A Finite-Time Analysis of Two Time-Scale Actor-Critic Methods

Yue Wu
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
ywu@cs.ucla.edu

Weitong Zhang
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
weightzero@cs.ucla.edu

Pan Xu
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
panxu@cs.ucla.edu

Quanquan Gu
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
qgu@cs.ucla.edu

Abstract

Actor-critic (AC) methods have exhibited great empirical success compared with other reinforcement learning algorithms, where the actor uses the policy gradient to improve the learning policy and the critic uses temporal difference learning to estimate the policy gradient. Under the two time-scale learning rate schedule, the asymptotic convergence of AC has been well studied in the literature. However, the non-asymptotic convergence and finite sample complexity of actor-critic methods are largely open. In this work, we provide a non-asymptotic analysis for two time-scale actor-critic methods under non-i.i.d. setting. We prove that the actor-critic method is guaranteed to find a first-order stationary point (i.e., $\|\nabla J(\boldsymbol{\theta})\|_2^2 \leq \epsilon$) of the non-concave performance function $J(\boldsymbol{\theta})$, with $\widetilde{\mathcal{O}}(\epsilon^{-2.5})$ sample complexity. To the best of our knowledge, this is the first work providing finite-time analysis and sample complexity bound for two time-scale actor-critic methods.

1 Introduction

Actor-Critic (AC) methods [2, 16] aim at combining the advantages of actor-only methods and critic-only methods, and have achieved great empirical success in reinforcement learning [31, 1]. Specifically, actor-only methods, such as policy gradient [28] and trust region policy optimization [24], utilize a parameterized policy function class and improve the policy by optimizing the parameters of some performance function using gradient ascent, whose exact form is characterized by the Policy Gradient Theorem [28]. Actor-only methods can be naturally applied to continuous setting but suffer from high variance when estimating the policy gradient. On the other hand, critic-only methods, such as temporal difference learning [26] and Q-learning [32], focus on learning a value function (expected cumulative rewards), and determine the policy based on the value function, which is recursively approximated based on the Bellman equation. Although the critic-only methods can efficiently learn a satisfying policy under tabular setting [14], they can diverge with function approximation under continuous setting [33]. Therefore, it is natural to combine actor and critic based methods to achieve the best of both worlds. The principal idea behind actor-critic methods is simple: the critic tries to learn the value function, given the policy from the actor, while the actor can estimate the policy gradient based on the approximate value function provided by the critic.

If the actor is fixed, the policy remains unchanged throughout the updates of the critic. Thus one can use policy evaluation algorithm such as temporal difference (TD) learning [27] to estimate the value function (critic). After many steps of the critic update, one can expect a good estimation of the value function, which in turn enables an accurate estimation of the policy gradient for the actor. A more favorable implementation is the so-called two time-scale actor-critic algorithm, where the actor and the critic are updated simultaneously at each iteration except that the actor changes more slowly (with a small step size) than the critic (with a large step size). In this way, one can hope the critic will be well approximated even after one step of update. From the theoretical perspective, the asymptotic analysis of two time-scale actor-critic methods has been established in [6, 16]. In specific, under the assumption that the ratio of the two time-scales goes to infinity (i.e. $\lim_{t\to\infty} \beta_t/\alpha_t = \infty$), the asymptotic convergence is guaranteed through the lens of the two time-scale ordinary differential equations(ODE), where the slower component is fixed and the faster component converges to its stationary point. This type of analysis was also applied in the context of generic two time-scale stochastic approximation [5].

However, finite-time analysis (non-asymptotic analysis) of two-time scale actor-critic is still largely missing in the literature, which is important because it can address the questions that how many samples are needed for two time-scale actor-critic to converge, and how to appropriately choose the different learning rates for the actor and the critic. Some recent work has attempted to provide the finite-time analysis for the "decoupled" actor-critic methods [18, 23]. The term "decoupled" means that before updating the actor at the t-th iteration, the critic starts from scratch to estimate the state-value (or Q-value) function. At each iteration, the "decoupled" setting requires the critic to perform multiple sampling and updating (often from another new sample trajectory). As we will see in the later comparison, this setting is sample-inefficient or even impractical. Besides, their analyses are based on either the i.i.d. assumption [18] or the partially i.i.d. assumption [23] (the actor receives i.i.d. samples), which is unrealistic in practice. In this paper, we present the first finite-time analysis on the convergence of the two time-scale actor-critic algorithm. We summarize our contributions as follows:

- We prove that, the actor in the two time-scale actor critic algorithm converges to an ϵ-approximate stationary point of the non-concave performance function J after accessing at most $\widetilde{\mathcal{O}}(\epsilon^{-2.5})$ samples. Compared with existing finite-time analysis of actor-critic methods [18, 23], the algorithm we analyzed is based on two time-scale update and therefore more practical and efficient than the "decoupled" version. Moreover, we do not need any i.i.d. data assumptions in the convergence analysis as required by Kumar et al. [18], Qiu et al. [23], which do not hold in real applications.

- From the technical viewpoint, we also present a new proof framework that can tightly characterize the estimation error in two time-scale algorithms. Compared with the proof technique used in [38], we remove the extra artificial factor $\mathcal{O}(t^\xi)$ in the convergence rate introduced by their "iterative refinement" technique. Therefore, our new proof technique may be of independent interest for analyzing the convergence of other two time-scale algorithms to get sharper rates.

Notation We use lower case letters to denote scalars, and use lower and upper case bold face letters to denote vectors and matrices respectively. For two sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = \mathcal{O}(b_n)$ if there exists an absolute constant C such that $a_n \leq Cb_n$. We use $\widetilde{\mathcal{O}}(\cdot)$ to further hide logarithm factors. Without other specification, $\|\cdot\|$ denotes the ℓ_2 norm of Euclidean vectors. $d_{TV}(P, Q)$ is the total variation norm between two probability measure P and Q, which is defined as $d_{TV}(P, Q) = 1/2 \int_{\mathcal{X}} |P(dx) - Q(dx)|$.

2 Related work

In this section, we briefly review and discuss existing work, which is mostly related to ours.

Stochastic bias characterization The main difficulty in analyzing reinforcement learning algorithms under non-i.i.d. data assumptions is that the samples and the trainable parameters are correlated, which makes the noise term biased. Bhandari et al. [3] used information-theoretical techniques to bound the Markovian bias and provide a simple and explicit analysis for the temporal difference learning. Similar techniques were also established in [25] through the lens of stochastic approximation methods. Gupta et al. [12], Xu et al. [38] applied such methods to deriving the non-asymptotic convergence of two time-scale temporal difference learning algorithms (TDC). Zou et al. [44], Chen et al. [10], Xu and Gu [35] further applied these analysis methods to on-policy learning algorithms including SARSA

and Q-learning. In addition, Hu and Syed [13] formulated a family of TD learning algorithms as a Markov jump linear systems and analyzed the evolution of the mean and covariance matrix of the estimation error. Cai et al. [7] studied TD learning with neural network approximation, and proved its global convergence.

Two time-scale reinforcement learning The two time-scale stochastic approximation can be seen as a general framework for analyzing reinforcement learning [5, 29, 17]. Recently, the finite-time analysis of two time-scale stochastic approximation has gained much interest. Dalal et al. [11] proved convergence rate for the two time-scale linear stochastic approximation under i.i.d. assumption. Gupta et al. [12] also provided finite-time analysis for the two time-scale linear stochastic approximation algorithms. Both can be applied to analyze two time-scale TD methods like GTD, GTD2 and TDC. Xu et al. [38] proved convergence rate and sample complexity for the TDC algorithm over Markovian samples. [15] further improved the convergence rate of two time-scale linear stochastic approximation and removed the projection step. However, since the update rule for the actor is generally not linear, we cannot apply these results to the actor-critic algorithms.

Analysis for actor-critic methods The asymptotic analysis of actor-critic methods has been well established. Konda and Tsitsiklis [16] proposed the actor-critic algorithm, and established the asymptotic convergence for the two time-scale actor-critic, with TD(λ) learning-based critic. Bhatnagar et al. [4] proved the convergence result for the original actor-critic and natural actor-critic methods. Castro and Meir [8] proposed a single time-scale actor-critic algorithm and proved its convergence. Recently, [43] proved convergence of two time-scale off-policy actor-critic with function approximation. Recently, there has emerged some works concerning the finite-time behavior of actor-critic methods. Yang et al. [41] studied the global convergence of actor-critic algorithms under the Linear Quadratic Regulator. Yang et al. [40] analyzed the finite-sample performance of batched actor-critic, where all samples are assumed i.i.d. and the critic performs several empirical risk minimization (ERM) steps. Qiu et al. [23] treated the actor-critic algorithms as a bilevel optimization problem and established a finite sample analysis under the "average-reward" setting, assuming that the actor has access to independent samples. Similar result has also been established by Kumar et al. [18], where they considered the sample complexity for the "decoupled" actor-critic methods under i.i.d. assumption. Wang et al. [30] also proved the global convergence of actor-critic algorithms with both actor and critic being approximated by overparameterized neural networks.

When we were preparing this work, we noticed that there is a concurrent and independent work [39] which also analyzes the non-asymptotic convergence of two time-scale actor-critic algorithms and achieves the same sample complexity, i.e., $\widetilde{\mathcal{O}}(\epsilon^{-2.5})$. However, there are two key differences between their work and ours. First, the two time-scale algorithms analyzed in both papers are very different. We analyze the classical two time-scale algorithm described in [27], where both actor and critic take one step update in each iteration. It is very easy to implement and has been widely used in practice, while the update rule in [39] for the critic needs to call a sub-algorithm, which involves generating a fresh episode to estimate the Q-function. Second, the analysis in [39] relies on the compatible function approximation [28], which requires the critic to be a specific linear function class, while our analysis does not require such specific approximation, and therefore is more general. This makes our analysis potentially extendable to non-linear function approximation such as neural networks [7].

3 Preliminaries

In this section, we present the background of the two time-scale actor-critic algorithm.

3.1 Markov decision processes

Reinforcement learning tasks can be modeled as a discrete-time Markov Decision Process (MDP) $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{P}, r\}$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces respectively. In this work we consider the finite action space $|\mathcal{A}| < \infty$. $\mathcal{P}(s'|s, a)$ is the transition probability that the agent transits to state s' after taking action a at state s. Function $r : \mathcal{S} \times \mathcal{A} \to [-U_r, U_r]$ emits a bounded reward after the agent takes action a at state s, where $U_r > 0$ is a constant. A policy parameterized by $\boldsymbol{\theta}$ at state s is a probability function $\pi_{\boldsymbol{\theta}}(a|s)$ over action space $\mathcal{A}$. $\mu_{\boldsymbol{\theta}}$ denotes the stationary distribution induced by the policy $\pi_{\boldsymbol{\theta}}$.

In this work we consider the "average reward" setting [28], where under the ergodicity assumption, the average reward over time eventually converges to the expected reward under the stationary

distribution:

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

To evaluate the overall rewards given a starting state s_0 and the behavior policy $\pi_{\boldsymbol{\theta}}$, we define the state-value function as

$$V^{\pi_{\boldsymbol{\theta}}}(\cdot) := \mathbb{E}\bigg[\sum_{t=0}^{\infty} \big(r(s_t, a_t) - r(\boldsymbol{\theta}) \big) | s_0 = \cdot \bigg],$$

where the action follows the policy $a_t \sim \pi_{\boldsymbol{\theta}}(\cdot | s_t)$ and the next state follows the transition probability $s_{t+1} \sim \mathcal{P}(\cdot | s_t, a_t)$. Another frequently used function is the state-action value function, also called Q-value function:

$$\begin{aligned} Q^{\pi_{\boldsymbol{\theta}}}(s, a) : &= \mathbb{E}\bigg[\sum_{t=0}^{\infty} \big(r(s_t, a_t) - r(\boldsymbol{\theta}) \big) | s_0 = s, a_0 = a \bigg] \\ &= r(s, a) - r(\boldsymbol{\theta}) + \mathbb{E}\big[V^{\pi_{\boldsymbol{\theta}}}(s') \big], \end{aligned}$$

where the expectation is taken over $s' \sim \mathcal{P}(\cdot | s, a)$.

Throughout this paper, we use O to denote the tuple $O = (s, a, s')$, some variants are like $O_t = (s_t, a_t, s_{t+1})$ and $\widetilde{O}_t = (\widetilde{s}_t, \widetilde{a}_t, \widetilde{s}_{t+1})$.

3.2 Policy gradient theorem

We define the performance function associated with policy $\pi_{\boldsymbol{\theta}}$ naturally as the expected reward under the stationary distribution $\mu_{\boldsymbol{\theta}}$ induced by $\pi_{\boldsymbol{\theta}}$, which takes the form

$$J(\boldsymbol{\theta}) : = r(\boldsymbol{\theta}). \tag{3.1}$$

To maximize the performance function with respect to the policy parameters, Sutton et al. [28] proved the following policy gradient theorem.

Lemma 3.1 (Policy Gradient). Consider the performance function defined in (3.1), its gradient takes the form

$$\nabla J(\boldsymbol{\theta}) = \mathbb{E}_{s \sim \mu_{\boldsymbol{\theta}}(\cdot)} \bigg[\sum_{a \in \mathcal{A}} Q^{\pi_{\boldsymbol{\theta}}}(s, a) \nabla \pi(a | s) \bigg].$$

The policy gradient also admits a neat form in expectation:

$$\nabla J(\boldsymbol{\theta}) = \mathbb{E}_{s \sim \mu_{\boldsymbol{\theta}}(\cdot), a \sim \pi_{\boldsymbol{\theta}}(\cdot | s)} \big[Q^{\pi_{\boldsymbol{\theta}}}(s, a) \nabla \log \pi_{\boldsymbol{\theta}}(a | s) \big].$$

A typical way to estimate the policy gradient $\nabla J(\boldsymbol{\theta})$ is by Monte Carlo method, namely using the summed return along the trajectory as the estimated Q-value, which is known as the "REINFORCE" method [34].

Remark 3.2. The problem formulation in this paper is what Sutton et al. [28] had defined as "average-reward" formulation. An alternative formulation is the "start-state" formulation, which avoids estimating the average reward, but gives a more complicated form for the policy-gradient algorithm and the AC algorithm.

3.3 REINFORCE with a baseline

Note that for any function $b(s)$ depending only on the state, which is usually called "baseline" function, we have

$$\sum_{a \in \mathcal{A}} b(s) \nabla \pi_{\boldsymbol{\theta}}(a | s) = b(s) \nabla \bigg(\sum_{a \in \mathcal{A}} \pi_{\boldsymbol{\theta}}(a | s) \bigg) = 0.$$

So we also have

$$\nabla J(\boldsymbol{\theta}) = \mathbb{E}\bigg[\sum_{a \in \mathcal{A}} \big(Q^{\pi_{\boldsymbol{\theta}}}(s, a) - b(s) \big) \nabla \pi_{\boldsymbol{\theta}}(a | s) \bigg].$$

A popular choice of $b(s)$ is $b(s) = V^{\pi_\theta}(s)$ and $\Delta^{\pi_\theta}(s, a) = Q^{\pi_\theta}(s, a) - V^{\pi_\theta}(s)$ is viewed as the advantage of taking a specific action a, compared with the expected reward at state s. Also note that the expectation form still holds:

$$\nabla J(\boldsymbol{\theta}) = \mathbb{E}_{s,a}\big[\Delta^{\pi_\theta}(s, a)\nabla \log \pi_{\boldsymbol{\theta}}(a|s)\big].$$

Based on this fact, Williams [34] also proposed a corresponding policy gradient algorithm named "REINFORCE with a baseline" which performs better due to the reduced variance.

In practice the policy gradient method could suffer from high variance. An alternative approach is to introduce another trainable model to approximate the state-value function, which is called the actor-critic methods.

3.4 The two time-scale actor-critic algorithm

In previous subsection, we have seen how the policy gradient theorem appears in the form of the advantage value instead of the Q-value. Assume the critic uses linear function approximation $\widehat{V}(\cdot; \boldsymbol{\omega}) = \boldsymbol{\phi}^\top(\cdot)\boldsymbol{\omega}$, and is updated by TD(0) algorithm, then this gives rise to Algorithm 1 that we are going to analyze.

Algorithm 1 has been proposed in many literature, and is clearly introduced in [27] as a classic on-line one-step actor-critic algorithm. It uses the advantage (namely temporal difference error) to update the critic and the actor simultaneously. Based on its on-line nature, this algorithm can be implemented both under episodic and continuing setting. In practice, the asynchronous variant of this algorithm, called Asynchronous Advantage Actor-Critic(A3C), is an empirically very successful parallel actor-critic algorithm.

Sometimes, Algorithm 1 is also called Advantage Actor-Critic (A2C) because it is the synchronous version of A3C and the name indicates its use of advantage instead of Q-value [20].

