
$$E[\Omega(O_t, \theta_t, v_t, L_t)] \leq (2U_wG + 4U_rU_wB)|L_t - L_{t-\tau}| + 8BU_r(G + U_w)\|v_t - v_{t-\tau}\|$$

$$+ M_1\|\theta_t - \theta_{t-\tau}\| + M_2\sum_{i=t-\tau}^{t} E\|\theta_i - \theta_{t-\tau}\| + M_3 bk^{\tau-1},$$

for some $M_1 > 0$, $M_2 > 0$ and $M_3 > 0$.

Proof. We can write $E[\Omega(O_t, \theta_t, v_t, L_t)]$ as
$$E[\Omega(O_t, \theta_t, v_t, L_t)]$$
$$= E[\Omega(O_t, \theta_t, v_t, L_t)] - E[\Omega(O_t, \theta_t, v_t, L_{t-\tau})] + E[\Omega(O_t, \theta_t, v_t, L_{t-\tau})]$$
$$- E[\Omega(O_t, \theta_t, v_{t-\tau}, L_{t-\tau})] + E[\Omega(O_t, \theta_t, v_{t-\tau}, L_{t-\tau})] - E[\Omega(O_t, \theta_{t-\tau}, v_{t-\tau}, L_{t-\tau})]$$
$$+ E[\Omega(O_t, \theta_{t-\tau}, v_{t-\tau}, L_{t-\tau})] - E[\Omega(\tilde{O}_t, \theta_{t-\tau}, v_{t-\tau}, L_{t-\tau})] + E[\Omega(\tilde{O}_t, \theta_{t-\tau}, v_{t-\tau}, L_{t-\tau})]$$
$$- E[\Omega(O'_t, \theta_{t-\tau}, v_{t-\tau}, L_{t-\tau})] + E[\Omega(O'_t, \theta_{t-\tau}, v_{t-\tau}, L_{t-\tau})]$$

In the above equality, $\tilde{O}_t$ represents the tuple $(\tilde{s}_t, \tilde{a}_t, \tilde{s}_{t+1})$, which is generated in the following manner :
$$s_{t-\tau} \xrightarrow{\theta_{t-\tau}} a_{t-\tau} \xrightarrow{P} s_{t-\tau+1} \xrightarrow{\theta_{t-\tau}} \tilde{a}_{t-\tau+1} \xrightarrow{P} \tilde{s}_{t-\tau+2} \xrightarrow{\theta_{t-\tau}} \tilde{a}_{t-\tau+2} \cdots \xrightarrow{P} \tilde{s}_t \xrightarrow{\theta_{t-\tau}} \tilde{a}_t \xrightarrow{P} \tilde{s}_{t+1}. \tag{11}$$
Thus, the policy parameter $\theta_{t-\tau}$ is held fixed for τ instants starting from the state $s_{t-\tau}$ in the original process. Here, for any time instant $k > t - \tau$, $\tilde{a}_k$ denotes the action taken under $\theta_{t-\tau}$. Similarly, for any time instant $l > t - \tau + 1$, $\tilde{s}_l$ denotes the state with actions chosen under the policy parameter $\theta_{t-\tau}$ held fixed. In this auxiliary chain, policy $\pi_{\theta_{t-\tau}}$ is repeatedly applied starting from state $s_{t-\tau}$.

Note that the original Markov chain has the following transitions:
$$s_{t-\tau} \xrightarrow{\theta_{t-\tau}} a_{t-\tau} \xrightarrow{P} s_{t-\tau+1} \xrightarrow{\theta_{t-\tau+1}} a_{t-\tau+1} \xrightarrow{P} s_{t-\tau+2} \xrightarrow{\theta_{t-\tau+2}} a_{t-\tau+2} \cdots \xrightarrow{P} s_t \xrightarrow{\theta_t} a_t \xrightarrow{P} s_{t+1}. \tag{12}$$

Moreover, $O'_t = (s'_t, a'_t, s'_{t+1})$, where $s'_t \sim \mu_{\theta_{t-\tau}}$, $a'_t \sim \pi_{\theta_{t-\tau}}$ and $s'_{t+1} \sim P(\cdot|s'_t, a'_t)$. The remainder of the proof of Lemma 2 is based on the results of the auxiliary lemmas 2.1- 2.5 below (that we now show).

We let $t \geq \tau > 0$ in the following auxiliary lemmas.

lemma 2.1.
$$E[\Omega(O_t, \theta_t, v_t, L_t)] - E[\Omega(O_t, \theta_t, v_t, L_{t-\tau})] \leq (2U_wG + 4U_rU_wB)|L_t - L_{t-\tau}|$$

Proof.

$$E[\Omega(O_t, \theta_t, v_t, L_t)] - E[\Omega(O_t, \theta_t, v_t, L_{t-\tau})]$$
$$= E[y_t\langle W(v_t, \theta_t), -N(O_t, \theta_t, v_t, L_t) + E_{\theta_t}[N(O_t, \theta_t, v_t, L_t)]\rangle$$
$$- (L_{t-\tau} - L(\theta_t))\langle W(v_t, \theta_t), -N(O_t, \theta_t, v_t, L_{t-\tau}) + E_{\theta_t}[N(O_t, \theta_t, v_t, L_{t-\tau})]\rangle]$$
$$= E[y_t\langle W(v_t, \theta_t), -N(O_t, \theta_t, v_t, L_t) + E_{\theta_t}[N(O_t, \theta_t, v_t, L_t)]\rangle$$
$$- (L_{t-\tau} - L(\theta_t))\langle W(v_t, \theta_t), -N(O_t, \theta_t, v_t, L_t) + E_{\theta_t}[N(O_t, \theta_t, v_t, L_t)]\rangle]$$
$$+ (L_{t-\tau} - L(\theta_t))\langle W(v_t, \theta_t), -N(O_t, \theta_t, v_t, L_t) + E_{\theta_t}[N(O_t, \theta_t, v_t, L_t)]\rangle]$$
$$- (L_{t-\tau} - L(\theta_t))\langle W(v_t, \theta_t), -N(O_t, \theta_t, v_t, L_{t-\tau}) + E_{\theta_t}[N(O_t, \theta_t, v_t, L_{t-\tau})]\rangle]$$
$$= E[(L_t - L_{t-\tau})\langle W(v_t, \theta_t), -N(O_t, \theta_t, v_t, L_t) + E_{\theta_t}[N(O_t, \theta_t, v_t, L_t)]\rangle]$$
$$+ E[(L_{t-\tau} - L(\theta_t))\langle W(v_t, \theta_t), N(O_t, \theta_t, v_t, L_{t-\tau}) - N(O_t, \theta_t, v_t, L_t)$$
$$+ E_{\theta_t}[N(O_t, \theta_t, v_t, L_t) - N(O_t, \theta_t, v_t, L_{t-\tau})]\rangle]$$
$$\leq 2U_wG|L_t - L_{t-\tau}| + 4U_rU_wB|L_t - L_{t-\tau}|$$
$$= 2U_w(G + 2U_rB)|L_t - L_{t-\tau}|.$$

The claim follows. □

lemma 2.2.

$$E[\Omega(O_t, \theta_t, v_t, L_{t-\tau})] - E[\Omega(O_t, \theta_t, v_{t-\tau}, L_{t-\tau})] \le 8BU_r(G + U_w)\|v_t - v_{t-\tau}\|.$$

Proof.

$$
\begin{aligned}
&E[\Omega(O_t, \theta_t, v_t, L_{t-\tau})] - E[\Omega(O_t, \theta_t, v_{t-\tau}, L_{t-\tau})] \\
&= E[(L_{t-\tau} - L(\theta_t))\langle W(v_t, \theta_t), -N(O_t, \theta_t, v_t, L_{t-\tau}) + E_{\theta_t}[N(O_t, \theta_t, v_t, L_{t-\tau})]\rangle] \\
&\quad - E[(L_{t-\tau} - L(\theta_t))\langle W(v_{t-\tau}, \theta_t), -N(O_t, \theta_t, v_{t-\tau}, L_{t-\tau}) \\
&\quad + E_{\theta_t}[N(O_t, \theta_t, v_{t-\tau}, L_{t-\tau})]\rangle] \\
&= E[(L_{t-\tau} - L(\theta_t))\langle W(v_t, \theta_t), -N(O_t, \theta_t, v_t, L_{t-\tau}) + E_{\theta_t}[N(O_t, \theta_t, v_t, L_{t-\tau})]\rangle] \\
&\quad - E[(L_{t-\tau} - L(\theta_t))\langle W(v_{t-\tau}, \theta_t), -N(O_t, \theta_t, v_t, L_{t-\tau}) \\
&\quad + E_{\theta_t}[N(O_t, \theta_t, v_t, L_{t-\tau})]\rangle] \\
&\quad + E[(L_{t-\tau} - L(\theta_t))\langle W(v_{t-\tau}, \theta_t), -N(O_t, \theta_t, v_t, L_{t-\tau}) + E_{\theta_t}[N(O_t, \theta_t, v_t, L_{t-\tau})]\rangle] \\
&\quad - E[(L_{t-\tau} - L(\theta_t))\langle W(v_{t-\tau}, \theta_t), -N(O_t, \theta_t, v_{t-\tau}, L_{t-\tau}) + E_{\theta_t}[N(O_t, \theta_t, v_{t-\tau}, L_{t-\tau})]\rangle] \\
&= E[(L_{t-\tau} - L(\theta_t))\langle W(v_t, \theta_t) - W(v_{t-\tau}, \theta_t), -N(O_t, \theta_t, v_t, L_{t-\tau}) \\
&\quad + E_{\theta_t}[N(O_t, \theta_t, v_t, L_{t-\tau})]\rangle] + E[(L_{t-\tau} - L(\theta_t))\langle W(v_{t-\tau}, \theta_t), N(O_t, \theta_t, v_{t-\tau}, L_{t-\tau}) \\
&\quad - N(O_t, \theta_t, v_t, L_{t-\tau}) + E_{\theta_t}[N(O_t, \theta_t, v_t, L_{t-\tau}) - N(O_t, \theta_t, v_{t-\tau}, L_{t-\tau})]\rangle] \\
&\le 4U_r G\|W(v_t, \theta_t) - W(v_{t-\tau}, \theta_t)\| + 4U_r U_w\|N(O_t, \theta_t, v_t, L_{t-\tau}) - N(O_t, \theta_t, v_{t-\tau}, L_{t-\tau})\| \\
&\le 8BU_r G\|v_t - v_{t-\tau}\| + 8U_r U_w B\|v_t - v_{t-\tau}\|.
\end{aligned}
$$

□

lemma 2.3.

$$E[\Omega(O_t, \theta_t, v_{t-\tau}, L_{t-\tau})] - E[\Omega(O_t, \theta_{t-\tau}, v_{t-\tau}, L_{t-\tau})] \le M_1\|\theta_t - \theta_{t-\tau}\|.$$

for some $M_1 > 0$.

Proof.

$$
\begin{aligned}
&E[\Omega(O_t, \theta_t, v_{t-\tau}, L_{t-\tau})] - E[\Omega(O_t, \theta_{t-\tau}, v_{t-\tau}, L_{t-\tau})] \\
&= E[(L_{t-\tau} - L(\theta_t))\langle W(v_{t-\tau}, \theta_t), -N(O_t, \theta_t, v_{t-\tau}, L_{t-\tau}) + E_{\theta_t}[N(O_t, \theta_t, v_{t-\tau}, L_{t-\tau})]\rangle] \\
&\quad - E[(L_{t-\tau} - L(\theta_{t-\tau}))\langle W(v_{t-\tau}, \theta_{t-\tau}), -N(O_t, \theta_{t-\tau}, v_{t-\tau}, L_{t-\tau}) \\
&\quad + E_{\theta_{t-\tau}}[N(O_t, \theta_{t-\tau}, v_{t-\tau}, L_{t-\tau})]\rangle] \\
&= E[(L_{t-\tau} - L(\theta_t))\langle W(v_{t-\tau}, \theta_t), -N(O_t, \theta_t, v_{t-\tau}, L_{t-\tau}) + E_{\theta_t}[N(O_t, \theta_t, v_{t-\tau}, L_{t-\tau})]\rangle] \\
&\quad - E[(L_{t-\tau} - L(\theta_{t-\tau}))\langle W(v_{t-\tau}, \theta_t), -N(O_t, \theta_t, v_{t-\tau}, L_{t-\tau}) + E_{\theta_t}[N(O_t, \theta_t, v_{t-\tau}, L_{t-\tau})]\rangle] \\
&\quad + E[(L_{t-\tau} - L(\theta_{t-\tau}))\langle W(v_{t-\tau}, \theta_t), -N(O_t, \theta_t, v_{t-\tau}, L_{t-\tau}) + E_{\theta_t}[N(O_t, \theta_t, v_{t-\tau}, L_{t-\tau})]\rangle] \\
&\quad - E[(L_{t-\tau} - L(\theta_{t-\tau}))\langle W(v_{t-\tau}, \theta_{t-\tau}), -N(O_t, \theta_{t-\tau}, v_{t-\tau}, L_{t-\tau}) \\
&\quad + E_{\theta_{t-\tau}}[N(O_t, \theta_{t-\tau}, v_{t-\tau}, L_{t-\tau})]\rangle] \\
&= E[(L(\theta_{t-\tau}) - L(\theta_t))\langle W(v_{t-\tau}, \theta_t), -N(O_t, \theta_t, v_{t-\tau}, L_{t-\tau}) + E_{\theta_t}[N(O_t, \theta_t, v_{t-\tau}, L_{t-\tau})]\rangle] \\
&\quad + E[(L_{t-\tau} - L(\theta_{t-\tau}))\langle W(v_{t-\tau}, \theta_t), -N(O_t, \theta_t, v_{t-\tau}, L_{t-\tau}) + E_{\theta_t}[N(O_t, \theta_t, v_{t-\tau}, L_{t-\tau})]\rangle] \\
&\quad - E[(L_{t-\tau} - L(\theta_{t-\tau}))\langle W(v_{t-\tau}, \theta_{t-\tau}), -N(O_t, \theta_t, v_{t-\tau}, L_{t-\tau}) \\
&\quad\quad + E_{\theta_t}[N(O_t, \theta_t, v_{t-\tau}, L_{t-\tau})]\rangle] \\
&\quad + E[(L_{t-\tau} - L(\theta_{t-\tau}))\langle W(v_{t-\tau}, \theta_{t-\tau}), -N(O_t, \theta_t, v_{t-\tau}, L_{t-\tau}) \\
&\quad\quad + E_{\theta_t}[N(O_t, \theta_t, v_{t-\tau}, L_{t-\tau})]\rangle] \\
&\quad - E[(L_{t-\tau} - L(\theta_{t-\tau}))\langle W(v_{t-\tau}, \theta_{t-\tau}), -N(O_t, \theta_{t-\tau}, v_{t-\tau}, L_{t-\tau}) \\
&\quad\quad + E_{\theta_{t-\tau}}[N(O_t, \theta_{t-\tau}, v_{t-\tau}, L_{t-\tau})]\rangle] \\
&\le 2GU_w\|L_{t-\tau} - L(\theta_{t-\tau})\| + 4U_r U_G\|W(v_{t-\tau}, \theta_t) - W(v_{t-\tau}, \theta_{t-\tau})\| \\
&\quad + 4U_r U_w\|N(O_t, \theta_t, v_{t-\tau}, L_{t-\tau}) - N(O_t, \theta_{t-\tau}, v_{t-\tau}, L_{t-\tau})\| \\
&= \mathcal{O}(\|\theta_t - \theta_{t-\tau}\|).
\end{aligned}
$$

The claim follows. □

lemma 2.4.

$$E[\Omega(O_t, \theta_{t-\tau}, v_{t-\tau}, L_{t-\tau})] - E[\Omega(\tilde{O}_t, \theta_{t-\tau}, v_{t-\tau}, L_{t-\tau})] \le M_2 \sum_{i=t-\tau}^{t} E\|\theta_i - \theta_{t-\tau}\|$$

for some $M_2 > 0$.

Proof. The proof is similar to that of Lemma D.10 in (Wu et al. 2022), hence we omit the same here. □

lemma 2.5.

$$E[\Omega(\tilde{O}_t, \theta_{t-\tau}, v_{t-\tau}, L_{t-\tau})] - E[\Omega(O'_t, \theta_{t-\tau}, v_{t-\tau}, L_{t-\tau})] \le M_3 b k^{\tau-1},$$

for some $M_3 > 0$.

Proof. The proof is similar to that of Lemma D.11 in (Wu et al. 2022), hence we omit the same here. □

It is clear from the definition in Section A that $E[\Omega(O'_t, \theta_{t-\tau}, v_{t-\tau}, L_{t-\tau})] = 0$. Now collecting the results from lemmas 2.1 - 2.5, we have,

$$E[\Omega(O_t, \theta_t, v_t, L_t)] \le (2U_w G + 4U_r U_w B)|L_t - L_{t-\tau}| + 8BU_r(G + U_w)\|v_t - v_{t-\tau}\|$$

$$+ M_1\|\theta_t - \theta_{t-\tau}\| + M_2 \sum_{i=t-\tau}^{t} E\|\theta_i - \theta_{t-\tau}\| + M_3 b k^{\tau-1}.$$

The claim in Lemma 2 now follows. □

Proof of convergence of the average reward estimate. From the update rule of the reward estimation recursion in Algorithm 1, we have

$$L_{t+1} - L(\theta_{t+1}) = L_t - L(\theta_t) + L(\theta_t) - L(\theta_{t+1}) + \gamma_t(r_t - L_t).$$

We then have

$$\begin{aligned}
y_{t+1}^2 &= (y_t + L(\theta_t) - L(\theta_{t+1}) + \gamma_t(r_t - L_t))^2 \\
&\le y_t^2 + 2y_t(L(\theta_t) - L(\theta_{t+1})) + 2\gamma_t y_t(r_t - L_t) + 2(L(\theta_t) - L(\theta_{t+1}))^2 + 2\gamma_t^2(r_t - L_t)^2 \\
&= (1 - 2\gamma_t)y_t^2 + 2\gamma_t y_t(r_t - L(\theta_t)) + 2y_t(L(\theta_t) - L(\theta_{t+1})) + 2(L(\theta_t) - L(\theta_{t+1}))^2 \\
&\quad + 2\gamma_t^2(r_t - L_t)^2.
\end{aligned}$$

Taking expectations, rearranging and summing from τ_t to t we obtain,

$$\sum_{k=\tau_t}^{t} \mathbb{E}[y_k^2] \le \underbrace{\sum_{t=\tau_t}^{t} \frac{1}{2\gamma_k}\mathbb{E}(y_k^2 - y_{k+1}^2)}_{I_1} + \underbrace{\sum_{k=\tau_t}^{t} \mathbb{E}[y_k(r_k - L(\theta_k))]}_{I_2} + \underbrace{\sum_{k=\tau_t}^{t} \frac{1}{\gamma_k}\mathbb{E}[y_k(L(\theta_k) - L(\theta_{k+1}))]}_{I_3}$$

$$+ \underbrace{\sum_{k=\tau_t}^{t} \frac{1}{\gamma_k}\mathbb{E}[(L(\theta_k) - L(\theta_{k+1}))^2]}_{I_4} + \underbrace{\sum_{k=\tau_t}^{t} \gamma_k \mathbb{E}[(r_k - L_k)^2]}_{I_5}.$$

For term I_1, from Abel summation by parts, we have

$$\begin{aligned}
I_1 &= \sum_{k=\tau_t}^{t} \frac{1}{2\gamma_k}(y_k^2 - y_{k+1}^2) \\
&= \sum_{k=\tau_t+1}^{t} y_k^2(\frac{1}{2\gamma_k} - \frac{1}{2\gamma_{k-1}}) + \frac{1}{2\gamma_{\tau_t}}y_{\tau_t}^2 - \frac{1}{\gamma_t}y_{t+1}^2 \\
&\le \frac{2U_r^2}{\gamma_t} \\
&= 2c_\gamma U_r^2(1 + t)^\nu.
\end{aligned}$$

For term I_2, we have

$$\sum_{k=\tau_t}^{t} \mathbb{E}[y_k(r_k - L(\theta_k))] = \mathcal{O}(\log^2 t \cdot t^{1-\nu}).$$

The analysis of part I_2 will be similar to the one of part I_2 in section C.2 of (Wu et al. 2022).

