A CONCENTRATION BOUND FOR TWO-TIMESCALE ACTOR-CRITIC ALGORITHM

Anonymous authors
Paper under double-blind review

ABSTRACT

Significant research effort has been directed in recent years towards establishing both asymptotic and non-asymptotic convergence guarantees for two-timescale actor–critic algorithms, where the actor recursion is run on a slower timescale than the critic recursion. This work derives a uniform all-time concentration bound for the actor-critic algorithm with function approximation in the long-run average-reward setting. This bound helps us analyze the behavior of the actor parameter with high probability. We show that, after some finite time, the actor parameter enters a safe region and remains within it thereafter with a high probability. Specifically, with probability at least $1 - \epsilon_1 - \epsilon_2$, the actor error $\|\theta_k - \theta^*\|$ is $O\left(\frac{n_0^{3/4}}{k}\frac{1}{\sqrt{\epsilon_2}} + \left(\frac{1}{n_0}\right)^{1/4}\log^{1/4}\left(\frac{1}{\epsilon_1}\right) + \left(\frac{1}{n_0}\right)^{1/4}\right)$ for all $k \geq n_0$ and sufficiently large n_0. We also present experimental results demonstrating that the aforementioned actor error diminishes with the number of actor-parameter updates.

1 INTRODUCTION

Actor-Critic (AC) methods have demonstrated strong performance across a wide range of reinforcement learning (RL) problems. Pure policy-based approaches, such as REINFORCE, often encounter high variance when estimating policy gradients, while value-based methods like Q-learning perform well in tabular environments but may become unstable or diverge when combined with function approximation. AC methods address these limitations by integrating policy-based and value-based learning. In this framework, the actor aims to learn an optimal policy using feedback from the critic, whereas the critic estimates the value function corresponding to the actor's current policy. The stability of these algorithms is typically achieved through a separation of learning timescales between the actor and the critic.

AC algorithms are based on coupled stochastic update rules that operate on two distinct timescales, with the actor's learning rate generally decreasing more rapidly than the critic's. This separation of timescales plays a fundamental role in maintaining the stability of the updates and guaranteeing their almost sure convergence. From the perspective of the faster timescale, the slower process appears nearly constant, whereas from the perspective of the slower timescale, the faster process can be viewed as having already converged. Such a mechanism allows AC algorithms to effectively mimic the behavior of the policy iteration procedure for numerically solving Markov decision processes (MDPs), thereby facilitating convergence toward an optimal policy. The asymptotic convergence properties of two-timescale AC methods are commonly analyzed using the ordinary differential equation (ODE) framework. Finite-time analysis and concentration bounds have recently gained prominence as approaches for establishing non-asymptotic convergence guarantees. Recent studies have also investigated the finite-time convergence properties of various actor–critic algorithms. Non-asymptotic (such as finite-time) analyses provide sample complexity guarantees and optimal learning rates, while concentration-bound analyses enable the derivation of high-probability convergence bounds.

In this paper, we derive a finite-time, high-probability convergence bound for the two-timescale actor-critic algorithm that characterises the actor parameter's behaviour with high probability. The bound is valid for all time after a given time instant. An all-time bound guarantees that once the algorithm enters a "safe zone" at time n_0, it will stay there forever with high probability. We analyse here a two-timescale actor–critic algorithm employing linear function approximation for the critic in

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the long-run average-reward setting and establish a lower bound on the probability that for a suitably chosen n_0 the actor parameter enters the safe region by time n_0 and remains within it thereafter.

2 Related Works

In Konda & Borkar (1999), several actor–critic algorithms based on lookup-table representations were introduced, along with the first asymptotic convergence analysis of such methods. This line of work was extended in Konda & Tsitsiklis (2003), where actor–critic algorithms employing function approximation through the Q-value function were proposed and their asymptotic convergence properties were analyzed. A natural-gradient-based actor–critic algorithm was subsequently developed in Kakade (2001). Later, works such as Castro & Meir (2009) and Zhang et al. (2020) further advanced the asymptotic analysis of actor–critic methods. In Bhatnagar et al. (2009), natural actor–critic algorithms incorporating bootstrapping in both the actor and critic updates were proposed, and a comprehensive asymptotic analysis, including stability guarantees, was established.

Over the past few years, there has been considerable interest in establishing finite-time convergence guarantees for reinforcement learning algorithms. In Wu et al. (2022), a finite-time analysis of a two-timescale actor–critic algorithm under Markovian sampling was carried out, yielding a sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-2.5})$ for convergence to an ϵ-approximate stationary point of the performance objective. Subsequently, Chen & Zhao (2023) analyzed a single-timescale actor–critic algorithm under Markovian sampling and established a sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-2})$. Finite-time convergence bounds for a natural policy gradient method applied to constrained discounted-cost Markov decision processes (MDPs) were derived in Ding et al. (2020). In Luo & Li (2023), the authors studied a decentralized single-timescale actor–critic algorithm and obtained a sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-2})$. Furthermore, the non-asymptotic convergence behavior of a two-timescale natural actor–critic algorithm was investigated in Khodadadian et al. (2023), where a sample complexity bound of $\tilde{\mathcal{O}}(\epsilon^{-6})$ was established. Panda & Bhatnagar (2024) carried out a non-asymptotic analysis of three-timescale constrained actor-critic and constrained natural actor-critic algorithms in a non-i.i.d. (Markovian) setting with a long-run average-cost criterion, where both the objective and constraint functions are suitable policy-dependent long-run averages of prescribed cost functions.

Korda & La (2015) provided high-probability bounds as well as bounds in expectation for the well-known temporal difference learning algorithm TD(0) with linear function approximators. Thoppe & Borkar (2019) developed a concentration bound for stochastic approximation that is both tighter and valid under substantially weaker assumptions. Notably, the bound remains applicable even when the step-size sequence is not square-summable. Also, Chandak & Borkar (2026) derived a uniform all-time concentration bound for TD(0) with linear function approximation.

3 Background

3.1 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 random single-stage reward that depends on the state s and action a at the given instant and with expected value $r(s, a)$. We let $|r(s, a)| \leq U_r$, $\forall s \in S, \forall a \in A$, where $U_r \in (0, \infty)$ 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):

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

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 differ-

Algorithm 1 Two Timescale Actor-Critic Algorithm

Input: initial average reward parameter L_0, initial actor parameter θ_0, initial critic parameter v_0, step-size β_t for actor, α_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 $r_t = r(s_t, a_t)$
  $L_{t+1} = L_t + \gamma_t(r_t - L_t)$
  $\delta_t = r_t - L_t + \phi(s_{t+1})^\top v_t - \phi(s_t)^\top v_t$
  $v_{t+1} = \Gamma(v_t + \alpha_t \delta_t \phi(s_t))$
  $\theta_{t+1} = \theta_t + \beta_t \delta_t \nabla_\theta \log \pi_{\theta_t}(a_t|s_t)$
end for

ential action-value (Q-value) function is defined as

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

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

The policy gradient theorem (Sutton et al., 1999; Sutton & 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.

3.2 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 & 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.

3.3 Two-Timescale Actor-Critic Algorithm

Algorithm 1 provides the two-timescale actor–critic (AC) algorithm, where the critic employs linear function approximation. The step-size sequences satisfy the standard Robbins–Monro conditions. Furthermore, they are chosen such that $\beta_t = o(\alpha_t)$ for all $t \geq 0$, and $\gamma_t = K\alpha_t$ for some constant $K > 0$. Consequently, the average reward and critic updates evolve on the faster timescale, while the actor updates operate on the slower timescale.

To ensure the boundedness of the critic estimates, we employ the projection operator $\Gamma(\cdot)$. Specifically, for any $x \in \mathbb{R}^{d_1}$, $\Gamma(x)$ denotes the projection of x onto a compact and convex set $C \subset \mathbb{R}^{d_1}$. Moreover, every vector $y \in C$ satisfies $\|y\| \leq U_v$, where $U_v > 0$ is a constant. Finally, as mentioned earlier, the single-stage reward is a function of the current state and the action selected.

4 Assumptions and Main Result

4.1 Assumptions

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

Assumption 1 is essential for establishing the various upper bounds in the analysis.

Assumption 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 $-\lambda \stackrel{\triangle}{=} \sup_\theta \lambda_\theta < 0$.*

This assumption helps give the existence and uniqueness of $v^*(\theta)$ because the following equations hold: For $s \sim \mu_\theta(\cdot), a \sim \pi_\theta(\cdot|s)$,

$$\begin{aligned} &\mathbf{A}v^*(\theta) + \mathbf{b} = 0, \\ &\mathbf{b} := \mathbb{E}_{s,a,s'}[(r(s,a) - L(\theta))\phi(s)]. \end{aligned} \tag{2}$$

The approximation error for the feature mapping can vary depending on its complexity. We define the approximation error that arises due to linear function approximation as follows.

$$\epsilon_{\text{app}}(\boldsymbol{\theta}, s) := |\phi(s)^{\top} v^*(\theta) - V^{\pi_{\boldsymbol{\theta}}}(s)|.$$

We assume that

$$\forall \boldsymbol{\theta}, \forall s,\ \epsilon_{\text{app}}(\boldsymbol{\theta}, s) \leq \epsilon_{\text{app}},$$

where $\epsilon_{\text{app}} \geq 0$ is some constant.

Assumption 3. *There exist $L, B, K > 0$ such that for all $s \in S$ and $a \in A$,*

(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 3(a) , 3(b), 3(c) are standard in the literature on policy gradient methods, see (Wu et al., 2022).

Assumption 4 (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:*

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

Also, $b \leq \epsilon_{app}$. In the above $d_{TV}(\mu, \nu)$ denotes the total variation distance between probability measures μ and ν.

Assumption 5 (Strong Concavity of the Performance Function)**.** *There exists a constant $m > 0$ such that the performance function $L : \mathbb{R}^d \to \mathbb{R}$ is m-strongly concave, i.e., for any $\theta_1, \theta_2 \in \mathbb{R}^d$,*

$$\langle \nabla L(\theta_1) - \nabla L(\theta_2), \theta_1 - \theta_2 \rangle \leq -m\|\theta_1 - \theta_2\|^2. \tag{3}$$

Assumption 5 plays a key role in establishing lemmas 7 and 2.

4.2 Concentration Bound Results

For our analysis, we define the learning rates as follows: $\alpha_t = c_\alpha/(1+t)^\nu, \beta_t = c_\beta/(1+t)^\sigma, \gamma_t = c_\gamma/(1+t)^\nu$, where $c_\alpha, c_\beta, c_\gamma$ are some positive constants and $0 < \nu < \sigma \leq 1$, $\forall t \geq 0$. We have θ^* as the fixed point of $G(\theta) = \sum_s \sum_a \mu_\theta(s)\pi_\theta(a|s)F(\theta, s, a)$ where $F(\theta, s, a) = (r(s,a) - L(\theta) + \sum_{s'} P(s^{'}|s,a)V^\theta(s^{'}) - V^\theta(s))\nabla \log \pi_\theta(a|s) + \theta$. We have $B < \dfrac{\sqrt{2m}}{L_v}$ and $U_r < \dfrac{\sqrt{2m} - L_v B}{\left(1 + \frac{2b}{1-k}\right)(L_\mu B + BL_\pi + K)}$, with constants $L_v, L_\mu, L_\pi > 0$ defined in lemma 4.

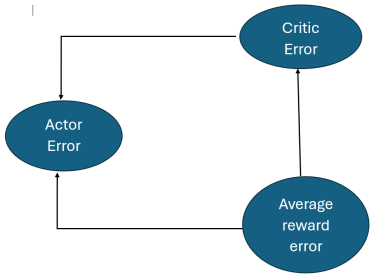

Figure 1: Dependency of errors in the analysis of actor critic algorithm for concentration bound

Theorem 1. *There exist finite positive constants* $C_1, C_2, C_3, C_4, C_5, D$, *such that for* $0 < \delta \leq 1, \varepsilon > 0$, *and* n_0 *sufficiently large such that* $2m\beta_{n_0} < 1$, *we have*

$$P\Big(\|\theta_m - \theta^*\| \leq e^{-(1-\alpha)b_{n_0}(m-1)}\varepsilon + \Delta(n_0,\delta), \forall m \geq n_0\Big)$$
$$\geq 1 - P\Big(\|\theta_{n_0} - \theta^*\| > \varepsilon\Big) - 8\sum_{m\geq n_0+1} e^{-\frac{D\delta^4}{\omega(m-1)}}.$$

where $\Delta(n_0,\delta) := \frac{4\delta + C_1\epsilon_{app} + C_2 n_0^{(\nu-\sigma)/2} + C_3 n_0^{-\nu/2} + C_4 n_0^{-\sigma} + C_5\sqrt{bk}}{1-\alpha}$ *and*

$$\omega(m) = \begin{cases} (n_0)^{c_\beta - \sigma} m^{-c_\beta}, & \text{if } c_\beta \leq \sigma, \\ m^{-\sigma}, & \text{otherwise.} \end{cases}$$

As illustrated in Figure 1, the actor error analysis depends on both the critic error and the error in the estimated average reward. Hence, we also analyzed the average reward estimation error and the critic error to establish the results presented in Theorem 1.

Corollary 1. *For* $\sigma = 1$, $\nu = 0.5$, $\varepsilon \geq 1$ *and sufficiently large* n_0 *such that* $2m\beta_{n_0} < 1$,

$$P\Bigg(\|\theta_k - \theta^*\| = O\Bigg(\frac{n_0^{3/4}}{k}\frac{1}{\sqrt{\epsilon_2}} + \left(\frac{1}{n_0}\right)^{1/4}\log^{1/4}\left(\frac{1}{\epsilon_1}\right) + \left(\frac{1}{n_0}\right)^{1/4}\Bigg)\Bigg)$$
$$\geq 1 - \epsilon_1 - \epsilon_2,$$

$\forall k \geq n_0$.

We provide below a proof sketch of Theorem 1 and Corollary 1. The proof begins by formulating the actor update as a stochastic approximation iteration. We then define an iterative sequence $\{z_n\}_{n\geq n_0}$ and use the decomposition

$$\|\theta_n - \theta^*\| \leq \|\theta_n - z_n\| + \|z_n - \theta^*\|.$$

Next, we derive probabilistic bounds for $\|\theta_n - z_n\|$ and $\|z_n - \theta^*\|$. Combining these bounds yields the desired probabilistic bound on $\|\theta_n - \theta^*\|$.

5 Proof Sketch of Main Result and Corollary

We first formulate the actor update as a stochastic approximation iteration as follows:

$$\begin{aligned}\theta_{t+1} = \theta_t + \beta_t(\sum_s \sum_a \mu_{\theta_t}(s)\pi_{\theta_t}(a|s)F(\theta_t, s, a) - \theta_t) + \underbrace{\beta_t M_{t+1}}_{I_1} \\ + \beta_t(\phi(s_{t+1})^\top v_t - V^{\theta_t}(s_{t+1}) - \phi(s_t)^\top v_t + V^{\theta_t}(s_t))\nabla \log \pi_{\theta_t}(a_t|s_t) \\ + \underbrace{\beta_t(F(\theta_t, s_t, a_t) - \sum_s \sum_a \mu_{\theta_t}(s)\pi_{\theta_t}(a|s)F(\theta_t, s, a))}_{I_2} \\ + \beta_t(L_t - L(\theta_t))\nabla \log \pi_{\theta_t}(a_t|s_t)\end{aligned}$$

where

$$F(\theta, s, a) = (r(s, a) - L(\theta) + \sum_{s'} P(s'|s, a)V^\theta(s') - V^\theta(s))\nabla \log \pi_\theta(a|s) + \theta,$$

and

$$M_{t+1} = (V^{\theta_t}(s_{t+1}) - \sum_{s'} P(s'|s_t, a_t)V^{\theta_t}(s'))\nabla \log \pi_{\theta_t}(a_t|s_t).$$

Define the family of $\sigma-$ fields $F_n := \sigma(\theta_m, s_m, a_m, m \leq n), n \geq 0$. Then $\{M_n\}$ is a martingale difference sequence with respect to F_n, and in particular, the following holds:

$$E[M_{n+1}|F_n] = 0, a.s., \forall n.$$

The term I_1 denotes the error because of the martingale noise term and term I_2 denotes the error because of Markov noise.

We define z_n for $n \geq n_0$ by

$$z_{n+1} = z_n + \beta_n\bigg(\sum_{s,a} \mu_{z_n}(s)\pi_{z_n}(a|s)F(z_n, s, a) - z_n\bigg),$$

where $z_{n_0} = \theta_{n_0}$. We assume n_0 is sufficiently large such that $2m\beta_{n_0} < 1$. Note that

$$\|\theta_n - \theta^*\| \leq \|\theta_n - z_n\| + \|z_n - \theta^*\|.$$

We will now determine bounds for $\|\theta_n - z_n\|$ and $\|z_n - \theta^*\|$. For the term $\|z_n - \theta^*\|$, we have,

$$\begin{aligned}\|z_{n+1} - \theta^*\| &= (1 - (1 - \alpha)\beta_n)\|z_n - \theta^*\| \\ &\leq e^{-(1-\alpha)b_{n_0}(n)}\|z_{n_0} - \theta^*\|,\end{aligned}$$

where $\alpha \in (0, 1)$ is the contraction factor for the function $G(\theta) = \sum_s \sum_a \mu_\theta(s)\pi_\theta(a|s)F(\theta, s, a)$ and $b_{n_0}(n) = \sum_{k=n_0}^{n} \beta_k$.

For $n, m \geq 0$, let $\xi(n, m) = \prod_{k=m}^{k=n}(1-\beta_k)$ if $n \geq m$ and 1 otherwise. Next, we give a probabilistic bound on the term $\|\theta_n - z_n\|$. We have ,

$$\begin{aligned}\|\theta_{m+1} - z_{m+1}\| \leq \bigg\| \sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k M_{k+1} \bigg\| + \bigg\| \sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k \tilde{U}_{k+1} \bigg\| + \alpha\theta_m' \\ + O(n_0^{-\sigma}) + O(\epsilon_{\text{app}}) + O(n_0^{(\nu-\sigma)/2}) + O(bk) + O(\sqrt{bk}) \\ + 2B\sqrt{\frac{2}{\lambda}\bigg|\sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k\langle z_k, \tilde{W}_{k+1}\rangle\bigg|} + O(n_0^{-\nu/2}) \\ + B\bigg(\sqrt{\frac{2}{\lambda}} + 1\bigg)\bigg|\sqrt{\sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k y_k \tilde{Z}_{k+1}}\bigg|\end{aligned}$$

where,

$$\theta_m^{'} = \sup_{n_0 \le k \le m} \|\theta_k - z_k\|$$

$$U(\theta, s, a) = (r(s,a) - L(\theta) + \sum_{s^{'}} P(s^{'}|s,a)V^{\theta}(s^{'}) - V^{\theta}(s))\nabla \log \pi_{\theta}(a|s)$$

$$W(s, a, s^{'}, L, v) = (r(s,a) - L + \phi(s^{'})^T v - \phi(s)^T v)\phi(s)$$

$$\tilde{U}_{k+1} = U(\theta_k, s_k, a_k) - \sum_a \pi_{\theta_k}(a|s_k)U(\theta_k, s_k, a) + \sum_a \pi_{\theta_k}(a|s_{k+1})U(\theta_k, s_{k+1}, a)$$
$$- \sum_s P^{\theta_k}(s|s_k) \sum_a \pi_{\theta_k}(a|s)U(\theta_k, s, a)$$

$$\tilde{W}_{k+1} = W(s_k, a_k, s_{k+1}, L_k, v_k) - \sum_a \sum_{s^{'}} \pi_{\theta_k}(a|s_k)P(s^{'}|s_k, a)W(s_k, a, s^{'}, L_k, v_k)$$
$$+ \sum_a \sum_{s^{'}} \pi_{\theta_k}(a|s_{k+1})P(s^{'}|s_{k+1}, a)W(s_{k+1}, a, s^{'}, L_k, v_k)$$
$$- \sum_s P^{\theta_k}(s|s_k) \sum_a \sum_{s^{'}} \pi_{\theta_k}(a|s)P(s^{'}|s, a)W(s, a, s^{'}, L_k, v_k)$$

$$\tilde{Z}_{k+1} = r(s_k, a_k) - \sum_a \pi_{\theta_k}(a|s_k)r(s_k, a) + \sum_a \pi_{\theta_k}(a|s_{k+1})r(s_{k+1}, a)$$
$$- \sum_s P^{\theta_k}(s|s_k) \sum_a \pi_{\theta_k}(a|s)r(s, a)$$

We define the family of σ- fields : $\mathcal{A}_n := \sigma(s_0, \theta_m, L_m, v_m, s_m, a_{m-1}, 1 \le m \le n)$. Then $\{\tilde{W}_n\}, \{\tilde{U}_n\}, \{\tilde{Z}_n\}$ are martingale difference sequences, with respect to $\{\mathcal{A}_n\}$.
Now we will obtain a bound on the probability:

$$P\bigg(\|\theta_m - \theta^*\| \le e^{-(1-\alpha)b_{n_0}(n)}\varepsilon + \Delta(n_0, \delta), \forall n_0 \le m \le n\bigg),$$

for some $\varepsilon > 0, \delta > 0$. We define,

$$\Delta(n_0, \delta) := \frac{4\delta + C_1\epsilon_{\text{app}} + C_2 n_0^{(\nu-\sigma)/2} + C_3 n_0^{-\nu/2} + C_4 n_0^{-\sigma} + C_5\sqrt{bk}}{1-\alpha}$$

where C_1, C_2, C_3, C_4, C_5 are positive constants.