Algorithm 1 Two Time-Scale Actor-Critic

1: **Input:** initial actor parameter $\boldsymbol{\theta}_0$, initial critic parameter $\boldsymbol{\omega}_0$, initial average reward estimator η_0, step size α_t for actor, β_t for critic and γ_t for the average reward estimator.
2: Draw s_0 from some initial distribution
3: **for** $t = 0, 1, 2, \ldots$ **do**
4:     Take the action $a_t \sim \pi_{\boldsymbol{\theta}_t}(\cdot|s_t)$
5:     Observe next state $s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)$ and the reward $r_t = r(s_t, a_t)$
6:     $\delta_t = r_t - \eta_t + \boldsymbol{\phi}(s_{t+1})^\top\boldsymbol{\omega}_t - \boldsymbol{\phi}(s_t)^\top\boldsymbol{\omega}_t$
7:     $\eta_{t+1} = \eta_t + \gamma_t(r_t - \eta_t)$
8:     $\boldsymbol{\omega}_{t+1} = \Pi_{R_\omega}\big(\boldsymbol{\omega}_t + \beta_t\delta_t\boldsymbol{\phi}(s_t)\big)$
9:     $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha_t\delta_t\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}_t}(a_t|s_t)$
10: **end for**

In Line 6 of Algorithm 1, the temporal difference error δ_t can be calculated based on the critic's estimation of the value function $\boldsymbol{\phi}(\cdot)^\top\boldsymbol{\omega}_t$, where $\boldsymbol{\omega}_t \in \mathbb{R}^d$ and $\boldsymbol{\phi}(\cdot) : \mathcal{S} \to \mathbb{R}^d$ is a known feature mapping. Then the critic will be updated using the semi-gradient from TD(0) method. Line 8 in Algorithm 1 also contains a projection operator. This is required to control the algorithm's convergence which also appears in some other literature [3, 38]. The actor uses the advantage δ_t (estimated by critic) and the samples to get an estimation of the policy gradient.

Algorithm 1 is more general and practical than the algorithms analyzed in many previous work [23, 18]. In our algorithm, there is no need for independent samples or samples from the stationary distribution. There is only one naturally generated sample path. Also, the critic inherits from last iteration and continuously updates its parameter, without requiring a restarted sample path (or a new episode).

4 Main theory

In this section, we first discuss on some standard assumptions used in the literature for deriving the convergence of reinforcement learning algorithms and then present our theoretical results for two time-scale actor-critic methods.

4.1 Assumptions and propositions

We consider the setting where the critic uses TD [27] with linear function approximation to estimate the state-value function, namely $\widehat{V}(\cdot; \boldsymbol{\omega}) = \boldsymbol{\phi}^\top(\cdot)\boldsymbol{\omega}$. We assume that the feature mapping has bounded norm $\|\boldsymbol{\phi}(\cdot)\| \leq 1$. Denote by $\boldsymbol{\omega}^*(\boldsymbol{\theta})$ the limiting point of TD(0) algorithms under the behavior policy $\pi_{\boldsymbol{\theta}}$, and define $\mathbf{A}$ and $\mathbf{b}$ as:

$$\mathbf{A} := \mathbb{E}_{s,a,s'}\big[\boldsymbol{\phi}(s)\big(\boldsymbol{\phi}(s') - \boldsymbol{\phi}(s)\big)^\top\big],$$
$$\mathbf{b} := \mathbb{E}_{s,a,s'}[(r(s,a) - r(\boldsymbol{\theta}))\boldsymbol{\phi}(s)],$$

where $s \sim \mu_{\boldsymbol{\theta}}(\cdot), a \sim \pi_{\boldsymbol{\theta}}(\cdot|s), s' \sim \mathcal{P}(\cdot|s,a)$. It is known that the TD limiting point satisfies:

$$\mathbf{A}\boldsymbol{\omega}^*(\boldsymbol{\theta}) + \mathbf{b} = \mathbf{0}.$$

In the sequel, when there is no confusion, we will use a shorthand notation $\boldsymbol{\omega}^*$ to denote $\boldsymbol{\omega}^*(\boldsymbol{\theta})$. Based on the complexity of the feature mapping, the approximation error of this function class can vary. The approximation error of the linear function class is defined as follows:

$$\epsilon_{\text{app}}(\boldsymbol{\theta}) := \sqrt{\mathbb{E}_{s\sim\mu_{\boldsymbol{\theta}}}\big(\boldsymbol{\phi}(s)^\top\boldsymbol{\omega}^*(\boldsymbol{\theta}) - V^{\pi_{\boldsymbol{\theta}}}(s)\big)^2}.$$

Throughout this paper, we assume the approximation error for all potential policies is uniformly bounded,

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

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

In the analysis of TD learning, the following assumption is often made to ensure the uniqueness of the limiting point of TD and the problem's solvability.

Assumption 4.1. For all potential policy parameters $\boldsymbol{\theta}$, the matrix $\mathbf{A}$ defined above is negative definite and has the maximum eigenvalues as $-\lambda$.

Assumption 4.1 is often made to guarantee the problem's solvability [3, 44, 38]. Note that Algorithm 1 contains a projection step at Line 8. To guarantee convergence it is required all $\boldsymbol{\omega}^*$ lie within this projection radius $R_{\boldsymbol{\omega}}$. Assumption 4.1 indicates that a sufficient condition is to set $R_{\omega} = 2U_r/\lambda$ because $\|\mathbf{b}\| \leq 2U_r$ and $\|\mathbf{A}^{-1}\| \leq \lambda^{-1}$.

The next assumption, first adopted by Bhandari et al. [3] in TD learning, addresses the issue of Markovian noise.

Assumption 4.2 (Uniform ergodicity). For a fixed $\boldsymbol{\theta}$, denote $\mu_{\boldsymbol{\theta}}(\cdot)$ as the stationary distribution induced by the policy $\pi_{\boldsymbol{\theta}}(\cdot|s)$ and the transition probability measure $\mathcal{P}(\cdot|s,a)$. Consider a Markov chain generated by the rule $a_t \sim \pi_{\boldsymbol{\theta}}(\cdot|s_t), s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)$. Then there exists $m > 0$ and $\rho \in (0,1)$ such that:

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

We also need some regularity assumptions on the policy.

Assumption 4.3. Let $\pi_{\boldsymbol{\theta}}(a|s)$ be a policy parameterized by $\boldsymbol{\theta}$. There exist constants $L, B, L_l > 0$ such that for all given state s and action a it holds

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

(b) $\big\|\nabla \log \pi_{\boldsymbol{\theta}_1}(a|s) - \nabla \log \pi_{\boldsymbol{\theta}_2}(a|s)\big\| \leq L_l\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|, \forall\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$,

(c) $\big|\pi_{\boldsymbol{\theta}_1}(a|s) - \pi_{\boldsymbol{\theta}_2}(a|s)\big| \leq L\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|, \forall\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$.

The first two inequalities are regularity conditions to guarantee actor's convergence in the literature of policy gradient [22, 42, 18, 36, 37]. The last inequality in Assumption 4.3 is also adopted by Zou et al. [44] when analyzing SARSA.

An important fact arises from our assumptions is that the limiting point $\boldsymbol{\omega}^*$ of TD(0), which can be viewed as a mapping of the policy's parameter $\boldsymbol{\theta}$, is Lipschitz.

Proposition 4.4. Under Assumptions 4.1 and 4.2, there exists a constant $L_* > 0$ such that

$$\left\|\boldsymbol{\omega}^*(\boldsymbol{\theta}_1) - \boldsymbol{\omega}^*(\boldsymbol{\theta}_2)\right\| \leq L_*\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|, \forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d.$$

Proposition 4.4 states that the target point $\boldsymbol{\omega}^*$ moves slowly compared with the actor's update on $\boldsymbol{\theta}$. This is an observation pivotal to the two time-scale analysis. Specifically, the two time-scale analysis can be informally described as "the actor moves slowly while the critic chases the slowly moving target determined by the actor".

Now we are ready to present the convergence result of two time-scale actor-critic methods. We first define an integer that depends on the learning rates α_t and β_t.

$$\tau_t := \min\left\{i \geq 0 \mid m\rho^{i-1} \leq \min\{\alpha_t, \beta_t\}\right\}, \tag{4.1}$$

where m, ρ are defined as in Assumption 4.2. By definition, τ_t is a mixing time of an ergodic Markov chain. We will use τ_t to control the Markovian noise encountered in the training process.

4.2 Convergence of the actor

At the k-th iteration of the actor's update, $\boldsymbol{\omega}_k$ is the critic parameter estimated by Line 7 of Algorithm 1 and $\boldsymbol{\omega}_k^*$ is the unknown parameter of value function $V^{\pi_{\boldsymbol{\theta}_k}}(\cdot)$ defined in Assumption 4.1. The following theorem gives the convergence rate of the actor when the averaged mean squared error between $\boldsymbol{\omega}_k$ and $\boldsymbol{\omega}_k^*$ and the error between η_k and $r(\boldsymbol{\theta}_k)$ from $k = \tau_t$ to $k = t$ are small.

Theorem 4.5. Suppose Assumptions 4.1-4.3 hold and we choose $\alpha_t = c_\alpha/(1+t)^\sigma$ in Algorithm 1, where $\sigma \in (0,1)$ and $c_\alpha > 0$ are constants. If we assume at the t-th iteration, the critic satisfies

$$\frac{8}{t}\sum_{k=1}^{t}\mathbb{E}\|\boldsymbol{\omega}_k - \boldsymbol{\omega}_k^*\|^2 + \frac{2}{t}\sum_{k=1}^{t}\mathbb{E}\big(\eta_k - r(\boldsymbol{\theta}_k)\big)^2 = \mathcal{E}(t), \tag{4.2}$$

where $\mathcal{E}(t)$ is a bounded sequence, then we have

$$\min_{0 \leq k \leq t}\mathbb{E}\left\|\nabla J(\boldsymbol{\theta}_k)\right\|^2 = \mathcal{O}(\epsilon_{\text{app}}) + \mathcal{O}\left(\frac{1}{t^{1-\sigma}}\right) + \mathcal{O}\left(\frac{\log^2 t}{t^\sigma}\right) + \mathcal{O}\big(\mathcal{E}(t)\big),$$

where $\mathcal{O}(\cdot)$ hides constants, whose exact forms can be found in the detailed proof in Appendix C.1.

Note that $\mathcal{E}(t)$ in Theorem 4.5 is the averaged estimation error made by the critic throughout the learning process, which will be bounded in the next Theorem 4.7.

Remark 4.6. Theorem 4.5 recovers the results for the decoupled case [23, 18] by setting $\sigma = 1/2$. Nevertheless, we are considering a much more practical and challenging case where the actor and critic are simultaneously updated under Markovian noises. It is worth noting that the non-i.i.d. data assumption leads to an additional logarithm term, which is also observed in [3, 44, 25, 10].

4.3 Convergence of the critic

The condition in (4.2) is guaranteed by the following theorem that characterizes the convergence of the critic.

Theorem 4.7. Suppose Assumptions 4.1-4.3 hold and we choose $\alpha_t = c_\alpha/(1+t)^\sigma$ and $\beta_t = c_\beta/(1+t)^\nu$ in Algorithm 1, where $0 < \nu < \sigma < 1$, c_α and $c_\beta \leq \lambda^{-1}$ are positive constants. Then we have

$$\frac{1}{1+t-\tau_t}\sum_{k=\tau_t}^{t}\mathbb{E}\|\boldsymbol{\omega}_k - \boldsymbol{\omega}_k^*\|^2 = \mathcal{O}\left(\frac{1}{t^{1-\nu}}\right) + \mathcal{O}\left(\frac{\log t}{t^\nu}\right) + \mathcal{O}\left(\frac{1}{t^{2(\sigma-\nu)}}\right), \tag{4.3}$$

$$\frac{1}{1+t-\tau_t}\sum_{k=\tau_t}^{t}\mathbb{E}\big(\eta_k - r(\boldsymbol{\theta}_k)\big)^2 = \mathcal{O}\left(\frac{1}{t^{1-\nu}}\right) + \mathcal{O}\left(\frac{\log t}{t^\nu}\right) + \mathcal{O}\left(\frac{1}{t^{2(\sigma-\nu)}}\right), \tag{4.4}$$

where $\mathcal{O}(\cdot)$ hides constants, whose exact forms can be found in the detailed proof in Appendix C.2 and C.3.

Remark 4.8. The first term $\mathcal{O}(t^{\nu-1})$ on the right hand side of (4.3) and (4.4) comes from loosely bounding the error's norm, and can be removed by applying the "iterative refinement" technique used in Xu et al. [38]. Using this technique, we can obtain a bound (also holds for η_t) $\mathbb{E}\|\boldsymbol{\omega}_t - \boldsymbol{\omega}_t^*\|^2 = \mathcal{O}(\log t/t^{\nu}) + \mathcal{O}(1/t^{2(\sigma-\nu)-\xi})$, where $\xi > 0$ is an arbitrarily small constant. The constant ξ is an artifact due to the the "iterative refinement" technique. Similar simplification can be done for (4.4). Nevertheless, if we plug (4.3) and (4.4) (after some transformation) into the result of Theorem 4.5, it is easy to see that the term $\mathcal{O}(1/t^{1-\nu})$ is actually dominated by the term $\mathcal{O}(1/t^{1-\sigma})$. Thus this term makes no difference in the total sample complexity of Algorithm 1 and we choose not to complicate the proof or introduce the extra artificial parameter ξ in the result of Theorem 4.7.

The second term in both (4.3) and (4.4) comes from the Markovian noise and the variance of the semi-gradient. The third term in these two equations comes from the slow drift of the actor. These two terms together can be interpreted as follows: if the actor moves much slower than the critic (i.e., $\sigma - \nu \gg \nu$), then the error is dominated by the Markovian noise and gradient variance; if the actor moves not too slowly compared with the critic (i.e. $\sigma - \nu \ll \nu$), then the critic's error is dominated by the slowly drifting effect of the actor.

4.4 Convergence rate and sample complexity

Combining Theorems 4.5 and 4.7 leads to the following convergence rate and sample complexity for Algorithm 1. Detailed proof is in Appendix C.4.

Corollary 4.9. Under the same assumptions of Theorems 4.5 and 4.7, we have

$$\min_{0 \le k \le t} \mathbb{E}\|\nabla J(\boldsymbol{\theta}_k)\|^2 = \mathcal{O}(\epsilon_{\text{app}}) + \mathcal{O}\left(\frac{1}{t^{1-\sigma}}\right) + \mathcal{O}\left(\frac{\log t}{t^{\nu}}\right) + \mathcal{O}\left(\frac{1}{t^{2(\sigma-\nu)}}\right).$$

If we set $\sigma = 3/5, \nu = 2/5$, leading to the actor step size $\alpha_t = O(1/t^{3/5})$ and the critic step size $\beta_t = O(1/t^{2/5})$, Algorithm 1 can find an ϵ-approximate stationary point of $J(\cdot)$ within T steps, namely,

$$\min_{0 \le k \le T} \mathbb{E}\|\nabla J(\boldsymbol{\theta}_k)\|^2 \le \mathcal{O}(\epsilon_{\text{app}}) + \epsilon,$$

where $T = \widetilde{\mathcal{O}}(\epsilon^{-2.5})$ is the total iteration number.

Corollary 4.9 combines the results of Theorems 4.5 and 4.7 and shows that the convergence rate of Algorithm 1 is $\widetilde{\mathcal{O}}(t^{-2/5})$. Since the per iteration sample is 1, the sample complexity of two time-scale actor-critic is $\widetilde{\mathcal{O}}(\epsilon^{-2.5})$.

Remark 4.10. We compare our results with existing results on the sample complexity of actor-critic methods in the literature. Kumar et al. [18] provided a general result that after $T = \mathcal{O}(\epsilon^{-2})$ updates for the actor, the algorithm can achieve $\min_{0 \le k \le T} \mathbb{E}\|\nabla J(\boldsymbol{\theta}_k)\|^2 \le \epsilon$, as long as the estimation error of the critic can be bounded by $\mathcal{O}(t^{-1/2})$ at the t-th actor's update. However, to ensure such a condition on the critic, they need to draw t samples to estimate the critic at the t-th actor's update. Therefore, the total number of samples drawn from the whole training process by the actor-critic algorithm in [18] is $\mathcal{O}(T^2)$, yielding a $\mathcal{O}(\epsilon^{-4})$ sample complexity. Under the similar setting, Qiu et al. [23] proved the same sample complexity $\widetilde{\mathcal{O}}(\epsilon^{-4})$ when TD(0) is used for estimating the critic. Thus Corollary 4.9 suggests that the sample complexity of Algorithm 1 is significantly better than the sample complexity presented in [18, 23] by a factor of $\mathcal{O}(\epsilon^{-1.5})$.

Remark 4.11. The gap between the "decoupled" actor-critic and the two time-scale actor-critic seems huge. Intuitively, this is due to the inefficient usage of the samples. At each iteration, the critic in the "decoupled" algorithm starts over to evaluate the policy's value function and discards the history information, regardless of the fact that the policy might only changed slightly. The two time-scale actor-critic keeps the critic's parameter and thus takes full advantage of each samples in the trajectory.