For I_3, if $y_t > 0$, from (9), we have

$$
\begin{aligned}
&y_t(L(\theta_t) - L(\theta_{t+1})) \\
&\leq y_t(\frac{L_{J'}}{2}\|\theta_t - \theta_{t+1}\|^2 + \langle \nabla L(\theta_t), \theta_t - \theta_{t+1}\rangle) \\
&\leq L_{J'}U_r\|\theta_t - \theta_{t+1}\|^2 + y_t\langle M(\theta_t, v_t), \theta_t - \theta_{t+1}\rangle \\
&\qquad + y_t\langle E_{\theta_t}[(V^{\theta_t}(s_{t+1}) - v(t)^T\phi(s_{t+1}) - V^{\theta_t}(s_t) + v(t)^T\phi(s_t))\nabla\log\pi_{\theta_t}(a_t|s_t)] \\
&\qquad , \theta_t - \theta_{t+1}\rangle
\end{aligned}
$$

If $y_t \leq 0$, from (10), we have

$$
\begin{aligned}
&y_t(L(\theta_t) - L(\theta_{t+1})) \\
&\leq y_t(-\frac{L_{J'}}{2}\|\theta_t - \theta_{t+1}\|^2 + \langle \nabla L(\theta_t), \theta_t - \theta_{t+1}\rangle) \\
&\leq L_{J'}U_r\|\theta_t - \theta_{t+1}\|^2 + y_t\langle M(\theta_t, v_t), \theta_t - \theta_{t+1}\rangle \\
&\qquad + y_t\langle E_{\theta_t}[(V^{\theta_t}(s_{t+1}) - v(t)^T\phi(s_{t+1}) - V^{\theta_t}(s_t) + v(t)^T\phi(s_t))\nabla\log\pi_{\theta_t}(a_t|s_t)], \theta_t - \theta_{t+1}\rangle.
\end{aligned}
$$

Overall, we get

$$I_3 = \sum_{k=\tau_t}^{t} \frac{1}{\gamma_k} \mathbb{E}[y_k(L(\theta_k) - L(\theta_{k+1}))]$$

$$\leq \sum_{k=\tau_t}^{t} \frac{1}{\gamma_k} \mathbb{E}[L_{J'} U_r \|\theta_k - \theta_{k+1}\|^2 + |y_k|\|\theta_k - \theta_{k+1}\|\|M(\theta_k, v_k)\|]$$

$$+ \sum_{k=\tau_t}^{t} \frac{1}{\gamma_k} E[y_k \langle E_{\theta_k}[(V^{\theta_k}(s_{k+1}) - v(k)^T \phi(s_{k+1}) - V^{\theta_k}(s_k)$$

$$+ v(k)^T \phi(s_k))\nabla \log \pi_{\theta_k}(a_k|s_k)], \theta_k - \theta_{k+1}\rangle]$$

$$\leq \sum_{k=\tau_t}^{t} \mathbb{E}[L_{J'} U_r G^2 \frac{\alpha_k^2}{\gamma_k} + G\frac{c_\alpha}{c_\gamma}|y_k|\|M(\theta_k, v_k)\|]$$

$$+ \sum_{k=\tau_t}^{t} \frac{1}{\gamma_k} E[y_k \langle E_{\theta_k}[(V^{\theta_k}(s_{k+1}) - v(k)^T \phi(s_{k+1}) - V^{\theta_k}(s_k)$$

$$+ v(k)^T \phi(s_k))\nabla \log \pi_{\theta_k}(a_k|s_k)], \theta_k - \theta_{k+1}\rangle]$$

$$\leq \frac{2L_{J'} U_r G^2 c_\alpha^2}{c_\gamma}(1 + t - \tau_t)^{1-\nu} + G\frac{c_\alpha}{c_\gamma}\Big(\sum_{k=\tau_t}^{t} \mathbb{E}y_t^2\Big)^{\frac{1}{2}}\Big(\sum_{k=\tau_t}^{t} \mathbb{E}\|M(\theta_k, v_k)\|^2\Big)^{\frac{1}{2}}$$

$$+ \sum_{k=\tau_t}^{t} \frac{1}{\gamma_k} E[y_k \langle E_{\theta_k}[(V^{\theta_k}(s_{k+1}) - v(k)^T \phi(s_{k+1}) - V^{\theta_k}(s_k)$$

$$+ v(k)^T \phi(s_k))\nabla \log \pi_{\theta_k}(a_k|s_k)], \theta_k - \theta_{k+1}\rangle]$$

$$= \frac{2L_{J'} U_r G^2 c_\alpha^2}{c_\gamma}(1 + t - \tau_t)^{1-\nu} + G\frac{c_\alpha}{c_\gamma}\Big(\sum_{k=\tau_t}^{t} \mathbb{E}y_t^2\Big)^{\frac{1}{2}}\Big(\sum_{k=\tau_t}^{t} \mathbb{E}\|M(\theta_k, v_k)\|^2\Big)^{\frac{1}{2}}$$

$$+ \underbrace{\sum_{k=\tau_t}^{t} \frac{c_\alpha}{c_\gamma} E[y_k \langle W(v_k, \theta_k), -\delta_k \nabla_\theta \log \pi_{\theta_k}(s_k|a_k) + E_{\theta_k}[\delta_k \nabla_\theta \log \pi_{\theta_k}(s_k|a_k)]\rangle]}_{I_a}$$

$$+ \underbrace{\sum_{k=\tau_t}^{t} \frac{c_\alpha}{c_\gamma} E[y_k \langle W(v_k, \theta_k), -E_{\theta_k}[\delta_k \nabla_\theta \log \pi_{\theta_k}(s_k|a_k)]\rangle]}_{I_b}$$

For term I_a, we have,

$$\frac{c_\alpha}{c_\gamma} \sum_{k=\tau_t}^{t} E[y_k \langle W(v_k, \theta_k), -N(O_k, \theta_k, v_k, L_k) + E_{\theta_k}[N(O_k, \theta_k, v_k, L_k)]\rangle]$$

$$= \frac{c_\alpha}{c_\gamma} \sum_{k=\tau_t}^{t} \Big((2U_w G + 4U_r U_w B)|L_k - L_{k-\tau}| + 8BU_r(G + U_w)\|v_k - v_{k-\tau}\|$$

$$+ M_1\|\theta_k - \theta_{k-\tau}\| + M_2 \sum_{i=k-\tau}^{k} E\|\theta_i - \theta_{k-\tau}\| + M_3 b k^{m-1}\Big).$$

Taking $\tau := \tau_t$, we have,

$$I_a = \mathcal{O}(\tau_t^2 \cdot t^{1-\nu}).$$

For term I_b, we have,

$$\sum_{k=\tau_t}^{t} \frac{c_\alpha}{c_\gamma} E[y_k \langle W(v_k, \theta_k), -E_{\theta_k}[\delta_k \nabla_\theta \log \pi_{\theta_k}(s_k|a_k)]\rangle]$$

$$= \frac{c_\alpha}{c_\gamma} \sum_{k=\tau_t}^{t} E[y_k \langle W(v_k, \theta_k), -M(\theta_k, v_k)\rangle] + \frac{c_\alpha}{c_\gamma} \sum_{k=\tau_t}^{t} E[y_k \langle W(v_k, \theta_k), y_k E_{\theta_k}[\nabla_\theta \log \pi_{\theta_k}(s_k|a_k)]\rangle]$$

$$\leq U_w \frac{c_\alpha}{c_\gamma} (\sum_{k=\tau_t}^{t} \mathbb{E}y_t^2)^{\frac{1}{2}} (\sum_{k=\tau_t}^{t} \mathbb{E}\|M(\theta_k, v_k)\|^2)^{\frac{1}{2}} + \frac{c_\alpha}{c_\gamma} \sum_{k=\tau_t}^{t} E[y_k^2 \langle W(v_k, \theta_k), E_{\theta_k}[\nabla_\theta \log \pi_{\theta_k}(s_k|a_k)]\rangle]$$

$$\leq U_w \frac{c_\alpha}{c_\gamma} (\sum_{k=\tau_t}^{t} \mathbb{E}y_t^2)^{\frac{1}{2}} (\sum_{k=\tau_t}^{t} \mathbb{E}\|M(\theta_k, v_k)\|^2)^{\frac{1}{2}} + \frac{c_\alpha}{c_\gamma} U_w B \sum_{k=\tau_t}^{t} E[y_k^2].$$

Hence collecting all the terms, we have,

$$I_3 = \frac{2L_{J'} U_r G^2 c_\alpha^2}{c_\gamma}(1 + t - \tau_t)^{1-\nu} + G\frac{c_\alpha}{c_\gamma}(\sum_{k=\tau_t}^{t} \mathbb{E}y_t^2)^{\frac{1}{2}}(\sum_{k=\tau_t}^{t} \mathbb{E}\|M(\theta_k, v_k)\|^2)^{\frac{1}{2}} + \mathcal{O}(\log^2 t \cdot t^{1-\nu})$$

$$+ U_w \frac{c_\alpha}{c_\gamma}(\sum_{k=\tau_t}^{t} \mathbb{E}y_t^2)^{\frac{1}{2}}(\sum_{k=\tau_t}^{t} \mathbb{E}\|M(\theta_k, v_k)\|^2)^{\frac{1}{2}} + \frac{c_\alpha}{c_\gamma} U_w B \sum_{k=\tau_t}^{t} E[y_k^2]$$

where $G = 2B(U_r + U_v)$.

For term I_4, we have

$$I_4 = \sum_{k=\tau_t}^{t} \frac{1}{\alpha_k} \mathbb{E}[(L(\theta_k) - L(\theta_{k+1}))^2]$$

$$\leq \sum_{k=\tau_t}^{t} \frac{1}{\alpha_k} L_J^2 \mathbb{E}\|\theta_k - \theta_{k+1}\|^2$$

$$\leq \sum_{k=\tau_t}^{t} \frac{1}{\alpha_k} L_J^2 G^2 \alpha_k^2$$

$$= L_J^2 G^2 \sum_{k=\tau_t}^{t} \alpha_k$$

$$\leq L_J^2 G^2 (1 + t)^{1-\nu}.$$

For term I_5, we have

$$I_5 = \sum_{k=\tau_t}^{t} \alpha_k \mathbb{E}[(r_k - L(\theta_k))^2]$$

$$\leq \sum_{k=\tau_t}^{t} 4U_r^2 \alpha_k$$

$$\leq 4U_r^2 (1 + t)^{1-\nu}.$$

After combining all of the terms, we have,

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

$$+ \frac{c_\alpha}{c_\gamma} U_w B \sum_{k=\tau_t}^{t} E[y_k^2].$$

After rearranging terms above, we obtain,

$$\left(1 - \frac{c_\alpha}{c_\gamma} U_w B\right) \sum_{k=\tau_t}^{t} \mathbb{E}[y_k^2] \leq \mathcal{O}(\log^2 t \cdot t^{1-\nu}) + \mathcal{O}(t^\nu)$$

$$+ (G + U_w) \frac{c_\alpha}{c_\gamma} \left(\sum_{k=\tau_t}^{t} \mathbb{E}[y_k^2]\right)^{\frac{1}{2}} \left(\sum_{k=\tau_t}^{t} \mathbb{E}\|M(\theta_k, v_k)\|^2\right)^{\frac{1}{2}}.$$

Now we require the condition $\left(1 - \frac{c_\alpha}{c_\gamma} U_w B\right) > 0$ to be satisfied for the left hand side term to be positive. Hence, we need to choose the values of c_α and c_γ such that $\left(1 - \frac{c_\alpha}{c_\gamma} U_w B\right) > 0$. We thus have,

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

After applying the squaring technique (see page 23 of (Wu et al. 2022)), we have,

$$\sum_{k=\tau_t}^{t} \mathbb{E}[y_k^2] \leq \mathcal{O}(\log^2 t \cdot t^{1-\nu}) + \mathcal{O}(t^\nu) + 2 \frac{(G + U_w)^2}{(1 - \frac{c_\alpha}{c_\gamma} U_w B)^2} \frac{c_\alpha^2}{c_\gamma^2} \sum_{k=\tau_t}^{t} \mathbb{E}\|M(\theta_k, v_k)\|^2. \tag{13}$$

Convergence of the actor   Notations used here:

$$O_t := (s_t, a_t, s_{t+1})$$
$$h(O_t, \theta_t, L_t, v_t) := (r(s_t, a_t) - L_t + \phi(s_{t+1})^\top v_t - \phi(s_t)^\top v_t) \nabla \log \pi_{\theta_t}(a_t | s_t)$$
$$I(O_t, L_t, \theta_t, v_t) := \langle \nabla L(\theta_t), h(O_t, \theta_t, L_t, v_t) - E_{s_t \sim \mu_{\theta_t}, a_t \sim \pi_{\theta_t}, s_{t+1} \sim p}[h(O_t, \theta_t, L_t, v_t)] \rangle$$
$$\bar{h}(O_t, \theta_t, v_t) := (r(s_t, a_t) - L(\theta_t) + \phi(s_{t+1})^\top v_t - \phi(s_t)^\top v_t) \nabla \log \pi_{\theta_t}(a_t | s_t) \tag{14}$$
$$M(\theta_t, v_t) := E_{s_t \sim \mu_{\theta_t}, a_t \sim \pi_{\theta_t}, s_{t+1} \sim p}[\bar{h}(O_t, \theta_t, v_t)]$$
$$\bar{W}(O_t, \theta_t, v_t) := (V^{\theta_t}(s_{t+1}) - \phi(s_{t+1})^T v_t - V^{\theta_t}(s_t) + \phi(s_t)^T v_t) \nabla \log \pi_{\theta_t}(a_t | s_t)$$
$$\Xi(O_t, \theta_t, v_t) := \langle E_{\theta_t}[\bar{W}(O_t, \theta_t, v_t)], E_{\theta_t}[\bar{h}(O_t, \theta_t, v_t)] \rangle - \langle \bar{W}(O_t, \theta_t, v_t), E_{\theta_t}[\bar{h}(O_t, \theta_t, v_t)] \rangle.$$

The following supporting lemmas will help in the proof.

Lemma 3. *Under assumptions 4.1, 4.3, 4.4, 4.6, for any $t \geq \tau > 0$, we have*

$$E[I(O_t, L_t, \theta_t, v_t)] \geq -(D_1(\tau + 1) \sum_{k=t-\tau+1}^{t} E\|\theta_k - \theta_{k-1}\| + D_2 bk^{\tau-1} + D_3 E\|v_t - v_{t-\tau}\|$$
$$+ D_4 E|L_t - L_{t-\tau}|),$$

for some $D_1 > 0, D_2 > 0, D_3 > 0$ and $D_4 > 0$.

Proof. We can decompose $E[I(O_t, L_t, \theta_t, v_t)]$ as:

$$E[I(O_t, L_t, \theta_t, v_t)]$$
$$= E[I(O_t, L_t, \theta_t, v_t) - I(O_t, L_t, \theta_t, v_{t-\tau})] + E[I(O_t, L_t, \theta_t, v_{t-\tau}) - I(O_t, L_t, \theta_{t-\tau}, v_{t-\tau})]$$
$$+ E[I(O_t, L_t, \theta_{t-\tau}, v_{t-\tau}) - I(O_t, L_{t-\tau}, \theta_{t-\tau}, v_{t-\tau})]$$
$$+ E[I(O_t, L_{t-\tau}, \theta_{t-\tau}, v_{t-\tau}) - I(\tilde{O}_t, L_{t-\tau}, \theta_{t-\tau}, v_{t-\tau})]$$
$$+ E[I(\tilde{O}_t, L_{t-\tau}, \theta_{t-\tau}, v_{t-\tau}) - I(O'_t, L_{t-\tau}, \theta_{t-\tau}, v_{t-\tau})]$$
$$+ E[I(O'_t, L_{t-\tau}, \theta_{t-\tau}, v_{t-\tau})].$$

In the above equality, $\tilde{O}_t = (\tilde{s}_t, \tilde{a}_t, \tilde{s}_{t+1})$ is from the auxiliary Markov chain defined in Equation (11) and $O'_t = (s'_t, a'_t, s'_{t+1})$ where $s'_t \sim \mu_{\theta_{t-\tau}}, a'_t \sim \pi_{\theta_{t-\tau}}$ and $s'_{t+1} \sim P(.|s'_t, a'_t)$. The proof of Lemma 3 is based on the results of lemmas 3.1 - 3.5 below that we now state and prove. In the auxiliary lemmas below, we let $t \geq \tau > 0$.

lemma 3.1.

$$|E[I(O_t, L_t, \theta_t, v_t) - I(O_t, L_t, \theta_t, v_{t-\tau})]| \leq 4BG\|v_t - v_{t-\tau}\|.$$

Proof.

$$|E[I(O_t, L_t, \theta_t, v_t) - I(O_t, L_t, \theta_t, v_{t-\tau})]|$$
$$= |E[\langle \nabla L(\theta_t), h(O_t, \theta_t, L_t, v_t) - h(O_t, \theta_t, L_t, v_{t-\tau})$$
$$- E_{s_t \sim \mu_{\theta_t}, a_t \sim \pi_{\theta_t}, s_{t+1} \sim p}[h(O_t, \theta_t, L_t, v_t) - h(O_t, \theta_t, L_t, v_{t-\tau})]\rangle]|.$$

Further,

$$\|h(O_t, \theta_t, L_t, v_t) - h(O_t, \theta_t, L_t, v_{t-\tau})\| = \|((\phi(s_{t+1})^\top - \phi(s_t)^\top)(v_t - v_{t-\tau}))\nabla \log \pi_{\theta_t}(a_t|s_t)\|$$
$$\leq 2B\|v_t - v_{t-\tau}\|.$$

Hence,

$$|E[I(O_t, L_t, \theta_t, v_t) - I(O_t, L_t, \theta_t, v_{t-\tau})]| \leq 4BG\|v_t - v_{t-\tau}\|.$$

The claim follows. $\qquad\qquad\square$

lemma 3.2.

$$|E[I(O_t, L_t, \theta_t, v_{t-\tau}) - I(O_t, L_t, \theta_{t-\tau}, v_{t-\tau})]| \leq \overline{C}\|\theta_t - \theta_{t-\tau}\|$$

for some $\overline{C} > 0$.