We have,

$$P\bigg(\|\theta_m - \theta^*\| \le e^{-(1-\alpha)b_{n_0}(n-1)}\varepsilon + \Delta(n_0, \delta), \forall n_0 \le m \le n\bigg)$$
$$\ge 1 - P\bigg(\{\|\theta_{n_0} - \theta^*\| > \varepsilon\} \bigcup \{\theta_n^{'} > \Delta(n_0, \delta)\}\bigg),$$

Now we have,

$$P(\theta_n^{'} > \Delta(n_0, \delta)) \le p_{n-1} + P(\theta_{n-1}^{'} > \Delta(n_0, \delta))$$

where,

$$p_{n-1} = P\bigg(\bigg\|\sum_{k=n_0}^{k=n-1} \xi(n-1, k+1)\beta_k M_{k+1}\bigg\| > \delta\bigg) + P\bigg(\bigg\|\sum_{k=n_0}^{k=n-1} \xi(n-1, k+1)\beta_k \tilde{U}_{k+1}\bigg\| > \delta\bigg)$$
$$+ P\bigg(2B\sqrt{\frac{2}{\lambda}\bigg|\sum_{k=n_0}^{k=n-1} \xi(n-1, k+1)\beta_k \langle z_k, \tilde{W}_{k+1}\rangle\bigg|} > \delta\bigg)$$
$$+ P\bigg(B\bigg(\sqrt{\frac{2}{\lambda}} + 1\bigg)\sqrt{\bigg|\sum_{k=n_0}^{k=n-1} \xi(n-1, k+1)\beta_k y_k \tilde{Z}_{k+1}\bigg|} > \delta\bigg)$$

Now we will use the following lemma.

Lemma 1. *There exists positive constant D such that for $0 < \delta \leq 1$*

$$p_{n-1} \leq 8e^{-\frac{D\delta^4}{\omega(n-1)}}$$

where ,

$$\omega(m) = \begin{cases} (n_0)^{c_\beta - \sigma} m^{-c_\beta}, & \text{if } c_\beta \leq \sigma, \\ m^{-\sigma}, & \text{if otherwise.} \end{cases}$$

Using the above lemma we get

$$P(\theta_n^{'} > \Delta(n_0, \delta)) \leq 8 \sum_{m=n_0+1}^{n} e^{-\frac{D\delta^4}{\omega(m-1)}}$$

which finally gives us

$$P\bigg(\|\theta_m - \theta^*\| \leq e^{-(1-\alpha)b_{n_0}(m-1)}\varepsilon + \Delta(n_0, \delta), \forall n_0 \leq m \leq n\bigg)$$
$$\geq 1 - P\bigg(\|\theta_{n_0} - \theta^*\| > \varepsilon\bigg) - 8 \sum_{m=n_0+1}^{n} e^{-\frac{D\delta^4}{\omega(m-1)}},$$

where D is some positive constant. Now, let A_n be the set

$$\bigg\{\|\theta_m - \theta^*\| \leq e^{-(1-\alpha)b_{n_0}(m-1)}\varepsilon + \Delta(n_0, \delta), \forall n_0 \leq m \leq n\bigg\}.$$

Then $\{A_n\}$ is a decreasing sequence of sets, i.e., $A_{n+1} \subseteq A_n$ for all $n \geq n_0$. Now let A be the set

$$\bigg\{\|\theta_m - \theta^*\| \leq e^{-(1-\alpha)b_{n_0}(m-1)}\varepsilon + \Delta(n_0, \delta), \forall m \geq n_0\bigg\}$$

Then $A = \cap_{n=0}^{\infty} A_n$. Hence $P(A) = \lim_{n\uparrow\infty} P(A_n)$. So we have,

$$P\bigg(\|\theta_m - \theta^*\| \leq e^{-(1-\alpha)b_{n_0}(m-1)}\varepsilon + \Delta(n_0, \delta), \forall m \geq n_0\bigg)$$
$$\geq 1 - P\bigg(\|\theta_{n_0} - \theta^*\| > \varepsilon\bigg) - 8 \sum_{m \geq n_0+1} e^{-\frac{D\delta^4}{\omega(m-1)}}$$

Now, we use $\varsigma_i, i = 1, 2, ...$ to denote different constants in the proof. For $\sigma = 1$, $\varepsilon \geq 1$ and c_β sufficiently large, we have $\omega(m) = 1/m$. Hence, $2 \sum_{m \geq n_0} e^{-\frac{D\delta^4}{\omega(m)}} = 2 \sum_{m \geq n_0} e^{-D_1 m \delta_1^2} \leq \varsigma_1 e^{-D_1 n_0 \delta_1^2}$.

Let $\epsilon_1 = \varsigma_1 e^{-Dn_0\delta^4}$ which gives us $\delta = \varsigma_2 n_0^{-1/4} \log^{1/4}\left(\frac{\varsigma_1}{\epsilon_1}\right)$. This choice of δ gives us $8 \sum_{m \geq n_0} e^{-\frac{D\delta^4}{\omega(m)}} \leq \epsilon_1$.

We also show that $\exp\bigg(-(1-\alpha)b_{n_0}(m-1)\bigg) \leq \frac{n_0+1}{m+1}$ under the assumption that $c_\beta(1-\alpha) > 1$.

Next, we let $\varepsilon = \frac{\mathbb{E}[\|\theta_{n_0} - \theta^*\|^2]}{\epsilon_2}$. Then we have , $P\bigg(\|\theta_{n_0} - \theta^*\| > \varepsilon\bigg) = P\bigg(\|\theta_{n_0} - \theta^*\|^2 > \varepsilon^2\bigg) \leq P\bigg(\|\theta_{n_0} - \theta^*\|^2 > \varepsilon\bigg) = P\bigg(\|\theta_{n_0} - \theta^*\|^2 > \frac{\mathbb{E}[\|\theta_{n_0} - \theta^*\|^2]}{\epsilon_2}\bigg) \leq \epsilon_2$.

So we have , $\forall k \geq n_0 > 0$,

$$P\bigg(\|\theta_k - \theta^*\| = O\bigg(\frac{n_0}{k}\frac{\mathbb{E}[\|\theta_{n_0} - \theta^*\|^2]}{\epsilon_2} + \left(\frac{1}{n_0}\right)^{1/4} \log^{1/4}\left(\frac{1}{\epsilon_1}\right) + n_0^{(\nu-1)/2} + n_0^{-\nu/2}$$
$$+ \epsilon_{\text{app}} + \sqrt{\epsilon_{\text{app}}}\bigg)\bigg) \geq 1 - \epsilon_1 - \epsilon_2.$$

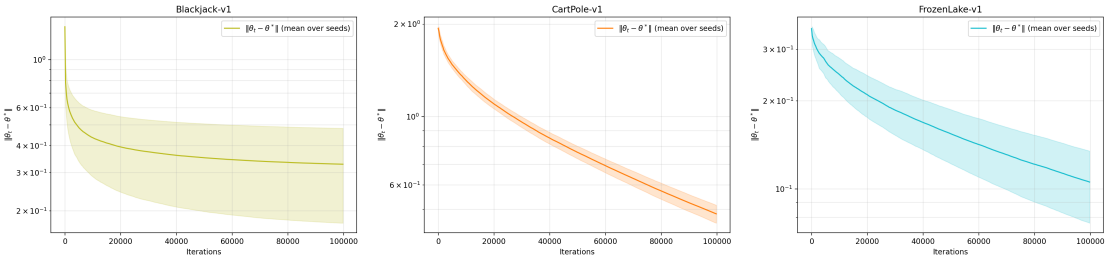

Figure 2: Convergence of $\|\theta_t - \theta^*\|$ averaged over ten independent runs

Now, we use the result of the following lemma.

Lemma 2. *For all $n \geq 1$ such that $2m\beta_n < 1$ and $\sigma = 1$, we have,*

$$\mathbb{E}[\|\theta_n - \theta^*\|^2] = \mathcal{O}(n^{-\nu/2}) + \mathcal{O}(n^{(\nu-1)/2}).$$

Optimising over the values of ν, we have $\nu = 0.5$.

Hence, (ignoring the approximation error) we conclude that for all $k \geq n_0$, where n_0 is chosen sufficiently large such that $2m\beta_{n_0} < 1$, the following holds.

$$P\left(\|\theta_k - \theta^*\| = O\left(\frac{n_0^{3/4}}{k}\frac{1}{\sqrt{\epsilon_2}} + \left(\frac{1}{n_0}\right)^{1/4}\log^{1/4}\left(\frac{1}{\epsilon_1}\right) + \left(\frac{1}{n_0}\right)^{1/4}\right)\right) \geq 1 - \epsilon_1 - \epsilon_2$$

6 Experimental Results

We may mention that the actor-critic algorithm is very well studied in the literature for its empirical properties and performance comparisons with other related algorithms. However, in order to empirically validate the convergence of Algorithm 1, we evaluate it on three different Open AI Gym environments, namely CartPole-v1, FrozenLake-v1, and Blackjack-v1, respectively, each reformulated as a continuing average-reward task by resetting the environment state to an initial state upon the termination of an episode. In other words, once the environment hits the goal state, rather than treating the latter as absorbing, we randomly pick another non-terminal state for a restart and continue.

Since the optimal parameter θ^* satisfying $\nabla L(\theta) = 0$ has no closed-form expression in these settings, we approximate it by running Algorithm 1 for a large number of iterations across several independent seeds and averaging the resulting parameters. Figure 2 plots $\|\theta_t - \theta^*\|$ against the iteration count t, averaged over ten independent training runs with a shaded band indicating one standard deviation. It can be seen that in all three environments, the actor parameter exhibits a smooth, monotonically decreasing trend, consistent with the theoretical two-timescale convergence guarantee. The hyperparameters used for these experiments are illustrated in Table 1.

7 Conclusions and Future Work

To the best of our knowledge, we obtained the first finite-time, high-probability convergence bound for the two-timescale actor-critic algorithm. Our bound characterises the behaviour of the actor parameter with high probability. The bound holds uniformly (over time instants) after a specified time instant. An interesting direction for future work would be to establish analogous concentration bounds for other classes of actor-critic algorithms, such as the two-timescale natural actor-critic and the three-timescale constrained actor-critic algorithms. This would provide insights into how learning rates, constraints and natural gradients affect the concentration bound.

AI USE STATEMENT

In this work, we used ChatGPT for the purpose of polishing the writing. We have not used any generative AI tools for retrieval and discovery, research ideation or execution, drafting sections of the paper, or proving mathematical claims as well as for generating synthetic datasets. We have also completely proofread the paper after using the tool for polishing the writing. In particular, we reviewed the rephrased sentences for grammatical correctness, clarity, and consistency with the intended meaning. We take full responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have provided detailed proofs of Theorem 1 and Corollary 1 in the appendix. We have also included proofs of all the supporting lemmas used to prove these results in the section titled "Proof of Lemmas." Additionally, we provide the hyperparameter details for all experiments towards the end of the appendix. We have also included the code required to reproduce the plots presented in our experiments in the supplementary material.

REFERENCES

Dimitri P. Bertsekas. *Optimal Control and Dynamic Programming*, volume 2. Athena Scientific, Belmont, MA, 4 edition, 2013.

Shalabh Bhatnagar, Richard S. Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor–critic algorithms. *Automatica*, 45(11):2471–2482, 2009. ISSN 0005-1098. doi: https://doi.org/10.1016/j.automatica.2009.07.008. URL https://www.sciencedirect.com/science/article/pii/S0005109809003549.

D. Di Castro and R. Meir. A convergent online single time scale actor critic algorithm, 2009.

Siddharth Chandak and Vivek S. Borkar. A concentration bound for TD(0) with function approximation. *Stochastic Systems*, 16(1):44–60, 2026. doi: 10.1287/stsy.2023.0055. URL https://doi.org/10.1287/stsy.2023.0055.

Xuyang Chen and Lin Zhao. Finite-time analysis of single-timescale actor-critic, 2023.

Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo Jovanovic. Natural policy gradient primal-dual method for constrained markov decision processes. In *Advances in Neural Information Processing Systems*, volume 33, pp. 8378–8390. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/5f7695debd8cde8db5abcb9f161b49ea-Paper.pdf.

Sham M Kakade. A natural policy gradient. In T. Dietterich, S. Becker, and Z. Ghahramani (eds.), *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. URL https://proceedings.neurips.cc/paper_files/paper/2001/file/4b86abe48d358ecf194c56c69108433e-Paper.pdf.

Sajad Khodadadian, Thinh T. Doan, Justin Romberg, and Siva Theja Maguluri. Finite-sample analysis of two-time-scale natural actor–critic algorithm. *IEEE Transactions on Automatic Control*, 68(6):3273–3284, 2023. doi: 10.1109/TAC.2022.3190032.

Vijay Konda and Vivek Borkar. Actor-critic–type learning algorithms for markov decision processes. *SIAM J. Control and Optimization*, 38:94–123, 12 1999. doi: 10.1137/S036301299731669X.

Vijay R. Konda and John N. Tsitsiklis. Onactor-critic algorithms. *SIAM Journal on Control and Optimization*, 42(4):1143–1166, 2003. doi: 10.1137/S0363012901385691. URL https://doi.org/10.1137/S0363012901385691.

Nathaniel Korda and Prashanth La. On td(0) with function approximation: Concentration bounds and a centered variant with exponential convergence. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 626–634, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/korda15.html.

Qijun Luo and Xiao Li. Finite-time analysis of fully decentralized single-timescale actor-critic, 2023.

Prashansa Panda and Shalabh Bhatnagar. Finite-time analysis of three-timescale constrained actor-critic and constrained natural actor-critic algorithms. In Negar Kiyavash and Joris M. Mooij (eds.), *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 2787–2834. PMLR, 15–19 Jul 2024. URL https://proceedings.mlr.press/v244/panda24a.html.

Prashansa Panda and Shalabh Bhatnagar. Two-timescale critic-actor for average reward mdps with function approximation, 2025. URL https://arxiv.org/abs/2402.01371.

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

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

Gugan Thoppe and Vivek S. Borkar. A concentration bound for stochastic approximation via alekseev's formula. *Stochastic Systems*, 9(1):1–26, 2019. doi: 10.1287/stsy.2018.0019.

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

Yue Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite time analysis of two time-scale actor critic methods, 2022.

Shangtong Zhang, Bo Liu, Hengshuai Yao, and Shimon Whiteson. Provably convergent two-timescale off-policy actor-critic with function approximation, 2020.

A Appendix

A.1 Proof of Main Result

For $n, m \geq 0$, let $\xi(n,m) = \prod_{k=m}^{k=n}(1-\beta_k)$ if $n \geq m$ and 1 otherwise.

There are several lemmas that will be used in the proof.

Lemma 3. *There exists a constant $M_L > 0$ such that for all $\theta_1, \theta_2 \in \mathbb{R}^d$, it holds that*

$$\|L(\theta_1) - L(\theta_2)\| \leq M_L \|\theta_1 - \theta_2\|.$$

Lemma 4. *There exists $L_v > 0, L_\mu > 0, L_\pi > 0$ such that such that for all $\theta_1, \theta_2 \in \mathbb{R}^d, s \in S$,*

$$\begin{aligned}
\sum_s |\mu_{\theta_1}(s) - \mu_{\theta_2}(s)| &\leq L_\mu \|\theta_1 - \theta_2\|, \\
\|V^{\theta_1}(s) - V^{\theta_2}(s)\| &\leq L_v \|\theta_1 - \theta_2\|, \\
\sum_a |\pi_{\theta_1}(a|s) - \pi_{\theta_2}(a|s)| &\leq L_\pi \|\theta_1 - \theta_2\|.
\end{aligned}$$

Lemma 5.

$$\begin{aligned}
&\sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k((L_k - L(\theta_k))^2 \leq \mathcal{O}(n_0^{\nu-\sigma}) + \mathcal{O}(n_0^{-\nu}) + \mathcal{O}(bk) \\
&\quad + \sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k(L_k - L(\theta_k))\left(r(s_k, a_k) - \sum_a \pi_{\theta_k}(a|s_k) r(s_k, a) + \bar{r}_{\theta_k}(s_{k+1}) - \sum_s P^{\theta_k}(s|s_k)\bar{r}_{\theta_k}(s)\right)
\end{aligned}$$

where $\bar{r}_\theta(s) = \sum_a \pi_\theta(a|s) r(s,a)$ and $P^{\theta_k}(s|s_k) = \sum_a P(s|s_k, a)\pi_{\theta_k}(a|s_k)$.

Lemma 6.

$$\begin{aligned}
&\sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k \|v_k - v^*(\theta_k)\|^2 \leq \mathcal{O}(n_0^{\nu-\sigma}) + \mathcal{O}(n_0^{-\nu}) + \mathcal{O}(bk) \\
&\quad + \frac{2}{\lambda} \sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k \langle z_k, W(s_k, a_k, s_{k+1}, L_k, v_k) - \sum_a \sum_{s'} \pi_{\theta_k}(a|s_k) P(s^{'}|s_k, a) W(s_k, a, s^{'}, L_k, v_k)\rangle \\
&\quad + \frac{2}{\lambda} \sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k \langle z_k, \bar{W}_{\theta_k}(s_{k+1}, L_k, v_k) - \sum_s P^{\theta_k}(s|s_k)\bar{W}_{\theta_k}(s, L_k, v_k)\rangle \\
&\quad + \frac{1}{2\lambda} \sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k y_k \left(r(s_k, a_k) - \sum_a \pi_{\theta_k}(a|s_k) r(s_k, a) + \bar{r}_{\theta_k}(s_{k+1}) - \sum_s P^{\theta_k}(s|s_k)\bar{r}_{\theta_k}(s)\right)
\end{aligned}$$

where $W(s, a, s^{'}, L, v) = (r(s,a) - L + \phi(s^{'})^T v - \phi(s)^T v)\phi(s)$, $\bar{W}_\theta(s, L, v) = \sum_a \sum_{s'} \pi_{\theta_k}(a|s) P(s^{'}|s, a) W(s, a, s^{'}, L, v)$ and $z_k = v_k - v^*(\theta_k)$.