Remark 4.12. According to [22], the sample complexity of policy gradient methods such as RE-INFORCE is $\mathcal{O}(\epsilon^{-2})$. As a comparison, if the critic converges faster than $\mathcal{O}(t^{-1/2})$, namely $\mathcal{E}(t) = \mathcal{O}(t^{-1/2})$, then Theorem 4.5 combined with Corollary 4.9 implies that the complexity of two time-scale actor-critic is $\widetilde{\mathcal{O}}(\epsilon^{-2})$, which matches the result of policy gradient methods [22] up to logarithmic factors. Nevertheless, as we have discussed in the previous remarks, a smaller

estimation error for critic often comes at the cost of more samples needed for the critic update [23, 18], which eventually increases the total sample complexity. Therefore, the $\widetilde{\mathcal{O}}(\epsilon^{-2.5})$ sample complexity in Corollary 4.9 is indeed the lowest we can achieve so far for classic two time-scale actor-critic methods. However, it is possible to further improve the sample complexity by using policy evaluation algorithms better than vanilla TD(0), such as GTD and TDC methods.

5 Conclusion and discussion

In this paper, we provided the first finite-time analysis of the two time-scale actor-critic methods, with non-i.i.d. Markovian samples and linear function approximation. The algorithm we analyzed is an on-line, one-step actor-critic algorithm which is practical and efficient. We proved its non-asymptotic convergence rate as well as its sample complexity. Our proof technique can be potentially extended to analyze other two time-scale reinforcement learning algorithms.

As one of the anonymous reviewers suggested, the compatible features are useful tools to address the function approximation error of the critic [16]. This can leads to finite-time analysis for the natural actor-critic algorithm [39], which also relates to the more general natural policy gradient methods [9]. Another possible improvement is to use regularization(e.g. ridge) for the critic to ensure the boundedness of the critic and remove the assumption on the maximum eigenvalue. The analysis can also be applied to the infinite-horizon discounted MDP, where the framework of analysis essentially remains the same.

Broader impact

This work could positively impact the industrial application of actor-critic algorithms and other reinforcement learning algorithms. The theorem exhibits the sample complexity of actor-critic algorithms, which could be used to estimate required training time of reinforcement learning models. Another direct application of our result is to set the learning rate according to the finite-time bound, by optimizing the constant factors of the dominant terms. In this sense, the result could potentially reduce the overhead of hyper-parameter tuning, thus saving both human and computational resources. Moreover, the new analysis in this paper can potentially help people in different fields to understand the broader class of two-time scale algorithms, in addition to actor-critic methods. To our knowledge, this algorithm and theory studied in our paper do not have any ethical issues.

Acknowledgement

References

[1] Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. An actor-critic algorithm for sequence prediction. *arXiv preprint arXiv:1607.07086*, 2016.

[2] A. G. Barto, R. S. Sutton, and C. W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(5):834–846, 1983.

[3] Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. *arXiv preprint arXiv:1806.02450*, 2018.

[4] Shalabh Bhatnagar, Richard S Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor–critic algorithms. *Automatica*, 45(11):2471–2482, 2009.

[5] Vivek S Borkar. Stochastic approximation with two time scales. *Systems & Control Letters*, 29 (5):291–294, 1997.

[6] Vivek S Borkar and Vijaymohan R Konda. The actor-critic algorithm as multi-time-scale stochastic approximation. *Sadhana*, 22(4):525–543, 1997.

[7] Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830*, 2019.

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

[9] Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *arXiv preprint arXiv:2007.06558*, 2020.

[10] Zaiwei Chen, Sheng Zhang, Thinh T Doan, Siva Theja Maguluri, and John-Paul Clarke. Performance of q-learning with linear function approximation: Stability and finite-time analysis. *arXiv preprint arXiv: 1905.11425*, 2019.

[11] Gal Dalal, Balazs Szorenyi, Gugan Thoppe, and Shie Mannor. Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. *arXiv preprint arXiv:1703.05376*, 2017.

[12] Harsh Gupta, R Srikant, and Lei Ying. Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 4706–4715, 2019.

[13] Bin Hu and Usman Syed. Characterizing the exact behaviors of temporal difference learning algorithms using markov jump linear system theory. In *Advances in Neural Information Processing Systems*, pages 8477–8488, 2019.

[14] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

[15] Maxim Kaledin, Eric Moulines, Alexey Naumov, Vladislav Tadic, and Hoi-To Wai. Finite time analysis of linear two-timescale stochastic approximation with markovian noise. *arXiv preprint arXiv:2002.01268*, 2020.

[16] Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, pages 1008–1014, 2000.

[17] Vijay R Konda, John N Tsitsiklis, et al. Convergence rate of linear two-time-scale stochastic approximation. *The Annals of Applied Probability*, 14(2):796–819, 2004.

[18] Harshat Kumar, Alec Koppel, and Alejandro Ribeiro. On the sample complexity of actor-critic method for reinforcement learning with function approximation. *arXiv preprint arXiv:1910.08412*, 2019.

[19] A Yu Mitrophanov. Sensitivity and convergence of uniformly ergodic markov chains. *Journal of Applied Probability*, 42(4):1003–1014, 2005.

[20] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937, 2016.

[21] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[22] Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. In *International Conference on Machine Learning*, pages 4023–4032, 2018.

[23] Shuang Qiu, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. On the finite-time convergence of actor-critic algorithm. *NeurIPS 2019 Optimization Foundations of Reinforcement Learning Workshop*, 2019.

[24] John Schulman, Sergey Levine, Pieter Abbeel, Michael I Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, volume 37, pages 1889–1897, 2015.

[25] R Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation andtd learning. In *Conference on Learning Theory*, pages 2803–2830, 2019.

[26] Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

[27] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[28] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, pages 1057–1063, 2000.

[29] Vladislav B Tadic and Sean P Meyn. Asymptotic properties of two time-scale stochastic approximation algorithms with constant step sizes. In *Proceedings of the 2003 American Control Conference, 2003.*, volume 5, pages 4426–4431. IEEE, 2003.

[30] Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*, 2020.

[31] Ziyu Wang, Victor Bapst, Nicolas Heess, Volodymyr Mnih, Remi Munos, Koray Kavukcuoglu, and Nando de Freitas. Sample efficient actor-critic with experience replay. *arXiv preprint arXiv:1611.01224*, 2016.

[32] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

[33] Marco A Wiering. Convergence and divergence in standard and averaging reinforcement learning. In *European Conference on Machine Learning*, pages 477–488. Springer, 2004.

[34] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256, 1992.

[35] Pan Xu and Quanquan Gu. A finite-time analysis of q-learning with neural network function approximation. *arXiv preprint arXiv:1912.04511*, 2019.

[36] Pan Xu, Felicia Gao, and Quanquan Gu. An improved convergence analysis of stochastic variance-reduced policy gradient. In *International Conference on Uncertainty in Artificial Intelligence*, 2019.

[37] Pan Xu, Felicia Gao, and Quanquan Gu. Sample efficient policy gradient methods with recursive variance reduction. In *International Conference on Learning Representations*, 2020.

[38] Tengyu Xu, Shaofeng Zou, and Yingbin Liang. Two time-scale off-policy td learning: Non-asymptotic analysis over markovian samples. In *Advances in Neural Information Processing Systems*, pages 10633–10643, 2019.

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

[40] Zhuoran Yang, Kaiqing Zhang, Mingyi Hong, and Tamer Başar. A finite sample analysis of the actor-critic algorithm. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 2759–2764. IEEE, 2018.

[41] Zhuoran Yang, Yongxin Chen, Mingyi Hong, and Zhaoran Wang. On the global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. In *Advances in Neural Information Processing Systems*, 2019.

[42] Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Başar. Global convergence of policy gradient methods to (almost) locally optimal policies. *arXiv preprint arXiv:1906.08383*, 2019.

[43] Shangtong Zhang, Bo Liu, Hengshuai Yao, and Shimon Whiteson. Provably convergent two-timescale off-policy actor-critic with function approximation. *arXiv*, pages arXiv–1911, 2019.

[44] Shaofeng Zou, Tengyu Xu, and Yingbin Liang. Finite-sample analysis for sarsa with linear function approximation. In *Advances in Neural Information Processing Systems*, pages 8665–8675, 2019.

A Proof Sketch

In this section, we provide the proof roadmap of the main theory. Detailed proofs can be found in Appendix C.

A.1 Proof Sketch of Theorem 4.5

The following lemma is important in that it enables the analysis of policy gradient method:

Lemma A.1 ([42]). For the performance function defined in (3.1), there exists a constant $L_J > 0$ such that for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$, it holds that

$$\big\|\nabla J(\boldsymbol{\theta}_1) - \nabla J(\boldsymbol{\theta}_2)\big\| \le L_J \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|,$$

which by the definition of smoothness [21] is also equivalent to

$$J(\boldsymbol{\theta}_2) \ge J(\boldsymbol{\theta}_1) + \big\langle \nabla J(\boldsymbol{\theta}_1), \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1 \big\rangle - \frac{L_J}{2} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2.$$

This lemma enables us to perform a gradient ascent style analysis on the non-concave function $J(\boldsymbol{\theta})$:

$$
\begin{aligned}
J(\boldsymbol{\theta}_{t+1}) \ge\ & J(\boldsymbol{\theta}_t) + \alpha_t \big\langle \nabla J(\boldsymbol{\theta}_t), \delta_t \nabla \log \pi_{\boldsymbol{\theta}_t}(a_t|s_t) \big\rangle - L_J \alpha_t^2 \big\| \delta_t \nabla \log \pi_{\boldsymbol{\theta}_t}(a_t|s_t) \big\|^2 \\
=\ & J(\boldsymbol{\theta}_t) + \alpha_t \big\langle \nabla J(\boldsymbol{\theta}_t), \Delta h(O_t, \eta_t, \boldsymbol{\omega}_t, \boldsymbol{\theta}_t) \big\rangle + \alpha_t \big\langle \nabla J(\boldsymbol{\theta}_t), \Delta h'(O_t, \boldsymbol{\theta}_t) \big\rangle \\
& + \alpha_t \Gamma(O_t, \boldsymbol{\theta}_t) + \alpha_t \big\| \nabla J(\boldsymbol{\theta}_t) \big\|^2 - L_J \alpha_t^2 \big\| \delta_t \nabla \log \pi_{\boldsymbol{\theta}_t}(a_t|s_t) \big\|^2,
\end{aligned} \tag{A.1}
$$

where $O_t = (s_t, a_t, s_{t+1})$ is a tuple of observations. The second term $\Delta h(O_t, \boldsymbol{\omega}_t, \boldsymbol{\theta}_t)$ on the right hand side of (A.1) is the bias introduced by the critic. The third term $\Delta h'(O_t, \boldsymbol{\theta}_t)$ is from the linear approximation error. The fourth term $\Gamma(O_t, \boldsymbol{\theta}_t)$ is due to the Markovian noise. The last term can be viewed as the variance of the stochastic gradient update. Please refer to (C.1) for the definition of each notation.

Now we bound each term's expectation in (A.1) respectively.

First, we have

$$\mathbb{E}\big\langle \nabla J(\boldsymbol{\theta}_t), \Delta h(O_t, \eta_t, \boldsymbol{\omega}_t, \boldsymbol{\theta}_t) \big\rangle \ge -B\sqrt{\mathbb{E}\big\|\nabla J(\boldsymbol{\theta}_t)\big\|^2} \sqrt{8\mathbb{E}\|\mathbf{z}_t\|^2 + 2\mathbb{E}[y_t^2]},$$

where $\mathbf{z}_t := \boldsymbol{\omega}_t - \boldsymbol{\omega}_t^*$ and $y_t := \eta_t - \eta_t^*$, and the inequality is due to Cauchy inequality and Lemma C.2.

Second, taking expectation over the approximation error term containing $\Delta h'$, we have

$$
\begin{aligned}
\mathbb{E}\big\langle \nabla J(\boldsymbol{\theta}_t), \Delta h'(O_t, \boldsymbol{\theta}_t) \big\rangle &\ge -G_{\boldsymbol{\theta}} \sqrt{\mathbb{E}\big\|\Delta h'(O_t, \boldsymbol{\theta}_t)\big\|^2} \\
&\ge -G_{\boldsymbol{\theta}} \cdot 2B\sqrt{\mathbb{E}\big(\phi(s)^\top \boldsymbol{\omega}_t^* - V^{\pi_{\boldsymbol{\theta}_t}}(s)\big)^2} \\
&\ge -2BG_{\boldsymbol{\theta}} \epsilon_{\text{app}},
\end{aligned}
$$

Third, we have

$$
\begin{aligned}
\mathbb{E}[\Gamma(O_t, \boldsymbol{\theta}_t)] &\ge -G_{\boldsymbol{\theta}}\bigg(D_1(\tau+1) \sum_{k=t-\tau+1}^{t} \mathbb{E}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}\| + D_2 m \rho^{\tau-1} \bigg), \\
&\ge -G_{\boldsymbol{\theta}}\bigg(D_1(\tau+1)G_{\boldsymbol{\theta}} \sum_{k=t-\tau+1}^{t-1} \alpha_k + D_2 m \rho^{\tau-1} \bigg),
\end{aligned}
$$

where the first inequality is due to Lemma C.3, and the second inequality is due to $\big\| \delta_t \nabla \log \pi_{\boldsymbol{\theta}_t}(a_t|s_t) \big\| \le G_{\boldsymbol{\theta}}$ by Lemma C.3.

Taking the expectation of (C.2), plugging the above terms back into it and rearranging give

$$
\begin{aligned}
\mathbb{E}\big\|\nabla J(\boldsymbol{\theta}_t)\big\|^2 \le\ & \frac{1}{\alpha_t}\big(\mathbb{E}[J(\boldsymbol{\theta}_{t+1})] - \mathbb{E}[J(\boldsymbol{\theta}_t)]\big) + B\sqrt{\mathbb{E}\big\|\nabla J(\boldsymbol{\theta}_t)\big\|^2}\sqrt{8\mathbb{E}\|\mathbf{z}_t\|^2 + 2\mathbb{E}[y_t^2]} \\
& + D_1 G_{\boldsymbol{\theta}}^2(\tau+1) \sum_{k=t-\tau}^{t-1} \alpha_k + D_2 G_{\boldsymbol{\theta}} m \rho^{\tau-1} + L_J G_{\boldsymbol{\theta}}^2 \alpha_t.
\end{aligned}
$$

13

Setting $\tau = \tau_t$ and summing over each term, and further dividing $(1 + t - \tau_t)$ at both sides and assuming $t > 2\tau_t - 1$, we can express the result as

$$\frac{1}{1+t-\tau_t} \sum_{k=\tau_t}^{t} \mathbb{E}\|\nabla J(\boldsymbol{\theta}_t)\|^2 \leq \mathcal{O}\left(\frac{1}{t^{1-\sigma}}\right) + \mathcal{O}\left(\frac{(\log t)^2}{t^\sigma}\right) + \mathcal{O}(\epsilon_{\text{app}})$$

$$+ \frac{2B}{1+t-\tau_t} \sum_{k=\tau_t}^{t} \sqrt{\mathbb{E}\|\nabla J(\boldsymbol{\theta}_t)\|^2} \sqrt{8\mathbb{E}\|\mathbf{z}_t\|^2 + 2\mathbb{E}[y_t^2]} \quad \text{(A.2)}$$

By Cauchy-Schwartz inequality, we have

$$\frac{1}{1+t-\tau_t} \sum_{k=\tau_t}^{t} \sqrt{\mathbb{E}\|\nabla J(\boldsymbol{\theta}_t)\|^2} \sqrt{8\mathbb{E}\|\mathbf{z}_t\|^2 + 2\mathbb{E}[y_t^2]}$$

$$\leq \left(\frac{1}{1+t-\tau_t} \sum_{k=\tau_t}^{t} \mathbb{E}\|\nabla J(\boldsymbol{\theta}_t)\|^2\right)^{\frac{1}{2}} \left(\frac{1}{1+t-\tau_t} \sum_{k=\tau_t}^{t} \left(8\mathbb{E}\|\mathbf{z}_t\|^2 + 2\mathbb{E}[y_t^2]\right)\right)^{\frac{1}{2}}.$$

Now, denote $F(t) := 1/(1 + t - \tau_t) \sum_{k=\tau_t}^{t} \mathbb{E}\|\nabla J(\boldsymbol{\theta}_k)\|^2$ and $Z(t) := 1/(1 + t - \tau_t) \sum_{k=\tau_t}^{t} \left(8\mathbb{E}\|\mathbf{z}_t\|^2 + 2\mathbb{E}[y_t^2]\right)$, and putting them back to (A.2) ($\mathcal{O}$-notation for simplicity):

$$F(t) \leq \mathcal{O}\left(\frac{1}{t^{1-\sigma}}\right) + \mathcal{O}\left(\frac{(\log t)^2}{t^\sigma}\right) + \mathcal{O}(\epsilon_{\text{app}}) + 2B\sqrt{F(t)} \cdot \sqrt{Z(t)},$$

which further gives

$$\left(\sqrt{F(t)} - B\sqrt{Z(t)}\right)^2 \leq \mathcal{O}\left(\frac{1}{t^{1-\sigma}}\right) + \mathcal{O}\left(\frac{(\log t)^2}{t^\sigma}\right) + \mathcal{O}(\epsilon_{\text{app}}) + B^2 Z(t).$$

Note that for a general function $H(t) = A(t) + B(t)$(with each positive), we have

$$H^2(t) = \mathcal{O}\left(A^2(t)\right) + \mathcal{O}\left(B^2(t)\right),$$
$$\sqrt{H(t)} = \mathcal{O}\left(\sqrt{A(t)}\right) + \mathcal{O}\left(\sqrt{B(t)}\right).$$

This means

$$\min_{0 \leq k \leq t} \mathbb{E}\|\nabla J(\boldsymbol{\theta}_k)\|^2 \leq \frac{1}{1+t-\tau_t} \sum_{k=\tau_t}^{t} \mathbb{E}\|\nabla J(\boldsymbol{\theta}_k)\|^2$$

$$= \mathcal{O}\left(\frac{1}{t^{1-\sigma}}\right) + \mathcal{O}\left(\frac{1}{t^\sigma}\right) + \mathcal{O}(\epsilon_{\text{app}}) + \mathcal{O}(\mathcal{E}(t)).$$

A.2 Proof Sketch of Theorem 4.7

The proof of Theorem 4.7 can be divided into the following two parts.