Proof.

$$|E[I(O_t, L_t, \theta_t, v_{t-\tau}) - I(O_t, L_t, \theta_{t-\tau}, v_{t-\tau})]|$$
$$= |E[\langle \nabla L(\theta_t), h(O_t, \theta_t, L_t, v_{t-\tau}) - E_{\theta_t}[h(O_t, \theta_t, L_t, v_{t-\tau})]\rangle$$
$$- \langle \nabla L(\theta_{t-\tau}), h(O_t, \theta_{t-\tau}, L_t, v_{t-\tau}) - E_{\theta_{t-\tau}}[h(O_t, \theta_{t-\tau}, L_t, v_{t-\tau})]\rangle]|$$
$$\leq |E[\langle \nabla L(\theta_t), h(O_t, \theta_t, L_t, v_{t-\tau}) - E_{\theta_t}[h(O_t, \theta_t, L_t, v_{t-\tau})]\rangle$$
$$- \langle \nabla L(\theta_{t-\tau}), h(O_t, \theta_t, L_t, v_{t-\tau}) - E_{\theta_t}[h(O_t, \theta_t, L_t, v_{t-\tau})]\rangle]|$$
$$+ |E[\langle \nabla L(\theta_{t-\tau}), h(O_t, \theta_t, L_t, v_{t-\tau}) - E_{\theta_t}[h(O_t, \theta_t, L_t, v_{t-\tau})]\rangle$$
$$- \langle \nabla L(\theta_{t-\tau}), h(O_t, \theta_{t-\tau}, L_t, v_{t-\tau}) - E_{\theta_{t-\tau}}[h(O_t, \theta_{t-\tau}, L_t, v_{t-\tau})]\rangle]|$$
$$= |E[\langle \nabla L(\theta_t) - \nabla L(\theta_{t-\tau}), h(O_t, \theta_t, L_t, v_{t-\tau}) - E_{\theta_t}[h(O_t, \theta_t, L_t, v_{t-\tau})]\rangle]|$$
$$+ |E[\langle \nabla L(\theta_{t-\tau}), h(O_t, \theta_t, L_t, v_{t-\tau}) - h(O_t, \theta_{t-\tau}, L_t, v_{t-\tau})$$
$$- E_{\theta_t}[h(O_t, \theta_t, L_t, v_{t-\tau})] + E_{\theta_{t-\tau}}[h(O_t, \theta_{t-\tau}, L_t, v_{t-\tau})]\rangle]|$$
$$\leq 2L_J G\|\theta_t - \theta_{t-\tau}\|$$
$$+ |E[\|\nabla L(\theta_{t-\tau})\|(\underbrace{\|h(O_t, \theta_t, L_t, v_{t-\tau}) - h(O_t, \theta_{t-\tau}, L_t, v_{t-\tau})\|}_{I_a}$$
$$+ \underbrace{\|E_{\theta_t}[h(O_t, \theta_t, L_t, v_{t-\tau})] - E_{\theta_{t-\tau}}[h(O_t, \theta_{t-\tau}, L_t, v_{t-\tau})]\|}_{I_b})]|,$$

where $E_{\theta_t}[.]$ denotes the expectation with respect to $s_t \sim \mu_{\theta_t}, a_t \sim \pi_{\theta_t}, s_{t+1} \sim p$.
Now, for the term I_a, note that

$$\|h(O_t, \theta_t, L_t, v_{t-\tau}) - h(O_t, \theta_{t-\tau}, L_t, v_{t-\tau})\|$$
$$= \|(r(s_t, a_t) - L_t + \phi(s_{t+1})^\top v_{t-\tau} - \phi(s_t)^\top v_{t-\tau})(\nabla \log \pi_{\theta_t}(a_t|s_t) - \nabla \log \pi_{\theta_{t-\tau}}(a_t|s_t))\|$$
$$\leq GK\|\theta_t - \theta_{t-\tau}\|.$$

For the term I_b, we have ,

$$\|E_{\theta_t}[h(O_t, \theta_t, L_t, v_{t-\tau})] - E_{\theta_{t-\tau}}[h(O_t, \theta_{t-\tau}, L_t, v_{t-\tau})]\|$$
$$\leq \|E_{\theta_t}[h(O_t, \theta_t, L_t, v_{t-\tau})] - E_{\theta_{t-\tau}}[h(O_t, \theta_t, L_t, v_{t-\tau})]\|$$
$$+ \|E_{\theta_{t-\tau}}[h(O_t, \theta_t, L_t, v_{t-\tau})] - E_{\theta_{t-\tau}}[h(O_t, \theta_{t-\tau}, L_t, v_{t-\tau})]\|$$
$$\leq 2Gd_{TV}(\mu_{\theta_t} \otimes \pi_{\theta_t}, \mu_{\theta_{t-\tau}} \otimes \pi_{\theta_{t-\tau}}) + GK\|\theta_t - \theta_{t-\tau}\|$$
$$\leq 2G|A|L\left(1 + \lceil\log_k b^{-1}\rceil + 1/(1-k)\right)\|\theta_t - \theta_{t-\tau}\| + GK\|\theta_t - \theta_{t-\tau}\| \qquad (15)$$
$$= C_1\|\theta_t - \theta_{t-\tau}\|$$

where $C_1 = 2G|A|L\left(1 + \lceil \log_k b^{-1} \rceil + 1/(1-k)\right) + GK$. The inequality in Equation (15) follows from Lemma B.1 of (Wu et al. 2022). Hence, after putting the results back, we obtain,

$$|E[I(O_t, L_t, \theta_t, v_{t-\tau}) - I(O_t, L_t, \theta_{t-\tau}, v_{t-\tau})]| \leq \overline{C}\|\theta_t - \theta_{t-\tau}\|$$

for some $\overline{C} > 0$. The claim follows. □

lemma 3.3.

$$|E[I(O_t, L_t, \theta_{t-\tau}, v_{t-\tau}) - I(O_t, L_{t-\tau}, \theta_{t-\tau}, v_{t-\tau})]| \leq 2BG|L_t - L_{t-\tau}|.$$

Proof.

$$
\begin{aligned}
&|E[I(O_t, L_t, \theta_{t-\tau}, v_{t-\tau}) - I(O_t, L_{t-\tau}, \theta_{t-\tau}, v_{t-\tau})]| \\
&= |E[\langle \nabla L(\theta_{t-\tau}), h(O_t, \theta_{t-\tau}, L_t, v_{t-\tau}) - h(O_t, \theta_{t-\tau}, L_{t-\tau}, v_{t-\tau}) \\
&\quad - E_{\theta_{t-\tau}}[h(O_t, \theta_{t-\tau}, L_t, v_{t-\tau})]\rangle] + E_{\theta_{t-\tau}}[h(O_t, \theta_{t-\tau}, L_{t-\tau}, v_{t-\tau})]\rangle]| \\
&\leq 2BG|L_t - L_{t-\tau}|.
\end{aligned}
$$

□

lemma 3.4.

$$|E[I(O_t, L_{t-\tau}, \theta_{t-\tau}, v_{t-\tau}) - I(\tilde{O}_t, L_{t-\tau}, \theta_{t-\tau}, v_{t-\tau})]| \leq \check{K} \sum_{i=t-\tau}^{t} E\|\theta_i - \theta_{t-\tau}\|.$$

for some $\check{K} > 0$.

Proof. The proof is as in lemma D.2 in (Wu et al. 2022). □

lemma 3.5.

$$|E[I(\tilde{O}_t, L_{t-\tau}, \theta_{t-\tau}, v_{t-\tau}) - I(O'_t, L_{t-\tau}, \theta_{t-\tau}, v_{t-\tau})]| \leq \overline{K}bk^{\tau-1}.$$

for some $\overline{K} > 0$.

Proof. The proof is as in lemma D.3 in (Wu et al. 2022). □

Now collecting the results of lemmas 3.1 - 3.5, we have,

$$
\begin{aligned}
&E[I(O_t, L_t, \theta_t, v_t)] \\
&\geq -4BG\|v_t - v_{t-\tau}\| - \overline{C}\|\theta_t - \theta_{t-\tau}\| - 2BG|L_t - L_{t-\tau}| - \check{K}\sum_{i=t-\tau}^{t} E\|\theta_i - \theta_{t-\tau}\| - \overline{K}bk^{\tau-1} \\
&\geq -4BG\|v_t - v_{t-\tau}\| - \overline{C}\|\theta_t - \theta_{t-\tau}\| - 2BG|L_t - L_{t-\tau}| - \check{K}(\tau+1)E\|\theta_t - \theta_{t-\tau}\| - \overline{K}bk^{\tau-1} \\
&\geq -4BG\|v_t - v_{t-\tau}\| - \overline{C}\|\theta_t - \theta_{t-\tau}\| - 2BG|L_t - L_{t-\tau}| \\
&\quad - \check{K}(\tau+1)\sum_{t-\tau+1}^{t} E\|\theta_k - \theta_{k-\tau}\| - \overline{K}bk^{\tau-1}.
\end{aligned}
$$

The claim of Lemma 3 now follows. □

Lemma 4. *Under assumptions 4.1, 4.3, 4.4, 4.6, for any $t \geq \tau > 0$, we have*

$$E[\Xi(O_t, \theta_t, v_t)] \geq -4B(G + B(\bar{U}_v + U_v))\|v_t - v_{t-\tau}\| - D\|\theta_t - \theta_{t-\tau}\|$$

$$- B_1 \sum_{i=t-\tau}^{t} E\|\theta_i - \theta_{t-\tau}\| - B_2 bk^{\tau-1},$$

where $D > 0, B_1 > 0$ and $B_2 > 0$ are constants.

Proof. We can decompose $E[\Xi(O_t, \theta_t, v_t)]$ as :

$$E[\Xi(O_t, \theta_t, v_t)] = \underbrace{E[\Xi(O_t, \theta_t, v_t) - \Xi(O_t, \theta_t, v_{t-\tau})]}_{I_a} + \underbrace{E[\Xi(O_t, \theta_t, v_{t-\tau}) - \Xi(O_t, \theta_{t-\tau}, v_{t-\tau})]}_{I_b}$$

$$+ \underbrace{E[\Xi(O_t, \theta_{t-\tau}, v_{t-\tau}) - \Xi(\tilde{O}_t, \theta_{t-\tau}, v_{t-\tau})]}_{I_c} + \underbrace{E[\Xi(\tilde{O}_t, \theta_{t-\tau}, v_{t-\tau})]}_{I_d}.$$

In the above equality, $\tilde{O}_t = (\tilde{s}_t, \tilde{a}_t, \tilde{s}_{t+1})$ is from the auxiliary Markov chain defined in Equation (11) and $O'_t = (s'_t, a'_t, s'_{t+1})$, where $s'_t \sim \mu_{\theta_{t-\tau}}, a'_t \sim \pi_{\theta_{t-\tau}}$ and $s'_{t+1} \sim P(.|s'_t, a'_t)$.

For the term I_a above, we have,

$$|\Xi(O_t, \theta_t, v_t) - \Xi(O_t, \theta_t, v_{t-\tau})|$$
$$= |\langle E_{\theta_t}[\bar{W}(O_t, \theta_t, v_t)], E_{\theta_t}[\bar{h}(O_t, \theta_t, v_t)]\rangle - \langle \bar{W}(O_t, \theta_t, v_t), E_{\theta_t}[\bar{h}(O_t, \theta_t, v_t)]\rangle$$
$$- \langle E_{\theta_t}[\bar{W}(O_t, \theta_t, v_{t-\tau})], E_{\theta_t}[\bar{h}(O_t, \theta_t, v_{t-\tau})]\rangle + \langle \bar{W}(O_t, \theta_t, v_{t-\tau}), E_{\theta_t}[\bar{h}(O_t, \theta_t, v_{t-\tau})]\rangle|$$
$$= |\langle E_{\theta_t}[\bar{W}(O_t, \theta_t, v_t)] - \bar{W}(O_t, \theta_t, v_t), E_{\theta_t}[\bar{h}(O_t, \theta_t, v_t)]\rangle$$
$$- \langle E_{\theta_t}[\bar{W}(O_t, \theta_t, v_{t-\tau})] - \bar{W}(O_t, \theta_t, v_{t-\tau}), E_{\theta_t}[\bar{h}(O_t, \theta_t, v_{t-\tau})]\rangle|$$
$$\leq |\langle E_{\theta_t}[\bar{W}(O_t, \theta_t, v_t)] - \bar{W}(O_t, \theta_t, v_t), E_{\theta_t}[\bar{h}(O_t, \theta_t, v_t)]\rangle$$
$$- \langle E_{\theta_t}[\bar{W}(O_t, \theta_t, v_{t-\tau})] - \bar{W}(O_t, \theta_t, v_{t-\tau}), E_{\theta_t}[\bar{h}(O_t, \theta_t, v_t)]\rangle|$$
$$+ |\langle E_{\theta_t}[\bar{W}(O_t, \theta_t, v_{t-\tau})] - \bar{W}(O_t, \theta_t, v_{t-\tau}), E_{\theta_t}[\bar{h}(O_t, \theta_t, v_t)]\rangle$$
$$- \langle E_{\theta_t}[\bar{W}(O_t, \theta_t, v_{t-\tau})] - \bar{W}(O_t, \theta_t, v_{t-\tau}), E_{\theta_t}[\bar{h}(O_t, \theta_t, v_{t-\tau})]\rangle|$$
$$= |\langle E_{\theta_t}[\bar{W}(O_t, \theta_t, v_t)] - E_{\theta_t}[\bar{W}(O_t, \theta_t, v_{t-\tau})] - \bar{W}(O_t, \theta_t, v_t)$$
$$+ \bar{W}(O_t, \theta_t, v_{t-\tau}), E_{\theta_t}[\bar{h}(O_t, \theta_t, v_t)]\rangle|$$
$$+ |\langle E_{\theta_t}[\bar{W}(O_t, \theta_t, v_{t-\tau})] - \bar{W}(O_t, \theta_t, v_{t-\tau}), E_{\theta_t}[\bar{h}(O_t, \theta_t, v_t)] - E_{\theta_t}[\bar{h}(O_t, \theta_t, v_{t-\tau})]\rangle|$$
$$\leq 4BG\|v_t - v_{t-\tau}\| + 4B^2(\bar{U}_v + U_v)\|v_t - v_{t-\tau}\|$$
$$= 4B(G + B(\bar{U}_v + U_v))\|v_t - v_{t-\tau}\|.$$

For term I_b, we have,

$$|\Xi(O_t, \theta_t, v_{t-\tau}) - \Xi(O_t, \theta_{t-\tau}, v_{t-\tau})|$$
$$= |\langle E_{\theta_t}[\bar{W}(O_t, \theta_t, v_{t-\tau})] - \bar{W}(O_t, \theta_t, v_{t-\tau}), E_{\theta_t}[\bar{h}(O_t, \theta_t, v_{t-\tau})]\rangle$$
$$- \langle E_{\theta_{t-\tau}}[\bar{W}(O_t, \theta_{t-\tau}, v_{t-\tau})] - \bar{W}(O_t, \theta_{t-\tau}, v_{t-\tau}), E_{\theta_{t-\tau}}[\bar{h}(O_t, \theta_{t-\tau}, v_{t-\tau})]\rangle|$$
$$= |\langle E_{\theta_t}[\bar{W}(O_t, \theta_t, v_{t-\tau})] - \bar{W}(O_t, \theta_t, v_{t-\tau}), E_{\theta_t}[\bar{h}(O_t, \theta_t, v_{t-\tau})]\rangle$$
$$- \langle E_{\theta_t}[\bar{W}(O_t, \theta_t, v_{t-\tau})] - \bar{W}(O_t, \theta_t, v_{t-\tau}), E_{\theta_{t-\tau}}[\bar{h}(O_t, \theta_{t-\tau}, v_{t-\tau})]\rangle$$
$$+ \langle E_{\theta_t}[\bar{W}(O_t, \theta_t, v_{t-\tau})] - \bar{W}(O_t, \theta_t, v_{t-\tau}), E_{\theta_{t-\tau}}[\bar{h}(O_t, \theta_{t-\tau}, v_{t-\tau})]\rangle$$
$$- \langle E_{\theta_{t-\tau}}[\bar{W}(O_t, \theta_{t-\tau}, v_{t-\tau})] - \bar{W}(O_t, \theta_{t-\tau}, v_{t-\tau}), E_{\theta_{t-\tau}}[\bar{h}(O_t, \theta_{t-\tau}, v_{t-\tau})]\rangle|$$
$$\leq \underbrace{|\langle E_{\theta_t}[\bar{W}(O_t, \theta_t, v_{t-\tau})] - \bar{W}(O_t, \theta_t, v_{t-\tau}), E_{\theta_t}[\bar{h}(O_t, \theta_t, v_{t-\tau})] - E_{\theta_{t-\tau}}[\bar{h}(O_t, \theta_{t-\tau}, v_{t-\tau})]\rangle|}_{I_{b1}}$$
$$+ \underbrace{|\langle E_{\theta_t}[\bar{W}(O_t, \theta_t, v_{t-\tau})] - E_{\theta_{t-\tau}}[\bar{W}(O_t, \theta_{t-\tau}, v_{t-\tau})]}_{I_{b2}}$$
$$\underbrace{- \bar{W}(O_t, \theta_t, v_{t-\tau}) + \bar{W}(O_t, \theta_{t-\tau}, v_{t-\tau}), E_{\theta_{t-\tau}}[\bar{h}(O_t, \theta_{t-\tau}, v_{t-\tau})]\rangle|}_{I_{b3}}.$$

For term I_{b1}, we have ,

$$|\langle E_{\theta_t}[\bar{W}(O_t, \theta_t, v_{t-\tau})] - \bar{W}(O_t, \theta_t, v_{t-\tau}), E_{\theta_t}[\bar{h}(O_t, \theta_t, v_{t-\tau})] - E_{\theta_{t-\tau}}[\bar{h}(O_t, \theta_{t-\tau}, v_{t-\tau})]\rangle|$$
$$\leq 4B(\overline{U_v} + U_v)\|E_{\theta_t}[\bar{h}(O_t, \theta_t, v_{t-\tau})] - E_{\theta_{t-\tau}}[\bar{h}(O_t, \theta_{t-\tau}, v_{t-\tau})]\|$$
$$\leq 4B(\overline{U_v} + U_v)\|E_{\theta_t}[\bar{h}(O_t, \theta_t, v_{t-\tau})] - E_{\theta_{t-\tau}}[\bar{h}(O_t, \theta_t, v_{t-\tau})]\|$$
$$+ 4B(\overline{U_v} + U_v)\|E_{\theta_{t-\tau}}[\bar{h}(O_t, \theta_t, v_{t-\tau})] - E_{\theta_{t-\tau}}[\bar{h}(O_t, \theta_{t-\tau}, v_{t-\tau})]\|$$
$$\leq 8GB(\overline{U_v} + U_v)d_{TV}(\mu_{\theta_t} \otimes \pi_{\theta_t}, \mu_{\theta_{t-\tau}} \otimes \pi_{\theta_{t-\tau}})$$
$$+ 4B(\overline{U_v} + U_v)E_{\theta_{t-\tau}}[\|\bar{h}(O_t, \theta_t, v_{t-\tau})] - E_{\theta_{t-\tau}}[\bar{h}(O_t, \theta_{t-\tau}, v_{t-\tau})]\|]$$
$$\leq 8GB(\overline{U_v} + U_v)|A|L\left(1 + \lceil\log_k b^{-1}\rceil + \frac{1}{1-k}\right)\|\theta_t - \theta_{t-\tau}\|$$
$$+ 8B(\overline{U_v} + U_v)(U_r + U_v)K\|\theta_t - \theta_{t-\tau}\|$$
$$= D_1\|\theta_t - \theta_{t-\tau}\|,$$

where $D_1 = 8B(\overline{U_v} + U_v)\left(G|A|L\left(1 + \lceil\log_k b^{-1}\rceil + \frac{1}{1-k}\right) + (U_r + U_v)K\right)$.
The last inequality above follows from Lemma B.1 in (Wu et al. 2022).

Next, for term I_{b2} + term I_{b3}, we have,

$$|\langle E_{\theta_t}[\bar{W}(O_t, \theta_t, v_{t-\tau})] - E_{\theta_{t-\tau}}[\bar{W}(O_t, \theta_{t-\tau}, v_{t-\tau})] - \bar{W}(O_t, \theta_t, v_{t-\tau})$$
$$+ \bar{W}(O_t, \theta_{t-\tau}, v_{t-\tau}), E_{\theta_{t-\tau}}[\bar{h}(O_t, \theta_{t-\tau}, v_{t-\tau})]\rangle|$$
$$\leq G(\|E_{\theta_t}[\bar{W}(O_t, \theta_t, v_{t-\tau})] - E_{\theta_{t-\tau}}[\bar{W}(O_t, \theta_{t-\tau}, v_{t-\tau})]\| + \|\bar{W}(O_t, \theta_t, v_{t-\tau})$$
$$- \bar{W}(O_t, \theta_{t-\tau}, v_{t-\tau})\|)$$
$$\leq G(\|E_{\theta_t}[\bar{W}(O_t, \theta_t, v_{t-\tau})] - E_{\theta_{t-\tau}}[\bar{W}(O_t, \theta_{t-\tau}, v_{t-\tau})]\| + 2(\overline{U}_v + U_v)K\|\theta_t - \theta_{t-\tau}\|)$$
$$\leq G(\|E_{\theta_t}[\bar{W}(O_t, \theta_t, v_{t-\tau})] - E_{\theta_{t-\tau}}[\bar{W}(O_t, \theta_t, v_{t-\tau})]\| + \|E_{\theta_{t-\tau}}[\bar{W}(O_t, \theta_t, v_{t-\tau})]$$
$$- E_{\theta_{t-\tau}}[\bar{W}(O_t, \theta_{t-\tau}, v_{t-\tau})]\|) + 2G(\overline{U}_v + U_v)K\|\theta_t - \theta_{t-\tau}\|$$
$$\leq D_1\|\theta_t - \theta_{t-\tau}\|,$$

where $D_1 > 0$.

The last inequality above again follows from Lemma B.1 in (Wu et al. 2022).
Hence after collecting the results of terms I_{b1} and I_{b2} we have,

$$I_b \geq -D\|\theta_t - \theta_{t-\tau}\|,$$

for some $D > 0$.

Now, for the term I_c, we have,

$$|E[\Xi(O_t, \theta_{t-\tau}, v_{t-\tau}) - \Xi(\tilde{O}_t, \theta_{t-\tau}, v_{t-\tau})]| \leq B_1 \sum_{i=t-\tau}^{t} E\|\theta_i - \theta_{t-\tau}\|,$$

for some $B_1 > 0$.

For term I_d, we have,

$$|E[\Xi(\tilde{O}_t, \theta_{t-\tau}, v_{t-\tau})]| \leq B_2 bk^{\tau-1}.$$

for some $B_2 > 0$. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

For analysis of terms I_c and I_d, please see lemmas D.10 and D.11 in (Wu et al. 2022).
Thus, after collecting all the terms, we have,

$$E[\Xi(O_t, \theta_t, v_t)] \geq -4B(G + B(\bar{U}_v + U_v))\|v_t - v_{t-\tau}\| - D\|\theta_t - \theta_{t-\tau}\|$$
$$- B_1 \sum_{i=t-\tau}^{t} E\|\theta_i - \theta_{t-\tau}\| - B_2 bk^{\tau-1},$$

where $D > 0, B_1 > 0$ and $B_2 > 0$.