Now we will formulate the actor update as a stochastic approximation iteration.

We have the following update for actor:-

$$
\begin{aligned}
\theta_{t+1} &= \theta_t + \beta_t \delta_t \nabla \log \pi_{\theta_t}(a_t|s_t) \\
&= \theta_t + \beta_t (r(s_t,a_t) - L_t + \phi(s_{t+1})^\top v_t - \phi(s_t)^\top v_t) \nabla \log \pi_{\theta_t}(a|s) \\
&= \theta_t + \beta_t (r(s_t,a_t) - L_t + V^{\theta_t}(s_{t+1}) - V^{\theta_t}(s_t)) \nabla \log \pi_{\theta_t}(a|s) \\
&\quad + \beta_t (\phi(s_{t+1})^\top v_t - V^{\theta_t}(s_{t+1}) - \phi(s_t)^\top v_t + V^{\theta_t}(s_t)) \nabla \log \pi_{\theta_t}(a|s) \\
&= \theta_t + \beta_t (r(s_t,a_t) - L(\theta_t) + \sum_{s'} P(s'|s_t,a_t) V^{\theta_t}(s') - V^{\theta_t}(s_t)) \nabla \log \pi_{\theta_t}(a|s) \\
&\quad + \beta_t (\phi(s_{t+1})^\top v_t - V^{\theta_t}(s_{t+1}) - \phi(s_t)^\top v_t + V^{\theta_t}(s_t)) \nabla \log \pi_{\theta_t}(a_t|s_t) \\
&\quad + \beta_t (V^{\theta_t}(s_{t+1}) - \sum_{s'} P(s'|s_t,a_t) V^{\theta_t}(s')) \nabla \log \pi_{\theta_t}(a_t|s_t) \\
&\quad + \beta_t (L_t - L(\theta_t)) \nabla \log \pi_{\theta_t}(a_t|s_t) \\
&= \theta_t + \beta_t (F(\theta_t, s_t, a_t) - \theta_t) + \beta_t M_{t+1} \\
&\quad + \beta_t (\phi(s_{t+1})^\top v_t - V^{\theta_t}(s_{t+1}) - \phi(s_t)^\top v_t + V^{\theta_t}(s_t)) \nabla \log \pi_{\theta_t}(a_t|s_t) \\
&\quad + \beta_t (L_t - L(\theta_t)) \nabla \log \pi_{\theta_t}(a_t|s_t) \\
&= \theta_t + \beta_t (\sum_s \sum_a \mu_{\theta_t}(s) \pi_{\theta_t}(a|s) F(\theta_t, s, a) - \theta_t) + \underbrace{\beta_t M_{t+1}}_{I_1} \\
&\quad + \beta_t (\phi(s_{t+1})^\top v_t - V^{\theta_t}(s_{t+1}) - \phi(s_t)^\top v_t + V^{\theta_t}(s_t)) \nabla \log \pi_{\theta_t}(a_t|s_t) \\
&\quad \underbrace{+ \beta_t (F(\theta_t, s_t, a_t) - \sum_s \sum_a \mu_{\theta_t}(s) \pi_{\theta_t}(a|s) F(\theta_t, s, a))}_{I_2} \\
&\quad + \beta_t (L_t - L(\theta_t)) \nabla \log \pi_{\theta_t}(a_t|s_t)
\end{aligned}
$$

where

$$F(\theta, s, a) = (r(s,a) - L(\theta) + \sum_{s'} P(s'|s,a) V^{\theta}(s') - V^{\theta}(s)) \nabla \log \pi_{\theta}(a|s) + \theta,$$

and

$$M_{t+1} = (V^{\theta_t}(s_{t+1}) - \sum_{s'} P(s'|s_t,a_t) V^{\theta_t}(s')) \nabla \log \pi_{\theta_t}(a_t|s_t)$$

Define the family of $\sigma-$ fields $F_n := \sigma(\theta_m, s_m, a_m, m \leq n), n \geq 0$. Then $\{M_n\}$ is a martingale difference sequence with respect to F_n; that is,

$$E[M_{n+1}|F_n] = 0, a.s. \forall n,$$

The term I_1 denotes the error because of the martingale noise term and term I_2 denotes the error because of Markov noise.

Lemma 7. *For any* $\theta_1, \theta_2 \in \mathbb{R}^d$,

$$\left\| \sum_s \sum_a \mu_{\theta_1}(s) \pi_{\theta_1}(a|s) F(\theta_1, s, a) - \sum_s \sum_a \mu_{\theta_2}(s) \pi_{\theta_2}(a|s) F(\theta_2, s, a) \right\| \leq \alpha \|\theta_1 - \theta_2\|$$

where $0 < \alpha < 1$.

Define z_n for $n \geq n_0$ by

$$z_{n+1} = z_n + \beta_n \left(\sum_{s,a} \mu_{z_n}(s) \pi_{z_n}(a|s) F(z_n, s, a) - z_n \right)$$

where $z_{n_0} = \theta_{n_0}$.

We assume $2m\beta_{n_0} < 1$. Note that $\|\theta_n - \theta^*\| \leq \|\theta_n - z_n\| + \|z_n - \theta^*\|$.
Now let us bound the second term. We have,

$$
\begin{aligned}
&z_{n+1} - \theta^* \\
&= z_n + \beta_n \Bigg(\sum_s \sum_a \mu_{z_n}(s)\pi_{z_n}(a|s)F(z_n, s, a) - z_n \Bigg) - \theta^* \\
&= (1-\beta_n)(z_n - \theta^*) + \beta_n \Bigg(\sum_s \sum_a \mu_{z_n}(s)\pi_{z_n}(a|s)F(z_n, s, a) \Bigg) - \beta_n \theta^* \\
&= (1-\beta_n)(z_n - \theta^*) + \beta_n \Bigg(\sum_s \sum_a \mu_{z_n}(s)\pi_{z_n}(a|s)F(z_n, s, a) - \theta^* \Bigg) \\
&= (1-\beta_n)(z_n - \theta^*) \\
&+ \beta_n \Bigg(\sum_s \sum_a \mu_{z_n}(s)\pi_{z_n}(a|s)F(z_n, s, a) - \sum_s \sum_a \mu_{\theta^*}(s)\pi_{\theta^*}(a|s)F(\theta^*, s, a)) \Bigg)
\end{aligned}
$$

Now we have ,

$$
\begin{aligned}
&\|z_{n+1} - \theta^*\| \\
&\leq (1-\beta_n)\|z_n - \theta^*\| \\
&\quad + \beta_n \Bigg\| \sum_s \sum_a \mu_{z_n}(s)\pi_{z_n}(a|s)F(z_n, s, a) - \sum_s \sum_a \mu_{\theta^*}(s)\pi_{\theta^*}(a|s)F(\theta^*, s, a)) \Bigg\| \\
&\leq (1-\beta_n)\|z_n - \theta^*\| + \alpha\beta_n\|z_n - \theta^*\| \\
&= (1 - (1-\alpha)\beta_n)\|z_n - \theta^*\| \qquad (4) \\
&\leq e^{-(1-\alpha)b_{n_0}(n)}\|z_{n_0} - \theta^*\| \qquad (5)
\end{aligned}
$$

where $b_{n_0}(n) = \sum_{k=n_0}^{n} \beta_k$.

Next we give a probabilistic bound on the term $\|\theta_n - z_n\|$. We have ,

$$
\begin{aligned}
&\theta_{t+1} - z_{t+1} \\
&= \theta_t + \beta_t(\sum_s \sum_a \mu_{\theta_t}(s)\pi_{\theta_t}(a|s)F(\theta_t, s, a) - \theta_t) + \beta_t M_{t+1} \\
&\quad + \beta_t(\phi(s_{t+1})^\top v_t - V^{\theta_t}(s_{t+1}) - \phi(s_t)^\top v_t + V^{\theta_t}(s_t))\nabla \log \pi_{\theta_t}(a|s) \\
&\quad + \beta_t(F(\theta_t, s_t, a_t) - \sum_s \sum_a \mu_{\theta_t}(s)\pi_{\theta_t}(a|s)F(\theta_t, s, a)) \\
&\quad + \beta_t(L_t - L(\theta_t))\nabla \log \pi_{\theta_t}(a_t|s_t) \\
&\quad - z_t - \beta_t\Big(\sum_{s,a} \mu_{z_t}(s)\pi_{z_t}(a|s)F(z_t, s, a) - z_t\Big) \\
&= (1-\beta_t)(\theta_t - z_t) + \beta_t M_{t+1} + \beta_t(F(\theta_t, s_t, a_t) - \sum_{s,a} \mu_{z_t}(s)\pi_{z_t}(a|s)F(z_t, s, a)) \\
&\quad + \beta_t(\phi(s_{t+1})^\top v_t - V^{\theta_t}(s_{t+1}) - \phi(s_t)^\top v_t + V^{\theta_t}(s_t))\nabla \log \pi_{\theta_t}(a|s) \\
&\quad + \beta_t(L_t - L(\theta_t))\nabla \log \pi_{\theta_t}(a_t|s_t) \\
&= (1-\beta_t)(\theta_t - z_t) + \beta_t M_{t+1} + \beta_t(\sum_{s,a} \mu_{\theta_t}(s)\pi_{\theta_t}(a|s)F(\theta_t, s, a) - \sum_{s,a} \mu_{z_t}(s)\pi_{z_t}(a|s)F(z_t, s, a)) \\
&\quad + \beta_t(\phi(s_{t+1})^\top v_t - V^{\theta_t}(s_{t+1}) - \phi(s_t)^\top v_t + V^{\theta_t}(s_t))\nabla \log \pi_{\theta_t}(a|s) \\
&\quad + \beta_t(F(\theta_t, s_t, a_t) - \sum_{s,a} \mu_{\theta_t}(s)\pi_{\theta_t}(a|s)F(\theta_t, s, a)) \\
&\quad + \beta_t(L_t - L(\theta_t))\nabla \log \pi_{\theta_t}(a_t|s_t) \\
&= (1-\beta_t)(\theta_t - z_t) + \beta_t M_{t+1} + \beta_t(\sum_{s,a} \mu_{\theta_t}(s)\pi_{\theta_t}(a|s)F(\theta_t, s, a) - \sum_{s,a} \mu_{z_t}(s)\pi_{z_t}(a|s)F(z_t, s, a)) \\
&\quad + \beta_t(F(\theta_t, s_t, a_t) - \sum_{s,a} \mu_{\theta_t}(s)\pi_{\theta_t}(a|s)F(\theta_t, s, a)) \\
&\quad + \beta_t(\phi(s_{t+1})^\top v_t - \phi(s_{t+1})^\top v^*(\theta_t) - \phi(s_t)^\top v_t + \phi(s_t)^\top v^*(\theta_t))\nabla \log \pi_{\theta_t}(a|s) \\
&\quad + \beta_t(\phi(s_{t+1})^\top v^*(\theta_t) - V^{\theta_t}(s_{t+1}) - \phi(s_t)^\top v^*(\theta_t) + V^{\theta_t}(s_t))\nabla \log \pi_{\theta_t}(a|s) \\
&\quad + \beta_t(L_t - L(\theta_t))\nabla \log \pi_{\theta_t}(a_t|s_t).
\end{aligned}
$$

For some $n \geq n_0$ we iterate the above for $n_0 \leq m \leq n$ to obtain

$$
\begin{aligned}
&\theta_{m+1} - z_{m+1} \\
&= \sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k M_{k+1} \\
&\quad + \sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k \Big(\sum_{s,a} \mu_{\theta_k}(s)\pi_{\theta_k}(a|s)F(\theta_k, s, a) - \sum_{s,a} \mu_{z_k}(s)\pi_{z_k}(a|s)F(z_k, s, a)\Big) \\
&\quad + \underbrace{\sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k \Big(F(\theta_k, s_k, a_k) - \sum_{s,a} \mu_{\theta_k}(s)\pi_{\theta_k}(a|s)F(\theta_k, s, a)\Big)}_{I_1} \\
&\quad + \underbrace{\sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k \big(\phi(s_{k+1})^\top v_k - \phi(s_{k+1})^\top v^*(\theta_k) - \phi(s_k)^\top v_k + \phi(s_k)^\top v^*(\theta_k)\big)\nabla \log \pi_{\theta_k}(a_k|s_k)}_{I_2} \\
&\quad + \underbrace{\sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k \big(\phi(s_{k+1})^\top v^*(\theta_k) - V^{\theta_k}(s_{k+1}) - \phi(s_k)^\top v^*(\theta_k) + V^{\theta_k}(s_k)\big)\nabla \log \pi_{\theta_k}(a_k|s_k)}_{I_3} \\
&\quad + \underbrace{\sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k \big((L_k - L(\theta_k))\nabla \log \pi_{\theta_k}(a_k|s_k)}_{I_4} \qquad (6)
\end{aligned}
$$

Here we use the definition that $z_{n_0} = \theta_{n_0}$.

For any $0 < k \leq m$,

$$\xi(m, k) + \xi(m, k+1)\beta_k = \xi(m, k+1),$$

and hence

$$\xi(m, n_0) + \sum_{k=n_0}^{m} \xi(m, k+1)\beta_k = \xi(m, m+1) = 1.$$

which implies

$$\sum_{k=n_0}^{m} \xi(m, k+1)\beta_k \leq 1$$

For term I_1 we have ,

$$
\begin{aligned}
I_1 &= \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k(F(\theta_k,s_k,a_k) - \sum_{s,a}\mu_{\theta_k}(s)\pi_{\theta_k}(a|s)F(\theta_k,s,a)) \\
&= \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k(U(\theta_k,s_k,a_k) - \sum_{s,a}\mu_{\theta_k}(s)\pi_{\theta_k}(a|s)U(\theta_k,s,a)) \\
&= \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k(U(\theta_k,s_k,a_k) - \sum_{a}\pi_{\theta_k}(a|s_k)U(\theta_k,s_k,a)) \\
&\quad + \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k(\sum_{a}\pi_{\theta_k}(a|s_k)U(\theta_k,s_k,a) - \sum_{s,a}\mu_{\theta_k}(s)\pi_{\theta_k}(a|s)U(\theta_k,s,a)) \\
&= \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k\Big(U(\theta_k,s_k,a_k) - \sum_{a}\pi_{\theta_k}(a|s_k)U(\theta_k,s_k,a)\Big) \\
&\quad + \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k(\bar{U}(\theta_k,s_k) - \sum_{s}P^{\theta_k}(s|s_k)\bar{U}(\theta_k,s)) \\
&\quad + \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k\Big(\sum_{s}(P^{\theta_k}(s|s_k) - \mu_{\theta_k}(s))\bar{U}(\theta_k,s)\Big) \\
&= \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k\Big(U(\theta_k,s_k,a_k) - \sum_{a}\pi_{\theta_k}(a|s_k)U(\theta_k,s_k,a)\Big) \\
&\quad + \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k(\bar{U}(\theta_k,s_{k+1}) - \sum_{s}P^{\theta_k}(s|s_k)\bar{U}(\theta_k,s)) \\
&\quad \underbrace{+ \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k(\bar{U}(\theta_k,s_k) - \bar{U}(\theta_k,s_{k+1}))}_{I_{1a}} \\
&\quad \underbrace{+ \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k\Big(\sum_{s}(P^{\theta_k}(s|s_k) - \mu_{\theta_k}(s))\bar{U}(\theta_k,s)\Big)}_{I_{1b}}
\end{aligned}
$$

where we define

$$
\begin{aligned}
U(\theta,s,a) &= F(\theta,s,a) - \theta, \\
\bar{U}(\theta,s) &= \sum_{a}\pi_\theta(a|s)U(\theta,s,a), \\
P^{\theta_k}(s|s_k) &= \sum_{a}P(s|s_k,a)\pi_{\theta_k}(a|s_k).
\end{aligned}
$$

Now for term I_{1a} we have,

$$
\begin{aligned}
&\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k(\bar{U}(\theta_k,s_k)-\bar{U}(\theta_k,s_{k+1})) \\
=&\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k(\bar{U}(\theta_k,s_k)-\bar{U}(\theta_{k+1},s_{k+1}))+\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k(\bar{U}(\theta_{k+1},s_{k+1})-\bar{U}(\theta_k,s_{k+1})) \\
=&\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k\bar{U}(\theta_k,s_k)-\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k\bar{U}(\theta_{k+1},s_{k+1}) \\
&\quad+\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k(\bar{U}(\theta_{k+1},s_{k+1})-\bar{U}(\theta_k,s_{k+1})) \\
=&\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k\bar{U}(\theta_k,s_k)-\sum_{k=n_0+1}^{k=m+1} \xi(m,k)\beta_{k-1}\bar{U}(\theta_k,s_k) \\
&\quad+\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k(\bar{U}(\theta_{k+1},s_{k+1})-\bar{U}(\theta_k,s_{k+1})) \\
=&\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k\bar{U}(\theta_k,s_k)-\sum_{k=n_0}^{k=m} \xi(m,k)\beta_{k-1}\bar{U}(\theta_k,s_k) \\
&\quad+\xi(m,n_0)\beta_{n_0-1}\bar{U}(\theta_{n_0},s_{n_0})-\xi(m,m+1)\beta_m\bar{U}(\theta_{m+1},s_{m+1}) \\
&\quad+\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k(\bar{U}(\theta_{k+1},s_{k+1})-\bar{U}(\theta_k,s_{k+1})) \\
=&\sum_{k=n_0}^{k=m} (\xi(m,k+1)\beta_k-\xi(m,k)\beta_{k-1})\bar{U}(\theta_k,s_k) \\
&\quad+\xi(m,n_0)\beta_{n_0-1}\bar{U}(\theta_{n_0},s_{n_0})-\xi(m,m+1)\beta_m\bar{U}(\theta_{m+1},s_{m+1}) \\
&\quad+\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k(\bar{U}(\theta_{k+1},s_{k+1})-\bar{U}(\theta_k,s_{k+1})) \\
=&\sum_{k=n_0}^{k=m} (\xi(m,k+1)\beta_k-\xi(m,k+1)\beta_{k-1})\bar{U}(\theta_k,s_k) \\
&\quad+\sum_{k=n_0}^{k=m} (\xi(m,k+1)\beta_{k-1})-\xi(m,k)\beta_{k-1})\bar{U}(\theta_k,s_k) \\
&\quad+\xi(m,n_0)\beta_{n_0-1}\bar{U}(\theta_{n_0},s_{n_0})-\xi(m,m+1)\beta_m\bar{U}(\theta_{m+1},s_{m+1}) \\
&\quad+\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k(\bar{U}(\theta_{k+1},s_{k+1})-\bar{U}(\theta_k,s_{k+1}))
\end{aligned}
$$

Hence we have ,

$$
\begin{aligned}
\|I_{1a}\| \leq & \Big\| \sum_{k=n_0}^{k=m} (\xi(m,k+1)\beta_k - \xi(m,k+1)\beta_{k-1})\bar{U}(\theta_k, s_k) \Big\| \\
& + \Big\| \sum_{k=n_0}^{k=m} (\xi(m,k+1)\beta_{k-1}) - \xi(m,k)\beta_{k-1})\bar{U}(\theta_k, s_k) \Big\| \\
& + \Big\| \xi(m,n_0)\beta_{n_0-1}\bar{U}(\theta_{n_0}, s_{n_0}) - \xi(m,m+1)\beta_m \bar{U}(\theta_{m+1}, s_{m+1}) \Big\| \\
& + \Big\| \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k(\bar{U}(\theta_{k+1}, s_{k+1}) - \bar{U}(\theta_k, s_{k+1})) \Big\| \\
\leq & \sum_{k=n_0}^{k=m} \xi(m,k+1)(\beta_{k-1} - \beta_k)U_{max} + \sum_{k=n_0}^{k=m} (\xi(m,k+1) - \xi(m,k))\beta_{k-1}U_{max} \\
& + 2\beta_{n_0-1}U_{max} + \mathcal{O}(\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k^2) \\
\leq & 4\beta_{n_0-1}U_{max} + \mathcal{O}(n_0^{-\sigma})
\end{aligned}
$$