A.2.1 Estimating the Average Reward η_k

We denote $y_k := \eta_k - r(\boldsymbol{\theta}_k)$. First, we shall mention that many components in this step is uses the same framework and partial result as the proof regarding $\boldsymbol{\omega}_t$ in the next part. Also, part of the proof is intriguingly similar with the proof of Theorem 4.5. For simplicity, here we only present the final result regarding η_k. Please refer to Section C.2 for the detailed proof. By setting $\gamma_k = (1 + t)^{-\nu}$, we have that

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

14

A.2.2 Approximating the TD Fixed Point

Step 1: decomposition of the estimation error. For simplicity, we denote $\mathbf{z}_t := \boldsymbol{\omega}_t - \boldsymbol{\omega}_t^*$, where the $\boldsymbol{\omega}_t^*$ denotes the exact parameter under policy $\pi_{\boldsymbol{\theta}_t}$. By the critic update in Line 7 of Algorithm 1, we have

$$
\|\mathbf{z}_{t+1}\|^2 = \|\mathbf{z}_t\|^2 + 2\beta_t\langle \mathbf{z}_t, \bar{g}(\boldsymbol{\omega}_t, \boldsymbol{\theta}_t)\rangle + 2\beta_t\Lambda(O_t, \boldsymbol{\omega}_t, \boldsymbol{\theta}_t) + 2\beta_t\langle \mathbf{z}_t, \Delta g(O_t, \eta_t, \boldsymbol{\theta}_t)\rangle
$$
$$
+ 2\langle \mathbf{z}_t, \boldsymbol{\omega}_t^* - \boldsymbol{\omega}_{t+1}^*\rangle + \left\|\beta_t(g(O_t, \boldsymbol{\omega}_t, \boldsymbol{\theta}_t) + \Delta g(O_t, \eta_t, \boldsymbol{\theta}_t)) + (\boldsymbol{\omega}_t^* - \boldsymbol{\omega}_{t+1}^*)\right\|^2. \tag{A.3}
$$

where $O_t := (s_t, a_t, s_{t+1})$ is a tuple of observations, $g(O_t, \boldsymbol{\omega}_t)$ and $\bar{g}(\boldsymbol{\theta}_t, \boldsymbol{\omega}_t)$ are the estimated gradient and the true gradient respectively. $\Lambda(O_t, \boldsymbol{\omega}_t, \boldsymbol{\theta}_t) := \langle \boldsymbol{\omega}_t - \boldsymbol{\omega}_t^*, g(O_t, \boldsymbol{\omega}_t) - \bar{g}(\boldsymbol{\theta}_t, \boldsymbol{\omega}_t)\rangle$ can be seen as the error induced by the Markovian noise. Please refer to (C.7) for formal definition of each notation.

The second term on the right hand side of (A.3) can be bounded by $-2\lambda\beta_t\|\mathbf{z}_t\|^2$ due to Assumption 4.1. The third term is a bias term caused by the Markovian noise. The fourth term $\Delta g(O_t, \eta_t, \boldsymbol{\theta}_t)$ is another bias term caused by inaccurate average reward estimator η_t. The fifth term is caused by the slowly drifting policy parameter $\boldsymbol{\theta}_t$. And the last term can be considered as the variance term.

Rewriting (A.3) and telescoping from $\tau = \tau_t$ to t, we have

$$
2\lambda \sum_{k=\tau_t}^t \mathbb{E}\|\mathbf{z}_k\|^2 \le \underbrace{\sum_{k=\tau_t}^t \frac{1}{\beta_k}\left(\mathbb{E}\|\mathbf{z}_k\|^2 - \mathbb{E}\|\mathbf{z}_{k+1}\|^2\right)}_{I_1} + 2\underbrace{\sum_{k=\tau_t}^t \mathbb{E}\Lambda(\boldsymbol{\theta}_k, \boldsymbol{\omega}_k, O_k)}_{I_2}
$$
$$
+ 2L_*G_{\boldsymbol{\theta}}\underbrace{\sum_{k=\tau_t}^t \frac{\alpha_k}{\beta_k}\sqrt{\mathbb{E}\|\mathbf{z}_k\|}}_{I_3} + \underbrace{\sum_{k=\tau_t}^t \sqrt{\mathbb{E}[y_k^2]}\cdot\sqrt{\mathbb{E}\|\mathbf{z}_k\|}}_{I_4} + C_q\underbrace{\sum_{k=\tau_t}^t \beta_k}_{I_5}. \tag{A.4}
$$

We will see that the Markovian noise I_2, the "slowly drifting policy" term I_3 and the estimation bias I_4 from η_t are significant, and bounding the Markovian term is another challenge.

Step 2: bounding the Markovian bias. We first decompose $\Lambda(\boldsymbol{\theta}_t, \boldsymbol{\omega}_t, O_t)$ as follows.

$$
\Lambda(\boldsymbol{\theta}_t, \boldsymbol{\omega}_t, O_t) = \left(\Lambda(\boldsymbol{\theta}_t, \boldsymbol{\omega}_t, O_t) - \Lambda(\boldsymbol{\theta}_{t-\tau}, \boldsymbol{\omega}_t, O_t)\right) + \left(\Lambda(\boldsymbol{\theta}_{t-\tau}, \boldsymbol{\omega}_t, O_t) - \Lambda(\boldsymbol{\theta}_{t-\tau}, \boldsymbol{\omega}_{t-\tau}, O_t)\right)
$$
$$
+ \left(\Lambda(\boldsymbol{\theta}_{t-\tau}, \boldsymbol{\omega}_{t-\tau}, O_t) - \Lambda(\boldsymbol{\theta}_{t-\tau}, \boldsymbol{\omega}_{t-\tau}, \widetilde{O}_t)\right) + \Lambda(\boldsymbol{\theta}_{t-\tau}, \boldsymbol{\omega}_{t-\tau}, \widetilde{O}_t). \tag{A.5}
$$

The motivation is to employ the uniform ergodicity defined by Assumption 4.2. This technique was first introduced by Bhandari et al. [3] to address the Markovian noise in policy evaluation. Zou et al. [44] extended to the Q-learning setting where the parameter itself both keeps updated and determines the behavior policy. In this work we take one step further to consider that the policy parameter $\boldsymbol{\theta}_t$ is changing, and the evaluation parameter $\boldsymbol{\omega}_t$ is updated. The analysis relies on the auxiliary Markov chain constructed by Zou et al. [44], which is obtained by repeatedly applying policy $\pi_{\boldsymbol{\theta}_{t-\tau}}$:

$$
s_{t-\tau} \xrightarrow{\boldsymbol{\theta}_{t-\tau}} a_{t-\tau} \xrightarrow{\mathcal{P}} s_{t-\tau+1} \xrightarrow{\boldsymbol{\theta}_{t-\tau}} \widetilde{a}_{t-\tau+1} \xrightarrow{\mathcal{P}} \widetilde{s}_{t-\tau+2} \xrightarrow{\boldsymbol{\theta}_{t-\tau}} \widetilde{a}_{t-\tau+2} \xrightarrow{\mathcal{P}} \cdots \xrightarrow{\mathcal{P}} \widetilde{s}_t \xrightarrow{\boldsymbol{\theta}_{t-\tau}} \widetilde{a}_t \xrightarrow{\mathcal{P}} \widetilde{s}_{t+1}.
$$

For reference, recall that the original Markov chain is given by:

$$
s_{t-\tau} \xrightarrow{\boldsymbol{\theta}_{t-\tau}} a_{t-\tau} \xrightarrow{\mathcal{P}} s_{t-\tau+1} \xrightarrow{\boldsymbol{\theta}_{t-\tau+1}} a_{t-\tau+1} \xrightarrow{\mathcal{P}} s_{t-\tau+2} \xrightarrow{\boldsymbol{\theta}_{t-\tau+2}} a_{t-\tau+2} \xrightarrow{\mathcal{P}} \cdots \xrightarrow{\mathcal{P}} s_t \xrightarrow{\boldsymbol{\theta}_t} a_t \xrightarrow{\mathcal{P}} s_{t+1}.
$$

By Lipschitz conditions, we can bound the first two terms in (A.5). The third term will be bounded by the total variation between s_k and $\widetilde{s}_k$, which is achieved by recursively bounding total variation between s_{k-1} and $\widetilde{s}_{k-1}$.

In fact, the Markovian noise $\Gamma(O_t, \boldsymbol{\theta}_t)$ in Section C.1 is obtained in a similar way. Due to the space limit, we only present how to bound the more complicated $\Lambda(\boldsymbol{\theta}_t, \boldsymbol{\omega}_t, O_t)$.
We have the final form as:

$$
\Lambda(\boldsymbol{\theta}_t, \boldsymbol{\omega}_t, O_t) \le C_1(\tau+1)\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-\tau}\| + C_2 m\rho^{\tau-1} + C_3\|\boldsymbol{\omega}_t - \boldsymbol{\omega}_{t-\tau}\|, \tag{A.6}
$$

where $C_1 = 2U_\delta^2|\mathcal{A}|L(1 + \lceil\log_\rho m^{-1}\rceil + 1/(1-\rho)) + 2U_\delta L_*, C_2 = 2U_\delta^2, C_3 = 4U_\delta$ are constants.

Step 3: integrating the results. By some calculation, terms I_1, I_2 and I_4 can be respectively bounded as follows (set $\tau = \tau_t$ defined in (4.1)). The detailed derivation can be found in Appendix C.3,

$$I_1 = 4R_{\boldsymbol{\omega}}^2 \frac{1}{\beta_t} = \mathcal{O}(t^{\nu}),$$

$$I_2 \leq C_1 G_{\boldsymbol{\theta}}(\tau_t + 1)^2 \sum_{k=0}^{t-\tau_t} \alpha_k + C_2(t - \tau_t + 1)\alpha_t + C_3 U_\delta \tau_t \sum_{k=0}^{t-\tau_t} \beta_k$$

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

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

$$I_5 = \sum_{k=0}^{t-\tau_t} \beta_k = \mathcal{O}(t^{1-\nu}).$$

The $\log t$ comes from $\tau_t = \mathcal{O}(\log t)$. Performing the same technique on I_3 as in Step 3 in the proof sketch of Theorem 4.5, we have

$$I_3 \leq \bigg(\sum_{k=0}^{t-\tau_t} \frac{\alpha_k^2}{\beta_k^2} \bigg)^{\frac{1}{2}} \bigg(\sum_{k=\tau_t}^{t} \mathbb{E}\|\mathbf{z}_k\|^2 \bigg)^{\frac{1}{2}},$$

$$I_4 \leq \bigg(\sum_{k=\tau_t}^{t} \mathbb{E}[y_k^2] \bigg)^{\frac{1}{2}} \bigg(\sum_{k=\tau_t}^{t} \mathbb{E}\|\mathbf{z}_k\|^2 \bigg)^{\frac{1}{2}}.$$

After plugging each term into (A.4), we have that

$$2\lambda \sum_{k=\tau_t}^{t} \mathbb{E}\|\mathbf{z}_k\|^2 \leq \mathcal{O}(t^{\nu}) + \mathcal{O}\big((\log t)t^{1-\nu}\big)$$

$$+ 2L_* G_{\boldsymbol{\theta}} \bigg(\sum_{k=0}^{t-\tau_t} \frac{\alpha_k^2}{\beta_k^2} \bigg)^{\frac{1}{2}} \bigg(\sum_{k=\tau_t}^{t} \mathbb{E}\|\mathbf{z}_k\|^2 \bigg)^{\frac{1}{2}} + \bigg(\sum_{k=0}^{t-\tau_t} \mathbb{E}[y_k^2] \bigg)^{\frac{1}{2}} \bigg(\sum_{k=\tau_t}^{t} \mathbb{E}\|\mathbf{z}_k\|^2 \bigg)^{\frac{1}{2}}.$$

This inequality actually resembles (A.2). Following the same procedure as the proof of Theorem 4.5, starting from (A.2), we can finally get

$$\frac{1}{1 + t - \tau_t} \sum_{k=\tau_t}^{t} \mathbb{E}\|\mathbf{z}_k\|^2 = \mathcal{O}\Big(\frac{1}{t^{1-\nu}}\Big) + \mathcal{O}\Big(\frac{\log t}{t^{\nu}}\Big) + \mathcal{O}\Big(\frac{1}{t^{2(\sigma-\nu)}}\Big).$$

Note that this requires the step sizes γ_t and β_t should be of the same order $\mathcal{O}(t^{-\nu})$.

B   Preliminary Lemmas

These useful lemmas are frequently applied throughout the proof.

B.1   Probabilistic Lemmas

The first two statements in the following lemma come from Zou et al. [44].

Lemma B.1. For any $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, it holds that

$$d_{TV}(\mu_{\boldsymbol{\theta}_1}, \mu_{\boldsymbol{\theta}_2}) \leq |\mathcal{A}|L\bigg(\lceil \log_\rho m^{-1} \rceil + \frac{1}{1-\rho} \bigg) \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|,$$

$$d_{TV}(\mu_{\boldsymbol{\theta}_1} \otimes \pi_{\boldsymbol{\theta}_1}, \mu_{\boldsymbol{\theta}_2} \otimes \pi_{\boldsymbol{\theta}_2}) \leq |\mathcal{A}|L\bigg(1 + \lceil \log_\rho m^{-1} \rceil + \frac{1}{1-\rho} \bigg) \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|,$$

$$d_{TV}(\mu_{\boldsymbol{\theta}_1} \otimes \pi_{\boldsymbol{\theta}_1} \otimes \mathcal{P}, \mu_{\boldsymbol{\theta}_2} \otimes \pi_{\boldsymbol{\theta}_2} \otimes \mathcal{P}) \leq |\mathcal{A}|L\bigg(1 + \lceil \log_\rho m^{-1} \rceil + \frac{1}{1-\rho} \bigg) \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.$$

Proof. The proof of the first two inequality is exactly the same as Lemma A.3 in Zou et al. [44], which mainly depends on Theorem 3.1 in Mitrophanov [19]. Here we provide the proof of the third

inequality. Note that

$$d_{TV}\left(\mu_{\boldsymbol{\theta}_1} \otimes \pi_{\boldsymbol{\theta}_1} \otimes \mathcal{P}, \mu_{\boldsymbol{\theta}_2} \otimes \pi_{\boldsymbol{\theta}_2} \otimes \mathcal{P}\right)$$

$$= \frac{1}{2} \int_{\mathcal{S}} \sum_{\mathcal{A}} \int_{\mathcal{S}} \left|\mu_{\boldsymbol{\theta}_1}(ds)\pi_{\boldsymbol{\theta}_1}(a|s)\mathcal{P}(ds'|s,a) - \mu_{\boldsymbol{\theta}_2}(ds)\pi_{\boldsymbol{\theta}_2}(a|s)\mathcal{P}(ds'|s,a)\right|$$

$$= \frac{1}{2} \int_{\mathcal{S}} \sum_{\mathcal{A}} \int_{\mathcal{S}} \mathcal{P}(ds'|s,a)\left|\mu_{\boldsymbol{\theta}_1}(ds)\pi_{\boldsymbol{\theta}_1}(a|s) - \mu_{\boldsymbol{\theta}_2}(ds)\pi_{\boldsymbol{\theta}_2}(a|s)\right|$$

$$= \frac{1}{2} \int_{\mathcal{S}} \sum_{\mathcal{A}} \left|\mu_{\boldsymbol{\theta}_1}(ds)\pi_{\boldsymbol{\theta}_1}(a|s) - \mu_{\boldsymbol{\theta}_2}(ds)\pi_{\boldsymbol{\theta}_2}(a|s)\right|$$

$$= d_{TV}\left(\mu_{\boldsymbol{\theta}_1} \otimes \pi_{\boldsymbol{\theta}_1}, \mu_{\boldsymbol{\theta}_2} \otimes \pi_{\boldsymbol{\theta}_2}\right), \tag{B.1}$$

so it has the same upper bound as the second inequality. □

Lemma B.2. Given time indexes t and τ such that $t \geq \tau > 0$, consider the auxiliary Markov chain starting from $s_{t-\tau}$. Conditioning on $s_{t-\tau+1}$ and $\boldsymbol{\theta}_{t-\tau}$, the Markov chain is obtained by repeatedly applying policy $\pi_{\boldsymbol{\theta}_{t-\tau}}$.