Proof of convergence of the actor. After applying Lemma 1 to the update rule of the actor, we have,

$$L(\theta_{t+1}) \geq L(\theta_t) + \alpha_t \langle \nabla L(\theta_t), \delta_t \nabla \log \pi_{\theta_t}(a_t|s_t) \rangle - M_L \alpha_t^2 \|\delta_t \nabla \log \pi_{\theta_t}(a_t|s_t)\|^2.$$

For the term $\langle \nabla L(\theta_t), \delta_t \nabla \log \pi_{\theta_t}(a_t|s_t) \rangle$, we have,

$$\langle \nabla L(\theta_t), \delta_t \nabla \log \pi_{\theta_t}(a_t|s_t) \rangle$$
$$= \langle \nabla L(\theta_t), (r(s_t, a_t) - L_t + \phi(s_{t+1})^\top v_t - \phi(s_t)^\top v_t) \nabla \log \pi_{\theta_t}(a_t|s_t) \rangle$$
$$= I(O_t, \theta_t, L_t, v_t) + \langle \nabla L(\theta_t), E_{s_t \sim \mu_{\theta_t}, a_t \sim \pi_{\theta_t}, s_{t+1} \sim p}[h(O_t, \theta_t, L_t, v_t)] \rangle.$$

Hence,

$$L(\theta_{t+1}) \tag{16}$$
$$\geq L(\theta_t) + \alpha_t I(O_t, \theta_t, L_t, v_t) + \alpha_t \langle \nabla L(\theta_t), M(\theta_t, v_t) \rangle$$
$$\quad + \alpha_t \langle \nabla L(\theta_t), E_{\theta_t}[(L(\theta_t) - L_t) \nabla \log \pi_{\theta_t}(a_t|s_t)] \rangle - M_L \alpha_t^2 \|\delta_t \nabla \log \pi_{\theta_t}(a_t|s_t)\|^2$$
$$= L(\theta_t) + \alpha_t I(O_t, \theta_t, L_t, v_t) + \alpha_t \|M(\theta_t, v_t)\|^2$$
$$\quad + \alpha_t \langle E_{\theta_t}[(V^{\theta_t}(s_{t+1}) - \phi(s_{t+1})^T v_t - V^{\theta_t}(s_t) + \phi(s_t)^T v_t) \nabla \log \pi_{\theta_t}(a_t|s_t)], \tag{17}$$
$$\qquad\qquad E_{\theta_t}[\bar{h}(O_t, \theta_t, v_t)] \rangle$$
$$\quad - \alpha_t \langle (V^{\theta_t}(s_{t+1}) - \phi(s_{t+1})^T v_t - V^{\theta_t}(s_t) + \phi(s_t)^T v_t) \nabla \log \pi_{\theta_t}(a_t|s_t), E_{\theta_t}[\bar{h}(O_t, \theta_t, v_t)] \rangle$$
$$\quad + \underbrace{\alpha_t \langle (V^{\theta_t}(s_{t+1}) - \phi(s_{t+1})^T v_t - V^{\theta_t}(s_t) + \phi(s_t)^T v_t) \nabla \log \pi_{\theta_t}(a_t|s_t), E_{\theta_t}[\bar{h}(O_t, \theta_t, v_t)] \rangle}_{I_1}$$
$$\quad + \alpha_t \langle \nabla L(\theta_t), E_{\theta_t}[(L(\theta_t) - L_t) \nabla \log \pi_{\theta_t}(a_t|s_t)] \rangle - M_L \alpha_t^2 \|\delta_t \nabla \log \pi_{\theta_t}(a_t|s_t)\|^2. \tag{18}$$

Now,

$$\alpha_t \langle (V^{\theta_t}(s_{t+1}) - \phi(s_{t+1})^T v_t) \nabla \log \pi_{\theta_t}(a_t|s_t), M(\theta_t, v_t) \rangle$$
$$= \alpha_t \langle (V^{\theta_t}(s_{t+1}) - V^{\theta_{t+1}}(s_{t+1}) + V^{\theta_{t+1}}(s_{t+1}) - \phi(s_{t+1})^T v_t) \nabla \log \pi_{\theta_t}(a_t|s_t), M(\theta_t, v_t) \rangle$$
$$= \alpha_t \langle (V^{\theta_t}(s_{t+1}) - V^{\theta_{t+1}}(s_{t+1})) \nabla \log \pi_{\theta_t}(a_t|s_t), M(\theta_t, v_t) \rangle$$
$$\quad + \alpha_t \langle (V^{\theta_{t+1}}(s_{t+1}) - \phi(s_{t+1})^T v_t) \nabla \log \pi_{\theta_t}(a_t|s_t), M(\theta_t, v_t) \rangle$$
$$= \alpha_t \langle (V^{\theta_t}(s_{t+1}) - V^{\theta_{t+1}}(s_{t+1})) \nabla \log \pi_{\theta_t}(a_t|s_t), M(\theta_t, v_t) \rangle$$
$$\quad + \alpha_t \langle (V^{\theta_{t+1}}(s_{t+1}) - \phi(s_{t+1})^T v_{t+1} + \phi(s_{t+1})^T v_{t+1} - \phi(s_{t+1})^T v_t) \nabla \log \pi_{\theta_t}(a_t|s_t)$$
$$\quad , M(\theta_t, v_t) \rangle$$
$$= \alpha_t \langle (V^{\theta_t}(s_{t+1}) - V^{\theta_{t+1}}(s_{t+1})) \nabla \log \pi_{\theta_t}(a_t|s_t), M(\theta_t, v_t) \rangle$$
$$\quad + \alpha_t \langle (\phi(s_{t+1})^T v_{t+1} - \phi(s_{t+1})^T v_t) \nabla \log \pi_{\theta_t}(a_t|s_t), M(\theta_t, v_t) \rangle$$
$$\quad + \alpha_t \langle (V^{\theta_{t+1}}(s_{t+1}) - \phi(s_{t+1})^T v_{t+1}) \nabla \log \pi_{\theta_t}(a_t|s_t), M(\theta_t, v_t) \rangle$$
$$= \alpha_t \langle (V^{\theta_t}(s_{t+1}) - V^{\theta_{t+1}}(s_{t+1})) \nabla \log \pi_{\theta_t}(a_t|s_t), M(\theta_t, v_t) \rangle$$
$$\quad + \alpha_t \langle (\phi(s_{t+1})^T v_{t+1} - \phi(s_{t+1})^T v_t) \nabla \log \pi_{\theta_t}(a_t|s_t), M(\theta_t, v_t) \rangle$$
$$\quad + \alpha_{t+1} \langle (V^{\theta_{t+1}}(s_{t+1}) - \phi(s_{t+1})^T v_{t+1}) \nabla \log \pi_{\theta_{t+1}}(a_{t+1}|s_{t+1}), M(\theta_{t+1}, v_{t+1}) \rangle$$
$$\quad + \alpha_t \langle (V^{\theta_{t+1}}(s_{t+1}) - \phi(s_{t+1})^T v_{t+1}) \nabla \log \pi_{\theta_t}(a_t|s_t), M(\theta_t, v_t) \rangle$$
$$\quad - \alpha_{t+1} \langle (V^{\theta_{t+1}}(s_{t+1}) - \phi(s_{t+1})^T v_{t+1}) \nabla \log \pi_{\theta_{t+1}}(a_{t+1}|s_{t+1}), M(\theta_{t+1}, v_{t+1}) \rangle.$$

Hence for the term I_1, we have,

$$I_1 = \alpha_t \langle (V^{\theta_t}(s_{t+1}) - V^{\theta_{t+1}}(s_{t+1})) \nabla \log \pi_{\theta_t}(a_t|s_t), M(\theta_t, v_t) \rangle$$
$$\quad + \alpha_t \langle (\phi(s_{t+1})^T v_{t+1} - \phi(s_{t+1})^T v_t) \nabla \log \pi_{\theta_t}(a_t|s_t), M(\theta_t, v_t) \rangle$$
$$\quad + \alpha_{t+1} \langle (V^{\theta_{t+1}}(s_{t+1}) - \phi(s_{t+1})^T v_{t+1}) \nabla \log \pi_{\theta_{t+1}}(a_{t+1}|s_{t+1}), M(\theta_{t+1}, v_{t+1}) \rangle$$
$$\quad + \alpha_t \langle (V^{\theta_{t+1}}(s_{t+1}) - \phi(s_{t+1})^T v_{t+1}) \nabla \log \pi_{\theta_t}(a_t|s_t), M(\theta_t, v_t) \rangle$$
$$\quad - \alpha_{t+1} \langle (V^{\theta_{t+1}}(s_{t+1}) - \phi(s_{t+1})^T v_{t+1}) \nabla \log \pi_{\theta_{t+1}}(a_{t+1}|s_{t+1}), M(\theta_{t+1}, v_{t+1}) \rangle$$
$$\quad + \alpha_t \langle (-V^{\theta_t}(s_t) + \phi(s_t)^T v_t) \nabla \log \pi_{\theta_t}(a_t|s_t), M(\theta_t, v_t) \rangle.$$

Putting this back in Equation (16), we obtain,

$$
\begin{aligned}
L(\theta_{t+1}) \geq{} & L(\theta_t) + \alpha_t I(O_t, \theta_t, L_t, v_t) + \alpha_t \|M(\theta_t, v_t)\|^2 + \alpha_t \Xi(O_t, \theta_t, v_t) \\
& + \alpha_t \langle (V^{\theta_t}(s_{t+1}) - V^{\theta_{t+1}}(s_{t+1}))\nabla \log \pi_{\theta_t}(a_t|s_t), M(\theta_t, v_t)\rangle \\
& + \alpha_t \langle (\phi(s_{t+1})^T v_{t+1} - \phi(s_{t+1})^T v_t)\nabla \log \pi_{\theta_t}(a_t|s_t), M(\theta_t, v_t)\rangle \\
& + \alpha_{t+1} \langle (V^{\theta_{t+1}}(s_{t+1}) - \phi(s_{t+1})^T v_{t+1})\nabla \log \pi_{\theta_{t+1}}(a_{t+1}|s_{t+1}), M(\theta_{t+1}, v_{t+1})\rangle \\
& + \alpha_t \langle (V^{\theta_{t+1}}(s_{t+1}) - \phi(s_{t+1})^T v_{t+1})\nabla \log \pi_{\theta_t}(a_t|s_t), M(\theta_t, v_t)\rangle \\
& - \alpha_{t+1} \langle (V^{\theta_{t+1}}(s_{t+1}) - \phi(s_{t+1})^T v_{t+1})\nabla \log \pi_{\theta_{t+1}}(a_{t+1}|s_{t+1}), M(\theta_{t+1}, v_{t+1})\rangle \\
& + \alpha_t \langle (-V^{\theta_t}(s_t) + \phi(s_t)^T v_t)\nabla \log \pi_{\theta_t}(a_t|s_t), M(\theta_t, v_t)\rangle \\
& + \alpha_t \langle \nabla L(\theta_t), E_{\theta_t}[(L(\theta_t) - L_t)\nabla \log \pi_{\theta_t}(a_t|s_t)]\rangle - M_L \alpha_t^2 \|\delta_t \nabla \log \pi_{\theta_t}(a_t|s_t)\|^2.
\end{aligned}
$$

$$
\begin{aligned}
\Rightarrow \|M(\theta_t, v_t)\|^2 \\
\leq{} & \frac{L(\theta_{t+1}) - L(\theta_t)}{\alpha_t} - I(O_t, \theta_t, L_t, v_t) - \Xi(O_t, \theta_t, v_t) \\
& - \langle (V^{\theta_t}(s_{t+1}) - V^{\theta_{t+1}}(s_{t+1}))\nabla \log \pi_{\theta_t}(a_t|s_t), M(\theta_t, v_t)\rangle \\
& - \langle (\phi(s_{t+1})^T v_{t+1} - \phi(s_{t+1})^T v_t)\nabla \log \pi_{\theta_t}(a_t|s_t), M(\theta_t, v_t)\rangle \\
& - \frac{1}{\alpha_t}\alpha_{t+1} \langle (V^{\theta_{t+1}}(s_{t+1}) - \phi(s_{t+1})^T v_{t+1})\nabla \log \pi_{\theta_{t+1}}(a_{t+1}|s_{t+1}), M(\theta_{t+1}, v_{t+1})\rangle \\
& - \frac{1}{\alpha_t}(\alpha_t \langle (V^{\theta_{t+1}}(s_{t+1}) - \phi(s_{t+1})^T v_{t+1})\nabla \log \pi_{\theta_t}(a_t|s_t), M(\theta_t, v_t)\rangle) \\
& + \frac{1}{\alpha_t}\alpha_{t+1} \langle (V^{\theta_{t+1}}(s_{t+1}) - \phi(s_{t+1})^T v_{t+1})\nabla \log \pi_{\theta_{t+1}}(a_{t+1}|s_{t+1}), M(\theta_{t+1}, v_{t+1})\rangle \\
& - \frac{1}{\alpha_t}(\alpha_t \langle (-V^{\theta_t}(s_t) + \phi(s_t)^T v_t)\nabla \log \pi_{\theta_t}(a_t|s_t), M(\theta_t, v_t)\rangle) \\
& - \langle \nabla L(\theta_t), (L(\theta_t) - L_t)\nabla \log \pi_{\theta_t}(a_t|s_t)\rangle + M_L \alpha_t \|\delta_t \nabla \log \pi_{\theta_t}(a_t|s_t)\|^2 \\
\leq{} & (L(\theta_{t+1}) - L(\theta_t) + Q_t - Q_{t+1})/\alpha_t - I(O_t, \theta_t, L_t, v_t) - \Xi(O_t, \theta_t, v_t) \\
& - \langle (V^{\theta_t}(s_{t+1}) - V^{\theta_{t+1}}(s_{t+1}))\nabla \log \pi_{\theta_t}(a_t|s_t), M(\theta_t, v_t)\rangle \\
& - \langle (\phi(s_{t+1})^T v_{t+1} - \phi(s_{t+1})^T v_t)\nabla \log \pi_{\theta_t}(a_t|s_t), M(\theta_t, v_t)\rangle \\
& - \frac{1}{\alpha_t}\alpha_t \langle (V^{\theta_{t+1}}(s_{t+1}) - \phi(s_{t+1})^T v_{t+1})\nabla \log \pi_{\theta_t}(a_t|s_t), M(\theta_t, v_t)\rangle \\
& + \frac{1}{\alpha_t}\alpha_{t+1} \langle (V^{\theta_{t+1}}(s_{t+1}) - \phi(s_{t+1})^T v_{t+1})\nabla \log \pi_{\theta_{t+1}}(a_{t+1}|s_{t+1}), M(\theta_{t+1}, v_{t+1})\rangle \\
& - \langle \nabla L(\theta_t), (L(\theta_t) - L_t)\nabla \log \pi_{\theta_t}(a_t|s_t)\rangle + M_L \alpha_t \|\delta_t \nabla \log \pi_{\theta_t}(a_t|s_t)\|^2,
\end{aligned}
$$

where, in the above, $Q_t = \alpha_t \langle (V^{\theta_t}(s_t) - \phi(s_t)^T v_t)\nabla \log \pi_{\theta_t}(a_t|s_t), M(\theta_t, v_t)\rangle$. Taking expectations on both sides and summing

from τ_t to t, we obtain,

$$\sum_{k=\tau_t}^{t} E\|M(\theta_k, v_k)\|^2$$

$$\leq \underbrace{\sum_{k=\tau_t}^{t} E[(L(\theta_{k+1}) - L(\theta_k) + Q_k - Q_{k+1})/\alpha_k]}_{I_1} - \underbrace{\sum_{k=\tau_t}^{t} E[I(O_k, \theta_k, L_k, v_k)]}_{I_2}$$

$$- \underbrace{\sum_{k=\tau_t}^{t} E[\Xi(O_k, \theta_k, v_k)]}_{I_3} - \underbrace{\sum_{k=\tau_t}^{t} E[\langle (V^{\theta_k}(s_{k+1}) - V^{\theta_{k+1}}(s_{k+1}))\nabla \log \pi_{\theta_k}(a_k|s_k), M(\theta_k, v_k)\rangle]}_{I_4}$$

$$- \underbrace{\sum_{k=\tau_t}^{t} E[\langle (\phi(s_{k+1})^T v_{k+1} - \phi(s_{k+1})^T v_k)\nabla \log \pi_{\theta_k}(a_k|s_k), M(\theta_k, v_k)\rangle]}_{I_5}$$

$$- \underbrace{\sum_{k=\tau_t}^{t} \frac{1}{\alpha_k} E[\alpha_k \langle (V^{\theta_{k+1}}(s_{k+1}) - \phi(s_{k+1})^T v_{k+1})\nabla \log \pi_{\theta_k}(a_k|s_k), M(\theta_k, v_k)\rangle]}_{I_6}$$

$$+ \underbrace{\sum_{k=\tau_t}^{t} \frac{1}{\alpha_k} E[\alpha_{k+1} \langle (V^{\theta_{k+1}}(s_{k+1}) - \phi(s_{k+1})^T v_{k+1})\nabla \log \pi_{\theta_{k+1}}(a_{k+1}|s_{k+1}), M(\theta_{k+1}, v_{k+1})\rangle]}_{I_7}$$

$$- \underbrace{\sum_{k=\tau_t}^{t} E[\langle \nabla L(\theta_k), (L(\theta_k) - L_k)\nabla \log \pi_{\theta_k}(a_k|s_k)]\rangle]}_{I_8} + \underbrace{\sum_{k=\tau_t}^{t} M_L \alpha_k E[\|\delta_k \nabla \log \pi_{\theta_k}(a_k|s_k)\|^2]}_{I_9}.$$

Now, for term I_1 we have,

$$\sum_{k=\tau_t}^{t} E[(L(\theta_{k+1}) - L(\theta_k) + Q_k - Q_{k+1})/\alpha_k]$$

$$= \sum_{k=\tau_t}^{t} E[(A_{k+1} - A_k)/\alpha_k]$$

$$= \mathcal{O}(t^\nu),$$

where $A_k = L(\theta_k) - Q_k$.

The analysis of term I_1 is similar to that of term I_1 in Section A.

For term I_2, we have,

$$-E[I(O_t, \theta_t, L_t, v_t)] \leq D_1(\tau + 1) \sum_{k=t-\tau+1}^{t} E\|\theta_k - \theta_{k-1}\| + D_2 bk^{\tau-1} + D_3 E\|v_t - v_{t-\tau}\|$$

$$+ D_4 E|L_t - L_{t-\tau}|.$$

This inequality comes from lemma 3.

After summing both sides from τ_t to t and taking $\tau = \tau_t$, we will get,

$$I_2 = \mathcal{O}(\log^2 t \cdot t^{1-\nu}).$$

For term I_3, we have,

$$-E[\Xi(O_t, \theta_t, v_t)] \le 4B(G + B(\bar{U}_v + U_v))\|v_t - v_{t-\tau}\| + D\|\theta_t - \theta_{t-\tau}\|$$

$$+ B_1 \sum_{i=t-\tau}^{t} E\|\theta_i - \theta_{t-\tau}\| + B_2 bk^{\tau-1}.$$

This inequality is a result of lemma 4.

After taking $\tau = \tau_t$ and summing the expectation on both the sides from τ_t to t, we get,

$$I_3 = \mathcal{O}(\log^2 t \cdot t^{1-\nu}),$$

where the $\log^2 t$ term arises here because of the definition of τ_t (see Equation (2)).

For term I_4, we have,

$$-\sum_{k=\tau_t}^{t} E[E_{\theta_k}[\langle(V^{\theta_k}(s_{k+1}) - V^{\theta_{k+1}}(s_{k+1}))\nabla \log \pi_{\theta_k}(a_k|s_k), M(\theta_k, v_k)\rangle]]$$

$$\le \sum_{k=\tau_t}^{t} E[E_{\theta_k}[\|(V^{\theta_k}(s_{k+1}) - V^{\theta_{k+1}}(s_{k+1}))\nabla \log \pi_{\theta_k}(a_k|s_k)\|\|M(\theta_k, v_k)\|]]$$

$$\le 4B^3(U_r + U_v)^2 L_v \sum_{k=\tau_t}^{t} \alpha_k.$$

The last inequality follows from Assumption 4.6.