For term I_{1b} we have,

$$
\begin{aligned}
& \Big\| \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k \Big(\sum_s (P^{\theta_k}(s|s_k) - \mu_{\theta_k}(s))\bar{U}(\theta_k, s) \Big) \Big\| \\
\leq & U_{max} \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k \sum_s |P^{\theta_k}(s|s_k) - \mu_{\theta_k}(s)| \\
\leq & 2U_{max} bk
\end{aligned}
$$

Hence we have ,

$$
\begin{aligned}
I_1 = & \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k \Big(U(\theta_k, s_k, a_k) - \sum_a \pi_{\theta_k}(a|s_k)U(\theta_k, s_k, a) + \bar{U}(\theta_k, s_{k+1}) - \sum_s P^{\theta_k}(s|s_k)\bar{U}(\theta_k, s) \Big) \\
& + \zeta(n_0, m)
\end{aligned}
$$

where

$$
\|\zeta(n_0, m)\| \leq 4\beta_{n_0-1}U_{max} + \mathcal{O}(n_0^{-\sigma}) + \mathcal{O}(bk)
$$

For term I_2 we have,

$$\begin{aligned}
&\left\| \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k(\phi(s_{k+1})^\top v_k - \phi(s_{k+1})^\top v^*(\theta_k) - \phi(s_k)^\top v_k + \phi(s_k)^\top v^*(\theta_k))\nabla \log \pi_{\theta_k}(a_k|s_k) \right\| \\
&\leq 2B \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k \|v_k - v^*(\theta_k)\| \\
&= 2B \sum_{k=n_0}^{k=m} \sqrt{\xi(m,k+1)^2\beta_k^2 \|v_k - v^*(\theta_k)\|^2} \\
&= 2B \sum_{k=n_0}^{k=m} \sqrt{\xi(m,k+1)\beta_k \|v_k - v^*(\theta_k)\|^2}\sqrt{\xi(m,k+1)\beta_k} \\
&\leq 2B \sqrt{\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k \|v_k - v^*(\theta_k)\|^2} \sqrt{\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k} \\
&\leq 2B \sqrt{\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k \|v_k - v^*(\theta_k)\|^2}
\end{aligned}$$

For term I_3 we have,

$$\left\| \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k(\phi(s_{k+1})^\top v^*(\theta_k) - V^{\theta_k}(s_{k+1}) - \phi(s_k)^\top v^*(\theta_k) + V^{\theta_k}(s_k))\nabla \log \pi_{\theta_k}(a_k|s_k) \right\| = \mathcal{O}(\epsilon_{\text{app}})$$

For term I_4 we have ,

$$\begin{aligned}
&\left\| \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k((L_k - L(\theta_k))\nabla \log \pi_{\theta_k}(a_k|s_k) \right\| \\
\leq & B \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k |L_k - L(\theta_k)| \\
\leq & B \sqrt{\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k |L_k - L(\theta_k)|^2}
\end{aligned}$$

Hence after gathering all the terms we have ,

$$
\begin{aligned}
&\|\theta_{m+1} - z_{m+1}\| \\
&\leq \left\| \sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k M_{k+1} \right\| \\
&+ \left\| \sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k \Big(\sum_{s,a} \mu_{\theta_k}(s)\pi_{\theta_k}(a|s)F(\theta_k, s, a) - \sum_{s,a} \mu_{z_k}(s)\pi_{z_k}(a|s)F(z_k, s, a) \Big) \right\| \\
&\quad + \left\| \sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k \left(U(\theta_k, s_k, a_k) - \sum_a \pi_{\theta_k}(a|s_k)U(\theta_k, s_k, a) + \bar{U}(\theta_k, s_{k+1}) - \sum_s P^{\theta_k}(s|s_k)\bar{U}(\theta_k, s) \right) \right\| \\
&\quad + 4\beta_{n_0-1}U_{max} + \mathcal{O}(n_0^{-\sigma}) + \mathcal{O}(\epsilon_{\text{app}}) \\
&\quad + 2B\sqrt{\sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k \|v_k - v^*(\theta_k)\|^2} \\
&\quad + B\sqrt{\sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k |L_k - L(\theta_k)|^2} \\
&\leq \left\| \sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k M_{k+1} \right\| \\
&\quad + \alpha \sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k \|\theta_k - z_k\| \\
&\quad + \left\| \sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k \left(U(\theta_k, s_k, a_k) - \sum_a \pi_{\theta_k}(a|s_k)U(\theta_k, s_k, a) + \bar{U}(\theta_k, s_{k+1}) - \sum_s P^{\theta_k}(s|s_k)\bar{U}(\theta_k, s) \right) \right\| \\
&\quad + \mathcal{O}(n_0^{-\sigma}) + \mathcal{O}(\epsilon_{\text{app}}) \\
&\quad + 2B\sqrt{\sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k \|v_k - v^*(\theta_k)\|^2} \\
&\quad + B\sqrt{\sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k |L_k - L(\theta_k)|^2}
\end{aligned}
$$

Now using the result of lemmas 5 and 6 we have ,

$$
\begin{aligned}
&\|\theta_{m+1} - z_{m+1}\| \\
&\leq \left\| \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k M_{k+1} \right\| \\
&\quad + \alpha \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k \|\theta_k - z_k\| \\
&\quad + \left\| \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k \left(U(\theta_k, s_k, a_k) - \sum_a \pi_{\theta_k}(a|s_k) U(\theta_k, s_k, a) + \bar{U}(\theta_k, s_{k+1}) - \sum_s P^{\theta_k}(s|s_k)\bar{U}(\theta_k, s) \right) \right\| \\
&\quad + \mathcal{O}(n_0^{-\sigma}) + \mathcal{O}(\epsilon_{\text{app}}) + \mathcal{O}(bk) + \mathcal{O}(\sqrt{bk}) \\
&\quad + \mathcal{O}(n_0^{(\nu-\sigma)/2}) + \mathcal{O}(n_0^{-\nu/2}) \\
&\quad + 2B \Bigg(\frac{2}{\lambda} \Bigg| \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k \langle z_k, W(s_k, a_k, s_{k+1}, L_k, v_k) - \sum_a \sum_{s'} \pi_{\theta_k}(a|s_k) P(s'|s_k, a) W(s_k, a, s', L_k, v_k) \rangle \\
&\quad + \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k \langle z_k, \bar{W}_{\theta_k}(s_{k+1}, L_k, v_k) - \sum_s P^{\theta_k}(s|s_k)\bar{W}_{\theta_k}(s, L_k, v_k) \rangle \Bigg| \Bigg)^{1/2} \\
&\quad + B\left(\sqrt{\frac{2}{\lambda}} + 1 \right) \sqrt{ \left| \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k y_k \left(r(s_k, a_k) - \sum_a \pi_{\theta_k}(a|s_k) r(s_k, a) + \bar{r}_{\theta_k}(s_{k+1}) - \sum_s P^{\theta_k}(s|s_k)\bar{r}_{\theta_k}(s) \right) \right| }
\end{aligned}
$$

Let

$$
\begin{aligned}
\tilde{U}_{k+1} &= U(\theta_k, s_k, a_k) - \sum_a \pi_{\theta_k}(a|s_k) U(\theta_k, s_k, a) + \bar{U}(\theta_k, s_{k+1}) - \sum_s P^{\theta_k}(s|s_k)\bar{U}(\theta_k, s) \\
\tilde{W}_{k+1} &= W(s_k, a_k, s_{k+1}, L_k, v_k) - \sum_a \sum_{s'} \pi_{\theta_k}(a|s_k) P(s'|s_k, a) W(s_k, a, s', L_k, v_k) \\
&\quad + \bar{W}_{\theta_k}(s_{k+1}, L_k, v_k) - \sum_s P^{\theta_k}(s|s_k)\bar{W}_{\theta_k}(s, L_k, v_k) \\
\tilde{Z}_{k+1} &= r(s_k, a_k) - \sum_a \pi_{\theta_k}(a|s_k) r(s_k, a) + \bar{r}_{\theta_k}(s_{k+1}) - \sum_s P^{\theta_k}(s|s_k)\bar{r}_{\theta_k}(s)
\end{aligned}
$$

We define the family of σ- fields : $\mathcal{A}_n := \sigma(s_0, \theta_m, L_m, v_m, s_m, a_{m-1}, 1 \leq m \leq n)$. Then $\{\tilde{W}_n\}, \{\tilde{U}_n\}, \{\tilde{Z}_n\}$ are martingale difference sequences, with respect to $\{\mathcal{A}_n\}$.

Hence we have,

$$
\begin{aligned}
&\|\theta_{m+1} - z_{m+1}\| \\
&\leq \left\| \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k M_{k+1} \right\| + \left\| \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k \tilde{U}_{k+1} \right\| + \alpha \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k \|\theta_k - z_k\| \\
&\quad + \mathcal{O}(n_0^{-\sigma}) + \mathcal{O}(\epsilon_{\text{app}}) + \mathcal{O}(n_0^{(\nu-\sigma)/2}) + \mathcal{O}(bk) + \mathcal{O}(\sqrt{bk}) \\
&\quad + 2B \sqrt{ \frac{2}{\lambda} \left| \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k \langle z_k, \tilde{W}_{k+1} \rangle \right| } + \mathcal{O}(n_0^{-\nu/2}) \\
&\quad + B\left(\sqrt{\frac{2}{\lambda}} + 1 \right) \left| \sqrt{ \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k y_k \tilde{Z}_{k+1} } \right|
\end{aligned}
$$

Now let us define $\theta_m^{'} = \sup_{n_0 \leq k \leq m} \|\theta_m - z_m\|$. Hence we have ,

$$\begin{aligned}\|\theta_{m+1} - z_{m+1}\| \leq & \left\| \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k M_{k+1} \right\| + \left\| \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k \tilde{U}_{k+1} \right\| + \alpha\theta_m^{'} \\ & + \mathcal{O}(n_0^{-\sigma}) + \mathcal{O}(\epsilon_{\text{app}}) + \mathcal{O}(n_0^{(\nu-\sigma)/2}) + \mathcal{O}(bk) + \mathcal{O}(\sqrt{bk}) \\ & + 2B\sqrt{\frac{2}{\lambda}\left| \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k \langle z_k, \tilde{W}_{k+1} \rangle \right|} + \mathcal{O}(n_0^{-\nu/2}) \\ & + B\left(\sqrt{\frac{2}{\lambda}} + 1\right) \left| \sqrt{\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k y_k \tilde{Z}_{k+1}} \right| \end{aligned} \tag{7}$$

Now we will obtain a bound on the probability :

$$P\left(\|\theta_m - \theta^*\| \leq e^{-(1-\alpha)b_{n_0}(m-1)}\varepsilon + \Delta(n_0,\delta), \forall n_0 \leq m \leq n \right),$$

for some $\varepsilon > 0, \delta > 0$. We define ,

$$\Delta(n_0,\delta) := \frac{4\delta + C_1\epsilon_{\text{app}} + C_2 n_0^{(\nu-\sigma)/2} + C_3 n_0^{-\nu/2} + C_4 n_0^{-\sigma} + C_5\sqrt{bk}}{1-\alpha}$$

where C_1, C_2, C_3, C_4, C_5 are positive constants.

From (5) we have ,

$$\|z_{n+1} - \theta^*\| \leq e^{-(1-\alpha)b_{n_0}(n)} \|\theta_{n_0} - \theta^*\|$$

and hence,

$$\|\theta_{n_0} - \theta^*\| \leq \varepsilon \implies \|z_n - \theta^*\| \leq e^{-(1-\alpha)b_{n_0}(n-1)}\varepsilon.$$

Also recall that $\theta_m^{'} = \sup_{n_0 \leq k \leq m} \|\theta_k - z_k\|$. Hence

$$\begin{aligned} & \{\|\theta_{n_0} - \theta^*\| \leq \varepsilon\} \bigcap \{\theta_n^{'} \leq \Delta(n_0,\delta)\} \\ & \subseteq \{\|\theta_m - \theta^*\| \leq e^{-(1-\alpha)b_{n_0}(m-1)}\varepsilon + \Delta(n_0,\delta), \forall n_0 \leq m \leq n\} \end{aligned}$$

This implies the following relation between the probabilities of the two sets.

$$\begin{aligned} & P\left(\|\theta_m - \theta^*\| \leq e^{-(1-\alpha)b_{n_0}(m-1)}\varepsilon + \Delta(n_0,\delta), \forall n_0 \leq m \leq n \right) \\ & \geq 1 - P\left(\{\|\theta_{n_0} - \theta^*\| > \varepsilon\} \bigcup \{\theta_n^{'} > \Delta(n_0,\delta)\} \right). \end{aligned} \tag{8}$$

Now we define $\xi = \{\theta_0, v_0, L_0, s_k, a_k, k \geq 0\}$ and the set B_m as

$$B_m = \left\{ \xi \middle| \theta_m^{'}(\xi) > \Delta(n_0,\delta) \right\}$$

For $\xi \notin B_{n-1}$, let us define $\bar{\theta}_{k,n-1}(\xi) = \theta_k(\xi)$ and $\bar{z}_{k,n-1}(\xi) = z_k(\xi)$ for all k. For $\xi \in B_{n-1}$, we define $\bar{\theta}_{k,n-1}(\xi) = \theta^*$ and $\bar{z}_{k,n-1}(\xi) = \theta^*$ for all k.

Also define $\bar{\theta}^{'}_{m,n-1}(\xi) = \sup_{n_0 \le k \le m} \|\bar{\theta}_{k,n-1}(\xi) - \bar{z}_{k,n-1}(\xi)\|$. Note that $\bar{\theta}^{'}_{m,n-1} = 0$ when $\xi \in B_{n-1}$ and $\bar{\theta}^{'}_{m,n-1} = \theta^{'}_{m} \le \Delta(n_0, \delta)$ when $\xi \notin B_{n-1}$.The intuition behind these definitions is that $\bar{\theta}^{'}_{m,n-1}$ is always bounded by $\Delta(n_0, \delta)$ for all $m \le n-1$.

Note that for $\xi \notin B_{n-1} \Rightarrow \bar{\theta}^{'}_{n,n-1}(\xi) = \theta^{'}_{n}(\xi)$ which implies $P(\bar{\theta}^{'}_{n,n-1}(\xi) \neq \theta^{'}_{n}(\xi)) \le P(B_{n-1})$. From this point onward, we suppress (ξ) in the notation, implicitly assuming that all random variables depend on (ξ).

Now,

$$
\begin{aligned}
P(\theta^{'}_{n} > \Delta(n_0, \delta)) &\le P(\bar{\theta}^{'}_{n,n-1} > \Delta(n_0, \delta) \bigcup \bar{\theta}^{'}_{n,n-1} \neq \theta^{'}_{n}) \\
&\le P(\bar{\theta}^{'}_{n,n-1} > \Delta(n_0, \delta)) + P(\bar{\theta}^{'}_{n,n-1} \neq \theta^{'}_{n}) \\
&\le P(\bar{\theta}^{'}_{n,n-1} > \Delta(n_0, \delta)) + P(B_{n-1}) \\
&= P(\bar{\theta}^{'}_{n,n-1} > \Delta(n_0, \delta)) + P(\theta^{'}_{n-1} > \Delta(n_0, \delta)) \qquad (9)
\end{aligned}
$$

Now we obtain a bound on $P(\bar{\theta}^{'}_{n,n-1} > \Delta(n_0, \delta))$ by induction. Note that $\bar{\theta}^{'}_{n-1,n-1}$ is bounded by $\Delta(n_0, \delta)$ by definition.

Hence $P\Big(\bar{\theta}^{'}_{n,n-1} > \Delta(n_0, \delta)\Big) = P\Big(\|\bar{\theta}_{n,n-1} - \bar{z}_{n,n-1}\| > \Delta(n_0, \delta)\Big)$.

Now restating (7) for $m = n-1$ we have,

$$
\begin{aligned}
\|\theta_n - z_n\| \le &\left\| \sum_{k=n_0}^{k=n-1} \xi(n-1, k+1)\beta_k M_{k+1} \right\| + \left\| \sum_{k=n_0}^{k=n-1} \xi(n-1, k+1)\beta_k \tilde{U}_{k+1} \right\| + \alpha\theta^{'}_{n-1} \\
&+ C_1\epsilon_{\text{app}} + C_2 n_0^{(\nu-\sigma)/2} + C_3 n_0^{-\nu/2} + C_4 n_0^{-\sigma} + C_5\sqrt{bk} \\
&+ 2B\sqrt{\frac{2}{\lambda}\left| \sum_{k=n_0}^{k=n-1} \xi(n-1, k+1)\beta_k \langle z_k, \tilde{W}_{k+1} \rangle \right|} \\
&+ B\left(\sqrt{\frac{2}{\lambda}} + 1\right)\sqrt{\left| \sum_{k=n_0}^{k=n-1} \xi(n-1, k+1)\beta_k y_k \tilde{Z}_{k+1} \right|}
\end{aligned}
$$

Now we have,

$$
\begin{aligned}
&\|\bar{\theta}_{n,n-1} - \bar{z}_{n,n-1}\| \\
=&\|\bar{\theta}_{n,n-1} - \bar{z}_{n,n-1}\| I\{\xi_{n-1} \in B_{n-1}\} + \|\bar{\theta}_{n,n-1} - \bar{z}_{n,n-1}\| I\{\xi_{n-1} \notin B_{n-1}\} \\
=&0 \times I\{\xi_{n-1} \in B_{n-1}\} + \|\bar{\theta}_{n,n-1} - \bar{z}_{n,n-1}\| I\{\xi_{n-1} \notin B_{n-1}\} \\
\leq& I\{\xi_{n-1} \notin B_{n-1}\} \Bigg(\left\| \sum_{k=n_0}^{k=n-1} \xi(n-1,k+1)\beta_k M_{k+1} \right\| + \left\| \sum_{k=n_0}^{k=n-1} \xi(n-1,k+1)\beta_k \tilde{U}_{k+1} \right\| + \alpha \theta^{'}_{n-1} \\
&+ C_1 \epsilon_{\text{app}} + C_2 n_0^{(\nu-\sigma)/2} + C_3 n_0^{-\nu/2} + C_4 n_0^{-\sigma} + C_5 \sqrt{bk} \\
&+ 2B \sqrt{\frac{2}{\lambda} \left| \sum_{k=n_0}^{k=n-1} \xi(n-1,k+1)\beta_k \langle z_k, \tilde{W}_{k+1} \rangle \right|} \\
&+ B\left(\sqrt{\frac{2}{\lambda}} + 1\right) \sqrt{\left| \sum_{k=n_0}^{k=n-1} \xi(n-1,k+1)\beta_k y_k \tilde{Z}_{k+1} \right|} \Bigg) \\
\leq& \left\| \sum_{k=n_0}^{k=n-1} \xi(n-1,k+1)\beta_k M_{k+1} \right\| + \left\| \sum_{k=n_0}^{k=n-1} \xi(n-1,k+1)\beta_k \tilde{U}_{k+1} \right\| + \alpha \Delta(n_0, \delta) \\
&+ C_1 \epsilon_{\text{app}} + C_2 n_0^{(\nu-\sigma)/2} + C_3 n_0^{-\nu/2} + C_4 n_0^{-\sigma} + C_5 \sqrt{bk} \\
&+ 2B \sqrt{\frac{2}{\lambda} \left| \sum_{k=n_0}^{k=n-1} \xi(n-1,k+1)\beta_k \langle z_k, \tilde{W}_{k+1} \rangle \right|} \\
&+ B\left(\sqrt{\frac{2}{\lambda}} + 1\right) \sqrt{\left| \sum_{k=n_0}^{k=n-1} \xi(n-1,k+1)\beta_k y_k \tilde{Z}_{k+1} \right|}
\end{aligned}
$$