$$s_{t-\tau} \xrightarrow{\boldsymbol{\theta}_{t-\tau}} a_{t-\tau} \xrightarrow{\mathcal{P}} s_{t-\tau+1} \xrightarrow{\boldsymbol{\theta}_{t-\tau}} \widetilde{a}_{t-\tau+1} \xrightarrow{\mathcal{P}} \widetilde{s}_{t-\tau+2} \xrightarrow{\boldsymbol{\theta}_{t-\tau}} \widetilde{a}_{t-\tau+2} \xrightarrow{\mathcal{P}} \cdots \xrightarrow{\mathcal{P}} \widetilde{s}_t \xrightarrow{\boldsymbol{\theta}_{t-\tau}} \widetilde{a}_t \xrightarrow{\mathcal{P}} \widetilde{s}_{t+1}.$$

For reference, recall that the original Markov chain is given as:

$$s_{t-\tau} \xrightarrow{\boldsymbol{\theta}_{t-\tau}} a_{t-\tau} \xrightarrow{\mathcal{P}} s_{t-\tau+1} \xrightarrow{\boldsymbol{\theta}_{t-\tau+1}} a_{t-\tau+1} \xrightarrow{\mathcal{P}} s_{t-\tau+2} \xrightarrow{\boldsymbol{\theta}_{t-\tau+2}} a_{t-\tau+2} \xrightarrow{\mathcal{P}} \cdots \xrightarrow{\mathcal{P}} s_t \xrightarrow{\boldsymbol{\theta}_t} a_t \xrightarrow{\mathcal{P}} s_{t+1}.$$

Throughout this lemma, we always condition the expectation on $s_{t-\tau+1}$ and $\boldsymbol{\theta}_{t-\tau}$ and omit this in order to simplify the presentation. Under the setting introduced above, we have:

$$d_{TV}\left(\mathbb{P}(s_{t+1} \in \cdot), \mathbb{P}(\widetilde{s}_{t+1} \in \cdot)\right) \leq d_{TV}\left(\mathbb{P}(O_t \in \cdot), \mathbb{P}(\widetilde{O}_t \in \cdot)\right), \tag{B.2}$$

$$d_{TV}\left(\mathbb{P}(O_t \in \cdot), \mathbb{P}(\widetilde{O}_t \in \cdot)\right) = d_{TV}\left(\mathbb{P}((s_t, a_t) \in \cdot), \mathbb{P}((\widetilde{s}_t, \widetilde{a}_t) \in \cdot)\right), \tag{B.3}$$

$$d_{TV}\left(\mathbb{P}((s_t, a_t) \in \cdot), \mathbb{P}((\widetilde{s}_t, \widetilde{a}_t) \in \cdot)\right) \leq d_{TV}\left(\mathbb{P}(s_t \in \cdot), \mathbb{P}((\widetilde{s}_t \in \cdot)\right) + \frac{1}{2}|\mathcal{A}|L\mathbb{E}\left[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-\tau}\|\right]. \tag{B.4}$$

Proof of (B.2). By the Law of Total Probability,

$$\mathbb{P}(s_{t+1} \in \cdot) = \int_{\mathcal{S}} \sum_{\mathcal{A}} \mathbb{P}(s_t = ds, a_t = a, s_{t+1} \in \cdot),$$

and a similar argument also holds for $\widetilde{O}_t$. Then we have

$$2d_{TV}\left(\mathbb{P}(s_{t+1} \in \cdot), \mathbb{P}(\widetilde{s}_{t+1} \in \cdot)\right)$$

$$= \int_{\mathcal{S}} \left|\int_{\mathcal{S}} \sum_{\mathcal{A}} \mathbb{P}(s_t = ds, a_t = a, s_{t+1} = ds') - \int_{\mathcal{S}} \sum_{\mathcal{A}} \mathbb{P}(s_t = ds, a_t = a, s_{t+1} = ds')\right|$$

$$\leq \int_{\mathcal{S}} \int_{\mathcal{S}} \sum_{\mathcal{A}} \left|\mathbb{P}(s_t = ds, a_t = a, s_{t+1} = ds') - \mathbb{P}(s_t = ds, a_t = a, s_{t+1} = ds')\right|$$

$$= \int_{\mathcal{S}} \int_{\mathcal{S}} \sum_{\mathcal{A}} \left|\mathbb{P}(O_t = (ds, a, ds')) - \mathbb{P}(\widetilde{O}_t = (ds, a, ds'))\right|$$

$$= 2d_{TV}\left(\mathbb{P}(O_t \in \cdot), \mathbb{P}(\widetilde{O}_t \in \cdot)\right).$$

The last equality requires exchange of integral, which should be guaranteed by the regularity. □

Proof of (B.3).

$$2d_{TV}\big(\mathbb{P}(O_t \in \cdot), \mathbb{P}(\widetilde{O}_t \in \cdot)\big)$$

$$= \int_{\mathcal{S}} \sum_{\mathcal{A}} \int_{\mathcal{S}} \big|\mathbb{P}(O_t = (ds, a, ds')) - \mathbb{P}(\widetilde{O}_t = (ds, a, ds'))\big|$$

$$= \int_{\mathcal{S}} \sum_{\mathcal{A}} \int_{\mathcal{S}} \big|\mathcal{P}(ds'|s, a)\mathbb{P}((s_t, a_t) = (ds, a)) - \mathcal{P}(ds'|s, a)\mathbb{P}((\widetilde{s}_t, \widetilde{a}_t) = (ds, a))\big|$$

$$= \int_{\mathcal{S}} \sum_{\mathcal{A}} \int_{\mathcal{S}} \mathcal{P}(ds'|s, a)\big|\mathbb{P}((s_t, a_t) = (ds, a)) - \mathbb{P}((\widetilde{s}_t, \widetilde{a}_t) = (ds, a))\big|$$

$$= \int_{\mathcal{S}} \sum_{\mathcal{A}} \big|\mathbb{P}((s_t, a_t) = (ds, a)) - \mathbb{P}((\widetilde{s}_t, \widetilde{a}_t) = (ds, a))\big|$$

$$= 2d_{TV}\big(\mathbb{P}((s_t, a_t) \in \cdot), \mathbb{P}((\widetilde{s}_t, \widetilde{a}_t) \in \cdot)\big).$$

$\square$

Proof of (B.4). Because $\boldsymbol{\theta}_t$ is also dependent on s_t, we make it clear here that

$$\mathbb{P}\big((s_t, a_t) = (ds, a)\big) = \int_{\boldsymbol{\theta} \in \mathbb{R}^d} \mathbb{P}(s_t = ds)\mathbb{P}(\boldsymbol{\theta}_t = d\boldsymbol{\theta}|s_t = ds)\mathbb{P}(a_t = a|s_t = ds, \boldsymbol{\theta}_t = d\boldsymbol{\theta})$$

$$= \int_{\boldsymbol{\theta} \in \mathbb{R}^d} \mathbb{P}(s_t = ds)\mathbb{P}(\boldsymbol{\theta}_t = d\boldsymbol{\theta}|s_t = ds)\pi_{\boldsymbol{\theta}_t}(a|ds)$$

$$= \mathbb{P}(s_t = ds) \int_{\boldsymbol{\theta} \in \mathbb{R}^d} \mathbb{P}(\boldsymbol{\theta}_t = d\boldsymbol{\theta}|s_t = ds)\pi_{\boldsymbol{\theta}_t}(a|ds)$$

$$= \mathbb{P}(s_t = ds)\mathbb{E}\big[\pi_{\boldsymbol{\theta}_t}(a|ds)|s_t = ds\big].$$

Therefore, the total variance can be bounded as

$$2d_{TV}\big(\mathbb{P}((s_t, a_t) \in \cdot), \mathbb{P}((\widetilde{s}_t, \widetilde{a}_t) \in \cdot)\big)$$

$$= \int_{\mathcal{S}} \sum_{\mathcal{A}} \big|\mathbb{P}(s_t = ds)\mathbb{E}[\pi_{\boldsymbol{\theta}_t}(a|ds)|s_t = ds] - \mathbb{P}(\widetilde{s}_t = ds)\pi_{\boldsymbol{\theta}_{t-\tau}}(a|ds)\big|$$

$$= \int_{\mathcal{S}} \sum_{\mathcal{A}} \big|\mathbb{P}(s_t = ds)\mathbb{E}[\pi_{\boldsymbol{\theta}_t}(a|ds)|s_t = ds] - \mathbb{P}(s_t = ds)\pi_{\boldsymbol{\theta}_{t-\tau}}(a|ds)\big|$$

$$\qquad + \int_{\mathcal{S}} \sum_{\mathcal{A}} \big|\mathbb{P}(s_t = ds)\pi_{\boldsymbol{\theta}_{t-\tau}}(a|ds) - \mathbb{P}(\widetilde{s}_t = ds)\pi_{\boldsymbol{\theta}_{t-\tau}}(a|ds)\big|$$

$$= \int_{\mathcal{S}} \mathbb{P}(s_t = ds) \sum_{\mathcal{A}} \big|\mathbb{E}[\pi_{\boldsymbol{\theta}_t}(a|ds)|s_t = ds] - \pi_{\boldsymbol{\theta}_{t-\tau}}(a|ds)\big|$$

$$\qquad + 2d_{TV}\big(\mathbb{P}(s_t \in \cdot), \mathbb{P}((\widetilde{s}_t \in \cdot)\big)$$

$$\leq |\mathcal{A}|L\mathbb{E}\big[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-\tau}\|\big] + 2d_{TV}\big(\mathbb{P}(s_t \in \cdot), \mathbb{P}((\widetilde{s}_t \in \cdot)\big),$$

where the inequality holds due to the Lipschitz continuity of the policy as in Assumption 4.3. $\square$

B.2 Lipschitzness of the Optimal Parameter

This section is used to present the proof of Proposition 4.4.

Proof of Proposition 4.4. Sutton and Barto [27] has proved in Chapter 9 the fact that the linear TD(0) will converge to the optimal point (w.r.t. Mean Square Projected Bellman Error) which satisfies

$$\mathbf{A}_i \boldsymbol{\omega}^*(\boldsymbol{\theta}_i) = \mathbf{b}_i,$$

where $\mathbf{A}_i := \mathbb{E}[\boldsymbol{\phi}(s)(\boldsymbol{\phi}(s) - \boldsymbol{\phi}(s'))^\top]$ and $\mathbf{b}_i := \mathbb{E}[(r(s, a) - r(\boldsymbol{\theta}_i))\boldsymbol{\phi}(s)]$. The expectation is taken over the stationary distribution $s \sim \mu_{\boldsymbol{\theta}_i}$, the action $a \sim \pi_{\boldsymbol{\theta}_i}(\cdot|s)$ and the transition probability matrix $s' \sim \mathcal{P}(\cdot|s, a)$.

Now we denote $\boldsymbol{\omega}_1^*, \boldsymbol{\omega}_2^*, \widehat{\boldsymbol{\omega}}_1$ as the unique solutions of the following equations respectively:
$$\mathbf{A}_1\boldsymbol{\omega}_1^* = \mathbf{b}_1,$$
$$\mathbf{A}_2\widehat{\boldsymbol{\omega}}_1 = \mathbf{b}_1,$$
$$\mathbf{A}_2\boldsymbol{\omega}_2^* = \mathbf{b}_2.$$
First we bound $\|\boldsymbol{\omega}_1^* - \widehat{\boldsymbol{\omega}}_1\|$. By definition, we have
$$\|\boldsymbol{\omega}_1^* - \widehat{\boldsymbol{\omega}}_1\| \le \|\mathbf{A}_1^{-1} - \mathbf{A}_2^{-1}\|\|\mathbf{b}_1\|.$$
It can be easily shown that
$$\mathbf{A}_1^{-1} - \mathbf{A}_2^{-1} = \mathbf{A}_1^{-1}(\mathbf{A}_2 - \mathbf{A}_1)\mathbf{A}_2^{-1},$$
which further gives
$$\|\boldsymbol{\omega}_1^* - \widehat{\boldsymbol{\omega}}_1\| \le \|\mathbf{A}_1^{-1}\|\|\mathbf{A}_1 - \mathbf{A}_2\|\|\mathbf{A}_2^{-1}\|\|\mathbf{b}_1\|.$$
Then we bound $\|\widehat{\boldsymbol{\omega}}_1 - \boldsymbol{\omega}_2^*\|$,
$$\|\widehat{\boldsymbol{\omega}}_1 - \boldsymbol{\omega}_2^*\| \le \|\mathbf{A}_2^{-1}\|\|\mathbf{b}_1 - \mathbf{b}_2\|.$$
By Assumption 4.1, the eigenvalues of $\mathbf{A}_i$ are bounded from below by $\lambda > 0$, therefore $\|\mathbf{A}_i^{-1}\| \le \lambda^{-1}$. Also $\|\mathbf{b}_1\| \le U_r$ due to the assumption that $|r(s,a)| \le U_r$ and $\|\boldsymbol{\phi}(s)\| \le 1$. To bound $\|\mathbf{A}_1 - \mathbf{A}_2\|$ and $\|\mathbf{b}_1 - \mathbf{b}_2\|$, we first note that
$$\|\mathbf{A}_1 - \mathbf{A}_2\|_2 \le \sup_{s,s'\in\mathcal{S}} \|\boldsymbol{\phi}(s)(\boldsymbol{\phi}(s) - \boldsymbol{\phi}(s'))^\top\|_2 \cdot 2d_{TV}\big(\mathbb{P}(O^1 \in \cdot), \mathbb{P}(O^2 \in \cdot)\big),$$
$$\le 4d_{TV}\big(\mathbb{P}(O^1 \in \cdot), \mathbb{P}(O^2 \in \cdot)\big)$$
$$\|\mathbf{b}_1 - \mathbf{b}_2\| \le \big\|\mathbb{E}[r(s^1,a^1)\boldsymbol{\phi}(s^1)] - \mathbb{E}[r(s^2,a^2)\boldsymbol{\phi}(s^2)]\big\| + \big\|r(\boldsymbol{\theta}_1)\mathbb{E}[\boldsymbol{\phi}(s^1)] - r(\boldsymbol{\theta}_2)\mathbb{E}[\boldsymbol{\phi}(s^2)]\big\|$$
$$\le 6U_r d_{TV}\big(\mathbb{P}(O^1 \in \cdot), \mathbb{P}(O^2 \in \cdot)\big),$$
where O^i is the tuple obtained by $s^i \sim \mu_{\boldsymbol{\theta}_i}(\cdot)$, $a^i \sim \pi_{\boldsymbol{\theta}_i}(\cdot|s^i)$ and $(s')^i \sim \mathcal{P}(\cdot|s^i, a^i)$. And the total variation norm can be bounded by Lemma B.1 as:
$$d_{TV}\big(\mathbb{P}(O^1 \in \cdot), \mathbb{P}(O^2 \in \cdot)\big) \le |\mathcal{A}|L\Big(1 + \lceil\log_\rho m^{-1}\rceil + \frac{1}{1-\rho}\Big)\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.$$
Collecting the results above gives
$$\|\boldsymbol{\omega}_1^* - \boldsymbol{\omega}_2^*\| \le \|\boldsymbol{\omega}_1^* - \widehat{\boldsymbol{\omega}}_1\| + \|\widehat{\boldsymbol{\omega}}_1 - \boldsymbol{\omega}_2^*\|$$
$$\le (2\lambda^{-2}U_r + 3\lambda^{-1}U_r)|\mathcal{A}|L\Big(1 + \lceil\log_\rho m^{-1}\rceil + \frac{1}{1-\rho}\Big)\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|,$$
and we set $L_* := (2\lambda^{-2}U_r + 3\lambda^{-1}U_r)|\mathcal{A}|L(1 + \lceil\log_\rho m^{-1}\rceil + 1/(1-\rho))$ to obtain the final result. $\qquad\square$

B.3 Asymptotic Equivalence

Lemma B.3. Suppose $\{a_i\}$ is a non-negative, bounded sequence, $\tau := C_1 + C_2 \log t (C_2 > 0)$, then for any large enough t such that $t \ge \tau > 0$, we have:
$$\frac{1}{1+t-\tau}\sum_{k=\tau}^t a_i = \mathcal{O}\Big(\frac{1}{t}\sum_{k=1}^t a_i\Big),$$
$$\frac{1}{t}\sum_{k=1}^t a_i = \mathcal{O}\Big(\frac{\log t}{t}\Big) + \mathcal{O}\Big(\frac{1}{1+t-\tau}\sum_{k=\tau}^t a_i\Big).$$

Proof. We know that $\tau = \mathcal{O}(\log t)$ and the sequence is bounded: $0 < a_i < B$. For the first equation, we have
$$\frac{1}{1+t-\tau}\sum_{k=\tau}^t a_i \le \frac{1}{1+t-\tau}\sum_{k=1}^t a_i \le \frac{t}{1+t-\tau}\cdot\frac{1}{t}\sum_{k=1}^t a_i \le \mathcal{O}\Big(\frac{1}{t}\sum_{k=1}^t a_i\Big),$$
and further assuming $t \ge 2\tau - 2$ gives a constant 2. For the second equation, we have
$$\frac{1}{t}\sum_{k=1}^t a_i \le \frac{1}{t}\Big((\tau-1)B + \sum_{k=\tau}^t a_i\Big) = \frac{\tau-1}{t}B + \frac{1}{t}\sum_{k=\tau}^t a_i = \mathcal{O}\Big(\frac{\log t}{t}\Big) + \mathcal{O}\Big(\frac{1}{1+t-\tau}\sum_{k=\tau}^t a_i\Big).$$
$$\square$$

19

C Proof of Main Theorems and Propositions

C.1 Proof of Theorem 4.5

We first define several notations to clarify the dependence:

$$O_t := (s_t, a_t, s_{t+1}),$$
$$\eta^* := \eta(\boldsymbol{\theta}) = \mathbb{E}_{s \sim \mu_{\boldsymbol{\theta}}, a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)}[r(s,a)]$$
$$\Delta h(O, \eta, \boldsymbol{\omega}, \boldsymbol{\theta}) := \Big(\eta(\boldsymbol{\theta}) - \eta + \big(\phi(s') - \phi(s)\big)^\top (\boldsymbol{\omega} - \boldsymbol{\omega}^*)\Big) \nabla \log \pi_{\boldsymbol{\theta}}(a|s),$$
$$\Delta h'(O, \boldsymbol{\theta}) := \Big(\big(\phi(s')^\top \boldsymbol{\omega}^* - V^{\pi_{\boldsymbol{\theta}}}(s')\big) - \big(\phi(s)^\top \boldsymbol{\omega}^* - V^{\pi_{\boldsymbol{\theta}}}(s)\big)\Big) \nabla \log \pi_{\boldsymbol{\theta}}(a|s),$$
$$h(O, \boldsymbol{\theta}) := \big(r(s,a) - \eta(\boldsymbol{\theta}) + V^{\pi_{\boldsymbol{\theta}}}(s') - V^{\pi_{\boldsymbol{\theta}}}(s)\big) \nabla \log \pi_{\boldsymbol{\theta}}(a|s),$$
$$\Gamma(O, \boldsymbol{\theta}) := \big\langle \nabla J(\boldsymbol{\theta}), h(O, \boldsymbol{\theta}) - \nabla J(\boldsymbol{\theta}) \big\rangle. \tag{C.1}$$

Note that Δh, $\Delta h'$ and h together gives a decomposition of the actual gradient. They each correspond to the error caused by the critic $\boldsymbol{\omega}_t$, the approximation error of the linear class and the stochastic policy gradient.