Next, for the term I_5, we have,

$$-\sum_{k=\tau_t}^{t} E[E_{\theta_k}[\langle(\phi(s_{k+1})^T v_{k+1} - \phi(s_{k+1})^T v_k)\nabla \log \pi_{\theta_k}(a_k|s_k), M(\theta_k, v_k)\rangle]]$$

$$\le \sum_{k=\tau_t}^{t} E[E_{\theta_k}[\|(\phi(s_{k+1})^T v_{k+1} - \phi(s_{k+1})^T v_k)\nabla \log \pi_{\theta_k}(a_k|s_k)\|\|M(\theta_k, v_k)\|]]$$

$$\le 4B^2(U_r + U_v)^2 \sum_{k=\tau_t}^{t} \beta_k.$$

For terms I_6 and I_7 summed together, we have,

$$I_6 + I_7$$

$$= \sum_{k=\tau_t}^{t} \frac{1}{\alpha_k} E[\alpha_{k+1}\langle(V^{\theta_{k+1}}(s_{k+1}) - \phi(s_{k+1})^T v_{k+1})\nabla\log\pi_{\theta_{k+1}}(a_{k+1}|s_{k+1}), M(\theta_{k+1}, v_{k+1})\rangle]$$

$$- \sum_{k=\tau_t}^{t} \frac{1}{\alpha_k} E[\alpha_k\langle(V^{\theta_{k+1}}(s_{k+1}) - \phi(s_{k+1})^T v_{k+1})\nabla\log\pi_{\theta_k}(a_k|s_k), M(\theta_k, v_k)\rangle]$$

$$= \sum_{k=\tau_t}^{t} \frac{1}{\alpha_k} E[\alpha_{k+1}\langle(V^{\theta_{k+1}}(s_{k+1}) - \phi(s_{k+1})^T v_{k+1})\nabla\log\pi_{\theta_{k+1}}(a_{k+1}|s_{k+1}), M(\theta_{k+1}, v_{k+1})\rangle]$$

$$- \sum_{k=\tau_t}^{t} \frac{1}{\alpha_k} E[\alpha_k\langle(V^{\theta_{k+1}}(s_{k+1}) - \phi(s_{k+1})^T v_{k+1})\nabla\log\pi_{\theta_{k+1}}(a_{k+1}|s_{k+1}), M(\theta_{k+1}, v_{k+1})\rangle]$$

$$+ \sum_{k=\tau_t}^{t} \frac{1}{\alpha_k} E[\alpha_k\langle(V^{\theta_{k+1}}(s_{k+1}) - \phi(s_{k+1})^T v_{k+1})\nabla\log\pi_{\theta_{k+1}}(a_{k+1}|s_{k+1}), M(\theta_{k+1}, v_{k+1})\rangle]$$

$$- \sum_{k=\tau_t}^{t} \frac{1}{\alpha_k} E[\alpha_k\langle(V^{\theta_{k+1}}(s_{k+1}) - \phi(s_{k+1})^T v_{k+1})\nabla\log\pi_{\theta_k}(a_k|s_k), M(\theta_{k+1}, v_{k+1})\rangle]$$

$$+ \sum_{k=\tau_t}^{t} \frac{1}{\alpha_k} E[\alpha_k\langle(V^{\theta_{k+1}}(s_{k+1}) - \phi(s_{k+1})^T v_{k+1})\nabla\log\pi_{\theta_k}(a_k|s_k), M(\theta_{k+1}, v_{k+1})\rangle]$$

$$- \sum_{k=\tau_t}^{t} \frac{1}{\alpha_k} E[\alpha_k\langle(V^{\theta_{k+1}}(s_{k+1}) - \phi(s_{k+1})^T v_{k+1})\nabla\log\pi_{\theta_k}(a_k|s_k), M(\theta_k, v_k)\rangle]$$

$$\leq GB(U_v + \bar{U}_v)\sum_{k=\tau_t}^{t} \frac{\alpha_{k+1} - \alpha_k}{\alpha_k} + \mathcal{O}\left(\sum_{k=\tau_t}^{t} \|\theta_k - \theta_{k+1}\|\right)$$

$$+ \mathcal{O}\left(\sum_{k=\tau_t}^{t} \|M(\theta_{k+1}, v_{k+1}) - M(\theta_k, v_k)\|\right)$$

$$= \mathcal{O}(t^{1-\nu}).$$

For term I_8, we have,

$$- \sum_{k=\tau_t}^{t} E[\langle\nabla L(\theta_k), (L(\theta_k) - L_k)\nabla\log\pi_{\theta_k}(a_k|s_k)]\rangle]$$

$$= \sum_{k=\tau_t}^{t} E[\langle E_{\theta_k}[(r(s,a) - L(\theta_k) + V^{\theta_k}(s^{'}) - V^{\theta_k}(s))\nabla\log\pi_{\theta_k}(a|s)] \tag{19}$$

$$, (L_k - L(\theta_k))\nabla\log\pi_{\theta_k}(a_k|s_k)]\rangle]$$

$$= \sum_{k=\tau_t}^{t} E[\langle E_{\theta_k}[(r(s,a) - L(\theta_k) + (\phi(s^{'}) - \phi(s))^T v(k))\nabla\log\pi_{\theta_k}(a|s)] \tag{20}$$

$$, (L_k - L(\theta_k))\nabla\log\pi_{\theta_k}(a_k|s_k)]\rangle]$$

$$+ \sum_{k=\tau_t}^{t} E[\langle E_{\theta_k}[(V^{\theta_k}(s^{'}) - \phi(s^{'})^T v_k + \phi(s)^T v_k - V^{\theta_k}(s))\nabla\log\pi_{\theta_k}(a|s)] \tag{21}$$

$$, (L_k - L(\theta_k))\nabla\log\pi_{\theta_k}(a_k|s_k)]\rangle]$$

$$\leq B\sqrt{\sum_{k=\tau_t}^{t} E\|M(\theta_k, v_k)\|^2}\sqrt{\sum_{k=\tau_t}^{t} E|L_k - L(\theta_k)|^2} + I_{8a}. \tag{22}$$

where

$$I_{8a} = \sum_{k=\tau_t}^{t} E[\langle E_{\theta_k}[(V^{\theta_k}(s^{'}) - \phi(s^{'})^T v_k + \phi(s)^T v_k - V^{\theta_k}(s))\nabla \log \pi_{\theta_k}(a|s)]$$
$$, (L_k - L(\theta_k))\nabla \log \pi_{\theta_k}(a_k|s_k)]\rangle]$$

Now, for the term I_{8a}, we have,

$$I_{8a} = I_{8a1} + I_{8a2}.$$

where,

$$I_{8a1} = \sum_{k=\tau_t}^{t} E[\langle E_{\theta_k}[\bar{W}(O_k, \theta_k, v_k)] - \bar{W}(O_k, \theta_k, v_k)], (L_k - L(\theta_k))\nabla \log \pi_{\theta_k}(a_k|s_k)]\rangle]$$

and,

$$I_{8a2} = \sum_{k=\tau_t}^{t} E[\langle (V^{\theta_k}(s_{k+1}) - \phi(s_{k+1})^T v_k + \phi(s_k)^T v_k - V^{\theta_k}(s_k))\nabla \log \pi_{\theta_k}(a_k|s_k)$$
$$, (L_k - L(\theta_k))\nabla \log \pi_{\theta_k}(a_k|s_k)]\rangle]$$

After analysing the term I_{8a1}, similar to lemma 4, we get,

$$I_{8a1} = \mathcal{O}(\log^2 t \cdot t^{1-\nu}).$$

For the term I_{8a2}, we have,

$$\sum_{k=\tau_t}^{t} E[\langle (V^{\theta_k}(s_{k+1}) - \phi(s_{k+1})^T v_k + \phi(s_k)^T v_k - V^{\theta_k}(s_k))\nabla \log \pi_{\theta_k}(a_k|s_k)$$

$$, (L_k - L(\theta_k))\nabla \log \pi_{\theta_k}(a_k|s_k)]\rangle]$$

$$= \sum_{k=\tau_t}^{t} E[\langle (V^{\theta_k}(s_{k+1}) - \phi(s_{k+1})^T v_k)\nabla \log \pi_{\theta_k}(a_k|s_k), (L_k - L(\theta_k))\nabla \log \pi_{\theta_k}(a_k|s_k)]\rangle]$$

$$- \sum_{k=\tau_t}^{t} E[\langle (V^{\theta_k}(s_k) - \phi(s_k)^T v_k)\nabla \log \pi_{\theta_k}(a_k|s_k), (L_k - L(\theta_k))\nabla \log \pi_{\theta_k}(a_k|s_k)]\rangle]$$

$$= \sum_{k=\tau_t}^{t} E[\langle (V^{\theta_k}(s_{k+1}) - \phi(s_{k+1})^T v_k)\nabla \log \pi_{\theta_k}(a_k|s_k), (L_k - L(\theta_k))\nabla \log \pi_{\theta_k}(a_k|s_k)]\rangle]$$

$$- \sum_{k=\tau_t}^{t} E[\langle (V^{\theta_{k+1}}(s_{k+1}) - \phi(s_{k+1})^T v_{k+1})\nabla \log \pi_{\theta_{k+1}}(a_{k+1}|s_{k+1}),$$

$$(L_{k+1} - L(\theta_{k+1}))\nabla \log \pi_{\theta_{k+1}}(a_{k+1}|s_{k+1})]\rangle]$$

$$+ \sum_{k=\tau_t}^{t} E[\langle (V^{\theta_{k+1}}(s_{k+1}) - \phi(s_{k+1})^T v_{k+1})\nabla \log \pi_{\theta_{k+1}}(a_{k+1}|s_{k+1}),$$

$$(L_{k+1} - L(\theta_{k+1}))\nabla \log \pi_{\theta_{k+1}}(a_{k+1}|s_{k+1})]\rangle]$$

$$- \sum_{k=\tau_t}^{t} E[\langle (V^{\theta_k}(s_k) - \phi(s_k)^T v_k)\nabla \log \pi_{\theta_k}(a_k|s_k), (L_k - L(\theta_k))\nabla \log \pi_{\theta_k}(a_k|s_k)]\rangle]$$

$$= \mathcal{O}(\sum_{k=\tau_t}^{t} \|\theta_{k+1} - \theta_k\|) + \mathcal{O}(\sum_{k=\tau_t}^{t} \|v_{k+1} - v_k\|) + E \sum_{k=\tau_t}^{t} (P_{k+1} - P_k)$$

$$= \mathcal{O}(\sum_{k=\tau_t}^{t} \alpha_k) + \mathcal{O}(\sum_{k=\tau_t}^{t} \beta_k) + E \sum_{k=\tau_t}^{t} (\alpha_k P_{k+1} - \alpha_k P_k)/\alpha_k$$

$$= \mathcal{O}(t^{1-\nu}) + E \sum_{k=\tau_t}^{t} (\alpha_{k+1} P_{k+1} - \alpha_k P_k)/\alpha_k + E \sum_{k=\tau_t}^{t} (\alpha_k - \alpha_{k+1})P_{k+1}/\alpha_k$$

$$= \mathcal{O}(t^{1-\nu}) + \mathcal{O}(t^{\nu}).$$

Hence, putting all these results back in Equation (22), we obtain,

$$I_8 \leq B\sqrt{\sum_{k=\tau_t}^{t} E\|M(\theta_k, v_k)\|^2}\sqrt{\sum_{k=\tau_t}^{t} E|L_k - L(\theta_k)|^2} + \mathcal{O}(\log^2 t \cdot t^{1-\nu}) + \mathcal{O}(t^{\nu}).$$

For term I_9, we have,

$$\sum_{k=\tau_t}^{t} M_L \alpha_k E[\|\delta_k \nabla \log \pi_{\theta_k}(a_k|s_k)\|^2] = \mathcal{O}(\sum_{k=\tau_t}^{t} \alpha_k)$$
$$= \mathcal{O}(t^{1-\nu}).$$

Hence after collecting all the terms, we obtain,

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

$$+ B\sqrt{\sum_{k=\tau_t}^{t} E\|M(\theta_k, v_k)\|^2}\sqrt{\sum_{k=\tau_t}^{t} E|L_k - L(\theta_k)|^2}.$$

After applying the squaring technique, we obtain,

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

$$\leq \mathcal{O}(t^\nu) + \mathcal{O}(\log^2 t \cdot t^{1-\nu}) + 4B^2 \frac{(G+U_w)^2}{(1 - \frac{c_\alpha}{c_\gamma}U_w B)^2} \frac{c_\alpha^2}{c_\gamma^2} \sum_{k=\tau_t}^{t} \mathbb{E}\|M(\theta_k, v_k)\|^2.$$

The last inequality follows from Equation (13).

Now if we select the values for c_α and c_γ such that $\dfrac{4B^2(G+U_w)^2}{(1 - \frac{c_\alpha}{c_\gamma}U_w B)^2}\dfrac{c_\alpha^2}{c_\gamma^2} < 1$, we shall obtain,

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

Dividing by $(1 + t - \tau_t)$ and assuming $t \geq 2\tau_t + 1$, we have,

$$\frac{1}{(1 + t - \tau_t)}\sum_{k=\tau_t}^{t} E\|M(\theta_k, v_k)\|^2 = \mathcal{O}(t^{\nu-1}) + \mathcal{O}(\log^2 t \cdot t^{-\nu}). \tag{23}$$

As seen earlier, the inequalities that need to be satisfied for the inequalities (13) and (23) to hold are the following:

$$\frac{c_\alpha}{c_\gamma} < \frac{1}{U_w B}, \tag{24}$$

$$\frac{2B(G+U_w)}{(1 - \frac{c_\alpha}{c_\gamma}U_w B)}\frac{c_\alpha}{c_\gamma} < 1. \tag{25}$$

Rearranging inequality (25), we get

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

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

$$\Rightarrow \frac{c_\alpha}{c_\gamma} < \frac{1}{2B(G+U_w) + U_w B}. \tag{26}$$

Now, from (24) and (26), we have,

$$\frac{c_\alpha}{c_\gamma} < \min\left(\frac{1}{2B(G+U_w) + U_w B}, \frac{1}{U_w B}\right).$$

Since $\dfrac{1}{2B(G+U_w) + U_w B} < \dfrac{1}{U_w B}$, we need to choose c_α and c_γ such that $\dfrac{c_\alpha}{c_\gamma} < \dfrac{1}{2B(G+U_w) + U_w B}$.

Now getting back to inequality (23), we can observe that $E\|M(\theta_k, v_k)\|^2 \to 0$ as $k \to \infty$. Finally, as the actor is on the faster timescale compared to the critic, the critic appears to be quasi-static from the viewpoint of the actor. Hence, we may let $v_t \equiv v, \forall t \geq 0$ (i.e., v independent of t). Therefore, the point of convergence of the actor parameter sequence θ_t will be $\theta(v)$ such that :

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

Convergence of the Critic   Recall that we have the following update rule for the critic:
$$v_{t+1} = v_t + \beta_t \delta_t \phi(s_t).$$

Notations:
$$
\begin{aligned}
O_t &:= (s_t, a_t, s_{t+1}) \\
z_t &:= v_t - v^*(\theta_t) \\
g(O_t, v_t, \theta_t) &:= (r_t - L(\theta_t) + \phi(s_{t+1})^\top v_t - \phi(s_t)^\top v_t)\phi(s_t) \\
\bar{g}(v_t, \theta_t) &:= E_{s \sim \mu_{\theta_t}, a \sim \pi_{\theta_t}, s' \sim p(.|s,a)}[(r(s,a) - L(\theta_t) + \phi(s')^\top v_t - \phi(s)^\top v_t)\phi(s)] \\
\bar{Q}(O_t, v_t, \theta_t) &:= \langle z_t, g(O_t, v_t, \theta_t) - \bar{g}(v_t, \theta_t) \rangle \\
\bar{U}(O_t, v_t, \theta_t) &:= (\nabla v_t^*)^T (r(s_t, a_t) - L(\theta_t) + \phi(s_{t+1})^\top v_t - \phi(s_t)^\top v_t)\nabla_\theta \log \pi_{\theta_t}(a_t|s_t) \\
\Psi(O_t, v_t, \theta_t) &:= \langle z_t, E_{\theta_t}[\bar{U}(O_t, v_t, \theta_t)] - \bar{U}(O_t, v_t, \theta_t) \rangle.
\end{aligned}
\tag{27}
$$

The proof of convergence of the critic is established through the following lemmas:

Lemma 5 ((Wu et al. 2022), Proposition 4.4). *There exists a constant $L_* > 0$ such that*
$$\|v^*(\theta_1) - v^*(\theta_2)\| \le L_*\|\theta_1 - \theta_2\|, \forall \theta_1, \theta_2 \in \mathbb{R}^d.$$

Lemma 6 ((Chen and Zhao 2023), Lemma B.4). *For any $\theta_1, \theta_2 \in \mathbb{R}^d$, we have*
$$\|\nabla v^*(\theta_1) - \nabla v^*(\theta_2)\| \le L_m\|\theta_1 - \theta_2\|,$$
where L_m is a positive constant.

Lemma 7. *Under assumptions 4.1, 4.3, 4.4, 4.6, for any $t \ge \tau > 0$, we have*

$$|E[\Psi(O_t, v_t, \theta_t)]| \le G_1\|\theta_t - \theta_{t-\tau}\| + G_2\|v_t - v_{t-\tau}\| + G_3 \sum_{i=t-\tau}^{t} E\|\theta_i - \theta_{t-\tau}\| + G_4 bk^{\tau-1},$$

where $G_1 > 0, G_2 > 0, G_3 > 0$ and $G_4 > 0$ are constants.

Proof. We can write $|E\Psi(O_t, v_t, \theta_t)|$ as follows:
$$
\begin{aligned}
&|E[\Psi(O_t, v_t, \theta_t)]| \\
&= |E[\Psi(O_t, v_t, \theta_t)] - E[\Psi(O_t, v_t, \theta_{t-\tau})] + E[\Psi(O_t, v_t, \theta_{t-\tau})] - E[\Psi(O_t, v_{t-\tau}, \theta_{t-\tau})] \\
&\quad + E[\Psi(O_t, v_{t-\tau}, \theta_{t-\tau})] - E[\Psi(\tilde{O}_t, v_{t-\tau}, \theta_{t-\tau})] + E[\Psi(\tilde{O}_t, v_{t-\tau}, \theta_{t-\tau})]| \\
&\le \underbrace{|E[\Psi(O_t, v_t, \theta_t)] - E[\Psi(O_t, v_t, \theta_{t-\tau})]|}_{I_a} + \underbrace{|E[\Psi(O_t, v_t, \theta_{t-\tau})] - E[\Psi(O_t, v_{t-\tau}, \theta_{t-\tau})]|}_{I_b} \\
&\quad + \underbrace{|E[\Psi(O_t, v_{t-\tau}, \theta_{t-\tau})] - E[\Psi(\tilde{O}_t, v_{t-\tau}, \theta_{t-\tau})]|}_{I_c} + \underbrace{|E[\Psi(\tilde{O}_t, v_{t-\tau}, \theta_{t-\tau})]|}_{I_d}
\end{aligned}
$$

In the above inequality, $\tilde{O}_t = (\tilde{s}_t, \tilde{a}_t, \tilde{s}_{t+1})$ is from the auxiliary Markov chain defined in Equation (11) and $O'_t = (s'_t, a'_t, s'_{t+1})$, where $s'_t \sim \mu_{\theta_{t-\tau}}, a'_t \sim \pi_{\theta_{t-\tau}}$ and $s'_{t+1} \sim P(.|s'_t, a'_t)$.