Now ,

$$
\begin{aligned}
&\left\| \sum_{k=n_0}^{k=n-1} \xi(n-1,k+1)\beta_k M_{k+1} \right\| \leq \delta \\
&2B \sqrt{\frac{2}{\lambda} \left| \sum_{k=n_0}^{k=n-1} \xi(n-1,k+1)\beta_k \langle z_k, \tilde{W}_{k+1} \rangle \right|} \leq \delta \\
&\left\| \sum_{k=n_0}^{k=n-1} \xi(n-1,k+1)\beta_k \tilde{U}_{k+1} \right\| \leq \delta \\
&B\left(\sqrt{\frac{2}{\lambda}} + 1\right) \sqrt{\left| \sum_{k=n_0}^{k=n-1} \xi(n-1,k+1)\beta_k y_k \tilde{Z}_{k+1} \right|} \leq \delta
\end{aligned}
$$

implies the below:

$$
\|\bar{\theta}_{n,n-1} - \bar{z}_{n,n-1}\| \leq \Delta(n_0, \delta)
$$

Hence we have ,

$$
\begin{aligned}
&P(\|\bar{\theta}_{n,n-1} - \bar{z}_{n,n-1}\| \leq \Delta(n_0,\delta)) \\
&\geq P\Bigg(\Bigg\|\sum_{k=n_0}^{k=n-1}\xi(n-1,k+1)\beta_k M_{k+1}\Bigg\| \leq \delta \bigcap \Bigg\|\sum_{k=n_0}^{k=n-1}\xi(n-1,k+1)\beta_k \tilde{U}_{k+1}\Bigg\| \leq \delta \\
&\quad \bigcap 2B\sqrt{\frac{2}{\lambda}\Bigg|\sum_{k=n_0}^{k=n-1}\xi(n-1,k+1)\beta_k\langle z_k, \tilde{W}_{k+1}\rangle\Bigg|} \leq \delta \\
&\quad \bigcap B\Bigg(\sqrt{\frac{2}{\lambda}}+1\Bigg)\sqrt{\Bigg|\sum_{k=n_0}^{k=n-1}\xi(n-1,k+1)\beta_k y_k \tilde{Z}_{k+1}\Bigg|} \leq \delta\Bigg) \\
\Rightarrow &P(\|\bar{\theta}_{n,n-1} - \bar{z}_{n,n-1}\| > \Delta(n_0,\delta)) \\
&\leq P\Bigg(\Bigg\|\sum_{k=n_0}^{k=n-1}\xi(n-1,k+1)\beta_k M_{k+1}\Bigg\| > \delta \bigcup \Bigg\|\sum_{k=n_0}^{k=n-1}\xi(n-1,k+1)\beta_k \tilde{U}_{k+1}\Bigg\| > \delta \\
&\quad \bigcup 2B\sqrt{\frac{2}{\lambda}\Bigg|\sum_{k=n_0}^{k=n-1}\xi(n-1,k+1)\beta_k\langle z_k, \tilde{W}_{k+1}\rangle\Bigg|} > \delta \\
&\quad \bigcup B\Bigg(\sqrt{\frac{2}{\lambda}}+1\Bigg)\sqrt{\Bigg|\sum_{k=n_0}^{k=n-1}\xi(n-1,k+1)\beta_k y_k \tilde{Z}_{k+1}\Bigg|} > \delta\Bigg) \\
\Rightarrow &P(\|\bar{\theta}_{n,n-1} - \bar{z}_{n,n-1}\| > \Delta(n_0,\delta)) \\
&\leq P\Bigg(\Bigg\|\sum_{k=n_0}^{k=n-1}\xi(n-1,k+1)\beta_k M_{k+1}\Bigg\| > \delta\Bigg) + P\Bigg(\Bigg\|\sum_{k=n_0}^{k=n-1}\xi(n-1,k+1)\beta_k \tilde{U}_{k+1}\Bigg\| > \delta\Bigg) \\
&\quad + P\Bigg(2B\sqrt{\frac{2}{\lambda}\Bigg|\sum_{k=n_0}^{k=n-1}\xi(n-1,k+1)\beta_k\langle z_k, \tilde{W}_{k+1}\rangle\Bigg|} > \delta\Bigg) \\
&\quad + P\Bigg(B\Bigg(\sqrt{\frac{2}{\lambda}}+1\Bigg)\sqrt{\Bigg|\sum_{k=n_0}^{k=n-1}\xi(n-1,k+1)\beta_k y_k \tilde{Z}_{k+1}\Bigg|} > \delta\Bigg)
\end{aligned}
$$

Let us define

$$
\begin{aligned}
p_{n-1} = &P\Bigg(\Bigg\|\sum_{k=n_0}^{k=n-1}\xi(n-1,k+1)\beta_k M_{k+1}\Bigg\| > \delta\Bigg) + P\Bigg(\Bigg\|\sum_{k=n_0}^{k=n-1}\xi(n-1,k+1)\beta_k \tilde{U}_{k+1}\Bigg\| > \delta\Bigg) \\
&+ P\Bigg(2B\sqrt{\frac{2}{\lambda}\Bigg|\sum_{k=n_0}^{k=n-1}\xi(n-1,k+1)\beta_k\langle z_k, \tilde{W}_{k+1}\rangle\Bigg|} > \delta\Bigg) \\
&+ P\Bigg(B\Bigg(\sqrt{\frac{2}{\lambda}}+1\Bigg)\sqrt{\Bigg|\sum_{k=n_0}^{k=n-1}\xi(n-1,k+1)\beta_k y_k \tilde{Z}_{k+1}\Bigg|} > \delta\Bigg)
\end{aligned}
$$

Hence we have from (9),

$$
P(\theta_n^{'} > \Delta(n_0,\delta)) \leq p_{n-1} + P(\theta_{n-1}^{'} > \Delta(n_0,\delta))
$$

Then repeating the same procedure for B_{n-2} we have,

$$
P(\theta_{n-1}^{'} > \Delta(n_0,\delta)) \leq p_{n-2} + P(\theta_{n-2}^{'} > \Delta(n_0,\delta))
$$

Iterating this for $n_0 + 1 \leq m \leq n$ we get

$$P(\theta_n^{'} > \Delta(n_0, \delta)) \leq \sum_{m=n_0+1}^{n} p_{m-1}$$

Now , going back to (8) we have,

$$\begin{aligned}
&P\bigg(\|\theta_m - \theta^*\| \leq e^{-(1-\alpha)b_{n_0}(m-1)}\varepsilon + \Delta(n_0, \delta), \forall n_0 \leq m \leq n\bigg) \\
&\geq 1 - P\bigg(\|\theta_{n_0} - \theta^*\| > \varepsilon\bigg) - P\bigg(\theta_n^{'} > \Delta(n_0, \delta)\bigg) \\
&\geq 1 - P\bigg(\|\theta_{n_0} - \theta^*\| > \varepsilon\bigg) - \sum_{m=n_0+1}^{n} p_{m-1} \\
&\geq 1 - P\bigg(\|\theta_{n_0} - \theta^*\| > \varepsilon\bigg) - 8\sum_{m=n_0+1}^{n} e^{-\frac{D\delta^4}{\omega(m-1)}}
\end{aligned}$$

We have used lemma 1 in the last inequality.

Now, let A_n be the set

$$\bigg\{\|\theta_m - \theta^*\| \leq e^{-(1-\alpha)b_{n_0}(m-1)}\varepsilon + \Delta(n_0, \delta), \forall n_0 \leq m \leq n\bigg\}.$$

Then $\{A_n\}$ is a decreasing sequence of sets, i.e., $A_{n+1} \subseteq A_n$ for all $n \geq n_0$. Now let A be the set

$$\bigg\{\|\theta_m - \theta^*\| \leq e^{-(1-\alpha)b_{n_0}(m-1)}\varepsilon + \Delta(n_0, \delta), \forall m \geq n_0\bigg\}$$

Then $A = \cap_{n=0}^{\infty} A_n$. Hence $P(A) = \lim_{n\uparrow\infty} P(A_n)$. So we have,

$$\begin{aligned}
&P\bigg(\|\theta_m - \theta^*\| \leq e^{-(1-\alpha)b_{n_0}(m-1)}\varepsilon + \Delta(n_0, \delta), \forall m \geq n_0\bigg) \\
&\geq 1 - P\bigg(\|\theta_{n_0} - \theta^*\| > \varepsilon\bigg) - 8\sum_{m\geq n_0+1} e^{-\frac{D\delta^4}{\omega(m-1)}}
\end{aligned}$$

A.2 Proof of Corollary

We use $\varsigma_i, i = 1, 2, ...$ to denote different constants in the proof. For $\sigma = 1$ and c_β sufficiently large, we have $\omega(m) = 1/m$. Hence,

$$2\sum_{m\geq n_0} e^{-\frac{D\delta^4}{\omega(m)}} = 2\sum_{m\geq n_0} e^{-D_1 m\delta_1^2} \leq \varsigma_1 e^{-D_1 n_0 \delta_1^2}$$

Let $\epsilon_1 = \varsigma_1 e^{-Dn_0\delta^4}$ which gives us $\delta = \varsigma_2 n_0^{-1/4}\log^{1/4}\left(\frac{\varsigma_1}{\epsilon_1}\right)$. This choice of δ gives us

$$8\sum_{m\geq n_0} e^{-\frac{D\delta^4}{\omega(m)}} \leq \epsilon_1$$

We assume $c_\beta(1-\alpha) > 1$.

Now,

$$\begin{aligned}
&\exp\Bigg(-(1-\alpha)b_{n_0}(m-1)\Bigg)\\
=&\exp\Bigg(-(1-\alpha)\sum_{k=n_0}^{k=m-1}\beta_k\Bigg)\\
\leq&\left(\frac{n_0+1}{m+1}\right)^{c_\beta(1-\alpha)}\\
\leq&\frac{n_0+1}{m+1}
\end{aligned}$$

Hence we have , $\forall m \geq n_0 > 0$ and $\beta_{n_0} < 1$,

$$\begin{aligned}
&P\Bigg(\|\theta_m-\theta^*\| = O\Bigg(\frac{n_0}{m}\varepsilon+\left(\frac{1}{n_0}\right)^{1/4}\log^{1/4}\left(\frac{1}{\epsilon_1}\right)+n_0^{(\nu-1)/2}+n_0^{-\nu/2}+\epsilon_{\text{app}}+\sqrt{bk}\Bigg)\Bigg)\\
&\geq 1-\epsilon_1-P\Bigg(\|\theta_{n_0}-\theta^*\|>\varepsilon\Bigg)
\end{aligned}$$

From assumption 4 we have $b \leq \epsilon_{\text{app}}$. Hence we have, $\forall m \geq n_0 > 0$,

$$\begin{aligned}
&P\Bigg(\|\theta_m-\theta^*\| = O\Bigg(\frac{n_0}{m}\varepsilon+\left(\frac{1}{n_0}\right)^{1/4}\log^{1/4}\left(\frac{1}{\epsilon_1}\right)+n_0^{(\nu-1)/2}+n_0^{-\nu/2}+\epsilon_{\text{app}}+\sqrt{\epsilon_{\text{app}}}\Bigg)\Bigg)\\
&\geq 1-\epsilon_1-P\Bigg(\|\theta_{n_0}-\theta^*\|>\varepsilon\Bigg)
\end{aligned}$$

Now , for $\varepsilon \geq 1$, we have $\varepsilon \leq \varepsilon^2$. Hence,

$$\|\theta_{n_0}-\theta^*\|^2 > \epsilon^2 \implies \|\theta_{n_0}-\theta^*\|^2 > \epsilon$$

which implies the below:

$$\begin{aligned}
&\Bigg\{\|\theta_{n_0}-\theta^*\|^2>\epsilon^2\Bigg\}\subseteq\Bigg\{\|\theta_{n_0}-\theta^*\|^2>\epsilon\Bigg\}\\
\Rightarrow&P\Bigg(\|\theta_{n_0}-\theta^*\|^2>\epsilon^2\Bigg)\leq P\Bigg(\|\theta_{n_0}-\theta^*\|^2>\epsilon\Bigg)
\end{aligned}$$

Now , let $\varepsilon = \frac{\mathbb{E}[\|\theta_{n_0}-\theta^*\|^2]}{\epsilon_2}$. Then we have ,

$$\begin{aligned}
P\Bigg(\|\theta_{n_0}-\theta^*\|>\varepsilon\Bigg)=&P\Bigg(\|\theta_{n_0}-\theta^*\|^2>\varepsilon^2\Bigg)\\
\leq&P\Bigg(\|\theta_{n_0}-\theta^*\|^2>\varepsilon\Bigg)\\
=&P\Bigg(\|\theta_{n_0}-\theta^*\|^2>\frac{\mathbb{E}[\|\theta_{n_0}-\theta^*\|^2]}{\epsilon_2}\Bigg)\leq\epsilon_2
\end{aligned}$$

So we have , $\forall k \geq n_0 > 0$,

$$\begin{aligned}
&P\Bigg(\|\theta_k-\theta^*\| = O\Bigg(\frac{n_0}{k}\frac{\mathbb{E}[\|\theta_{n_0}-\theta^*\|^2]}{\epsilon_2}+\left(\frac{1}{n_0}\right)^{1/4}\log^{1/4}\left(\frac{1}{\epsilon_1}\right)+n_0^{(\nu-1)/2}+n_0^{-\nu/2}\\
&\quad+\epsilon_{\text{app}}+\sqrt{\epsilon_{\text{app}}}\Bigg)\Bigg)\geq 1-\epsilon_1-\epsilon_2
\end{aligned}$$

Now, we use the result of lemma 2.

Optimising over the values of ν, we have $\nu = 0.5$.

Hence finally we have, (ignoring the approximation error) $\forall k \geq n_0$, and sufficiently large n_0 such that $2m\beta_{n_0} < 1$,

$$P\Bigg(\|\theta_k - \theta^*\| = O\bigg(\frac{n_0^{3/4}}{k}\frac{1}{\sqrt{\epsilon_2}} + \bigg(\frac{1}{n_0}\bigg)^{1/4}\log^{1/4}\bigg(\frac{1}{\epsilon_1}\bigg) + \bigg(\frac{1}{n_0}\bigg)^{1/4}\bigg)\Bigg)$$
$$\geq 1 - \epsilon_1 - \epsilon_2$$

A.3 Proof of Lemmas

A.3.1 Proof of lemma 1

We have

$$p_{n-1} = P\bigg(\bigg\|\sum_{k=n_0}^{k=n-1}\xi(n-1,k+1)\beta_k M_{k+1}\bigg\| > \delta\bigg) + P\bigg(\bigg\|\sum_{k=n_0}^{k=n-1}\xi(n-1,k+1)\beta_k \tilde{U}_{k+1}\bigg\| > \delta\bigg)$$
$$+ P\bigg(2B\sqrt{\frac{2}{\lambda}\bigg|\sum_{k=n_0}^{k=n-1}\xi(n-1,k+1)\beta_k\langle z_k, \tilde{W}_{k+1}\rangle\bigg|} > \delta\bigg)$$
$$+ P\bigg(B\bigg(\sqrt{\frac{2}{\lambda}}+1\bigg)\sqrt{\bigg|\sum_{k=n_0}^{k=n-1}\xi(n-1,k+1)\beta_k y_k \tilde{Z}_{k+1}\bigg|} > \delta\bigg)$$

Now ,

$$\bigg\|\sum_{k=n_0}^{k=n-1}\xi(n-1,k+1)\beta_k M_{k+1}\bigg\|$$
$$\leq \sqrt{d}\max_{1\leq l\leq d}\bigg|\sum_{k=n_0}^{k=n-1}\xi(n-1,k+1)\beta_k (M_{k+1})^{(l)}\bigg|$$

Now note that

$$(M_{k+1})^{(l)} \leq 2U_v B$$

We let $C = \exp(2U_v B), \varepsilon = 1, \zeta_{k,m} = \xi(m,k+1)\beta_k, \gamma_1 = 1$. Also,

$$\xi(m,k+1) = \prod_{i=k+1}^{m}(1-\beta_i) \leq \prod_{i=k+1}^{m} e^{-\beta_i} = exp\bigg(-\sum_{k+1}^{m}\beta_i\bigg)$$
$$= exp\bigg(-\sum_{k+1}^{m}\frac{c_\beta}{(1+i)^\sigma}\bigg) \leq exp\bigg(-\sum_{k+1}^{m}\frac{c_\beta}{(1+i)}\bigg)$$
$$\leq exp\bigg(-\int_{k+1}^{m+1}\frac{c_\beta}{(y+1)}\,dy\bigg) = exp\bigg(c_\beta(\log(k+2)-\log(m+2))\bigg)$$
$$= exp\bigg(c_\beta\log\frac{k+2}{m+2}\bigg) = \bigg(\frac{k+2}{m+2}\bigg)^{c_\beta}$$

Hence we have,

$$\max_{n_0\leq k\leq m}\xi(m,k+1)\beta_k \leq \max_{n_0\leq k\leq m}\bigg(\frac{k+2}{m+2}\bigg)^{c_\beta}\frac{c_\beta}{(1+k)^\sigma} \leq \max_{n_0\leq k\leq m}\bigg(\frac{3k}{m}\bigg)^{c_\beta}\frac{c_\beta}{(1+k)^\sigma}$$
$$\leq \max_{n_0\leq k\leq m}\bigg(\frac{3k}{m}\bigg)^{c_\beta}\frac{c_\beta}{(k)^\sigma}$$

Now if $c_\beta \leq \sigma$, we have,

$$\max_{n_0 \leq k \leq m} \xi(m, k+1)\beta_k \leq 3^{c_\beta} c_\beta (n_0)^{c_\beta - \sigma} m^{-c_\beta}$$

Otherwise we have,

$$\max_{n_0 \leq k \leq m} \xi(m, k+1)\beta_k \leq 3^{c_\beta} c_\beta m^{-\sigma}$$

Let $\gamma_2 = 3^{c_\beta} c_\beta$. Also,

$$\omega(m) = \begin{cases} (n_0)^{c_\beta - \sigma} m^{-c_\beta}, & \text{if } c_\beta \leq \sigma, \\ m^{-\sigma}, & \text{if } otherwise. \end{cases}$$

So we have,

$$\max_{n_0 \leq k \leq m} \zeta_{k,m} \leq \gamma_2 \omega(m)$$

So if we have, $\gamma_1 = 1, \epsilon = 1, C = \exp(2U_v B)$ and if we apply the result of theorem 2 of Chandak & Borkar (2026), then ,

$$P\left(\left| \sum_{k=n_0}^{m} \xi(m, k+1)\beta_k \left(M_{k+1}\right)^{(l)} \right| > \delta\right) \leq 2e^{-\frac{D\delta_1^2}{\omega(m)}}, if \delta \in \left(0, C\gamma_1/\epsilon\right]$$
$$2e^{-\frac{D\delta_1}{\omega(m)}} otherwise.$$

Now,

$$\begin{aligned} &\left\{ \left\| \sum_{k=n_0}^{m} \xi(m, k+1)\beta_k \left(M_{k+1}\right) \right\| > \delta \right\} \\ &\quad \subseteq \left\{ \max_{1 \leq l \leq d} \left| \sum_{k=n_0}^{m} \xi(m, k+1)\beta_k \left(M_{k+1}\right)^{(l)} \right| > \frac{\delta}{\sqrt{d}} \right\} \\ \Rightarrow &P\left(\left\| \sum_{k=n_0}^{m} \xi(m, k+1)\beta_k \left(M_{k+1}\right) \right\| > \delta \right) \\ &\quad \leq P\left(\max_{1 \leq l \leq d} \left| \sum_{k=n_0}^{m} \xi(m, k+1)\beta_k \left(M_{k+1}\right)^{(l)} \right| > \frac{\delta}{\sqrt{d}} \right) \\ &\quad \leq 2e^{-\frac{D\delta_1^2}{d\omega(m)}}, if \delta \in \left(0, C\gamma_1\sqrt{d}/\epsilon\right] \\ &\qquad 2e^{-\frac{D\delta}{\sqrt{d}\omega(m)}} otherwise \end{aligned}$$

Under the assumption that $\delta \leq 1$ we have,

$$P\left(\left\| \sum_{k=n_0}^{m} \xi(m, k+1)\beta_k \left(M_{k+1}\right) \right\| > \delta \right) \leq 2e^{-\frac{D_1\delta^2}{\omega(m)}}$$

for some constant $D_1 > 0$.