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

Lemma C.1. For the performance function defined in (3.1), there exists a constant $L_J > 0$ such that for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$, it holds that

$$\big\|\nabla J(\boldsymbol{\theta}_1) - \nabla J(\boldsymbol{\theta}_2)\big\| \le L_J \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|,$$

which by the definition of smoothness [21] implies

$$J(\boldsymbol{\theta}_2) \ge J(\boldsymbol{\theta}_1) + \big\langle \nabla J(\boldsymbol{\theta}_1), \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1 \big\rangle - \frac{L_J}{2} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2.$$

The following two lemmas characterize the bias introduced by the critic's approximation and the Markovian noise.

Lemma C.2. For any $t \ge 0$,

$$\big\|\Delta h(O_t, \eta_t, \boldsymbol{\omega}_t, \boldsymbol{\theta}_t)\big\|^2 \le B^2 \big(8\|\boldsymbol{\omega}_t - \boldsymbol{\omega}_t^*\|^2 + 2(\eta_t - \eta_t^*)^2\big).$$

Lemma C.3. For any $\boldsymbol{\theta} \in \mathbb{R}^d$, we have $\|\delta \nabla \log \pi_{\boldsymbol{\theta}}(a|s)\| \le G_{\boldsymbol{\theta}} := U_\delta \cdot B$, where $U_\delta = 2U_r + 2R_{\boldsymbol{\omega}}$. Furthermore, for any $t \ge 0$, it holds that

$$\mathbb{E}\big[\Gamma(O_t, \boldsymbol{\theta}_t)\big] \ge -G_{\boldsymbol{\theta}}\Big(D_1(\tau+1) \sum_{k=t-\tau+1}^{t} \mathbb{E}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}\| + D_2 m \rho^{\tau-1}\Big),$$

where $D_1 = \max\{(U_\delta L_l + 2L_* B + 3L_J), 2U_\delta B|\mathcal{A}|L\}$ and $D_2 = 4U_\delta B$.

Proof of Theorem 4.5. Under the update rule of Algorithm 1, we have by Lemma C.1

$$
\begin{aligned}
J(\boldsymbol{\theta}_{t+1}) &\ge J(\boldsymbol{\theta}_t) + \alpha_t \big\langle \nabla J(\boldsymbol{\theta}_t), \delta_t \nabla \log \pi_{\boldsymbol{\theta}_t}(a_t|s_t) \big\rangle - L_J \alpha_t^2 \big\|\delta_t \nabla \log \pi_{\boldsymbol{\theta}_t}(a_t|s_t)\big\|^2 \\
&= J(\boldsymbol{\theta}_t) + \alpha_t \big\langle \nabla J(\boldsymbol{\theta}_t), \Delta h(O_t, \eta_t, \boldsymbol{\omega}_t, \boldsymbol{\theta}_t) \big\rangle + \alpha_t \big\langle \nabla J(\boldsymbol{\theta}_t), \Delta h'(O_t, \boldsymbol{\theta}_t) \big\rangle \\
&\quad + \alpha_t \big\langle \nabla J(\boldsymbol{\theta}_t), h(O_t, \boldsymbol{\theta}_t) \big\rangle - L_J \alpha_t^2 \big\|\delta_t \nabla \log \pi_{\boldsymbol{\theta}_t}(a_t|s_t)\big\|^2 \\
&= J(\boldsymbol{\theta}_t) + \alpha_t \big\langle \nabla J(\boldsymbol{\theta}_t), \Delta h(O_t, \eta_t, \boldsymbol{\omega}_t, \boldsymbol{\theta}_t) \big\rangle + \alpha_t \big\langle \nabla J(\boldsymbol{\theta}_t), \Delta h'(O_t, \boldsymbol{\theta}_t) \big\rangle \\
&\quad + \alpha_t \Gamma(O_t, \boldsymbol{\theta}_t) + \alpha_t \big\|\nabla J(\boldsymbol{\theta}_t)\big\|^2 - L_J \alpha_t^2 \big\|\delta_t \nabla \log \pi_{\boldsymbol{\theta}_t}(a_t|s_t)\big\|^2. \tag{C.2}
\end{aligned}
$$

We will bound the expectation of each term on the right hand side of (C.2) as follows. First, we have

$$\mathbb{E}\big\langle \nabla J(\boldsymbol{\theta}_t), \Delta h(O_t, \eta_t, \boldsymbol{\omega}_t, \boldsymbol{\theta}_t) \big\rangle \ge -B\sqrt{\mathbb{E}\big\|\nabla J(\boldsymbol{\theta}_t)\big\|^2} \sqrt{8\mathbb{E}\|\mathbf{z}_t\|^2 + 2\mathbb{E}[y_t^2]},$$

where $\mathbf{z}_t := \boldsymbol{\omega}_t - \boldsymbol{\omega}_t^*$ and $y_t := \eta_t - \eta_t^*$, and the inequality is due to Cauchy inequality and Lemma C.2.

20

Second, we have

$$\mathbb{E}[\Gamma(O_t, \boldsymbol{\theta}_t)] \geq -G_{\boldsymbol{\theta}}\left(D_1(\tau+1)\sum_{k=t-\tau+1}^{t}\mathbb{E}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}\| + D_2 m\rho^{\tau-1}\right),$$

$$\geq -G_{\boldsymbol{\theta}}\left(D_1(\tau+1)G_{\boldsymbol{\theta}}\sum_{k=t-\tau+1}^{t-1}\alpha_k + D_2 m\rho^{\tau-1}\right),$$

where the first inequality is due to Lemma C.3, and the second inequality is due to $\left\|\delta_t \nabla \log \pi_{\boldsymbol{\theta}_t}(a_t|s_t)\right\| \leq G_{\boldsymbol{\theta}}$ by Lemma C.3.

Third, taking expectation over the approximation error term containing $\Delta h'$, we have

$$\mathbb{E}\langle \nabla J(\boldsymbol{\theta}_t), \Delta h'(O_t, \boldsymbol{\theta}_t)\rangle \geq -G_{\boldsymbol{\theta}}\sqrt{\mathbb{E}\|\Delta h'(O_t, \boldsymbol{\theta}_t)\|^2}$$

$$\geq -G_{\boldsymbol{\theta}} \cdot 2B\sqrt{\mathbb{E}\left(\phi(s)^\top \boldsymbol{\omega}_t^* - V^{\pi_{\boldsymbol{\theta}_t}}(s)\right)^2}$$

$$\geq -2BG_{\boldsymbol{\theta}}\epsilon_{\text{app}},$$

Taking the expectation of (C.2) and plugging the above terms back into it gives

$$\mathbb{E}[J(\boldsymbol{\theta}_{t+1})] \geq \mathbb{E}[J(\boldsymbol{\theta}_t)] - \alpha_t B\sqrt{\mathbb{E}\|\nabla J(\boldsymbol{\theta}_t)\|^2}\sqrt{8\mathbb{E}\|\mathbf{z}_t\|^2 + 2\mathbb{E}[y_t^2]} - 2BG_{\boldsymbol{\theta}}\epsilon_{\text{app}}\alpha_t$$

$$- \alpha_t G_{\boldsymbol{\theta}}\left(D_1(\tau+1)G_{\boldsymbol{\theta}}\sum_{k=t-\tau}^{t-1}\alpha_k + D_2 m\rho^{\tau-1}\right) + \alpha_t \mathbb{E}\|\nabla J(\boldsymbol{\theta}_t)\|^2 - L_J G_{\boldsymbol{\theta}}^2\alpha_t^2.$$

Rearranging the above inequality gives

$$\mathbb{E}\|\nabla J(\boldsymbol{\theta}_t)\|^2 \leq \frac{1}{\alpha_t}\left(\mathbb{E}[J(\boldsymbol{\theta}_{t+1})] - \mathbb{E}[J(\boldsymbol{\theta}_t)]\right) + B\sqrt{\mathbb{E}\|\nabla J(\boldsymbol{\theta}_t)\|^2}\sqrt{8\mathbb{E}\|\mathbf{z}_t\|^2 + 2\mathbb{E}[y_t^2]}$$

$$+ D_1 G_{\boldsymbol{\theta}}^2(\tau+1)\sum_{k=t-\tau}^{t-1}\alpha_k + D_2 G_{\boldsymbol{\theta}}m\rho^{\tau-1} + L_J G_{\boldsymbol{\theta}}^2\alpha_t.$$

By setting $\tau = \tau_t$, we get

$$\mathbb{E}\|\nabla J(\boldsymbol{\theta}_t)\|^2 \leq \frac{1}{\alpha_t}\left(\mathbb{E}[J(\boldsymbol{\theta}_{t+1})] - \mathbb{E}[J(\boldsymbol{\theta}_t)]\right) + B\sqrt{\mathbb{E}\|\nabla J(\boldsymbol{\theta}_t)\|^2}\sqrt{8\mathbb{E}\|\mathbf{z}_t\|^2 + 2\mathbb{E}[y_t^2]}$$

$$+ 2BG_{\boldsymbol{\theta}}\epsilon_{\text{app}} + D_1 G_{\boldsymbol{\theta}}^2(\tau_t+1)^2\alpha_{t-\tau_t} + D_2 G_{\boldsymbol{\theta}}\alpha_t + L_J G_{\boldsymbol{\theta}}^2\alpha_t.$$

Summing over k from τ_t to t gives

$$\sum_{k=\tau_t}^{t}\mathbb{E}\|\nabla J(\boldsymbol{\theta}_t)\|^2 \leq \underbrace{\sum_{k=\tau_t}^{t}\frac{1}{\alpha_k}\left(\mathbb{E}[J(\boldsymbol{\theta}_{k+1})] - \mathbb{E}[J(\boldsymbol{\theta}_k)]\right)}_{I_1} + B\sum_{k=\tau_t}^{t}\sqrt{\mathbb{E}\|\nabla J(\boldsymbol{\theta}_t)\|^2}\sqrt{8\mathbb{E}\|\mathbf{z}_t\|^2 + 2\mathbb{E}[y_t^2]}$$

$$+ \underbrace{\sum_{k=\tau_t}^{t}D_1 G_{\boldsymbol{\theta}}^2(\tau_t+1)^2\alpha_{k-\tau_t} + \sum_{k=\tau_t}^{t}(D_2 G_{\boldsymbol{\theta}} + L_J G_{\boldsymbol{\theta}}^2)\alpha_k}_{I_2} + 2BG_{\boldsymbol{\theta}}\epsilon_{\text{app}}(t - \tau_t + 1).$$

For the term I_1, we have,

$$\sum_{k=\tau_t}^{t}\frac{1}{\alpha_k}\left(J(\boldsymbol{\theta}_{k+1}) - J(\boldsymbol{\theta}_k)\right) = \sum_{k=\tau_t}^{t}\left(\frac{1}{\alpha_{k-1}} - \frac{1}{\alpha_k}\right)\mathbb{E}[J(\boldsymbol{\theta}_k)] - \frac{1}{\alpha_{\tau_t-1}}\mathbb{E}[J(\boldsymbol{\theta}_{\tau_t})] + \frac{1}{\alpha_t}\mathbb{E}[J(\boldsymbol{\theta}_{t+1})]$$

$$\leq \sum_{k=\tau_t}^{t}\left(\frac{1}{\alpha_k} - \frac{1}{\alpha_{k-1}}\right)U_r + \frac{1}{\alpha_{\tau_t-1}}U_r + \frac{1}{\alpha_t}U_r$$

$$= U_r\left[\sum_{k=\tau_t}^{t}\left(\frac{1}{\alpha_k} - \frac{1}{\alpha_{k-1}}\right) + \frac{1}{\alpha_{\tau_t-1}} + \frac{1}{\alpha_t}\right]$$

$$= 2U_r\alpha_t^{-1},$$

where the inequality holds due to $|\mathbb{E}[J(\boldsymbol{\theta})]| \le U_r$.

For the term I_2, we have

$$
\begin{aligned}
\sum_{k=\tau_t}^{t} D_1 G_{\boldsymbol{\theta}}^2 (\tau_t + 1)^2 \alpha_{k-\tau_t} &= D_1 G_{\boldsymbol{\theta}}^2 (\tau_t + 1)^2 \sum_{k=\tau_t}^{t} \alpha_{k-\tau_t} \\
&= D_1 G_{\boldsymbol{\theta}}^2 (\tau_t + 1)^2 \sum_{k=0}^{t-\tau_t} \alpha_k \\
&= D_1 G_{\boldsymbol{\theta}}^2 (\tau_t + 1)^2 c_\alpha \sum_{k=0}^{t-\tau_t} \frac{1}{(1+k)^\sigma},
\end{aligned}
$$

and

$$
\begin{aligned}
\sum_{k=\tau_t}^{t} (D_2 G_{\boldsymbol{\theta}} + L_J G_{\boldsymbol{\theta}}^2) \alpha_k &= (D_2 G_{\boldsymbol{\theta}} + L_J G_{\boldsymbol{\theta}}^2) \sum_{k=\tau_t}^{t} \alpha_k \\
&\le (D_2 G_{\boldsymbol{\theta}} + L_J G_{\boldsymbol{\theta}}^2) \sum_{k=0}^{t-\tau_t} \alpha_k \\
&= (D_2 G_{\boldsymbol{\theta}} + L_J G_{\boldsymbol{\theta}}^2) c_\alpha \sum_{k=0}^{t-\tau_t} \frac{1}{(1+k)^\sigma}.
\end{aligned}
$$

Note that both upper bounds rely on the summation $\sum_{k=0}^{t-\tau_t} 1/(1+k)^\sigma \le \int_0^{t-\tau_t+1} x^{-\sigma} dx = 1/(1-\sigma)(t-\tau_t+1)^{1-\sigma}$. Combining the results for terms I_1 and I_2, we have

$$
\begin{aligned}
\sum_{k=\tau_t}^{t} \mathbb{E}\big\|\nabla J(\boldsymbol{\theta}_t)\big\|^2 \le\ & \frac{2U_r}{c_\alpha}(1+t)^\sigma \\
& + \big(D_1 G_{\boldsymbol{\theta}}^2 (\tau_t + 1)^2 + D_2 G_{\boldsymbol{\theta}} + L_J G_{\boldsymbol{\theta}}^2\big) \frac{c_\alpha}{1-\sigma}(t - \tau_t + 1)^{1-\sigma} \\
& + B \sum_{k=\tau_t}^{t} \sqrt{\mathbb{E}\big\|\nabla J(\boldsymbol{\theta}_t)\big\|^2} \sqrt{8\mathbb{E}\|\mathbf{z}_t\|^2 + 2\mathbb{E}[y_t^2]} \\
& + 2BG_{\boldsymbol{\theta}} \epsilon_{\mathrm{app}}(t - \tau_t + 1).
\end{aligned}
$$