For term I_a, we have,
$$
|E[\Psi(O_t, v_t, \theta_t)] - E[\Psi(O_t, v_t, \theta_{t-\tau})]|
$$
$$
= |E[\langle z_t, E_{\theta_t}[\bar{U}(O_t, v_t, \theta_t)] - \bar{U}(O_t, v_t, \theta_t)\rangle] - E[\langle v_t - v^*(\theta_{t-\tau})), E_{\theta_{t-\tau}}[\bar{U}(O_t, v_t, \theta_{t-\tau}))] \tag{28}
$$
$$
- \bar{U}(O_t, v_t, \theta_{t-\tau}))\rangle]|
$$
$$
\le |E[\langle z_t, E_{\theta_t}[\bar{U}(O_t, v_t, \theta_t)] - \bar{U}(O_t, v_t, \theta_t)\rangle] - E[\langle v_t - v^*(\theta_{t-\tau}), E_{\theta_t}[\bar{U}(O_t, v_t, \theta_t)] \tag{29}
$$
$$
- \bar{U}(O_t, v_t, \theta_t)\rangle]|
$$
$$
+ |E[\langle v_t - v^*(\theta_{t-\tau}), E_{\theta_t}[\bar{U}(O_t, v_t, \theta_t)] - \bar{U}(O_t, v_t, \theta_t)\rangle] \tag{30}
$$
$$
- E[\langle v_t - v^*(\theta_{t-\tau})), E_{\theta_{t-\tau}}[\bar{U}(O_t, v_t, \theta_{t-\tau}))] - \bar{U}(O_t, v_t, \theta_{t-\tau}))\rangle]|
$$
$$
\le |E[\langle v^*(\theta_{t-\tau}) - v^*(\theta_t), E_{\theta_t}[\bar{U}(O_t, v_t, \theta_t)] - \bar{U}(O_t, v_t, \theta_t)\rangle]|
$$
$$
+ |E[\langle v_t - v^*(\theta_{t-\tau}), E_{\theta_t}[\bar{U}(O_t, v_t, \theta_t)] - E_{\theta_{t-\tau}}[\bar{U}(O_t, v_t, \theta_{t-\tau}))] - \bar{U}(O_t, v_t, \theta_t) \tag{31}
$$
$$
+ \bar{U}(O_t, v_t, \theta_{t-\tau}))\rangle]|
$$
$$
\le |E[\langle v^*(\theta_{t-\tau}) - v^*(\theta_t), E_{\theta_t}[\bar{U}(O_t, v_t, \theta_t)] - \bar{U}(O_t, v_t, \theta_t)\rangle]|
$$
$$
+ 2U_v|E[\|E_{\theta_t}[\bar{U}(O_t, v_t, \theta_t)] - E_{\theta_{t-\tau}}[\bar{U}(O_t, v_t, \theta_{t-\tau}))]\|] \tag{32}
$$
$$
+ \|\bar{U}(O_t, v_t, \theta_t) - \bar{U}(O_t, v_t, \theta_{t-\tau})\|]|. \tag{33}
$$

Now,

$$\|\bar{U}(O_t, v_t, \theta_t) - \bar{U}(O_t, v_t, \theta_{t-\tau})\| \leq A_1\|\theta_t - \theta_{t-\tau}\|,$$

where $A_1 > 0$ is some constant. This inequality follows from Lemma 6 as well as Lemma B.1 of (Wu et al. 2022). Also,

$$\|E_{\theta_t}[\bar{U}(O_t, v_t, \theta_t)] - E_{\theta_{t-\tau}}[\bar{U}(O_t, v_t, \theta_{t-\tau}))]\| \leq A_2\|\theta_t - \theta_{t-\tau}\|,$$

for some $A_2 > 0$. Lemma B.1 of (Wu et al. 2022) is used in obtaining the above inequality.
Hence putting these results back in inequality (33), we have,

$$
\begin{aligned}
I_a &\leq 2LG\|v^*(\theta_{t-\tau}) - v^*(\theta_t)\| + 2U_v(A_1 + A_2)\|\theta_t - \theta_{t-\tau}\| \\
&\leq (2LGL_* + 2U_v(A_1 + A_2))\|\theta_t - \theta_{t-\tau}\|.
\end{aligned}
$$

The last inequality is because of Lemma 5.
For term I_b, we have,

$$
\begin{aligned}
&|E[\Psi(O_t, v_t, \theta_{t-\tau})] - E[\Psi(O_t, v_{t-\tau}, \theta_{t-\tau})]| \\
&= |E[\langle v_t - v^*(\theta_{t-\tau}), E_{\theta_{t-\tau}}[\bar{U}(O_t, v_t, \theta_{t-\tau})] - \bar{U}(O_t, v_t, \theta_{t-\tau})\rangle] \\
&\quad - E[\langle v_{t-\tau} - v^*(\theta_{t-\tau}), E_{\theta_{t-\tau}}[\bar{U}(O_t, v_{t-\tau}, \theta_{t-\tau})] - \bar{U}(O_t, v_{t-\tau}, \theta_{t-\tau})\rangle]| \\
&= |E[\langle v_t - v^*(\theta_{t-\tau}), E_{\theta_{t-\tau}}[\bar{U}(O_t, v_t, \theta_{t-\tau})] - \bar{U}(O_t, v_t, \theta_{t-\tau})\rangle] \\
&\quad - E[\langle v_{t-\tau} - v^*(\theta_{t-\tau}), E_{\theta_{t-\tau}}[\bar{U}(O_t, v_t, \theta_{t-\tau})] - \bar{U}(O_t, v_t, \theta_{t-\tau})\rangle] \\
&\quad + E[\langle v_{t-\tau} - v^*(\theta_{t-\tau}), E_{\theta_{t-\tau}}[\bar{U}(O_t, v_t, \theta_{t-\tau})] - \bar{U}(O_t, v_t, \theta_{t-\tau})\rangle] \\
&\quad - E[\langle v_{t-\tau} - v^*(\theta_{t-\tau}), E_{\theta_{t-\tau}}[\bar{U}(O_t, v_{t-\tau}, \theta_{t-\tau})] - \bar{U}(O_t, v_{t-\tau}, \theta_{t-\tau})\rangle]| \\
&\leq |E[\langle v_t - v_{t-\tau}, E_{\theta_{t-\tau}}[\bar{U}(O_t, v_t, \theta_{t-\tau})] - \bar{U}(O_t, v_t, \theta_{t-\tau})\rangle]| \\
&\quad + |E[\langle v_{t-\tau} - v^*(\theta_{t-\tau}), E_{\theta_{t-\tau}}[\bar{U}(O_t, v_t, \theta_{t-\tau})] - E_{\theta_{t-\tau}}[\bar{U}(O_t, v_{t-\tau}, \theta_{t-\tau})] \\
&\quad - \bar{U}(O_t, v_t, \theta_{t-\tau}) + \bar{U}(O_t, v_{t-\tau}, \theta_{t-\tau})\rangle]| \\
&\leq 2L_*G\|v_t - v_{t-\tau}\| + 8BL_*U_v\|v_t - v_{t-\tau}\|.
\end{aligned}
$$

For term I_c, we have,

$$|E[\Psi(O_t, v_{t-\tau}, \theta_{t-\tau})] - E[\Psi(\tilde{O}_t, v_{t-\tau}, \theta_{t-\tau})]| \leq M_1 \sum_{i=t-\tau}^{t} E\|\theta_i - \theta_{t-\tau}\|,$$

for some $M_1 > 0$.

For term I_d, we have,

$$|E[\Psi(\tilde{O}_t, v_{t-\tau}, \theta_{t-\tau})]| \leq M_2 bk^{\tau-1},$$

for some $M_2 > 0$.
For an analysis of terms I_c and I_d, see Lemmas D.10 and D.11 in (Wu et al. 2022). Hence, after collecting all the terms, we have,

$$|E[\Psi(O_t, v_t, \theta_t)]| \leq G_1\|\theta_t - \theta_{t-\tau}\| + G_2\|v_t - v_{t-\tau}\| + G_3 \sum_{i=t-\tau}^{t} E\|\theta_i - \theta_{t-\tau}\| + G_4 bk^{\tau-1},$$

where $G_1 > 0, G_2 > 0, G_3 > 0$ and $G_4 > 0$ are constants. $\qquad \square$

Proof of convergence of critic From the critic update rule, we have,

$$
\begin{aligned}
\|z_{t+1}\|^2 &= \|v_{t+1} - v^*(\theta_{t+1})\|^2 \\
&= \|\Gamma(v_t + \beta_t \delta_t \phi(s_t)) - v^*(\theta_{t+1})\|^2 \\
&\leq \|v_t + \beta_t \delta_t \phi(s_t) - v^*(\theta_{t+1})\|^2 \\
&= \|z_t + \beta_t \delta_t \phi(s_t) + v^*(\theta_t) - v^*(\theta_{t+1})\|^2 \\
&\leq \|z_t\|^2 + 2\beta_t \langle z_t, \delta_t \phi(s_t)\rangle + 2\langle z_t, v^*(\theta_t) - v^*(\theta_{t+1})\rangle + 2\beta_t^2 \delta_t^2 \|\phi(s_t)\|^2 \\
&\quad + 2\|v^*(\theta_t) - v^*(\theta_{t+1})\|^2 \\
&= \|z_t\|^2 + 2\beta_t \langle z_t, \delta_t \phi(s_t) - E_{\theta_t}[\delta_t \phi(s_t)]\rangle + 2\beta_t \langle z_t, E_{\theta_t}[\delta_t \phi(s_t)]\rangle \\
&\quad + 2\langle z_t, v^*(\theta_t) - v^*(\theta_{t+1})\rangle + 2\beta_t^2 \delta_t^2 \|\phi(s_t)\|^2 + 2\|v^*(\theta_t) - v^*(\theta_{t+1})\|^2 \\
&\leq \|z_t\|^2 + 2\beta_t \langle z_t, \delta_t \phi(s_t) - E_{\theta_t}[\delta_t \phi(s_t)]\rangle - 2\beta_t \lambda \|z_t\|^2 + 2\langle z_t, v^*(\theta_t) - v^*(\theta_{t+1})\rangle \\
&\quad + 2\beta_t^2 \delta_t^2 \|\phi(s_t)\|^2 + 2\|v^*(\theta_t) - v^*(\theta_{t+1})\|^2.
\end{aligned}
$$

We assume here that the projection set C is large enough so that $v^*(\theta_{t+1})$ lies within the set. Also, C being both compact and convex guarantees that the point within C where the update with an increment is projected is not only the closest but also unique. The last inequality follows from Assumption 4.2. Here, $-\lambda = \sup_\theta \lambda_\theta$, where λ_θ is an upper bound on the largest eigenvalue of $\mathbf{A}$ defined in 1. After rearranging the terms we obtain,

$$
\begin{aligned}
\lambda \|z_t\|^2 &\leq \frac{1}{2\beta_t}(\|z_t\|^2 - \|z_{t+1}\|^2) + \langle z_t, \delta_t \phi(s_t) - E_{\theta_t}[\delta_t \phi(s_t)]\rangle + \frac{1}{\beta_t}\langle z_t, v^*(\theta_t) - v^*(\theta_{t+1})\rangle \\
&\quad + (\nabla v_t^*)^T(\theta_{t+1} - \theta_t)\rangle + \frac{1}{\beta_t}\langle z_t, (\nabla v_t^*)^T(\theta_t - \theta_{t+1})\rangle + \beta_t \delta_t^2 \|\phi(s_t)\|^2 \\
&\quad + \frac{1}{\beta_t}\|v^*(\theta_t) - v^*(\theta_{t+1})\|^2.
\end{aligned}
$$

Taking summation of terms from indices τ_t to t we have,

$$
\begin{aligned}
\lambda \sum_{k=\tau_t}^t E\|z_k\|^2 &\leq \underbrace{\sum_{k=\tau_t}^t \frac{1}{2\beta_k} E[\|z_k\|^2 - \|z_{k+1}\|^2]}_{I_1} + \underbrace{\sum_{k=\tau_t}^t E[\langle z_k, \delta_t \phi(s_k) - E_{\theta_k}[\delta_k \phi(s_k)]\rangle]}_{I_2} \\
&\quad + \underbrace{\sum_{k=\tau_t}^t \frac{1}{\beta_k} E\langle z_k, v^*(\theta_k) - v^*(\theta_{k+1}) + (\nabla v_k^*)^T(\theta_{k+1} - \theta_k)\rangle}_{I_3} \\
&\quad + \underbrace{\sum_{k=\tau_t}^t \frac{1}{\beta_k} E\langle z_k, (\nabla v_k^*)^T(\theta_k - \theta_{k+1})\rangle}_{I_4} + \underbrace{\sum_{k=\tau_t}^t \beta_k E[\delta_k^2 \|\phi(s_k)\|^2]}_{I_5} \\
&\quad + \underbrace{\sum_{k=\tau_t}^t \frac{1}{\beta_k} E\|v^*(\theta_k) - v^*(\theta_{k+1})\|^2}_{I_6}
\end{aligned}
$$

For term I_1, we have,

$$
\sum_{k=\tau_t}^t \frac{1}{2\beta_k} E[\|z_k\|^2 - \|z_{k+1}\|^2] = \mathcal{O}(t^\sigma)
$$

The analysis of I_1 is similar to that of the term I_1 in Section A. For term I_2 here, we have,

$$\sum_{k=\tau_t}^{t} E[\langle z_k, \delta_t \phi(s_k) - E_{\theta_k}[\delta_k \phi(s_k)]\rangle] = \mathcal{O}(\log^2 t \cdot t^{1-\nu}).$$

For a detailed analysis of the term I_2, see the analysis of term I_2 in (Chen and Zhao 2023).

For term I_3 above, we have,

$$\begin{aligned}
\sum_{k=\tau_t}^{t} \frac{1}{\beta_k} E\langle z_k, v^*(\theta_k) - v^*(\theta_{k+1}) + (\nabla v_k^*)^T(\theta_{k+1} - \theta_k)\rangle &\leq \frac{L_m}{2} \sum_{k=\tau_t}^{t} \frac{1}{\beta_k} E\|z_k\|\|\theta_{k+1} - \theta_k\|^2 \\
&= \mathcal{O}(\sum_{k=\tau_t}^{t} \frac{\alpha_k^2}{\beta_k}) \\
&= \mathcal{O}(t^{\sigma-2\nu+1}).
\end{aligned}$$

The above inequality follows from the L_m-smoothness of v^* in Lemma 6. For term I_4, we have,

$$\sum_{k=\tau_t}^{t} \frac{1}{\beta_k} E\langle z_k, (\nabla v_k^*)^T (\theta_k - \theta_{k+1})\rangle$$

$$= -\sum_{k=\tau_t}^{t} \frac{1}{\beta_k} E\langle z_k, (\nabla v_k^*)^T \alpha_k \delta_k \nabla_\theta \log \pi_{\theta_k}(a_k|s_k)\rangle$$

$$= -\sum_{k=\tau_t}^{t} \frac{1}{\beta_k} E\langle z_k, (\nabla v_k^*)^T \alpha_k (r(s_k, a_k) - L_k + \phi(s_{k+1})^\top v_k - \phi(s_k)^\top v_k)\nabla_\theta \log \pi_{\theta_k}(a_k|s_k)\rangle$$

$$= -\sum_{k=\tau_t}^{t} \frac{1}{\beta_k} E\langle z_k, (\nabla v_k^*)^T \alpha_k (r(s_k, a_k) - L(\theta_k) + \phi(s_{k+1})^\top v_k - \phi(s_k)^\top v_k)\nabla_\theta \log \pi_{\theta_k}(a_k|s_k)\rangle$$

$$- \sum_{k=\tau_t}^{t} \frac{1}{\beta_k} E\langle z_k, (\nabla v_k^*)^T \alpha_k (L(\theta_k) - L_k)\nabla_\theta \log \pi_{\theta_k}(a_k|s_k)\rangle$$

$$= -\sum_{k=\tau_t}^{t} \frac{\alpha_k}{\beta_k} E\langle z_k, (\nabla v_k^*)^T (r(s_k, a_k) - L(\theta_k) + \phi(s_{k+1})^\top v_k - \phi(s_k)^\top v_k)\nabla_\theta \log \pi_{\theta_k}(a_k|s_k)\rangle$$

$$+ \sum_{k=\tau_t}^{t} \frac{\alpha_k}{\beta_k} E\langle z_k, (\nabla v_k^*)^T E_{\theta_k}[(r(s_k, a_k) - L(\theta_k) + \phi(s_{k+1})^\top v_k - \phi(s_k)^\top v_k)\nabla_\theta \log \pi_{\theta_k}(a_k|s_k)]\rangle$$

$$- \sum_{k=\tau_t}^{t} \frac{\alpha_k}{\beta_k} E\langle z_k, (\nabla v_k^*)^T E_{\theta_k}[(r(s_k, a_k) - L(\theta_k) + \phi(s_{k+1})^\top v_k - \phi(s_k)^\top v_k)\nabla_\theta \log \pi_{\theta_k}(a_k|s_k)]\rangle$$

$$- \sum_{k=\tau_t}^{t} \frac{\alpha_k}{\beta_k} E\langle z_k, (\nabla v_k^*)^T (L(\theta_k) - L_k)\nabla_\theta \log \pi_{\theta_k}(a_k|s_k)\rangle$$

$$= \sum_{k=\tau_t}^{t} E[\frac{\alpha_k}{\beta_k} \Psi(O_k, v_k, \theta_k)]$$

$$- \sum_{k=\tau_t}^{t} \frac{\alpha_k}{\beta_k} E\langle z_k, (\nabla v_k^*)^T E_{\theta_k}[(r(s_k, a_k) - L(\theta_k) + \phi(s_{k+1})^\top v_k - \phi(s_k)^\top v_k)\nabla_\theta \log \pi_{\theta_k}(a_k|s_k)]\rangle$$

$$- \sum_{k=\tau_t}^{t} \frac{\alpha_k}{\beta_k} E\langle z_k, (\nabla v_k^*)^T (L(\theta_k) - L_k)\nabla_\theta \log \pi_{\theta_k}(a_k|s_k)\rangle$$

$$\leq \frac{c_\alpha}{c_\beta} \sum_{k=\tau_t}^{t} E[(1+k)^{\sigma-\nu}\Psi(O_k, v_k, \theta_k)] + L_* \sqrt{\sum_{k=\tau_t}^{t} E\|z_k\|^2} \sqrt{\sum_{k=\tau_t}^{t} E[\frac{\alpha_k^2}{\beta_k^2}\|M(\theta_k, v_k)\|^2]}$$

$$+ L_* B \sqrt{\sum_{k=\tau_t}^{t} E\|z_k\|^2} \sqrt{\sum_{k=\tau_t}^{t} E[\frac{\alpha_k^2}{\beta_k^2}(L(\theta_k) - L_k)^2]}$$

$$\leq \frac{c_\alpha}{c_\beta}(1+t)^{\sigma-\nu} \sum_{k=\tau_t}^{t} |E[\Psi(O_k, v_k, \theta_k)]| + L_* \sqrt{\sum_{k=\tau_t}^{t} E\|z_k\|^2} \sqrt{\sum_{k=\tau_t}^{t} E[\frac{\alpha_k^2}{\beta_k^2}\|M(\theta_k, v_k)\|^2]}$$

$$+ L_* B \sqrt{\sum_{k=\tau_t}^{t} E\|z_k\|^2} \sqrt{\sum_{k=\tau_t}^{t} E[\frac{\alpha_k^2}{\beta_k^2}(L(\theta_k) - L_k)^2]}$$

$$\leq \mathcal{O}(\log^2 t \cdot t^{\sigma-2\nu+1}) + L_* \sqrt{\sum_{k=\tau_t}^{t} E\|z_k\|^2} \sqrt{\sum_{k=\tau_t}^{t} E[\frac{\alpha_k^2}{\beta_k^2}\|M(\theta_k, v_k)\|^2]}$$

$$+ L_* B \sqrt{\sum_{k=\tau_t}^{t} E\|z_k\|^2} \sqrt{\sum_{k=\tau_t}^{t} E[\frac{\alpha_k^2}{\beta_k^2}(L(\theta_k) - L_k)^2]}.$$

The last inequality follows from Lemma 7.

For the term I_5, we have,

$$\sum_{k=\tau_t}^{t} \beta_k E[\delta_k^2 \|\phi(s_k)\|^2] = \mathcal{O}(t^{1-\sigma}).$$

Next, for the term I_6, we have,

$$\sum_{k=\tau_t}^{t} \frac{1}{\beta_k} E\|v^*(\theta_k) - v^*(\theta_{k+1})\|^2 = \mathcal{O}(t^{1-2\nu+\sigma}).$$

For detailed analysis of terms I_5 and I_6, see section C.2 of (Chen and Zhao 2023). Thus, after collecting all the terms we have,

$$\lambda \sum_{k=\tau_t}^{t} E\|z_k\|^2 \leq \mathcal{O}(t^\sigma) + \mathcal{O}(\log^2 t \cdot t^{1-\nu}) + \mathcal{O}(\log^2 t \cdot t^{\sigma-2\nu+1})$$

$$+ L_* \sqrt{\sum_{k=\tau_t}^{t} E\|z_k\|^2} \sqrt{\sum_{k=\tau_t}^{t} E[\frac{\alpha_k^2}{\beta_k^2} M(\theta_k, v_k)^2]}$$

$$+ L_* B \sqrt{\sum_{k=\tau_t}^{t} E\|z_k\|^2} \sqrt{\sum_{k=\tau_t}^{t} E[\frac{\alpha_k^2}{\beta_k^2}(L(\theta_k) - L_k)^2]}$$

$$+ \mathcal{O}(t^{1-\sigma}).$$

After applying the squaring technique, we have,

$$\sum_{k=\tau_t}^{t} E\|z_k\|^2 = \mathcal{O}(t^\sigma) + \mathcal{O}(\log^2 t \cdot t^{1-\nu}) + \mathcal{O}(\log^2 t \cdot t^{1+\sigma-2\nu})$$

$$+ \mathcal{O}(\sum_{k=\tau_t}^{t} E[\frac{\alpha_k^2}{\beta_k^2} M(\theta_k, v_k)^2]) + \mathcal{O}(\sum_{k=\tau_t}^{t} E[\frac{\alpha_k^2}{\beta_k^2}(L(\theta_k) - L_k)^2])$$

$$= \mathcal{O}(t^\sigma) + \mathcal{O}(\log^2 t \cdot t^{1-\nu}) + \mathcal{O}(\log^2 t \cdot t^{1+\sigma-2\nu}) + \mathcal{O}(t^{2\sigma-\nu})$$

$$+ \mathcal{O}(\log^2 t \cdot t^{1-3\nu+2\sigma})$$

$$= \mathcal{O}(\log^2 t \cdot t^{1+\sigma-2\nu}) + \mathcal{O}(t^{2\sigma-\nu}) + \mathcal{O}(\log^2 t \cdot t^{1-3\nu+2\sigma}).$$

The second equality above comes from the result of Theorems A and 4.8. Assuming $t \geq 2\tau_t - 1$, we have,

$$\frac{1}{1+t-\tau_t} \sum_{k=\tau_t}^{t} E\|z_k\|^2 = \mathcal{O}(\log^2 t \cdot t^{\sigma-2\nu}) + \mathcal{O}(t^{2\sigma-\nu-1}) + \mathcal{O}(\log^2 t \cdot t^{-3\nu+2\sigma}).$$

So, we can observe that $E\|z_t\|^2 \to 0$ as $t \to \infty$, if the following conditions are satisfied:

$$2\sigma - \nu < 1,$$
$$2\sigma < 3\nu.$$

Hence $(v_t - v^*(\theta_t)) \to 0$ as $t \to \infty$.