Similarly under the assumption that $\delta \leq 1$ we have ,

$$P\left(2B\sqrt{\frac{2}{\lambda}\left| \sum_{k=n_0}^{k=n-1} \xi(n-1, k+1)\beta_k \langle z_k, \tilde{W}_{k+1} \rangle \right|} > \delta \right) \leq 2e^{-\frac{D_2\delta^4}{\omega(n-1)}}$$

for some constant $D_2 > 0$,

$$P\left(\left\| \sum_{k=n_0}^{k=n-1} \xi(n-1,k+1)\beta_k \tilde{U}_{k+1} \right\| > \delta\right) \leq 2e^{-\frac{D_3\delta^2}{\omega(n-1)}}$$

for some constant $D_3 > 0$,

$$P\left(B\left(\sqrt{\frac{2}{\lambda}}+1\right)\sqrt{\left| \sum_{k=n_0}^{k=n-1} \xi(n-1,k+1)\beta_k y_k \tilde{Z}_{k+1} \right|} > \delta\right) \leq 2e^{-\frac{D_4\delta^4}{\omega(n-1)}}$$

for some constant $D_4 > 0$.

Hence we have ,

$$\begin{aligned} p_{n-1} \leq & 2e^{-\frac{D_1\delta^2}{\omega(n-1)}} + 2e^{-\frac{D_2\delta^4}{\omega(n-1)}} + 2e^{-\frac{D_3\delta^2}{\omega(n-1)}} + 2e^{-\frac{D_4\delta^4}{\omega(n-1)}} \\ \leq & 2e^{-\frac{D_1\delta^4}{\omega(n-1)}} + 2e^{-\frac{D_2\delta^4}{\omega(n-1)}} + 2e^{-\frac{D_3\delta^4}{\omega(n-1)}} + 2e^{-\frac{D_4\delta^4}{\omega(n-1)}} \\ \leq & 8e^{-\frac{D\delta^4}{\omega(n-1)}} \end{aligned}$$

where $D = \min\{D_1, D_2, D_3, D_4\}$.

A.3.2 Proof of lemma 2

We have,

$$
\begin{aligned}
&\|\theta_{t+1}-\theta^*\|^2\\
=&\|\theta_t+\beta_t\delta_t\nabla\log\pi_{\theta_t}(a_t|s_t)-\theta^*\|^2\\
=&\|\theta_t-\theta^*\|^2+2\beta_t\langle\theta_t-\theta^*,\delta_t\nabla\log\pi_{\theta_t}(a_t|s_t)\rangle+\beta_t^2\|\delta_t\nabla\log\pi_{\theta_t}(a_t|s_t)\|^2\\
=&\|\theta_t-\theta^*\|^2+2\beta_t\langle\theta_t-\theta^*,(r_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\\
&\quad+\beta_t^2\|\delta_t\nabla\log\pi_{\theta_t}(a_t|s_t)\|^2\\
=&\|\theta_t-\theta^*\|^2+2\beta_t\langle\theta_t-\theta^*,(r_t-L(\theta_t)+V^{\theta_t}(s_{t+1})-V^{\theta_t}(s_t))\nabla\log\pi_{\theta_t}(a_t|s_t)\rangle\\
&\quad+\beta_t^2\|\delta_t\nabla\log\pi_{\theta_t}(a_t|s_t)\|^2\\
&\quad+2\beta_t\langle\theta_t-\theta^*,(L(\theta_t)-L_t+\phi(s_{t+1})^\top v_t-V^{\theta_t}(s_{t+1})-\phi(s_t)^\top v_t+V^{\theta_t}(s_t))\nabla\log\pi_{\theta_t}(a_t|s_t)\rangle\\
=&\|\theta_t-\theta^*\|^2+2\beta_t\langle\theta_t-\theta^*,(r_t-L(\theta_t)+V^{\theta_t}(s_{t+1})-V^{\theta_t}(s_t))\nabla\log\pi_{\theta_t}(a_t|s_t)\rangle\\
&\quad+\beta_t^2\|\delta_t\nabla\log\pi_{\theta_t}(a_t|s_t)\|^2\\
&\quad+2\beta_t\langle\theta_t-\theta^*,(L(\theta_t)-L_t+\phi(s_{t+1})^\top(v_t-v^*(\theta_t))-\phi(s_t)^\top(v_t-v^*(\theta_t)))\nabla\log\pi_{\theta_t}(a_t|s_t)\rangle\\
&\quad+2\beta_t\langle\theta_t-\theta^*,(\phi(s_{t+1})^\top v^*(\theta_t)-V^{\theta_t}(s_{t+1})-\phi(s_t)^\top v^*(\theta_t)+V^{\theta_t}(s_t))\nabla\log\pi_{\theta_t}(a_t|s_t)\rangle\\
=&\|\theta_t-\theta^*\|^2+2\beta_t\langle\theta_t-\theta^*,\mathbb{E}_{\theta_t}[(r_t-L(\theta_t)+V^{\theta_t}(s_{t+1})-V^{\theta_t}(s_t))\nabla\log\pi_{\theta_t}(a_t|s_t)]\rangle\\
&\quad+2\beta_t\langle\theta_t-\theta^*,H(O_t,\theta_t)-\mathbb{E}_{\theta_t}[H(O,\theta_t)]\rangle\\
&\quad+\beta_t^2\|\delta_t\nabla\log\pi_{\theta_t}(a_t|s_t)\|^2\\
&\quad+2\beta_t\langle\theta_t-\theta^*,(L(\theta_t)-L_t+\phi(s_{t+1})^\top(v_t-v^*(\theta_t))-\phi(s_t)^\top(v_t-v^*(\theta_t)))\nabla\log\pi_{\theta_t}(a_t|s_t)\rangle\\
&\quad+2\beta_t\langle\theta_t-\theta^*,(\phi(s_{t+1})^\top v^*(\theta_t)-V^{\theta_t}(s_{t+1})-\phi(s_t)^\top v^*(\theta_t)+V^{\theta_t}(s_t))\nabla\log\pi_{\theta_t}(a_t|s_t)\rangle\\
=&\|\theta_t-\theta^*\|^2+2\beta_t\langle\theta_t-\theta^*,\nabla L(\theta_t)-\nabla L(\theta^*)\rangle\\
&\quad+2\beta_t\langle\theta_t-\theta^*,H(O_t,\theta_t)-\mathbb{E}_{\theta_t}[H(O,\theta_t)]\rangle\\
&\quad+\beta_t^2\|\delta_t\nabla\log\pi_{\theta_t}(a_t|s_t)\|^2\\
&\quad+2\beta_t\langle\theta_t-\theta^*,(L(\theta_t)-L_t+\phi(s_{t+1})^\top(v_t-v^*(\theta_t))-\phi(s_t)^\top(v_t-v^*(\theta_t)))\nabla\log\pi_{\theta_t}(a_t|s_t)\rangle\\
&\quad+2\beta_t\langle\theta_t-\theta^*,(\phi(s_{t+1})^\top v^*(\theta_t)-V^{\theta_t}(s_{t+1})-\phi(s_t)^\top v^*(\theta_t)+V^{\theta_t}(s_t))\nabla\log\pi_{\theta_t}(a_t|s_t)\rangle\\
\le&\|\theta_t-\theta^*\|^2-2m\beta_t\|\theta_t-\theta^*\|^2+2\beta_t\langle\theta_t-\theta^*,H(O_t,\theta_t)-\mathbb{E}_{\theta_t}[H(O,\theta_t)]\rangle+\beta_t^2\|\delta_t\nabla\log\pi_{\theta_t}(a_t|s_t)\|^2\\
&\quad+2\beta_tB\|\theta_t-\theta^*\||L(\theta_t)-L_t|+4\beta_tB\|\theta_t-\theta^*\|\|v_t-v^*(\theta_t)\|+4\beta_tB\|\theta_t-\theta^*\|\epsilon_{\text{app}}\\
\le&(1-2m\beta_t)\|\theta_t-\theta^*\|^2+2\beta_t\langle\theta_t-\theta^*,H(O_t,\theta_t)-\mathbb{E}_{\theta_t}[H(O,\theta_t)]\rangle+\beta_t^2U_\delta^2B^2\\
&\quad+\frac{2}{m}\beta_tB^2U_\delta|L(\theta_t)-L_t|+\frac{4}{m}\beta_tB^2U_\delta\|v_t-v^*(\theta_t)\|+\frac{4}{m}\beta_tB^2U_\delta\epsilon_{\text{app}}
\end{aligned}
$$

Hence we have ,

$$
\begin{aligned}
&\mathbb{E}[\|\theta_{t+1}-\theta^*\|^2]\\
\le&(1-2m\beta_t)\mathbb{E}[\|\theta_t-\theta^*\|^2]+2\beta_t\mathbb{E}\langle\theta_t-\theta^*,H(O_t,\theta_t)-\mathbb{E}_{\theta_t}[H(O,\theta_t)]\rangle+\beta_t^2U_\delta^2B^2\\
&\quad+\frac{2}{m}\beta_tB^2U_\delta\mathbb{E}|L(\theta_t)-L_t|+\frac{4}{m}\beta_tB^2U_\delta\mathbb{E}\|v_t-v^*(\theta_t)\|+\frac{4}{m}\beta_tB^2U_\delta\epsilon_{\text{app}}
\end{aligned}
$$

Now, analysing similar to lemma C.3 in Wu et al. (2022), we have,

$$
|\mathbb{E}\langle\theta_t-\theta^*,H(O_t,\theta_t)-\mathbb{E}_{\theta_t}[H(O,\theta_t)]\rangle|=\mathcal{O}(1/t)+\mathcal{O}(\epsilon_{\text{app}})
$$

Hence we have ,

$$
\begin{aligned}
&\mathbb{E}[\|\theta_{t+1}-\theta^*\|^2]\\
\leq&(1-2m\beta_t)\mathbb{E}[\|\theta_t-\theta^*\|^2]+2\beta_t(C_1/t+C_2\epsilon_{\text{app}})+\beta_t^2U_\delta^2B^2\\
&+\frac{2}{m}\beta_tB^2U_\delta\mathbb{E}|L(\theta_t)-L_t|+\frac{4}{m}\beta_tB^2U_\delta\mathbb{E}\|v_t-v^*(\theta_t)\|+\frac{4}{m}\beta_tB^2U_\delta\epsilon_{\text{app}}\\
\Rightarrow&\mathbb{E}[\|\theta_{t+1}-\theta^*\|^2]\\
\leq&\underbrace{\xi(t,l)\mathbb{E}\|\theta_l-\theta^*\|}_{I_1}+\underbrace{2\sum_{k=l}^{k=t}\Xi(t,k+1)\beta_k(C_1/k+C_2\epsilon_{\text{app}})}_{I_2}+\underbrace{U_\delta^2B^2\sum_{k=l}^{k=t}\Xi(t,k+1)\beta_k^2}_{I_3}\\
&+\underbrace{\frac{2}{m}B^2U_\delta\sum_{k=l}^{k=t}\Xi(t,k+1)\beta_k\mathbb{E}|L(\theta_t)-L_t|}_{I_4}+\underbrace{\frac{4}{m}B^2U_\delta\sum_{k=l}^{k=t}\Xi(t,k+1)\beta_k\mathbb{E}\|v_t-v^*(\theta_t)\|}_{I_5}\\
&+\underbrace{\frac{4}{m}B^2U_\delta\epsilon_{\text{app}}\sum_{k=l}^{k=t}\Xi(t,k+1)\beta_k}_{I_6}
\end{aligned}
$$

We have $l=\min\{i\geq 1|2m\beta_i<1\}$.

For term I_1 we have,

$$
\begin{aligned}
&\Xi(t,l)\mathbb{E}\|\theta_0-\theta^*\|\\
=&\prod_{k=l}^{k=t}(1-2m\beta_k)\mathbb{E}\|\theta_0-\theta^*\|\\
\leq&\frac{U_\delta B}{m}\prod_{k=l}^{k=t}\exp(-2m\beta_k)\\
=&\frac{U_\delta B}{m}\exp(-2m\sum_{k=l}^{k=t}\beta_k)\\
=&\frac{U_\delta B}{m}\exp\Big(-2mc_\beta\sum_{k=l}^{k=t}\frac{1}{1+k}\Big)\\
\leq&\frac{U_\delta B}{m}\exp\Big(-2mc_\beta\int_{k=l}^{k=t+1}\frac{1}{1+k}\Big)\\
=&\frac{U_\delta B}{m}\exp\Big(2mc_\beta(\log(l+1)-\log(t+2))\Big)\\
=&\frac{U_\delta B}{m}\exp\Big(2mc_\beta\log\Big(\frac{1+l}{t+2}\Big)\Big)\\
=&\frac{U_\delta B}{m}\Big(\frac{1+l}{t+2}\Big)^{2mc_\beta}\\
\leq&\frac{U_\delta B}{m}\Big(\frac{1+l}{t+2}\Big)
\end{aligned}
$$

Now we have $2mc_\beta>l$. Hence $I_1=\mathcal{O}(1/t)$.

For term I_2 we have,

$$
I_2=2\sum_{k=l}^{k=t}\Xi(t,k+1)\beta_k(C_1/k+C_2\epsilon_{\text{app}})
$$

Now, $\Xi(t, k+1) \leq \frac{k+2}{t}$. Hence we have,

$$\begin{aligned} I_2 =& \frac{2C_1}{t} \sum_{k=l}^{k=t} (k+2)\beta_k/k + \mathcal{O}(\epsilon_{\text{app}}) \\ =& \frac{2C_1 c_\beta}{t} \sum_{k=l}^{k=t} (k+2)/(k(1+k)) + \mathcal{O}(\epsilon_{\text{app}}) \\ =& \mathcal{O}\left(\frac{\log t}{t}\right) + \mathcal{O}(\epsilon_{\text{app}}) \end{aligned}$$

For term I_3 we have,

$$\begin{aligned} & U_\delta^2 B^2 \sum_{k=l}^{k=t} \Xi(t, k+1)\beta_k^2 \\ \leq & \frac{U_\delta^2 B^2 c_\beta^2}{t} \sum_{k=l}^{k=t} (k+2)/(k+1)^2 \\ =& \mathcal{O}\left(\frac{\log t}{t}\right) \end{aligned}$$

For term I_4 we have,

$$\begin{aligned} I_4 &= \frac{2}{m} B^2 U_\delta \sum_{k=l}^{k=t} \Xi(t, k+1)\beta_k \mathbb{E}|L(\theta_k) - L_k| \\ &= \frac{2}{m} B^2 U_\delta \sum_{k=l}^{k=t} \sqrt{\Xi(t, k+1)^2 \beta_k^2 (\mathbb{E}|L(\theta_k) - L_k|)^2} \\ &= \frac{2}{m} B^2 U_\delta \sum_{k=l}^{k=t} \sqrt{\Xi(t, k+1)\beta_k (\mathbb{E}|L(\theta_k) - L_k|)^2} \sqrt{\Xi(t, k+1)\beta_k} \\ &\leq \frac{2}{m} B^2 U_\delta \sqrt{\sum_{k=l}^{k=t} \Xi(t, k+1)\beta_k (\mathbb{E}|L(\theta_k) - L_k|)^2} \sqrt{\sum_{k=l}^{k=t} \Xi(t, k+1)\beta_k} \\ &\leq \frac{2c_\beta}{m\sqrt{m}} B^2 U_\delta \sqrt{\sum_{k=l}^{k=t} \frac{k+2}{t+2} \frac{1}{1+k} \mathbb{E}|L(\theta_k) - L_k|^2} \\ &\leq \frac{4c_\beta}{m\sqrt{m}} B^2 U_\delta \sqrt{\frac{1}{t} \sum_{k=l}^{k=t} \mathbb{E}|L(\theta_k) - L_k|^2} \end{aligned}$$

Carrying out analysis similar to (11), we have,

$$\sum_{k=l}^{k=t}\mathbb{E}y_k^2 \leq \underbrace{\sum_{k=l}^{k=t}\mathbb{E}\left[\frac{y_k^2-y_{k+1}^2}{2\gamma_k}\right]}_{I_a}+\underbrace{\sum_{k=l}^{k=t}\mathbb{E}[y_k(r_k-L(\theta_k))]}_{I_b}+\underbrace{\sum_{k=l}^{k=t}\mathbb{E}\left[\frac{y_k}{\gamma_k}(L(\theta_k)-L(\theta_{k+1}))\right]}_{I_c}$$
$$+\underbrace{\sum_{k=l}^{k=t}\mathbb{E}\left[\frac{1}{\gamma_k}(L(\theta_k)-L(\theta_{k+1}))^2\right]}_{I_d}$$
$$+\underbrace{\sum_{k=l}^{k=t}\mathbb{E}[\gamma_k(r_k-L_k)^2]}_{I_e} \tag{10}$$

Now, analysing terms $I_a - I_e$ similar to the analysis in section C.2 in Wu et al. (2022), we have,

$$I_4 = \mathcal{O}(t^{-\nu/2}) + \mathcal{O}(t^{(\nu-1)/2}) + \mathcal{O}(\sqrt{\epsilon_{\text{app}}})$$

Similarly, after analysing term I_5 similar to section C.3 in Wu et al. (2022), we have,

$$I_5 = \mathcal{O}(t^{-\nu/2}) + \mathcal{O}(t^{(\nu-1)/2}) + \mathcal{O}(\sqrt{\epsilon_{\text{app}}})$$

We take $\tau = 1$ for the results of lemmas C.5 and C.7 in Wu et al. (2022).
For term I_6, we have,

$$\frac{4}{m}B^2U_\delta\epsilon_{\text{app}}\sum_{k=l}^{k=t}\Xi(t,k+1)\beta_k = \mathcal{O}(\epsilon_{\text{app}})$$

Hence we have,

$$\mathbb{E}[\|\theta_{t+1}-\theta^*\|^2] = \mathcal{O}\left(\frac{\log t}{t}\right) + \mathcal{O}(t^{-\nu/2}) + \mathcal{O}(t^{(\nu-1)/2}) + \mathcal{O}(\epsilon_{\text{app}}) + \mathcal{O}(\sqrt{\epsilon_{\text{app}}})$$