Dividing $(1 + t - \tau_t)$ at both sides and assuming $t > 2\tau_t - 1$, we can express the result as

$$
\begin{aligned}
\frac{1}{1+t-\tau_t} \sum_{k=\tau_t}^{t} \mathbb{E}\big\|\nabla J(\boldsymbol{\theta}_t)\big\|^2 \le\ & \frac{4U_r}{c_\alpha}\frac{1}{(t+1)^{1-\sigma}} \\
& + \big(D_1 G_{\boldsymbol{\theta}}^2 (\tau_t + 1)^2 + D_2 G_{\boldsymbol{\theta}} + L_J G_{\boldsymbol{\theta}}^2\big) \frac{c_\alpha}{1-\sigma}\frac{1}{(t - \tau_t + 1)^\sigma} \\
& + \frac{2B}{1+t-\tau_t} \sum_{k=\tau_t}^{t} \sqrt{\mathbb{E}\big\|\nabla J(\boldsymbol{\theta}_t)\big\|^2} \sqrt{8\mathbb{E}\|\mathbf{z}_t\|^2 + 2\mathbb{E}[y_t^2]} \\
& + 2BG_{\boldsymbol{\theta}} \epsilon_{\mathrm{app}}. \quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad (\text{C.3})
\end{aligned}
$$

By Cauchy-Schwartz inequality, we have

$$
\begin{aligned}
& \frac{1}{1+t-\tau_t} \sum_{k=\tau_t}^{t} \sqrt{\mathbb{E}\big\|\nabla J(\boldsymbol{\theta}_t)\big\|^2} \sqrt{8\mathbb{E}\|\mathbf{z}_t\|^2 + 2\mathbb{E}[y_t^2]} \\
& \le \left(\frac{1}{1+t-\tau_t} \sum_{k=\tau_t}^{t} \mathbb{E}\big\|\nabla J(\boldsymbol{\theta}_t)\big\|^2\right)^{\frac{1}{2}} \left(\frac{1}{1+t-\tau_t} \sum_{k=\tau_t}^{t} \big(8\mathbb{E}\|\mathbf{z}_t\|^2 + 2\mathbb{E}[y_t^2]\big)\right)^{\frac{1}{2}}.
\end{aligned}
$$

Now, denote $F(t) := 1/(1 + t - \tau_t) \sum_{k=\tau_t}^{t} \mathbb{E}\|\nabla J(\boldsymbol{\theta}_k)\|^2$ and $Z(t) := 1/(1 + t - \tau_t) \sum_{k=\tau_t}^{t} \left(8\mathbb{E}\|\mathbf{z}_t\|^2 + 2\mathbb{E}[y_t^2]\right)$, and putting them back to (C.3) ($\mathcal{O}$-notation for simplicity):

$$F(t) \leq \mathcal{O}\left(\frac{1}{t^{1-\sigma}}\right) + \mathcal{O}\left(\frac{(\log t)^2}{t^\sigma}\right) + \mathcal{O}(\epsilon_{\text{app}}) + 2B\sqrt{F(t)} \cdot \sqrt{Z(t)},$$

which further gives

$$\left(\sqrt{F(t)} - B\sqrt{Z(t)}\right)^2 \leq \mathcal{O}\left(\frac{1}{t^{1-\sigma}}\right) + \mathcal{O}\left(\frac{(\log t)^2}{t^\sigma}\right) + \mathcal{O}(\epsilon_{\text{app}}) + B^2 Z(t). \qquad \text{(C.4)}$$

Note that for a general function $H(t) \leq A(t) + B(t)$(with each positive), we have

$$H^2(t) \leq 2A^2(t) + 2B^2(t),$$
$$\sqrt{H(t)} \leq \sqrt{A(t)} + \sqrt{B(t)}.$$

This means (C.4) implies

$$\sqrt{F(t)} - B\sqrt{Z(t)} \leq \sqrt{A(t)} + B\sqrt{Z(t)},$$
$$\sqrt{F(t)} \leq \sqrt{A(t)} + 2B\sqrt{Z(t)},$$
$$F(t) \leq 2A(t) + 8B^2 Z(t).$$

By Lemma B.3, assuming $t \geq 2\tau_t - 1$, it holds that

$$Z(t) = \frac{1}{1 + t - \tau_t} \sum_{k=\tau_t}^{t} 8\mathbb{E}\|\mathbf{z}_k\|^2 + 2\mathbb{E}[y_t^2] \leq \frac{2}{t} \sum_{k=1}^{t} 8\mathbb{E}\|\mathbf{z}_k\|^2 + 2\mathbb{E}[y_t^2] = 2\mathcal{E}(t).$$

And finally, we have

$$\begin{aligned}
\min_{0 \leq k \leq t} \mathbb{E}\|\nabla J(\boldsymbol{\theta}_k)\|^2 &\leq \frac{1}{1 + t - \tau_t} \sum_{k=\tau_t}^{t} \mathbb{E}\|\nabla J(\boldsymbol{\theta}_k)\|^2 \\
&\leq \frac{8U_r}{c_\alpha} \frac{1}{(t+1)^{1-\sigma}} \\
&\quad + \left(D_1 G_{\boldsymbol{\theta}}^2 (\tau_t + 1)^2 + D_2 G_{\boldsymbol{\theta}} + L_J G_{\boldsymbol{\theta}}^2\right) \frac{2c_\alpha}{1 - \sigma} \frac{1}{(t - \tau_t + 1)^\sigma} \\
&\quad + 4BG_{\boldsymbol{\theta}} \epsilon_{\text{app}} \\
&\quad + 16B^2 \mathcal{E}(t) \\
&= \mathcal{O}\left(\frac{1}{t^{1-\sigma}}\right) + \mathcal{O}\left(\frac{1}{t^\sigma}\right) + \mathcal{O}(\epsilon_{\text{app}}) + \mathcal{O}(\mathcal{E}(t)).
\end{aligned}$$

$\qquad \square$

C.2 Proof of Theorem 4.7: Estimating the Average Reward

The two time-scale analysis with Markovian noise and moving behavior policy can be complicated, so we define some useful notations here that could hopefully clarify the probabilistic dependency.

$$\begin{aligned}
O_t &:= (s_t, a_t, s_{t+1}), \\
\eta_t^* &:= \eta^*(\boldsymbol{\theta}_t) = J(\boldsymbol{\theta}_t), \\
y_t &:= \eta_t - \eta_t^*, \\
\Xi(O, \eta, \boldsymbol{\theta}) &:= y_t(r_t - \eta_t^*).
\end{aligned} \qquad \text{(C.5)}$$

We also write $J(\boldsymbol{\theta}_t) = r(\boldsymbol{\theta}_t)$ sometimes in the proof.

Lemma C.4. For any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$, we have

$$\left|J(\boldsymbol{\theta}_1) - J(\boldsymbol{\theta}_2)\right| \leq C_J \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|,$$

where $C_J = 2U_r |\mathcal{A}| L(1 + \lceil \log_\rho m^{-1} \rceil + 1/(1 - \rho))$.

Lemma C.5. Given the definition of $\Xi(O_t, \eta_t, \boldsymbol{\theta}_t)$, for any $t > 0$, we have

$$\mathbb{E}[\Xi(O_t, \eta_t, \boldsymbol{\theta}_t)] \leq 4U_r C_J \|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-\tau}\| + 2U_r |\eta_t - \eta_{t-\tau}| + 2U_r^2 |\mathcal{A}| L \sum_{i=t-\tau}^{t} \mathbb{E}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{t-\tau}\|. + 4U_r^2 m \rho^{\tau-1}.$$

Proof. From the definition, η_t is the average reward estimator, $\eta_t^* = J(\boldsymbol{\theta}_t) = \mathbb{E}[r(s, a)]$ is the average reward under the stationary distribution $\mu_{\boldsymbol{\theta}_t} \otimes \pi_{\boldsymbol{\theta}_t}$, and $y_t = \eta_t - \eta_t^*$. From the algorithm we have the update rule as

$$\eta_{t+1} := \eta_t + \gamma_t\big(r(s_t, a_t) - \eta_t\big),$$

where we leave the step size γ_t unspecified for now. Unrolling the recursive definition we have

$$\begin{aligned}
y_{t+1}^2 &= \big(y_t + \eta_t^* - \eta_{t+1}^* + \gamma_t(r_t - \eta_t)\big)^2 \\
&\leq y_t^2 + 2\gamma_t y_t(r_t - \eta_t) + 2y_t(\eta_t^* - \eta_{t+1}^*) + 2(\eta_t^* - \eta_{t+1}^*)^2 + 2\gamma_t^2(r_t - \eta_t)^2 \\
&= (1 - 2\gamma_t)y_t^2 + 2\gamma_t y_t(r_t - \eta_t^*) + 2y_t(\eta_t^* - \eta_{t+1}^*) + 2(\eta_t^* - \eta_{t+1}^*)^2 + 2\gamma_t^2(r_t - \eta_t)^2 \\
&= (1 - 2\gamma_t)y_t^2 + 2\gamma_t \Xi(O_k, \eta_k, \boldsymbol{\theta}_k) + 2y_t(\eta_t^* - \eta_{t+1}^*) + 2(\eta_t^* - \eta_{t+1}^*)^2 + 2\gamma_t^2(r_t - \eta_t)^2.
\end{aligned}$$

Rearranging and summing from τ_t to t, we have

$$\sum_{k=\tau_t}^{t} \mathbb{E}[y_k^2] \leq \underbrace{\sum_{k=\tau_t}^{t} \frac{1}{2\gamma_k}\mathbb{E}(y_k^2 - y_{k+1}^2)}_{I_1} + \underbrace{\sum_{k=\tau_t}^{t} \mathbb{E}[\Xi(O_k, \eta_k, \boldsymbol{\theta}_k)]}_{I_2}$$

$$+ \underbrace{\sum_{k=\tau_t}^{t} \frac{1}{\gamma_k}\mathbb{E}[y_k(\eta_k^* - \eta_{k+1}^*)]}_{I_3} + \underbrace{\sum_{k=\tau_t}^{t} \frac{1}{\gamma_k}\mathbb{E}[(\eta_k^* - \eta_{k+1}^*)^2]}_{I_4} + \underbrace{\sum_{k=\tau_t}^{t} \gamma_k \mathbb{E}[(r_k - \eta_k)^2]}_{I_5}.$$

For I_1, following the Abel summation formula, we have

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

For I_2, from Lemma C.5, we have

$$\mathbb{E}[\Xi(O_t, \eta_t, \boldsymbol{\theta}_t)] \leq 4U_r C_J \|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-\tau}\| + 2U_r |\eta_t - \eta_{t-\tau}| + 2U_r^2 |\mathcal{A}| L \sum_{i=t-\tau}^{t} \mathbb{E}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{t-\tau}\|. + 4U_r^2 m \rho^{\tau-1}$$

$$\leq 4U_r C_J G_{\boldsymbol{\theta}} \tau \alpha_{t-\tau} + 4U_r^2 \tau \gamma_{t-\tau} + 2U_r^2 |\mathcal{A}| L \tau (\tau + 1) G_{\boldsymbol{\theta}} \alpha_{t-\tau} + 4U_r^2 m \rho^{\tau-1}$$

$$\leq C_1 \tau^2 \alpha_{t-\tau} + C_2 \tau \gamma_{t-\tau} + C_3 m \rho^{\tau-1}.$$

By the choice of τ_t, we have

$$I_2 = \sum_{k=\tau_t}^{t} \mathbb{E}[\Xi(O_k, \eta_k, \boldsymbol{\theta}_k)] \leq (C_1 \tau_t^2 + C_3) \sum_{k=\tau_t}^{t} \alpha_k + C_2 \tau_t \sum_{k=\tau_t}^{t} \gamma_k.$$

For I_3, we have

$$I_3 \leq \left(\sum_{k=\tau_t}^{t} \mathbb{E}[y_k^2]\right)^{1/2} \left(C_J^2 G_{\boldsymbol{\theta}}^2 \sum_{k=\tau_t}^{t} \frac{\alpha_k^2}{\gamma_k^2}\right)^{1/2},$$

which is because by Lemma C.4, $(\eta_k^* - \eta_{k+1}^*)$ can be linearly bounded by $\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k+1}\| \le G_{\boldsymbol{\theta}} \cdot \alpha_k$. For I_4, by the same argument it holds that

$$
\begin{aligned}
I_4 &= \sum_{k=\tau_t}^{t} \frac{1}{\gamma_k} \mathbb{E}[(\eta_k^* - \eta_{k+1}^*)^2] \\
&= \sum_{k=\tau_t}^{t} \frac{1}{\gamma_k} \mathbb{E}\big[\big(J(\boldsymbol{\theta}_k) - J(\boldsymbol{\theta}_{k+1})\big)^2\big] \\
&\le \sum_{k=\tau_t}^{t} \frac{1}{\gamma_k} C_J^2 \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k+1}\|^2 \\
&\le \sum_{k=\tau_t}^{t} \frac{1}{\gamma_k} C_J^2 G_{\boldsymbol{\theta}}^2 \alpha_k^2 \\
&= \mathcal{O}\bigg(\sum_{k=\tau_t}^{t} \frac{\alpha_k^2}{\gamma_k} \bigg).
\end{aligned}
$$

For I_5, we have

$$
\begin{aligned}
I_5 &= \sum_{k=\tau_t}^{t} \gamma_k \mathbb{E}[(r_k - \eta_k)^2] \\
&\le \sum_{k=\tau_t}^{t} 4 U_r^2 \gamma_k \\
&= \mathcal{O}\bigg(\sum_{k=\tau_t}^{t} \gamma_k \bigg),
\end{aligned}
$$

by bounding the expectation uniformly.

Now, we set $\gamma_k = 1/(1+t)^\nu$ and combine all the terms together to get

$$
\begin{aligned}
\sum_{k=\tau_t}^{t} \mathbb{E}[y_k^2] &\le 2 U_r^2 (1+t)^\nu + (C_1 \tau_t^2 + C_3) c_\alpha \sum_{k=\tau_t}^{t} (1+k)^{-\sigma} + C_2 \tau_t \sum_{k=\tau_t}^{t} (1+k)^{-\nu} \\
&\quad + C_J G_{\boldsymbol{\theta}} c_\alpha \bigg(\sum_{k=\tau_t}^{t} \mathbb{E}[y_k^2] \bigg)^{1/2} \bigg(\sum_{k=\tau_t}^{t} (1+k)^{-2(\sigma-\nu)} \bigg)^{1/2} \\
&\quad + C_J^2 G_{\boldsymbol{\theta}}^2 c_\alpha^2 \sum_{k=\tau_t}^{t} (1+k)^{\nu-2\sigma} + 4 U_r^2 \sum_{k=\tau_t}^{t} (1+k)^{-\nu} \\
&\le 2 U_r^2 (1+t)^\nu + \big[(C_1 \tau^2 + C_3) c_\alpha + C_2 \tau_t + C_J^2 G_{\boldsymbol{\theta}}^2 c_\alpha^2 + 4 U_r^2 \big] \sum_{k=\tau_t}^{t} (1+k)^{-\nu} \\
&\quad + C_J G_{\boldsymbol{\theta}} c_\alpha \bigg(\sum_{k=\tau_t}^{t} \mathbb{E}[y_k^2] \bigg)^{1/2} \bigg(\sum_{k=\tau_t}^{t} (1+k)^{-2(\sigma-\nu)} \bigg)^{1/2} \\
&\le 2 U_r^2 (1+t)^\nu + \big[(C_1 \tau^2 + C_3) c_\alpha + C_2 \tau_t + C_J^2 G_{\boldsymbol{\theta}}^2 c_\alpha^2 + 4 U_r^2 \big] \frac{(1+t-\tau_t)^{1-\nu}}{1-\nu} \\
&\quad + C_J G_{\boldsymbol{\theta}} c_\alpha \bigg(\sum_{k=\tau_t}^{t} \mathbb{E}[y_k^2] \bigg)^{1/2} \bigg(\frac{(1+t-\tau_t)^{1-2(\sigma-\nu)}}{1-2(\sigma-\nu)} \bigg)^{1/2}
\end{aligned}
$$

By applying the squaring technique already stated in the proof of Theorem 4.5, we have that

$$\sum_{k=\tau_t}^{t} \mathbb{E}[y_k^2] \leq 4U_r^2(1+t)^\nu + 2\big[(C_1\tau_t^2 + C_3)c_\alpha + C_2\tau_t + C_J^2 G_{\boldsymbol\theta}^2 c_\alpha^2 + 4U_r^2\big]\frac{(1+t-\tau_t)^{1-\nu}}{1-\nu}$$

$$+ 8C_J^2 G_{\boldsymbol\theta}^2 c_\alpha^2 \frac{(1+t-\tau_t)^{1-2(\sigma-\nu)}}{1-2(\sigma-\nu)} \tag{C.6}$$

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

$\square$

C.3 Proof of Theorem 4.7: Approximating the TD Fixed Point

Now we deal with the critic's parameter $\boldsymbol\omega_t$. The two time-scale analysis with Markovian noise and moving behavior policy can be complicated, so we define some useful notations here that could hopefully clarify the probabilistic dependency.

$$O_t := (s_t, a_t, s_{t+1}),$$
$$g(O, \boldsymbol\omega, \boldsymbol\theta) := [r(s,a) - J(\boldsymbol\theta) + (\boldsymbol\phi(s') - \boldsymbol\phi(s))^\top\boldsymbol\omega]\boldsymbol\phi(s),$$
$$\Delta g(O, \eta, \boldsymbol\theta) := [J(\boldsymbol\theta) - \eta]\boldsymbol\phi(s),$$
$$\bar{g}(\boldsymbol\omega, \boldsymbol\theta) := \mathbb{E}_{s\sim\mu_{\boldsymbol\theta}, a\sim\pi_{\boldsymbol\theta}, s'\sim\mathcal{P}}\Big[\big[r(s,a) - J(\boldsymbol\theta) + (\boldsymbol\phi(s') - \boldsymbol\phi(s))^\top\boldsymbol\omega\big]\boldsymbol\phi(s)\Big],$$
$$\boldsymbol\omega_t^* := \boldsymbol\omega^*(\boldsymbol\theta_t),$$
$$\eta_t^* := \eta^*(\boldsymbol\theta_t) = J(\boldsymbol\theta_t)$$
$$\Lambda(O, \boldsymbol\omega, \boldsymbol\theta) := \big\langle\boldsymbol\omega - \boldsymbol\omega^*(\boldsymbol\theta), g(O, \boldsymbol\omega, \boldsymbol\theta) - \bar{g}(\boldsymbol\omega, \boldsymbol\theta)\big\rangle,$$
$$\mathbf{z}_t := \boldsymbol\omega_t - \boldsymbol\omega_t^*$$
$$y_t := \eta_t - \eta_t^*. \tag{C.7}$$

A bounded lemma is used frequently in this section.