Now as actor is on the faster timescale compared to critic, we can say that when observed from the timescale of critic, θ_t closely tracks $\theta(v_t)$. Therefore we can conclude that v_t converges to a point w such that $w - v^*(\theta(w)) = 0$.

Optimising over the values of ν and σ 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 :-

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

Now,

$$2\beta > 0$$
$$\Rightarrow 2\beta - 0.5 > -0.5$$
$$\Rightarrow \frac{1}{2\beta - 0.5} < -2$$

So we can write $\frac{1}{2\beta - 0.5} = -2 - \delta$ where $\delta > 0$ is arbitrarily close to zero as $\beta > 0$ is made arbitrarily close to zero.

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.

Remark A.2. It is important to note here that unlike (Wu et al. 2022), none of our results have terms corresponding to function approximation errors as we do not encounter the term $\Delta h^{'}(O, \theta)$ in our analysis. For the definition of $\Delta h^{'}(O, \theta)$ please refer the proof of Theorem 4.5 of (Wu et al. 2022).

Asymptotic Analysis

We now analyse Algorithm 1 which represents the two-timescale CA algorithm involving linear function approximation for its asymptotic convergence. To begin with, we first present the asymptotic analysis of (almost sure) convergence by using two projection operators Γ_1 and Γ_2 respectively. Let C and D be compact subsets of $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, respectively. Then, $\Gamma_1 : \mathbb{R}^{d_1} \to C$ and $\Gamma_2 : \mathbb{R}^{d} \to D$ are the two projection operators. These operators ensure that the algorithm remains stable throughout the run.

Subsequently, in Section A, we remove the projection operator Γ_1 and consider an un-projected critic that can take values in the whole of $\mathbb{R}^{d_1}$. We also prove the stability of the recursion in this case in addition to asymptotic convergence. In other words, we show that $\sup_t \|v_t\| < \infty$ w.p.1. Such results on asymptotic stability and almost sure convergence are not available in many non-asymptotic (finite-time) analyses of algorithms in the literature. The resulting scheme that we analyse is then similar to AC algorithms in the literature, where one projects the actor but not the critic, see for instance, (Bhatnagar et al. 2009), except that now the time scales of the 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 ordinary differential equation (ODE) approach to stochastic approximation recursions and is necessitated because we allow for multiple local maxima for the actor-recursion for any given critic update that result in an underlying DI instead of an ODE. In the context of AC or CA algorithms, ours is the first analysis that incorporates this level of sophistication and generality.

As noted in Assumption 5.2, 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 Case of Projected Critic   To begin with, we consider the case where the critic recursion is projected using a projection operator $\Gamma_1(\cdot)$ to a compact and convex set $C \subset \mathbb{R}^{d_1}$. Thus, for any $x \in \mathbb{R}^{d_1}$, $\Gamma_1(x) \in C$. For any vector $y \in C$, we have $\|y\| \leq U_\theta$, where $U_\theta > 0$ is a constant. As mentioned earlier, the single-stage reward is a function of the current state and action taken.

Asymptotic Convergence of average reward estimate and the actor   We begin here with the case where both critic and actor recursions are projected to certain compact and convex sets. In this case, we rewrite the critic recursion as follows:

$$\begin{align}
v_{t+1} &= \Gamma_1(v_t + \beta_t \delta_t \phi(s_t)), \tag{34}\\
&= \Gamma_1\left(v_t + \alpha_t \left(\frac{\beta_t}{\alpha_t}\right) \delta_t \phi(s_t)\right), \tag{35}
\end{align}$$

where

$$\delta_t = r_t - L_t + \phi(s_{t+1})^\top v_t - \phi(s_t)^\top v_t. \tag{36}$$

Observe now that since $\beta_t = o(\alpha_t)$, the above asymptotically tracks the ordinary differential equation (ODE)

$$\dot{v}(t) = 0,$$

which means we may let $v(t) \equiv v$ (i.e., independent of t). Now consider the average reward recursion, viz.,

$$L_{t+1} = L_t + \gamma_t(r_t - L_t), \tag{37}$$

where $\gamma_t = K\alpha_t$. For simplicity, let $K = 1$. In other words, the timescales of recursions governed by γ_t and α_t are the same and faster while the recursion governed by β_t is on the slower timescale. Recall also the (faster) θ-update rule here:

$$\theta_{t+1} = \Gamma_2(\theta_t + \alpha_t \delta_t \psi_{s_t a_t}), \tag{38}$$

where $\psi_{s_t a_t} = \nabla_\theta \log \pi_{\theta_t}(a_t|s_t)$.

Consider now the following system of ODEs associated with all the above recursions:

$$\dot{v} = 0, \tag{39}$$

$$\dot{L} = -L + \sum_s \mu_\theta(s) \sum_a \pi^\theta(a|s)R(s,a), \tag{40}$$

$$\dot{\theta} = \hat{\Gamma}_2 \left(E[\delta_t \nabla_\theta \log \pi^\theta(a_t|s_t)|\theta] \right), \tag{41}$$

where the operator $\hat{\Gamma}_2(\cdot)$ is defined as

$$\hat{\Gamma}_2(v(y)) = \lim_{0 < \eta \to \infty} \left(\frac{\Gamma_2(y + \eta v(y)) - y}{\eta} \right).$$

Here, as before, $\mu_\theta(s), s \in S$ denotes the stationary distribution of the state-valued Markov chain $\{s_t\}$ when actions in the MDP are chosen as per the parameterized policy π^θ.

The system of ODEs (39)-(41) has the following set of equilibria:

$$\{(v, L^\theta, \theta)|\theta \in \theta^*(v), v \in \mathbb{R}^{d_1}\},$$

where for any $v \in \mathbb{R}^{d_1}, \theta \in \theta^*(v)$,

$$L^\theta = \sum_s \mu_\theta(s) \sum_a \pi^\theta(s,a)R(s,a), \tag{42}$$

is the unique globally asymptotically stable equilibrium of the ODE (40). Note here that for any $v \in \mathbb{R}^{d_1}, \theta^*(v)$ is the set of asymptotically stable equilibria of the ODE (41). Here $R(s,a)$ is the expected single-stage reward when state is s and action a is chosen. Thus, $R(s,a) = E[r_t|s_t = s, a_t = a]$.

By Assumption 4.4, π^θ is continuously differentiable. Let

$$\check{p}_\theta(s,s') \triangleq \sum_{a \in A(s)} \pi^\theta(s,a)p(s,a,s').$$

Then, from Assumption 4.4, $\check{p}_\theta(s,s')$ are continuously differentiable (in θ) transition probabilities of the resulting Markov chain, $\forall s, s' \in S$. We have the following preliminary result:

Lemma 8. *Under Assumptions 4.4 and 5.1, L^θ, $\theta \in C$ is continuously differentiable in θ.*

Proof. By Theorem A.1, ν_θ is continuously differentiable in θ. Recall that Assumption 4.4 has been used in that proof. Now since the set of states and actions is finite, from the definition of L^θ in (42), it is easy to see that L^θ is continuously differentiable as well. $\square$

Notice that the equilibrium of (41) depends on v since δ_t depends on v. Also, in general, for any given v, $\theta^*(v)$ will not be a unique point. In fact, this is the set of local minima of the associated performance objective under function approximation.

For $\theta \in \theta^*(v)$, let

$$\delta^\theta = r_t - L^\theta + \phi(s_{t+1})^T v - \phi(s_t)^T v.$$

Lemma 9. *Under Assumptions 4.4 and 5.1, for $\theta \in \theta^*(v)$, we have*

$$E[\delta^\theta \psi_{s_t a_t}|\theta] = \nabla_\theta L^\theta + \sum_s \mu_\theta(s)\nabla \bar{V}^{\pi^\theta}(s),$$

where

$$\bar{V}^{\pi^\theta}(s) = \sum_a \pi^\theta(s,a) \left[R(s,a) - L^{\pi^\theta} + \sum_{s'} p(s,a,s') v^T \phi(s') \right].$$

Proof. Note that for $\theta \in \theta^*(v)$,

$$\nabla_\theta \bar{V}^{\pi^\theta}(s) = \sum_a \pi^\theta(s,a)(-\nabla_\theta L^{\pi^\theta})$$

$$+ \sum_a \nabla_\theta \pi^\theta(s,a) \left(R(s,a) - L^{\pi^\theta} + \sum_{s'} p(s,a,s') v^T \phi(s') \right).$$

Hence,

$$\sum_s \mu_\theta(s) \nabla_\theta \bar{V}^{\pi^\theta}(s) = -\sum_s \mu_\theta(s) \sum_a \pi^\theta(s,a) \nabla_\theta L^{\pi^\theta}$$

$$+ \sum_s \mu_\theta(s) \sum_a \nabla_\theta \pi^\theta(s,a) \left(R(s,a) - L^{\pi^\theta} + \sum_{s'} p(s,a,s') v^T \phi(s') \right).$$

$$= -\nabla_\theta L^{\pi^\theta} + E[\delta^\theta \psi_{s_t a_t} | \theta].$$

The claim now follows by rearranging the terms above. $\qquad \square$

Remark A.3. Lemma 9 is similar to Lemma 4 of (Bhatnagar et al. 2009). The main difference between these two results is that since in (Bhatnagar et al. 2009), the critic is faster than the actor, the actor parameter θ is held fixed while analysing the critic recursion. As a result the critic parameter v there depends on θ unlike here where the critic parameter v is held fixed. As a result, there is an additional term that appears on the RHS of Lemma 4 of (Bhatnagar et al. 2009) contributing to the bias in the estimator that is not present in Lemma 9. Note also that Lemma 9 tells us that for $\theta \in \theta^*(v)$, $\delta^\theta \psi_{s_t a_t}$ is a biased gradient estimator of L^θ with $\sum_s \mu_\theta(s) \nabla \bar{V}^{\pi^\theta}(s)$ as the bias term.

We now proceed with the analysis of the faster recursions.

Proposition A.4. *Under Assumptions 4.4, 5.1 and 5.2, for any given $v \in \mathbb{R}^{d_1}$, and L_t, θ_t updated as in (37)-(38), respectively, we have $L_t \to L^\theta$ and $\theta_t \to \theta$, $\theta \in \theta^*(v)$, where*

$$L^\theta = \sum_s \mu_\theta(s) \sum_a \pi^\theta(s,a) R(s,a),$$

and $\theta^*(v)$ is the set of points θ where $\hat{\Gamma}(E[\delta^\theta \psi_{s_t a_t} | \theta]) = 0$.

Proof. As described previously, the ODEs associated with (37)-(38) are (40)-(41). Consider (40) first and let

$$f(L) = -L + \sum_s \mu_\theta(s) \sum_a \pi^\theta(a|s) R(s,a),$$

i.e., the RHS of (40). For any integer $c > 0$, let $f_c(L) \triangleq f(cL)/c$. Since $\max_{s,a} |R(s,a)| < \infty$ as single-stage rewards are finite, $f_c(L) \to f_\infty(L) = -L$ as $c \to \infty$. The ODE

$$\dot{L} = f_\infty(L) = -L$$

is clearly globally asymptotically stable to the origin. Thus, from Chapter 6 of (Borkar 2023), it follows that $\sup_t |L_t| < \infty$ and that $L_t \to L^\theta$ almost surely as $t \to \infty$.

Recall now the θ-update (38):

$$\theta_{t+1} = \Gamma_2(\theta_t + \alpha_t \delta_t \psi_{s_t a_t}).$$

We rewrite the same as

$$\theta_{t+1} = \Gamma_2(\theta_t + \alpha_t E[\delta^{\theta^*(v)} \psi_{s_t a_t} | \mathcal{F}_1(t)] + \alpha_t \gamma_1(t) + \alpha_t \gamma_2(t)),$$

where $\gamma_1(t) = \delta_t \psi_{s_t a_t} - E[\delta_t \psi_{s_t a_t} | \mathcal{F}_1(t)]$ and (for $\theta \in \theta^*(v)$), $\gamma_2(t) = E[(\delta_t \psi_{s_t a_t} - \delta^\theta \psi_{s_t a_t}) | \mathcal{F}_1(t)]$, respectively, and with $\mathcal{F}_1(t) = \sigma(\theta_k, s_k, a_k, k \leq t), t \geq 0$. Observe that $(\gamma_1(t), \mathcal{F}_1(t)), t \geq 0$, is a martingale difference sequence and is uniformly bounded for any given $v \in \mathbb{R}^{d_1}$. This is because rewards are bounded and so is the L_t-sequence as argued above. Moreover, since state and action spaces are finite, we have that $\max_{s,a} |\psi_{sa}| \leq \check{K} < \infty$, for some constant $\check{K} < \infty$. If we now form the sequence

$$Z(t) = \sum_{k=0}^{t-1} \alpha_k \gamma_1(k), \ t \geq 0,$$

with $Z(0) = 0$, then it is easy to see that $(Z(t), \mathcal{F}_1(t)), t \geq 0$, forms a martingale. Moreover, from the square summability of the step-size sequence $\{\alpha_k\}$ and the fact that $\gamma_1(t)$ is uniformly bounded (see above), it can be shown from a routine application of the martingale convergence theorem for square integrable martingales (cf. (Borkar 2012)), that $\sup_t |Z(t)| < \infty$ with probability one, or in other words, the martingale converges almost surely. Now observe that $\gamma_2(t) \to 0$ as $t \to \infty$ a.s. by Assumption 4.3.

Now (for $\theta \in \theta^*(v)$), let

$$e^{\pi^\theta} \triangleq \sum_s \mu_\theta(s) \nabla \bar{V}^{\pi^\theta}(s).$$

Consider the ODE:

$$\dot{\theta} = \hat{\Gamma}_2 \left(\nabla L^\theta + e^{\pi^\theta} \right). \tag{43}$$

It follows from Theorem 5.3.1 of (Kushner and Clark 2012) that $\theta_t \to \theta \in \theta^*(v)$ almost surely as $t \to \infty$. The claim follows. $\qquad \square$

Remark A.5. The claim in Proposition A.4 indicates that the recursion (38) converges to the stationary points of the ODE (43). One can now argue as in (Bhatnagar et al. 2009), that the convergence is indeed to local maxima of the ODE (43) which are also stationary points of the above ODE.

Asymptotic Convergence of the Critic   We now turn to the analysis of the slower (critic) recursion (34). Similar to the actor recursion, we define the following projection operator that will be used for the critic ODE, where the operator $\hat{\Gamma}_2(\cdot)$ is defined as

$$\hat{\Gamma}_1(w(x)) = \lim_{0 < \beta \to \infty} \left(\frac{\Gamma_1(x + \beta w(x)) - x}{\beta} \right).$$

Let

$$Y_t = \delta_t \phi(s_t) - E[\delta_t \phi(s_t)|\mathcal{F}_2(t)],$$

where $\mathcal{F}_2(t) = \sigma(v_r, \theta_r, s_r, a_r, r \leq t), t \geq 0$, is a sequence of associated sigma fields. Then $(Y_t, \mathcal{F}_2(t)), t \geq 0$ is a martingale difference sequence. Consider now the associated process

$$\check{Z}(t) = \sum_{m=0}^{t-1} \alpha_m Y_m,$$

$t \geq 0$ with $\check{Z}(0) = 0$. Then $(\check{Z}(t), \mathcal{F}_2(t)), t \geq 0$ is a martingale sequence. Now because of the fact that single-stage rewards are uniformly bounded, the set of states and actions is finite and hence the state features are uniformly bounded, the fact that $v_n \in C$, $\forall n$, a compact set, and $\sum_m \alpha_m^2 < \infty$, it follows from the martingale convergence theorem for square integrable martingales (cf. Chapter 3 of (Borkar 2012)) that $\check{Z}(t), t \geq 0$ is an almost surely convergent martingale sequence.

We rewrite (34) as

$$v_{t+1} = \Gamma_1(v_t + \beta_t(y_t + Y_t + \kappa_t)), \tag{44}$$

where

$$y_t = \sum_s \mu_{\theta_t}(s) \sum_a \pi^{\theta_t}(s, a)(R(s, a) - L^{\theta_t} + v_t^T \sum_{s'} p(s, a, s')\phi(s') - v_t^T \phi(s))\phi(s),$$

$$\kappa_t = E[(R(s_t, a_t) - L^{\theta_t} + v_t^T \sum_{s_{t+1}} p(s_t, a_t, s_{t+1})\phi(s_{t+1}) - v_t^T \phi(s_t))\phi(s_t)|\mathcal{F}_2(t)] - y_t,$$

respectively, with $\theta_t \in \theta^*(v)$ (cf. Remark A.8). Now from Assumption 4.3, it follows that $\kappa_t \to 0$ almost surely as $t \to \infty$. Proceeding as in Chapter 6 of (Borkar 2023), one can rewrite (44) as

$$v_{t+1} = v_t + \beta_t \left(\frac{\Gamma_1(v_t + \beta_t(y_t + Y_t + \kappa_t)) - v_t}{\beta_t} \right) \tag{45}$$

$$= v_t + \beta_t(\check{\gamma}_1(v_t; y_t + Y_t) + \xi_t, \tag{46}$$

where

$$\check{\gamma}_1(v; y) = \lim_{\eta \to 0} \left(\frac{\Gamma_1(v + \eta y) - v}{\eta} \right),$$

is the directional derivative of Γ_1 at v in the direction y. Note that in (46) and in what follows, we drop κ_t since as explained before, $\kappa_t \to 0$ a.s. as $t \to \infty$. Thus, in the limiting system, this term will not appear. From the definition of $\check{\gamma}_1(v, y)$, note that if $v \in C^o$ (i.e., the interior of C), then for any $y \in \mathbb{R}^{d_1}$, $\Gamma_1(v + \eta y) = v + \eta y$ for $\eta > 0$ 'small enough'. Thus, $\check{\gamma}_1(v; y) = y$ if $v \in C^o$.

Suppose now

$$z_t \triangleq E[\check{\gamma}_1(v_t; y_t + Y_t)|\mathcal{F}_2(t)]$$

and

$$\check{Y}_t \triangleq \check{\gamma}_1(v_t; y_t + Y_t) - z_t.$$

Thus, (46) can be rewritten as

$$v_{t+1} = v_t + \beta_t(z_t + \check{Y}_t + \xi_t). \qquad (47)$$

Let $h(v)$ denote the set-valued map

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

$$-v^T\phi(s))\phi(s)|\theta \in \bar{\theta}^*(v)\},$$

where $\bar{\theta}^*(v)$ is the closure of the set $\theta^*(v)$. Now since $\mu_\theta(s)$, $\pi^\theta(s,a)$, L^θ are all continuous functions of θ and $\bar{\theta}^*(v)$ is a compact set for any v, it follows that $h(v)$ is a compact set.

Now from the definition of Y_t, it can be seen that

$$Y_t \in \{v_t^T(\phi(s_{t+1}) - \sum_{s',a_t} \pi^\theta(s_t,a_t)p(s_t,a_t,s')\phi(s'))\phi(s_t)|\theta \in \bar{\theta}^*(v)\}.$$

Since $\{\sum_{s',a_t} \pi^\theta(s_t,a_t)p(s_t,a_t,s')|\theta \in \bar{\theta}^*(v)\}$ is compact for every v and s_t takes only finitely many values, the conditional distribution of Y_t given $\mathcal{F}_2(t)$ has a compact support $\mathcal{B}(v_t)$.