After simplification we have (ignoring the approximation error),

$$\mathbb{E}[\|\theta_{t+1}-\theta^*\|^2] = \mathcal{O}(t^{-\nu/2}) + \mathcal{O}(t^{(\nu-1)/2})$$

A.3.3 Proof of Lemma 3

We have,

$$
\begin{aligned}
\nabla L(\theta) &= \sum_s \sum_a \mu_\theta(s)\pi_\theta(a|s)(r(s,a) - L(\theta) + \sum_{s'} P(s'|s,a)V^\theta(s') - V^\theta(s))\nabla \log \pi_\theta(a|s) \\
&= \sum_s \sum_a \mu_\theta(s)\pi_\theta(a|s)(r(s,a) + \sum_{s'} P(s'|s,a)V^\theta(s'))\nabla \log \pi_\theta(a|s) \\
&\quad - \sum_s \sum_a \mu_\theta(s)\pi_\theta(a|s)(L(\theta) + V^\theta(s))\nabla \log \pi_\theta(a|s) \\
&= \sum_s \sum_a \mu_\theta(s)\pi_\theta(a|s)(r(s,a) + \sum_{s'} P(s'|s,a)V^\theta(s'))\nabla \log \pi_\theta(a|s) \\
&\quad - \sum_s \mu_\theta(s)(L(\theta) + V^\theta(s)) \sum_a \pi_\theta(a|s)\nabla \log \pi_\theta(a|s) \\
&= \sum_s \sum_a \mu_\theta(s)\pi_\theta(a|s)(r(s,a) + \sum_{s'} P(s'|s,a)V^\theta(s'))\nabla \log \pi_\theta(a|s) \\
&\quad - \sum_s \mu_\theta(s)(L(\theta) + V^\theta(s)) \sum_a \nabla \pi_\theta(a|s) \\
&= \sum_s \sum_a \mu_\theta(s)\pi_\theta(a|s)(r(s,a) + \sum_{s'} P(s'|s,a)V^\theta(s'))\nabla \log \pi_\theta(a|s) \\
&\quad - \sum_s \mu_\theta(s)(L(\theta) + V^\theta(s)) \times 0 \\
&= \sum_s \sum_a \mu_\theta(s)\pi_\theta(a|s)(r(s,a) + \sum_{s'} P(s'|s,a)V^\theta(s'))\nabla \log \pi_\theta(a|s) \\
&= \sum_s \sum_a \mu_\theta(s)(r(s,a) + \sum_{s'} P(s'|s,a)V^\theta(s'))\nabla \pi_\theta(a|s)
\end{aligned}
$$

Now,

$$
\begin{aligned}
&\|\nabla L(\theta_1) - L(\theta_2)\| \\
=&\Big\| \sum_s \sum_a \mu_{\theta_1}(s)\Big(r(s,a) + \sum_{s'} P(s'|s,a)V^{\theta_1}(s')\Big)\nabla \pi_{\theta_1}(a|s) \\
&\quad - \sum_s \sum_a \mu_{\theta_2}(s)\Big(r(s,a) + \sum_{s'} P(s'|s,a)V^{\theta_2}(s')\Big)\nabla \pi_{\theta_2}(a|s)\Big\|
\end{aligned}
$$

Let $X_\theta(s,a) = r(s,a) + \sum_{s'} P(s'|s,a)V^\theta(s')$.

Hence we have ,

$$
\begin{aligned}
&\|\nabla L(\theta_1) - L(\theta_2)\| \\
=&\left\| \sum_s \sum_a \mu_{\theta_1}(s) X_{\theta_1}(s,a) \nabla \pi_{\theta_1}(a|s) - \sum_s \sum_a \mu_{\theta_2}(s) X_{\theta_2}(s,a) \nabla \pi_{\theta_2}(a|s) \right\| \\
=&\Bigg\| \sum_s \sum_a \mu_{\theta_1}(s) X_{\theta_1}(s,a) \nabla \pi_{\theta_1}(a|s) - \sum_s \sum_a \mu_{\theta_2}(s) X_{\theta_1}(s,a) \nabla \pi_{\theta_1}(a|s) \\
&\quad + \sum_s \sum_a \mu_{\theta_2}(s) X_{\theta_1}(s,a) \nabla \pi_{\theta_1}(a|s) - \sum_s \sum_a \mu_{\theta_2}(s) X_{\theta_2}(s,a) \nabla \pi_{\theta_2}(a|s) \Bigg\| \\
=&\Bigg\| \sum_s \sum_a \mu_{\theta_1}(s) X_{\theta_1}(s,a) \nabla \pi_{\theta_1}(a|s) - \sum_s \sum_a \mu_{\theta_2}(s) X_{\theta_1}(s,a) \nabla \pi_{\theta_1}(a|s) \\
&\quad + \sum_s \sum_a \mu_{\theta_2}(s) X_{\theta_1}(s,a) \nabla \pi_{\theta_1}(a|s) - \sum_s \sum_a \mu_{\theta_2}(s) X_{\theta_2}(s,a) \nabla \pi_{\theta_1}(a|s) \\
&\quad + \sum_s \sum_a \mu_{\theta_2}(s) X_{\theta_2}(s,a) \nabla \pi_{\theta_1}(a|s) - \sum_s \sum_a \mu_{\theta_2}(s) X_{\theta_2}(s,a) \nabla \pi_{\theta_2}(a|s) \Bigg\| \\
\leq&\left\| \sum_s \sum_a \mu_{\theta_1}(s) X_{\theta_1}(s,a) \nabla \pi_{\theta_1}(a|s) - \sum_s \sum_a \mu_{\theta_2}(s) X_{\theta_1}(s,a) \nabla \pi_{\theta_1}(a|s) \right\| \\
&\quad + \left\| \sum_s \sum_a \mu_{\theta_2}(s) X_{\theta_1}(s,a) \nabla \pi_{\theta_1}(a|s) - \sum_s \sum_a \mu_{\theta_2}(s) X_{\theta_2}(s,a) \nabla \pi_{\theta_1}(a|s) \right\| \\
&\quad + \left\| \sum_s \sum_a \mu_{\theta_2}(s) X_{\theta_2}(s,a) \nabla \pi_{\theta_1}(a|s) - \sum_s \sum_a \mu_{\theta_2}(s) X_{\theta_2}(s,a) \nabla \pi_{\theta_2}(a|s) \right\| \\
=&\left\| \sum_s \sum_a (\mu_{\theta_1}(s) - \mu_{\theta_2}(s)) \pi_{\theta_1}(a|s) X_{\theta_1}(s,a) \nabla \log \pi_{\theta_1}(a|s) \right\| \\
&\quad + \left\| \sum_s \sum_a \mu_{\theta_2}(s) (X_{\theta_1}(s,a) - X_{\theta_2}(s,a)) \nabla \pi_{\theta_1}(a|s) \right\| \\
&\quad + \left\| \sum_s \sum_a \mu_{\theta_2}(s) X_{\theta_2}(s,a) (\nabla \pi_{\theta_1}(a|s) - \nabla \pi_{\theta_2}(a|s)) \right\| \\
=&\underbrace{\left\| \sum_s (\mu_{\theta_1}(s) - \mu_{\theta_2}(s)) \sum_a \pi_{\theta_1}(a|s) X_{\theta_1}(s,a) \nabla \log \pi_{\theta_1}(a|s) \right\|}_{I_a} \\
&\quad + \underbrace{\left\| \sum_s \sum_a \mu_{\theta_2}(s) (X_{\theta_1}(s,a) - X_{\theta_2}(s,a)) \nabla \pi_{\theta_1}(a|s) \right\|}_{I_b} \\
&\quad + \underbrace{\left\| \sum_s \sum_a \mu_{\theta_2}(s) X_{\theta_2}(s,a) (\nabla \pi_{\theta_1}(a|s) - \nabla \pi_{\theta_2}(a|s)) \right\|}_{I_c}
\end{aligned}
$$

For term I_a, we have ,

$$\begin{aligned}
&\left\| \sum_s (\mu_{\theta_1}(s) - \mu_{\theta_2}(s)) \sum_a \pi_{\theta_1}(a|s) X_{\theta_1}(s,a) \nabla \log \pi_{\theta_1}(a|s) \right\| \\
\leq &\left\| \sum_s (\mu_{\theta_1}(s) - \mu_{\theta_2}(s)) \sum_a \pi_{\theta_1}(a|s) X_{\theta_1}(s,a) \nabla \log \pi_{\theta_1}(a|s) \right\| \\
\leq &BU_X \sum_s |\mu_{\theta_1}(s) - \mu_{\theta_2}(s)| \\
\leq &L_\mu B U_X \|\theta_1 - \theta_2\|
\end{aligned}$$

For term I_b we have ,

$$\begin{aligned}
&\left\| \sum_s \sum_a \mu_{\theta_2}(s) (X_{\theta_1}(s,a) - X_{\theta_2}(s,a)) \pi_{\theta_1}(s|a) \nabla \log \pi_{\theta_1}(a|s) \right\| \\
\leq &L_v B \|\theta_1 - \theta_2\|
\end{aligned}$$

For term I_c we have ,

$$\begin{aligned}
&\left\| \sum_s \sum_a \mu_{\theta_2}(s) X_{\theta_2}(s,a) (\nabla \pi_{\theta_1}(a|s) - \nabla \pi_{\theta_2}(a|s)) \right\| \\
= &\left\| \sum_s \sum_a \mu_{\theta_2}(s) X_{\theta_2}(s,a) (\pi_{\theta_1}(a|s) \nabla \log \pi_{\theta_1}(a|s) - \pi_{\theta_2}(a|s) \nabla \log \pi_{\theta_2}(a|s)) \right\| \\
\leq &\left\| \sum_s \sum_a \mu_{\theta_2}(s) X_{\theta_2}(s,a) (\pi_{\theta_1}(a|s) \nabla \log \pi_{\theta_1}(a|s) - \pi_{\theta_2}(a|s) \nabla \log \pi_{\theta_1}(a|s)) \right\| \\
&+ \left\| \sum_s \sum_a \mu_{\theta_2}(s) X_{\theta_2}(s,a) (\pi_{\theta_2}(a|s) \nabla \log \pi_{\theta_1}(a|s) - \pi_{\theta_2}(a|s) \nabla \log \pi_{\theta_2}(a|s)) \right\| \\
\leq &BU_X L_\pi \|\theta_1 - \theta_2\| + U_X K \|\theta_1 - \theta_2\|
\end{aligned}$$

Hence we have ,

$$\|\nabla L(\theta_1) - L(\theta_2)\| \leq (L_\mu B U_X + L_v B + B U_X L_\pi + U_X K) \|\theta_1 - \theta_2\|$$

Now , we have ,

$$\begin{aligned}
V^\theta(s) &= \mathbb{E}\left[\sum_{t=0}^{\infty} (r(s_t, a_t) - L(\theta)) | s_0 = s \right] \\
&= \sum_{t=0}^{\infty} \mathbb{E}[r(s_t, a_t) - L(\theta) | s_0 = s] \\
&\leq \frac{2U_r b}{1-k}
\end{aligned}$$

Hence we have $U_X = U_r + \frac{2U_r b}{1-k}$.

So finally we have,

$$\|\nabla L(\theta_1) - L(\theta_2)\| \leq M_L \|\theta_1 - \theta_2\|$$

where

$$\begin{aligned}
M_L &= L_\mu B U_X + L_v B + B U_X L_\pi + U_X K \\
&= U_X (L_\mu B + B L_\pi + K) + L_v B \\
&= U_r \left(1 + \frac{2b}{1-k}\right) (L_\mu B + B L_\pi + K) + L_v B
\end{aligned}$$

A.3.4 PROOF OF LEMMA 4

The first inequality follows from lemma B.1 of Wu et al. (2022). The proof for second inequality is as follows :

For finite state and action MDPs, under bounded costs and with policy continuously differentiable and which selects each action with a positive probability, one can show that the stationary distribution and so the average reward is continuously differentiable and bounded, making it Lipschitz (see Theorem A.1, Panda & Bhatnagar (2025)) . The differential value function can be seen to then be differentiable from the fact that

$$\sum_{k=0}^{K} P_\pi^k (H^\pi - Q(\pi)) = V^\pi - P_\pi^{K+1} V^\pi,$$

where $H^\pi = E[r|X_0 = i, \pi]$ and $Q(\pi) = J(\pi)$, see p.284,Bertsekas (2013), upon noting that $\lim_{K\to\infty} P_\pi^{K+1} V^\pi = 0$.
The third inequality follows from assumption 3(c)

A.3.5 PROOF OF LEMMA 5

Let us denote $y_n = L_n - L(\theta_n)$.

So we have,

$$\begin{aligned}
&y_{n+1} = y_n + L(\theta_n) - L(\theta_{n+1}) + \gamma_n(r_n - L_n) \\
\Rightarrow &y_{n+1}^2 = (y_n + L(\theta_n) - L(\theta_{n+1}) + \gamma_n(r_n - L_n))^2 \\
\Rightarrow &y_{n+1}^2 \leq y_n^2 + 2y_n(L(\theta_n) - L(\theta_{n+1})) + 2\gamma_n y_n(r_n - L_n) + 2(L(\theta_n) - L(\theta_{n+1}))^2 + 2\gamma_n^2(r_n - L_n)^2 \\
\Rightarrow &y_{t+1}^2 \leq (1 - 2\gamma_t)y_t^2 + 2\gamma_t y_t(r_t - L(\theta_t)) + 2y_t(L(\theta_t) - L(\theta_{t+1})) + 2(L(\theta_t) - L(\theta_{t+1}))^2 \\
&\quad + 2\gamma_t^2(r_t - L_t)^2.
\end{aligned}$$

where $r_n = r(s_n, a_n)$.

After rearranging the terms we have,

$$\begin{aligned}
y_m^2 \leq \frac{y_m^2 - y_{m+1}^2}{2\gamma_m} + y_m(r_m - L(\theta_m)) + \frac{y_m}{\gamma_m}(L(\theta_m) - L(\theta_{m+1})) + \frac{1}{\gamma_m}(L(\theta_m) - L(\theta_{m+1}))^2 \\
+ \gamma_m(r_m - L_m)^2
\end{aligned}$$

Now we have,

$$\begin{aligned}
\sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k y_k^2 \leq &\underbrace{\sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k \frac{y_k^2 - y_{k+1}^2}{2\gamma_k}}_{I_a} + \underbrace{\sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k y_k(r_k - L(\theta_k))}_{I_b} \\
&+ \underbrace{\sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k \frac{y_k}{\gamma_k}(L(\theta_k) - L(\theta_{k+1}))}_{I_c} \\
&+ \underbrace{\sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k \frac{1}{\gamma_k}(L(\theta_k) - L(\theta_{k+1}))^2}_{I_d} \\
&+ \underbrace{\sum_{k=n_0}^{k=m} \xi(m, k+1)\beta_k \gamma_k (r_k - L_k)^2}_{I_e} \qquad (11)
\end{aligned}$$

For term I_a we have,

$$
\begin{aligned}
I_a &= \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k \frac{y_k^2 - y_{k+1}^2}{2\gamma_k} \\
&= \sum_{k=n_0}^{k=m} \xi(m,k+1)\frac{y_k^2 - y_{k+1}^2}{2\frac{\gamma_k}{\beta_k}} \\
&= \sum_{k=n_0}^{k=m} \frac{\xi(m,k+1)y_k^2 - \xi(m,k+1)y_{k+1}^2}{2\frac{\gamma_k}{\beta_k}} \\
&= \sum_{k=n_0}^{k=m} \frac{\xi(m,k)y_k^2 - \xi(m,k+1)y_{k+1}^2}{2\frac{\gamma_k}{\beta_k}} + \sum_{k=n_0}^{k=m} \frac{(\xi(m,k+1) - \xi(m,k))y_k^2}{2\frac{\gamma_k}{\beta_k}} \\
&= \underbrace{\sum_{k=n_0}^{k=m} \frac{\xi(m,k)y_k^2 - \xi(m,k+1)y_{k+1}^2}{2\frac{\gamma_k}{\beta_k}}}_{I_{a1}} + \underbrace{\mathcal{O}\Bigg(\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k\frac{\beta_k}{\gamma_k}\Bigg)}_{I_{a2}} \\
&= \mathcal{O}\left(\frac{\beta_{n_0-1}}{\gamma_{n_0-1}}\right) + \mathcal{O}(n_0^{\nu-\sigma})
\end{aligned}
$$

Term I_{a1} can be analysed similar to term I_1 in page number 22 of Wu et al. (2022). Term I_{a2} can be analysed similar to term I_2 in page number 22 of Wu et al. (2022).

For term I_b we have,

$$
\begin{aligned}
&\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k y_k (r(s_k,a_k) - L(\theta_k)) \\
&= \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k y_k (r(s_k,a_k) - \sum_a \pi_{\theta_k}(a|s_k) r(s_k,a)) \\
&\quad + \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k y_k (\sum_a \pi_{\theta_k}(a|s_k) r(s_k,a) - \sum_{s,a} \mu_{\theta_k}(s)\pi_{\theta_k}(a|s) r(s,a)) \\
&= \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k y_k (r(s_k,a_k) - \sum_a \pi_{\theta_k}(a|s_k) r(s_k,a)) \\
&\quad + \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k y_k (\bar{r}_{\theta_k}(s_k) - \sum_s \mu_{\theta_k}(s)\bar{r}_{\theta_k}(s)) \\
&\leq \sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k y_k \left(r(s_k,a_k) - \sum_a \pi_{\theta_k}(a|s_k) r(s_k,a) + \bar{r}_{\theta_k}(s_{k+1}) - \sum_s P^{\theta_k}(s|s_k)\bar{r}_{\theta_k}(s)\right) \\
&\quad + \mathcal{O}(\beta_{n_0}) + \mathcal{O}(n_0^{-\nu}) + \mathcal{O}(bk)
\end{aligned}
$$

where we define $\bar{r}_\theta(s) = \sum_a \pi_\theta(a|s) r(s,a)$.

Term I_b above can be analysed similar to term I_1 in (6).