Let

$$\hat{\Gamma}_v(h(v)) \triangleq \cap_{\epsilon>0}\bar{co}\left(\cup_{\|w-v\|<\epsilon}\{\gamma_1(w;y+Y)|y \in h(v), Y \in \mathcal{B}(v)\}\right),$$

where $\bar{co}(\cdot)$ is the closed convex hull of '$\cdot$'. We now have the following result.

Lemma 10. *We have*

(i) $\hat{\Gamma}_v(h(v))$ *is a compact and convex set for any $v \in C$.*
(ii) *For all $v \in C$,*

$$\sup_{w \in \hat{\Gamma}_v(h(v))} \|w\| \leq M(1 + \|w\|),$$

for some $M > 0$.
(iii) $\hat{\Gamma}_v(h(v))$ *is upper semi-continuous, i.e., if $v_n \to v$ and $w_n \to w$ with $w_n \in \hat{\Gamma}_{v_n}(h(v_n))$ for all n, then $w \in \hat{\Gamma}_v(h(v))$.*

Proof. (i) By definition $\hat{\Gamma}_v(h(v))$ is convex and closed. It is also easily seen to be bounded and hence is compact.
(ii) Let

$$Q \triangleq \sup_{v \in C} \sup_{w \in \hat{\Gamma}_v(h(v))} \|w\|$$

Then since C is compact and $\hat{\Gamma}_v(h(v))$ is compact as well from the foregoing, $0 < Q < \infty$. The claim then follows by letting $M = Q$.
(iii) Let

$$g(v) \triangleq \cup\{\gamma_1(w;y+Y)|y \in h(v), Y \in \mathcal{B}(v)\}.$$

Then,

$$\hat{\Gamma}_v(h(v)) = \cap_{\epsilon>0}\bar{co}(\{g(v)|\|w-v\|<\epsilon\}).$$

Note now that the family of sets $H(v,\epsilon) \triangleq \bar{co}(\{g(v)|\|w-v\|<\epsilon\})$ with $\epsilon > 0$ is a family of diminishing sets in ϵ and $H(v,\epsilon) \downarrow \hat{\Gamma}_v(h(v))$ as $\epsilon \downarrow 0$.

Let $w_n \in \hat{\Gamma}_{v_n}(h(v_n))$ for n large enough so that $\|v_n - v\| < \epsilon/3$ for some $\epsilon > 0$. It is easy to find such a v_n since $v_n \to v$ as $n \to \infty$. Then, $w_n \in H(v_n, \frac{\epsilon}{3})$. It is then easy to see that $H(v_n, \frac{\epsilon}{3}) \subset H(v, \frac{2\epsilon}{3})$ for n large. Thus, given ϵ, there exists $N_0 > 0$ such that for all $n > N_0$, $w_n \in H(v, \frac{2\epsilon}{3})$. Further, $w_n \to w$ as $n \to \infty$. Thus, $w \in H(v, \frac{2\epsilon}{3})$ since it is a closed set. The claim now follows since $H(v, \frac{2\epsilon}{3}) \downarrow \hat{\Gamma}_v(h(v))$ as $\epsilon \downarrow 0$ implying that $w \in \hat{\Gamma}_v(h(v))$.
Thus, all three claims are verified. $\square$

Remark A.6. Any set-valued map $H : \mathcal{R}^l \to \{$ subsets of $\mathcal{R}^m$ and satisfying the three claims in Lemma 10 is called a Marchaud or Peano map, cf. (Aubin and Frankowska 2009). If such a map is used for the driving vector field of a differential inclusion, the trajectories of the inclusion are absolutely continuous functions.

Define now a sequence of time points $\{t(n)\}$, $n \geq 0$ in the following manner: $t(0) = 0$ and for $n \geq 1$, $t(n) = \sum_{\tau=0}^{n-1} \beta_\tau$. We now define a process $\bar{v}(t)$ obtained from the iterate sequence $\{v_n\}$ as given by the recursion (34) as follows: $v(t(n)) = v_n$, $\forall n$ and for $t \in [t(n), t(n+1)]$, $\bar{v}(t)$ is obtained as

$$\bar{v}(t) = \left(\frac{t(n+1) - t}{\beta_n}\right) v_n + \left(\frac{t - t(n)}{\beta_n}\right) v_{n+1}.$$

Let

$$G = \cap_{t \geq 0} \overline{\{\bar{v}(t+s) | s \geq 0\}}. \tag{48}$$

Consider now the differential inclusion (DI):

$$\dot{v}(t) \in \hat{\Gamma}_v(h(v(t))). \tag{49}$$

Any solution to (49) as mentioned in Remark A.6, is guaranteed to be absolutely continuous and satisfy (49) almost everywhere.

We recall Definition II of (Benaïm, Hofbauer, and Sorin 2005). By a perturbed solution $\mathbf{y} : [0, \infty) \to \mathbb{R}^{d_1}$ to the DI (49), we mean that $\mathbf{y}$ is absolutely continuous. In addition, there exists a locally integrable function $t \mapsto U(t)$ such that

$$\lim_{t \to \infty} \sup_{0 \leq v \leq T} \| \int_t^{t+v} U(s) ds \| = 0.$$

Also, $\mathbf{y}$ and U together satisfy the DI

$$\dot{\mathbf{y}}(t) - U(t) \in \hat{\Gamma}_v^{\delta(t)}(h(v(t))),$$

for almost every $t > 0$ and where $\delta(t) \to 0$ as $t \to \infty$. Here,

$$\hat{\Gamma}_v^{\delta(t)}(h(v(t))) = \{y \in \mathbb{R}^{d_1} | \exists z : \|z - x\| < \delta, \ d(y, \hat{\Gamma}_v^{\delta(t)}(h(z))) < \delta\}.$$

We have the following result on the convergence of (34):

Theorem A.7. *Under Assumptions 4.4, 4.2, 5.1 and 5.2, the iterates $\{v_t, t \geq 0\}$ obtained according to (34) satisfy $v_n \to G$ almost surely, where G is as in (48). Further, G is a closed connected internally chain recurrent invariant set of the DI (49).*

Proof. Note again that as a consequence of the projection operator Γ_1, $\sup_n \|v_n\| < \infty$ w.p.1. From the foregoing, the process $\bar{v}(t)$ can be shown as in Proposition 1.3 of (Benaïm, Hofbauer, and Sorin 2005) to be a bounded perturbation of the DI. The claim now follows from Theorem 3.6 and Lemma 3.8 of (Benaïm, Hofbauer, and Sorin 2005). $\square$

Remark A.8. Recall that $\theta^*(v)$ is the set of stable equilibria of the ODE (43. Consider now another associated ODE for the faster (actor) recursion:

$$\dot{\theta} = \hat{\Gamma}_2(\nabla L^\theta). \tag{50}$$

Let the set of asymptotically stable equilibria (i.e., the maxima of L^θ) correspond to $\check{\theta}(v)$. It follows as a consequence of Proposition A.4 and the continuity of ∇L^θ and e^{π^θ} that given $\epsilon > 0$, there exists $\delta > 0$ such that if $\|e^{\pi^\theta}\| < \delta$, then $\theta_t \to \check{\theta}(v)^\epsilon \triangleq \{\theta | \|\theta - \theta_0\| < \epsilon, \theta_0 \in \check{\theta}(v)\}$.

As a consequence of Remark A.8, we may let $h(v)$ be defined as

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

$$-v^T \phi(s))\phi(s) | \theta \in \bar{\check{\theta}}(v)^\epsilon\},$$

where $\bar{\check{\theta}}(v)^\epsilon$ is the closure of $\check{\theta}(v)^\epsilon$.

Let $R^{\pi^\theta} = (\sum_a \pi^\theta(s, a)R(s, a), s \in S)^T$ be a column vector of the size of the state space. Also, recall (see Lemma 8) that P_θ denotes the transition probability matrix under policy π^θ. For $J \in \mathbb{R}^{|S|}$, let $T_\theta : \mathbb{R}^{|S|} \to \mathbb{R}^{|S|}$ be the Bellman operator under policy π^θ defined according to

$$T_\theta(J) = R^{\pi^\theta} - L^\theta e + P_\theta J,$$

where $e \in \mathbb{R}^{|S|}$ is the unit vector with all entries one. Let Φ denote the $|S| \times d_1$ feature matrix with d_1-dimensional rows $\phi(s)^T$, $s \in S$. Then, in vector-matrix notation, we have

$$h(v) = \Phi^T D^\theta (T(\Phi v) - \Phi v).$$

Notice that $h(v)$ is a set-valued map for any v and v_0 will be an equilibrium for the DI (49) if $0 \in \hat{\Gamma}_v(h(v_0))$. Now as per Remark A.8, if $\|e^{\pi^\theta}\| < \delta$, then $\theta \in N^\epsilon(\bar{\check{\theta}}(v))$, an ϵ-neighborhood of the set of local maxima of the function L^θ. We now have the following useful result:

Theorem A.9. *Consider a solution $v(\cdot)$ to the differential inclusion (49). Suppose $\lim_{t\to\infty} v(t) = \hat{v}$. Then $\hat{v}$ is an equilibrium of the DI (49), i.e., $0 \in \hat{\Gamma}_v(h(\hat{v}))$.*

Proof. Recall that by Lemma 10, the set-valued map $\hat{\Gamma}_v(h(v))$ is a Peano map. The claim now follows by Theorem 10.1.12 of (Aubin and Frankowska 2009). $\qquad\square$

Theorem A.10 gives the main result. For any function f, we say that it's local maxima are isolated if around each such maximum, one can construct an open ball of small enough radius so that any two balls do not intersect.

Theorem A.10. *Suppose the ODE (43) has isolated local maxima θ^*. Correspondingly suppose $v^* \in C^o$ (the interior of C) is a limit point of the solution to the DI (49). Then under Assumptions 4.4, 4.2, 5.1 and 5.2, $\{(v_t, \theta_t)\}$ governed according to (34)-(38) satisfy $(v_t, \theta_t) \to (v^*, \theta^*)$ almost surely, where θ^* is a local maximum of (43) and v^* is the unique solution to the projected Bellman equation corresponding to the policy π^{θ^*}, i.e., the two together satisfy*

$$\Phi^T D^{\theta^*} \Phi v^* = \Phi^T D^{\theta^*} T_{\theta^*}(\Phi v^*). \tag{51}$$

Proof. Recall that for any v, $\theta_t \to \theta^*$ almost surely according to Proposition A.4-Remark A.5. Moreover, $v^* \in C^o$ would mean $0 \in h(v^*)$ and in fact since $\theta_t \to \theta^*$ almost surely, as $t \to \infty$, we have that corresponding to θ^*, $h(v)$ would be a point-to-point map and will no longer be a (nontrivial) point-to-set map as before. Setting $h(v) = 0$ in this case would imply the Bellman equation (51) for policy π^{θ^*}. The claim follows. $\qquad\square$

Remark A.11. In Theorem A.10, (51) may or may not hold if $v^* \in \partial C$ (the boundary of C). This is because projection set boundaries themselves can induce spurious attractors, see (Kushner and Yin 2003). We analyze below the case where the critic is not projected to the set C but can take any value in $\mathbb{R}^{d_1}$. We prove the stability of the critic recursion in this case, i.e., that $\sup_t \|v_t\| < \infty$ w.p.1 and Theorem 5.4 presents the main (general) result on convergence of the joint iterate process in this case of unprojected critic iterates.

Remark A.12. It has been argued in the case of the discounted cost CA algorithm in (Bhatnagar, Borkar, and Guin 2023) for the look-up table setting that the algorithm there mimics value iteration unlike AC that mimics the policy iteration procedure. Now by iterating θ on a faster timescale as opposed to the v-update, the resulting scheme is seen to resemble projected value iteration, see (Bertsekas 2017).

The Case of Critic without Projection We now consider the case when the critic recursion is unconstrained though the actor recursion continues to be constrained (using the projection operator Γ_2 as before). The AC analog of this algorithm has for instance been analysed for its asymptotic convergence in (Bhatnagar et al. 2009). An important observation here is that since there is no projection now on the critic, one needs to establish explicitly that the critic recursion remains uniformly bounded almost surely. In this case, we rewrite (34) as

$$v_{t+1} = v_t + \beta_t(y_t + Y_t + \kappa_t), \tag{52}$$

where,

$$y_t = \sum_s \mu_{\theta_t}(s) \sum_a \pi^{\theta_t}(s,a)(R(s,a) - L^{\theta_t} + v_t^T \sum_{s'} p(s,a,s')\phi(s') - v_t^T\phi(s))\phi(s),$$

$$Y_t = \delta_t \phi(s_t) - E[\delta_t \phi(s_t)|\mathcal{F}_2(t)],$$

$$\kappa_t = E[(R(s_t, a_t) - L^{\theta_t} + v_t^T \sum_{s_{t+1}} p(s_t, a_t, s_{t+1})\phi(s_{t+1}) - v_t^T\phi(s_t))\phi(s_t)|\mathcal{F}_2(t)] - y_t,$$

as before. Thus, $(Y_t, \mathcal{F}_2(t)), t \geq 0$ is a martingale difference sequence and $\kappa_t, t \geq 0$ constitutes the Markov noise. Now from Assumption 5.1, it follows that $\kappa_t \to 0$ almost surely as $t \to \infty$. We now prove Theorem 5.3 concerning the stability of the recursion (52):

Proof of Theorem 5.3:

Note that (53) denotes the differential inclusion (5) associated with the critic update.

$$\dot{v} \in h(v), \tag{53}$$

where recall that

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

$$-v^T\phi(s))\phi(s)|\theta \in \bar{\theta}^*(v)\},$$

Table 3: Hyper-parameters used for the Average Reward CA Algorithm

Hyper-parameter	Value/Description
Number of hidden layers in actor NN	1
Number of hidden layers in critic NN	1
Number of nodes in hidden layer of actor NN	64
Number of nodes in hidden layer of critic NN	64
Activation Function used in critic NN	ReLU
Activation Function used in actor NN	softmax
Learning rate for actor update (at time instant t)	$\frac{1.5}{(1+t)^{0.5}}$
Learning rate for critic update (at time instant t)	$\frac{1.5}{(1+t)^{0.51}}$

Table 4: Hyper-parameters used for the Average Reward AC Algorithm

Hyper-parameter	Value/Description
Number of hidden layers in actor NN	1
Number of hidden layers in critic NN	1
Number of nodes in hidden layer of actor NN	64
Number of nodes in hidden layer of critic NN	64
Activation Function used in critic NN	ReLU
Activation Function used in actor NN	softmax
Learning rate for actor update (at time instant t)	$\frac{1.5}{(1+t)^{0.6}}$
Learning rate for critic update (at time instant t)	$\frac{1.5}{(1+t)^{0.4}}$

and where $\bar{\theta}^*(v)$ is the closure of the set $\theta^*(v)$, that in turn is convex and bounded. Thus, $\bar{\theta}^*(v)$ is a convex and compact set. Observe that μ_θ, π^θ and L^θ are bounded and Lipschitz continuous. Moreover, letting $h_\infty(v) = \lim_{c \to \infty} \frac{h(cv)}{c}$, we have

$$h_\infty(v) = \{\sum_s \mu_\theta(s) \sum_a \pi^\theta(s,a)(v^T \sum_{s'} p(s,a,s')\phi(s') - v^T\phi(s))\phi(s)|\theta \in \check{\theta}^*\}, \tag{54}$$

where $\check{\theta}^* = \lim_{c \to \infty} \bar{\theta}^*(cv)$. By Assumption 4.2, for all potential policy parameters θ, the matrix $\mathbf{A}$ defined as under is negative definite: $\mathbf{A} := \mathbb{E}_{s,a,s'}\left[\phi(s)\big(\phi(s') - \phi(s)\big)^\top\right]$, where $s \sim \mu_\theta(\cdot)$, $a \sim \pi^\theta(s,\cdot)$, $s' \sim p(s,a,\cdot)$. Consider now the DI:

$$\dot{v} \in \{\mathbf{A}v|\theta \in \check{\theta}^*\}.$$

Since $\mathbf{A}$ is negative definite regardless of θ, the above DI will have the origin as an asymptotically stable attractor with a unit ball around the origin as its fundamental neighborhood. The claim now follows from Theorem 1 of (Ramaswamy and Bhatnagar 2018). □

The conclusions of Theorem A.7-A.9 shown above continue to hold. We finally come to the proof of Theorem A.10.

Proof of Theorem 5.4:

The proof now follows from Theorem 2 of (Ramaswamy and Bhatnagar 2018).

Hyper-parameters and Compute time

We describe in Tables 3-8 below the hyper-parameters used for each of the algorithms and subsequently the training time for the various algorithms is given in Table 9. The learning rates used for our CA algorithm are those that have been found optimal (see the finite-time analysis). Similarly, the rates used in the AC algorithm are those that have been found optimal in (Wu et al. 2022). These rates have also been used in the PPO-CA and PPO-AC algorithms respectively.

Observe from Table 9 that CA shows better results than AC. In particular, on Pendulum and Frozen Lake environments, CA requires nearly half the training time of AC while showing better average reward performance (see Table 2. Similarly, amongst PPO-CA and PPO-AC, the former requires lower training time. Amongst all algorithms, DQN has the lowest training time even though it does not perform as well as the CA algorithm in terms of average reward performance, see Table 2. We have used the pytorch library in our code.

We observed that when the number of hidden layers in both the actor NN and critic NN was more than 1, the performance of the algorithms was affected negatively. Hence the number of hidden layers in both the actor NN and critic NN was chosen to be

Table 5: Hyper-parameters used for the Average Reward Single timescale AC Algorithm

Hyper-parameter	Value/Description
Number of hidden layers in actor NN	1
Number of hidden layers in critic NN	1
Number of nodes in hidden layer of actor NN	64
Number of nodes in hidden layer of critic NN	64
Activation Function used in critic NN	ReLU
Activation Function used in actor NN	softmax
Learning rate for actor update (at time instant t)	$\frac{1.5}{(1+t)^{0.6}}$
Learning rate for critic update (at time instant t)	$\frac{1.5}{(1+t)^{0.6}}$

Table 6: Hyper-parameters used for the Average Reward DQN Algorithm

Hyper-parameter	Value/Description
Number of hidden layers in QNN	1
Number of nodes in hidden layer of QNN	64
Activation Function used in QNN	ReLU
Learning rate	0.5

1. Similarly the performance was observed to be the best when 64 nodes were used in the hidden layer. This was obtained by trying out over the full range of values 1-100. For finding the best activation function, we tried Linear, Tanh, ReLU and Softmax for each network, in the various algorithms. The best in each case was selected for each network (actor/critic) and these are provided in Tables 3–8, respectively.

Finally, Table 10 shows the CPU configuration of the server on which the various simulations were conducted.

Table 7: Hyper-parameters used for the Average Reward PPO-AC Algorithm

Hyper-parameter	Value/Description
Number of hidden layers in actor NN	1
Number of hidden layers in critic NN	1
Number of nodes in hidden layer of actor NN	64
Number of nodes in hidden layer of critic NN	64
Activation Function used in critic NN	ReLU
Activation Function used in actor NN	softmax
Learning rate for actor update (at time instant t)	$\frac{1.5}{(1+t)^{0.6}}$
Learning rate for critic update (at time instant t)	$\frac{1.5}{(1+t)^{0.4}}$
Batch length	50

Table 8: Hyper-parameters used for the Average Reward PPO-CA Algorithm

Hyper-parameter	Value/Description
Number of hidden layers in actor NN	1
Number of hidden layers in critic NN	1
Number of nodes in hidden layer of actor NN	64
Number of nodes in hidden layer of critic NN	64
Activation Function used in critic NN	ReLU
Activation Function used in actor NN	softmax
Learning rate for actor update (at time instant t)	$\frac{1.5}{(1+t)^{0.5}}$
Learning rate for critic update (at time instant t)	$\frac{1.5}{(1+t)^{0.51}}$
Batch length	50

Table 9: Training Time (in seconds) Averaged over 10 Seeds for 10,000 Iterations.

Environment	CA	AC	DQN	PPO-AC	PPO-CA	Single Timescale AC
Frozen Lake	21.82	22.66	14.26	53.01	51.72	23.77
Pendulum	13.31	26.87	9.51	61.84	61.55	27.91
Mountain Car Continous	12.09	24.2	7.77	53.86	53.13	26.28

Table 10: CPU Configuration of the Server used for the Experiments

Architecture	x86_64
CPU op-mode(s)	32-bit, 64-bit
Byte Order	Little Endian
Address sizes	46 bits physical, 48 bits virtual
CPU(s)	40
On-line CPU(s) list	0-39
CPU family	6
Model name	Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20GHz
CPU MHz	2759.590
OS	Linux