For term I_c we have,

$$\begin{aligned}&\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\frac{y_k}{\gamma_k}(L(\theta_k)-L(\theta_{k+1}))\\=&\mathcal{O}\Big(\sum_{k=n_0}^{k=m}\frac{\xi(m,k+1)\beta_k^2}{\gamma_k}\Big)\\=&\mathcal{O}(n_0^{\nu-\sigma})\end{aligned}$$

For term I_d we have,

$$\begin{aligned}&\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\frac{1}{\gamma_k}(L(\theta_k)-L(\theta_{k+1}))^2\\&=\mathcal{O}\Big(\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\frac{\beta_k^2}{\gamma_k}\Big)\\&=\mathcal{O}(n_0^{\nu-2\sigma}).\end{aligned}$$

For term I_e we have,

$$\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\gamma_k(r_k-L_k)^2=\mathcal{O}\Big(\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\gamma_k\Big)=\mathcal{O}(n_0^{-\nu})$$

Hence after gathering all the terms we have ,

$$\begin{aligned}&\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k y_k^2\\&\leq\mathcal{O}(n_0^{\nu-\sigma})+\mathcal{O}(n_0^{-\nu})+\mathcal{O}(bk)\\&\quad+\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k y_k\Big(r(s_k,a_k)-\sum_a\pi_{\theta_k}(a|s_k)r(s_k,a)+\bar{r}_{\theta_k}(s_{k+1})-\sum_s P^{\theta_k}(s|s_k)\bar{r}_{\theta_k}(s)\Big)\end{aligned}$$

A.3.6 Proof of Lemma 6

We have the following update rule for critic :

$$v_{t+1}=\Gamma(v_t+\alpha_t\delta_t\phi(s_t))$$

Let $z_t=v_t-v^*(\theta_t)$. So we have ,

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

Now we have ,

$$
\begin{aligned}
&\langle z_t, \mathbb{E}_{\theta_t}[\delta_t \phi(s_t)]\rangle \\
=&\langle z_t, \mathbb{E}_{\theta_t}[(r(s_t,a_t) - L_t + \phi(s_{t+1})^T v_t - \phi(s_t)^T v_t)\phi(s_t)]\rangle \\
=&\langle z_t, \mathbb{E}_{\theta_t}[(r(s_t,a_t) - L(\theta_t) + \phi(s_{t+1})^T v_t - \phi(s_t)^T v_t)\phi(s_t)]\rangle + \langle z_t, (L(\theta_t) - L_t)\mathbb{E}_{\theta_t}[\phi(s_t)]\rangle \\
=&\langle z_t, \mathbb{E}_{\theta_t}[(r(s_t,a_t) - L(\theta_t) + \phi(s_{t+1})^T v_t - \phi(s_t)^T v_t)\phi(s_t)]\rangle \\
&\quad + \langle z_t, \mathbb{E}_{\theta_t}[(r(s_t,a_t) - L(\theta_t) + \phi(s_{t+1})^T v^*(\theta_t) - \phi(s_t)^T v^*(\theta_t))\phi(s_t)]\rangle \\
&\quad + \langle z_t, (L(\theta_t) - L_t)E_{\theta_t}[\phi(s_t)]\rangle \\
=&\langle z_t, \mathbb{E}[(\phi(s_{t+1})^T - \phi(s_t)^T)(v_t - v^*(\theta_t))\phi(s_t)]\rangle + \langle z_t, (L(\theta_t) - L_t)E_{\theta_t}[\phi(s_t)]\rangle \\
\leq& -\lambda\|z_t\|^2 + \langle z_t, (L(\theta_t) - L_t)E_{\theta_t}[\phi(s_t)]\rangle
\end{aligned}
$$

So finally we have ,

$$
\begin{aligned}
\|z_{t+1}\|^2 \leq \|z_t\|^2 &+ 2\alpha_t\langle z_t, \delta_t\phi(s_t) - E_{\theta_t}[\delta_t\phi(s_t)]\rangle - 2\alpha_t\lambda\|z_t\|^2 + 2\alpha_t\langle z_t, (L(\theta_t) - L_t)E_{\theta_t}[\phi(s_t)]\rangle \\
&+ 2\langle z_t, v^*(\theta_t) - v^*(\theta_{t+1})\rangle + 2\alpha_t^2\|\delta_t\phi(s_t)\|^2 + 2\|v^*(\theta_t) - v^*(\theta_{t+1})\|^2
\end{aligned}
$$

After rearranging the terms we have ,

$$
\begin{aligned}
2\lambda\|z_t\|^2 &\leq \frac{\|z_t\|^2 - \|z_{t+1}\|^2}{\alpha_t} + 2\langle z_t, \delta_t\phi(s_t) - E_{\theta_t}[\delta_t\phi(s_t)]\rangle + 2\langle z_t, (L(\theta_t) - L_t)E_{\theta_t}[\phi(s_t)]\rangle \\
&\quad + \frac{2}{\alpha_t}\langle z_t, v^*(\theta_t) - v^*(\theta_{t+1})\rangle + 2\alpha_t\|\delta_t\phi(s_t)\|^2 + \frac{2}{\alpha_t}\|v^*(\theta_t) - v^*(\theta_{t+1})\|^2 \\
&\leq \frac{\|z_t\|^2 - \|z_{t+1}\|^2}{\alpha_t} + 2\langle z_t, \delta_t\phi(s_t) - E_{\theta_t}[\delta_t\phi(s_t)]\rangle + 2\|z_t\|\|L(\theta_t) - L_t\| \\
&\quad + C_1\frac{\beta_t}{\alpha_t}\|z_t\| + 2\alpha_t\|\delta_t\phi(s_t)\|^2 + C_2\frac{\beta_t^2}{\alpha_t}
\end{aligned}
$$

Hence we have ,

$$
\begin{aligned}
\Rightarrow 2\lambda\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\|z_k\|^2 \leq &\underbrace{\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\frac{\|z_k\|^2 - \|z_{k+1}\|^2}{\alpha_k}}_{I_1} \\
&+ \underbrace{2\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\langle z_k, \delta_k\phi(s_k) - E_{\theta_k}[\delta_k\phi(s_k)]\rangle}_{I_2} \\
&+ \underbrace{2\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\|z_k\||L(\theta_k) - L_k|}_{I_3} \\
&+ \underbrace{C_1\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\frac{\beta_k}{\alpha_k}\|z_k\|}_{I_4} + \underbrace{2\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\alpha_k\|\delta_k\phi(s_k)\|^2}_{I_5} \\
&+ \underbrace{C_2\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\frac{\beta_k^2}{\alpha_k}}_{I_6}
\end{aligned}
$$

For term I_1 we have,

$$\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k \frac{\|z_k\|^2 - \|z_{k+1}\|^2}{\alpha_k}$$
$$= \sum_{k=n_0}^{k=m} \xi(m,k+1) \frac{\|z_k\|^2 - \|z_{k+1}\|^2}{\frac{\alpha_k}{\beta_k}}$$

Hence we have

$$I_1 = \mathcal{O}\left(\frac{\beta_{n_0-1}}{\alpha_{n_0-1}}\right) + \mathcal{O}(n_0^{\nu-\sigma})$$

Term I_1 can be analysed similar to term I_a in subsection A.3.5.

For term I_2 we have,

$$I_2 = 2\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k \langle z_k, \delta_k \phi(s_k) - E_{\theta_k}[\delta_k \phi(s_k)]\rangle$$

Let $W(s,a,s^{'},L,v) = (r(s,a) - L + \phi(s^{'})^T v - \phi(s)^T v)\phi(s)$.

Hence we have,

I_2

$$= 2\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k \langle z_k, W(s_k,a_k,s_{k+1},L_k,v_k) - \sum_s \sum_a \sum_{s^{'}} \mu_{\theta_k}(s)\pi_{\theta_k}(a|s)P(s^{'}|s,a)W(s,a,s^{'},L_k,v_k)\rangle$$

$$= 2\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k \langle z_k, W(s_k,a_k,s_{k+1},L_k,v_k) - \sum_a \sum_{s^{'}} \pi_{\theta_k}(a|s_k)P(s^{'}|s_k,a)W(s_k,a,s^{'},L_k,v_k)\rangle$$
$$+ 2\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k \langle z_k, \sum_a \sum_{s^{'}} \pi_{\theta_k}(a|s_k)P(s^{'}|s_k,a)W(s_k,a,s^{'},L_k,v_k)$$
$$- \sum_s \sum_a \sum_{s^{'}} \mu_{\theta_k}(s)\pi_{\theta_k}(a|s)P(s^{'}|s,a)W(s,a,s^{'},L_k,v_k)\rangle$$

$$= 2\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k \langle z_k, W(s_k,a_k,s_{k+1},L_k,v_k) - \sum_a \sum_{s^{'}} \pi_{\theta_k}(a|s_k)P(s^{'}|s_k,a)W(s_k,a,s^{'},L_k,v_k)\rangle$$
$$+ 2\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k \langle z_k, \bar{W}_{\theta_k}(s_k,L_k,v_k) - \sum_s \mu_{\theta_k}(s)\bar{W}_{\theta_k}(s,L_k,v_k)\rangle$$

where we define $\bar{W}_\theta(s,L,v) = \sum_a \sum_{s^{'}} \pi_\theta(a|s)P(s^{'}|s,a)W(s,a,s^{'},L,v)$.

After analysing term I_2 similar to term I_1 in (6).

I_2

$$\leq 2\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k \langle z_k, W(s_k,a_k,s_{k+1},L_k,v_k) - \sum_a \sum_{s^{'}} \pi_{\theta_k}(a|s_k)P(s^{'}|s_k,a)W(s_k,a,s^{'},L_k,v_k)\rangle$$
$$+ 2\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k \langle z_k, \bar{W}_{\theta_k}(s_{k+1},L_k,v_k) - \sum_s P^{\theta_k}(s|s_k)\bar{W}_{\theta_k}(s,L_k,v_k)\rangle$$
$$+ 4\beta_{n_0-1}W_{max} + \mathcal{O}(n_0^{-\nu}) + \mathcal{O}(bk)$$

For term I_3 we have ,

$$\begin{aligned}
&2\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k|L(\theta_k)-L_k|\|z_k\| \\
=&2\sum_{k=n_0}^{k=m}\sqrt{\xi(m,k+1)\beta_k(L(\theta_k)-L_k)^2}\sqrt{\xi(m,k+1)\beta_k\|z_k\|^2} \\
\leq&\sqrt{\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k(L(\theta_k)-L_k)^2}\sqrt{\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\|z_k\|^2}
\end{aligned}$$

For term I_4 we have ,

$$C_1\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\frac{\beta_k}{\alpha_k}\|z_k\|=\mathcal{O}(n_0^{\nu-\sigma})$$

For term I_5 we have,

$$2\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\alpha_k\|\delta_k\phi(s_k)\|^2=\mathcal{O}(n_0^{-\nu})$$

For term I_6 we have,

$$C_2\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\frac{\beta_k^2}{\alpha_k}=\mathcal{O}(n_0^{\nu-2\sigma})$$

Hence after gathering all the terms we have,

$$\begin{aligned}
&\Rightarrow 2\lambda\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\|z_k\|^2 \\
&\leq\mathcal{O}(n_0^{\nu-\sigma}) \\
&\quad+2\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\langle z_k,W(s_k,a_k,s_{k+1},L_k,v_k)-\sum_a\sum_{s'}\pi_{\theta_k}(a|s_k)P(s'|s_k,a)W(s_k,a,s',L_k,v_k)\rangle \\
&\quad+2\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\langle z_k,\bar{W}_{\theta_k}(s_{k+1},L_k,v_k)-\sum_s P^{\theta_k}(s|s_k)\bar{W}_{\theta_k}(s,L_k,v_k)\rangle \\
&\quad+4\beta_{n_0-1}W_{max}+\mathcal{O}(n_0^{-\nu})+\mathcal{O}(bk) \\
&\quad+\sqrt{\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k(L(\theta_k)-L_k)^2}\sqrt{\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\|z_k\|^2}
\end{aligned}$$

Let

$$\begin{aligned}
A(m,n_0)=&\mathcal{O}(n_0^{\nu-\sigma}) \\
&+2\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\langle z_k,W(s_k,a_k,s_{k+1},L_k,v_k)-\sum_a\sum_{s'}\pi_{\theta_k}(a|s_k)P(s'|s_k,a)W(s_k,a,s',L_k,v_k)\rangle \\
&+2\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\langle z_k,\bar{W}_{\theta_k}(s_{k+1},L_k,v_k)-\sum_s P^{\theta_k}(s|s_k)\bar{W}_{\theta_k}(s,L_k,v_k)\rangle \\
&+4\beta_{n_0-1}W_{max}+\mathcal{O}(n_0^{-\nu})+\mathcal{O}(bk)
\end{aligned}$$

and

$$B(m,n_0)=\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k(L(\theta_k)-L_k)^2$$

Hence we have ,

$$\begin{aligned}
&2\lambda\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\|z_k\|^2\leq A(m,n_0)+\sqrt{B(m,n_0)}\sqrt{\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\|z_k\|^2}\\
&\Rightarrow\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\|z_k\|^2-\frac{1}{2\lambda}\sqrt{B(m,n_0)}\sqrt{\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\|z_k\|^2}\leq\frac{A(m,n_0)}{2\lambda}\\
&\Rightarrow\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\|z_k\|^2-\frac{2}{4\lambda}\sqrt{B(m,n_0)}\sqrt{\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\|z_k\|^2}+\frac{B(m,n_0)}{16\lambda^2}\\
&\quad\leq\frac{A(m,n_0)}{2\lambda}+\frac{B(m,n_0)}{16\lambda^2}\\
&\Rightarrow\left(\sqrt{\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\|z_k\|^2}-\frac{\sqrt{B(m,n_0)}}{4\lambda}\right)^2\leq\frac{A(m,n_0)}{2\lambda}+\frac{B(m,n_0)}{16\lambda^2}\\
&\Rightarrow\sqrt{\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\|z_k\|^2}-\frac{\sqrt{B(m,n_0)}}{4\lambda}\leq\sqrt{\frac{A(m,n_0)}{2\lambda}}+\sqrt{\frac{B(m,n_0)}{16\lambda^2}}\\
&\Rightarrow\sqrt{\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\|z_k\|^2}\leq\sqrt{\frac{A(m,n_0)}{2\lambda}}+2\sqrt{\frac{B(m,n_0)}{16\lambda^2}}\\
&\Rightarrow\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\|z_k\|^2\leq\frac{A(m,n_0)}{\lambda}+\frac{1}{2\lambda}B(m,n_0)\\
&\Rightarrow\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\|z_k\|^2\leq\mathcal{O}(n_0^{\nu-\sigma})\\
&\quad+\frac{2}{\lambda}\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\langle z_k,W(s_k,a_k,s_{k+1},L_k,v_k)-\sum_a\sum_{s'}\pi_{\theta_k}(a|s_k)P(s'|s_k,a)W(s_k,a,s',L_k,v_k)\rangle\\
&\quad+\frac{2}{\lambda}\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\langle z_k,\bar{W}_{\theta_k}(s_{k+1},L_k,v_k)-\sum_s P^{\theta_k}(s|s_k)\bar{W}_{\theta_k}(s,L_k,v_k)\rangle\\
&\quad+\mathcal{O}(bk)+\mathcal{O}(n_0^{-\nu})+\frac{1}{2\lambda}\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k(L(\theta_k)-L_k)^2
\end{aligned}$$

Now putting the results of lemma 5 in the above inequality we have ,

$$\sum_{k=n_0}^{k=m} \xi(m,k+1)\beta_k\|z_k\|^2 \leq \mathcal{O}(n_0^{\nu-\sigma})$$
$$+\frac{2}{\lambda}\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\langle z_k, W(s_k,a_k,s_{k+1},L_k,v_k)-\sum_a\sum_{s'}\pi_{\theta_k}(a|s_k)P(s^{'}|s_k,a)W(s_k,a,s^{'},L_k,v_k)\rangle$$
$$+\frac{2}{\lambda}\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k\langle z_k, \bar{W}_{\theta_k}(s_{k+1},L_k,v_k)-\sum_s P^{\theta_k}(s|s_k)\bar{W}_{\theta_k}(s,L_k,v_k)\rangle$$
$$+\mathcal{O}(bk)+\mathcal{O}(n_0^{-\nu})$$
$$+\frac{1}{2\lambda}\sum_{k=n_0}^{k=m}\xi(m,k+1)\beta_k y_k\left(r(s_k,a_k)-\sum_a\pi_{\theta_k}(a|s_k)r(s_k,a)+\bar{r}_{\theta_k}(s_{k+1})-\sum_s P^{\theta_k}(s|s_k)\bar{r}_{\theta_k}(s)\right)$$

A.3.7 Proof of lemma 7

We have ,

$$\begin{aligned}
&\sum_s\sum_a \mu_\theta(s)\pi_\theta(a|s)F(\theta,s,a)\\
=&\sum_s\sum_a \mu_\theta(s)\pi_\theta(a|s)((r(s,a)-L(\theta)+\sum_{s'}P(s^{'}|s,a)V^\theta(s^{'})-V^\theta(s))\nabla\log\pi_\theta(a|s)+\theta)\\
=&\sum_s\sum_a \mu_\theta(s)\pi_\theta(a|s)(r(s,a)-L(\theta)+\sum_{s'}P(s^{'}|s,a)V^\theta(s^{'})-V^\theta(s))\nabla\log\pi_\theta(a|s)+\theta\\
=&\nabla L(\theta)+\theta
\end{aligned}$$

Now ,

$$\begin{aligned}
&\left\|\sum_s\sum_a \mu_{\theta_1}(s)\pi_{\theta_1}(a|s)F(\theta_1,s,a)-\sum_s\sum_a \mu_{\theta_2}(s)\pi_{\theta_2}(a|s)F(\theta_2,s,a)\right\|^2\\
=&\left\|\nabla L(\theta_1)+\theta_1-\nabla L(\theta_2)-\theta_2\right\|^2\\
=&\|\nabla L(\theta_1)-\nabla L(\theta_2)\|^2+\|\theta_1-\theta_2\|^2+2\langle\nabla L(\theta_1)-\nabla L(\theta_2),\theta_1-\theta_2\rangle\\
\leq&M_L^2\|\theta_1-\theta_2\|^2+\|\theta_1-\theta_2\|^2-m\|\theta_1-\theta_2\|^2\\
=&(1-2m+M_L^2)\|\theta_1-\theta_2\|^2
\end{aligned}$$

Now we have ,

$$B<\frac{\sqrt{2m}}{L_v}$$
$$U_r<\frac{\sqrt{2m}-L_vB}{\left(1+\frac{2b}{1-k}\right)(L_\mu B+BL_\pi+K)}$$

Hence ,

$$\begin{aligned}
&U_r\left(1+\frac{2b}{1-k}\right)(L_\mu B+BL_\pi+K)+L_vB<\sqrt{2m}\\
\Rightarrow&M_L<\sqrt{2m}\\
\Rightarrow&M_L^2<2m\\
\Rightarrow&1-2m+M_L^2<1
\end{aligned}$$

Table 1: Actor and critic architecture and step-size (learning-rate) schedule used in each environment.

	CartPole-v1	**FrozenLake-v1**	**Blackjack-v1**
Actions $\|A\|$	2 (push left / right)	4 (up/down/left/right)	2 (stick / hit)
Actor parameterization	binary logistic: $\pi(a{=}0\|s) = \sigma(\theta^\top \phi(s))$ $\theta \in \mathbb{R}^{11}$	tabular softmax, one logit anchored per state $\theta \in \mathbb{R}^{3\times 16}$	binary logistic $\theta \in \mathbb{R}^7$
Critic parameter v	$v \in \mathbb{R}^{11}$	$v \in \mathbb{R}^{16}$	$v \in \mathbb{R}^7$
Feature map $\phi(s)$	bias, 4 normalized state dims, their squares, 2 cross terms	exact one-hot over the 16 discrete states	bias, normalized player sum, dealer card, ace indicator, their squares, 1 cross term
Approximation	linear FA	tabular	linear FA
Critic step-size α_t	$5.0/(t+50)^{0.5}$	$5.0/(t+50)^{0.5}$	$1.0/(t+1)^{0.5}$
Actor step-size β_t	$5.0/(t+50)^{1.0}$	$5.0/(t+50)^{1.0}$	$2.0/(t+1)^{1.0}$
Avg.-reward step-size γ_t	$K\alpha_t,\ K{=}1.0$	$K\alpha_t,\ K{=}1.0$	$K\alpha_t,\ K{=}1.0$
Critic projection radius U_v	50.0	50.0	50.0

A.4 HYPERPARAMETERS

Table 1 gives us the information about the actor and critic network used in the experiments along with the learning rates.

For CartPole, the feature map $\phi(s)$ consists of a constant bias term, the four normalized state variables, the squared values of these state variables, and two selected pairwise interaction terms. This results in a feature vector of dimension 11. The same feature map is used by both the actor and the critic. For FrozenLake, we use an exact one-hot representation of the state. Since there are 16 discrete states, each state is represented by a 16-dimensional feature vector with a value of one corresponding to the current state and zeros elsewhere. Consequently, the critic has one parameter for each state. For Blackjack, the feature map consists of a bias term, the normalized player sum, the normalized dealer card, an indicator for whether the player has a usable ace, the squared values of these state variables, and one selected interaction term. This gives a 7-dimensional feature vector, which is used by both the actor and the critic.