Lemma C.6. Under Assumption 4.3, for any $\boldsymbol\theta$, $\boldsymbol\omega$, $O = (s, a, s')$ such that $\|\boldsymbol\omega\| \leq R_{\boldsymbol\omega}$,

$$\big\|g(O, \boldsymbol\omega, \boldsymbol\theta)\big\| \leq U_\delta := 2U_r + 2R_{\boldsymbol\omega},$$
$$\big\|\Delta g(O, \eta, \boldsymbol\theta)\big\| \leq 2U_r,$$
$$\big|\Lambda(O, \boldsymbol\omega, \boldsymbol\theta)\big| \leq 2R_{\boldsymbol\omega} \cdot 2U_\delta \leq 2U_\delta^2.$$

The following lemma is used to control the bias due to Markovian noise.

Lemma C.7. Given the definition of $\Lambda(\boldsymbol\theta_t, \boldsymbol\omega_t, O_t)$, for any $0 \leq \tau \leq t$, we have

$$\mathbb{E}[\Lambda(O_t, \boldsymbol\omega_t, \boldsymbol\theta_t)] \leq C_1(\tau+1)\|\boldsymbol\theta_t - \boldsymbol\theta_{t-\tau}\| + C_2 m\rho^{\tau-1} + C_3\|\boldsymbol\omega_t - \boldsymbol\omega_{t-\tau}\|,$$

where $C_1 = 2U_\delta^2|\mathcal{A}|L(1 + \lceil\log_\rho m^{-1}\rceil + 1/(1-\rho)) + 2U_\delta L_*, C_2 = 2U_\delta^2, C_3 = 4U_\delta$ are constants.

Proof of Theorem 4.7. By the updating rule of $\boldsymbol\omega_t$ in Algorithm 1, unrolling and decomposing the squared error gives

$$\|\mathbf{z}_{t+1}\|^2 = \big\|\mathbf{z}_t + \beta_t(g(O_t, \boldsymbol\omega_t, \boldsymbol\theta_t) + \Delta g(O_t, \eta_t, \boldsymbol\theta_t)) + (\boldsymbol\omega_t^* - \boldsymbol\omega_{t+1}^*)\big\|^2$$
$$= \|\mathbf{z}_t\|^2 + 2\beta_t\big\langle\mathbf{z}_t, g(O_t, \boldsymbol\omega_t, \boldsymbol\theta_t)\big\rangle + 2\beta_t\big\langle\mathbf{z}_t, \Delta g(O_t, \eta_t, \boldsymbol\theta_t)\big\rangle$$
$$\quad + 2\big\langle\mathbf{z}_t, \boldsymbol\omega_t^* - \boldsymbol\omega_{t+1}^*\big\rangle + \big\|\beta_t(g(O_t, \boldsymbol\omega_t, \boldsymbol\theta_t) + \Delta g(O_t, \eta_t, \boldsymbol\theta_t)) + (\boldsymbol\omega_t^* - \boldsymbol\omega_{t+1}^*)\big\|^2$$
$$= \|\mathbf{z}_t\|^2 + 2\beta_t\big\langle\mathbf{z}_t, \bar{g}(\boldsymbol\omega_t, \boldsymbol\theta_t)\big\rangle + 2\beta_t\Lambda(O_t, \boldsymbol\omega_t, \boldsymbol\theta_t) + 2\beta_t\big\langle\mathbf{z}_t, \Delta g(O_t, \eta_t, \boldsymbol\theta_t)\big\rangle$$
$$\quad + 2\big\langle\mathbf{z}_t, \boldsymbol\omega_t^* - \boldsymbol\omega_{t+1}^*\big\rangle + \big\|\beta_t(g(O_t, \boldsymbol\omega_t, \boldsymbol\theta_t) + \Delta g(O_t, \eta_t, \boldsymbol\theta_t)) + (\boldsymbol\omega_t^* - \boldsymbol\omega_{t+1}^*)\big\|^2$$
$$\leq \|\mathbf{z}_t\|^2 + 2\beta_t\big\langle\mathbf{z}_t, \bar{g}(\boldsymbol\omega_t, \boldsymbol\theta_t)\big\rangle + 2\beta_t\Lambda(O_t, \boldsymbol\omega_t, \boldsymbol\theta_t) + 2\beta_t\big\langle\mathbf{z}_t, \Delta g(O_t, \eta_t, \boldsymbol\theta_t)\big\rangle$$
$$\quad + 2\big\langle\mathbf{z}_t, \boldsymbol\omega_t^* - \boldsymbol\omega_{t+1}^*\big\rangle + 2\beta_t^2\big\|g(O_t, \boldsymbol\omega_t, \boldsymbol\theta_t) + \Delta g(O_t, \eta_t, \boldsymbol\theta_t)\big\|^2 + 2\|\boldsymbol\omega_t^* - \boldsymbol\omega_{t+1}^*\|^2$$
$$\leq \|\mathbf{z}_t\|^2 + 2\beta_t\big\langle\mathbf{z}_t, \bar{g}(\boldsymbol\omega_t, \boldsymbol\theta_t)\big\rangle + 2\beta_t\Lambda(O_t, \boldsymbol\omega_t, \boldsymbol\theta_t) + 2\beta_t\big\langle\mathbf{z}_t, \Delta g(O_t, \eta_t, \boldsymbol\theta_t)\big\rangle$$
$$\quad + 2\big\langle\mathbf{z}_t, \boldsymbol\omega_t^* - \boldsymbol\omega_{t+1}^*\big\rangle + 2U_\delta^2\beta_t^2 + 2\|\boldsymbol\omega_t^* - \boldsymbol\omega_{t+1}^*\|^2,$$

where the first inequality is due to $\|\mathbf{x}+\mathbf{y}\|^2 \le 2\|\mathbf{x}\|^2+2\|\mathbf{y}\|^2$ and the second is due to $\|g(O_t,\boldsymbol{\omega}_t,\boldsymbol{\theta}_t)+\Delta g(O_t,\eta_t,\boldsymbol{\theta}_t)\| \le U_\delta$. First, note that due to Assumption 4.1, we have

$$
\begin{aligned}
\langle \mathbf{z}_t, \bar{g}(\boldsymbol{\omega}_t,\boldsymbol{\theta}_t)\rangle &= \langle \mathbf{z}_t, \bar{g}(\boldsymbol{\omega}_t,\boldsymbol{\theta}_t) - \bar{g}(\boldsymbol{\omega}_t^*,\boldsymbol{\theta}_t)\rangle \\
&= \Big\langle \mathbf{z}_t, \mathbb{E}\big[(\boldsymbol{\phi}(s') - \boldsymbol{\phi}(s))^\top (\boldsymbol{\omega}_t - \boldsymbol{\omega}_t^*)\boldsymbol{\phi}(s)\big]\Big\rangle \\
&= \mathbf{z}_t^\top \mathbb{E}\big[\boldsymbol{\phi}(s)\big(\boldsymbol{\phi}(s') - \boldsymbol{\phi}(s)\big)^\top\big]\mathbf{z}_t \\
&= \mathbf{z}_t^\top \mathbf{A} \mathbf{z}_t \\
&\le -\lambda \|\mathbf{z}_t\|^2,
\end{aligned}
$$

where the first equation is due to the fact that $\bar{g}(\boldsymbol{\omega}^*,\boldsymbol{\theta}) = 0$ [27]. Taking expectation up to s_{t+1}, we have

$$
\begin{aligned}
\mathbb{E}\|\mathbf{z}_{t+1}\|^2 &\le \mathbb{E}\|\mathbf{z}_t\|^2 + 2\beta_t\mathbb{E}\langle \mathbf{z}_t, \bar{g}(\boldsymbol{\omega}_t,\boldsymbol{\theta}_t)\rangle + 2\beta_t\mathbb{E}\Lambda(O_t,\boldsymbol{\omega}_t,\boldsymbol{\theta}_t) + 2\beta_t\mathbb{E}\langle \mathbf{z}_t, \Delta g(O_t,\eta_t,\boldsymbol{\theta}_t)\rangle \\
&\quad + 2\mathbb{E}\langle \mathbf{z}_t, \boldsymbol{\omega}_t^* - \boldsymbol{\omega}_{t+1}^*\rangle + 2U_\delta^2\beta_t^2 + 2\mathbb{E}\|\boldsymbol{\omega}_t^* - \boldsymbol{\omega}_{t+1}^*\|^2 \\
&\le (1 - 2\lambda\beta_t)\mathbb{E}\|\mathbf{z}_t\|^2 + 2\beta_t\mathbb{E}\Lambda(O_t,\boldsymbol{\omega}_t,\boldsymbol{\theta}_t) + 2\beta_t\mathbb{E}\langle \mathbf{z}_t, \Delta g(O_t,\eta_t,\boldsymbol{\theta}_t)\rangle \\
&\quad + 2\mathbb{E}\langle \mathbf{z}_t, \boldsymbol{\omega}_t^* - \boldsymbol{\omega}_{t+1}^*\rangle + 2U_\delta^2\beta_t^2 + 2\mathbb{E}\|\boldsymbol{\omega}_t^* - \boldsymbol{\omega}_{t+1}^*\|^2.
\end{aligned}
$$

Based on the result above, we can further rewrite it as:

$$
\begin{aligned}
\mathbb{E}\|\mathbf{z}_{t+1}\|^2 &\le (1 - 2\lambda\beta_t)\mathbb{E}\|\mathbf{z}_t\|^2 + 2\beta_t\mathbb{E}\Lambda(O_t,\boldsymbol{\omega}_t,\boldsymbol{\theta}_t) + 2\beta_t\mathbb{E}\|\mathbf{z}_t\| \cdot |y_t| \\
&\quad + 2L_*\mathbb{E}\|\mathbf{z}_t\| \cdot \|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}\| + 2U_\delta^2\beta_t^2 + 2L_*^2\mathbb{E}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}\|^2 \\
&\le (1 - 2\lambda\beta_t)\mathbb{E}\|\mathbf{z}_t\|^2 + 2\beta_t\mathbb{E}\Lambda(O_t,\boldsymbol{\omega}_t,\boldsymbol{\theta}_t) + 2\beta_t\mathbb{E}\|\mathbf{z}_t\| \cdot |y_t| \\
&\quad + 2L_*G_{\boldsymbol{\theta}}\alpha_t\mathbb{E}\|\mathbf{z}_t\| + 2U_\delta^2\beta_t^2 + 2L_*^2G_{\boldsymbol{\theta}}^2\alpha_t^2 \\
&\le (1 - 2\lambda\beta_t)\mathbb{E}\|\mathbf{z}_t\|^2 + 2\beta_t\mathbb{E}\Lambda(O_t,\boldsymbol{\omega}_t,\boldsymbol{\theta}_t) + 2\beta_t\mathbb{E}\|\mathbf{z}_t\| \cdot |y_t| \\
&\quad + 2L_*G_{\boldsymbol{\theta}}\alpha_t\mathbb{E}\|\mathbf{z}_t\| + \left(2U_\delta^2 + 2L_*^2G_{\boldsymbol{\theta}}^2\Big(\max_t \frac{\alpha_t}{\beta_t}\Big)^2\right)\beta_t^2 \\
&= (1 - 2\lambda\beta_t)\mathbb{E}\|\mathbf{z}_t\|^2 + 2\beta_t\mathbb{E}\Lambda(O_t,\boldsymbol{\omega}_t,\boldsymbol{\theta}_t) + 2\beta_t\mathbb{E}\|\mathbf{z}_t\| \cdot |y_t| + 2L_*G_{\boldsymbol{\theta}}\alpha_t\mathbb{E}\|\mathbf{z}_t\| + C_q\beta_t^2,
\end{aligned}
$$

where we denote the constant coefficient before the quadratic stepsize β_t^2 as C_q at the last step. The first inequality is due to Proposition 4.4 and Cauchy-Schwartz inequality. The second inequality is due to the update of $\boldsymbol{\theta}_t$ is bounded by $G_{\boldsymbol{\theta}}\alpha_t$. The third inequality is from employing the fact that $\sigma > \nu$ so α_t/β_t is bounded. Rearranging the inequality yields

$$
\begin{aligned}
2\lambda\mathbb{E}\|\mathbf{z}_t\|^2 &\le \frac{1}{\beta_t}\big(\mathbb{E}\|\mathbf{z}_t\|^2 - \mathbb{E}\|\mathbf{z}_{t+1}\|^2\big) + 2\mathbb{E}\Lambda(O_t,\boldsymbol{\omega}_t,\boldsymbol{\theta}_t) + \mathbb{E}\|\mathbf{z}_t\| \cdot |y_t| + 2L_*G_{\boldsymbol{\theta}}\frac{\alpha_t}{\beta_t}\mathbb{E}\|\mathbf{z}_t\| + C_q\beta_t \\
&\le \frac{1}{\beta_t}\big(\mathbb{E}\|\mathbf{z}_t\|^2 - \mathbb{E}\|\mathbf{z}_{t+1}\|^2\big) + 2\mathbb{E}\Lambda(O_t,\boldsymbol{\omega}_t,\boldsymbol{\theta}_t) + \sqrt{\mathbb{E}y_t^2} \cdot \sqrt{\mathbb{E}\|\mathbf{z}_t\|^2} + 2L_*G_{\boldsymbol{\theta}}\frac{\alpha_t}{\beta_t}\sqrt{\mathbb{E}\|\mathbf{z}_t\|^2} + C_q\beta_t,
\end{aligned}
$$

where the second inequality is due to the concavity of square root function. Telescoping from τ_t to t gives:

$$
2\lambda\sum_{k=\tau_t}^{t}\mathbb{E}\|\mathbf{z}_k\|^2 \le \underbrace{\sum_{k=\tau_t}^{t}\frac{1}{\beta_k}\big(\mathbb{E}\|\mathbf{z}_k\|^2 - \mathbb{E}\|\mathbf{z}_{k+1}\|^2\big)}_{I_1} + 2\underbrace{\sum_{k=\tau_t}^{t}\mathbb{E}\Lambda(\boldsymbol{\theta}_k,\boldsymbol{\omega}_k,O_k)}_{I_2}
$$

$$
+ 2L_*G_{\boldsymbol{\theta}}\underbrace{\sum_{k=\tau_t}^{t}\frac{\alpha_k}{\beta_k}\sqrt{\mathbb{E}\|\mathbf{z}_k\|^2}}_{I_3} + \underbrace{\sum_{k=\tau_t}^{t}\sqrt{\mathbb{E}y_k^2} \cdot \sqrt{\mathbb{E}\|\mathbf{z}_k\|^2}}_{I_4} + C_q\underbrace{\sum_{k=\tau_t}^{t}\beta_k}_{I_5}. \quad \text{(C.8)}
$$

From (C.8), we can see the proof of the critic again shares the same spirit with the proof of Theorem 4.5. For term I_1, we have

$$I_1 := \sum_{k=\tau_t}^{t} \frac{1}{\beta_k} (\mathbb{E}\|\mathbf{z}_k\|^2 - \mathbb{E}\|\mathbf{z}_{k+1}\|^2)$$

$$= \sum_{k=\tau_t}^{t} \left(\frac{1}{\beta_k} - \frac{1}{\beta_{k-1}} \right) \mathbb{E}\|\mathbf{z}_k\|^2 + \frac{1}{\beta_{\tau_t-1}} \mathbb{E}\|\mathbf{z}_{\tau_t}\|^2 - \frac{1}{\beta_t} \mathbb{E}\|\mathbf{z}_{t+1}\|^2$$

$$\leq \sum_{k=\tau_t}^{t} \left(\frac{1}{\beta_k} - \frac{1}{\beta_{k-1}} \right) \mathbb{E}\|\mathbf{z}_k\|^2 + \frac{1}{\beta_{\tau_t-1}} \mathbb{E}\|\mathbf{z}_{\tau_t}\|^2$$

$$\leq 4R_{\boldsymbol{\omega}}^2 \left(\sum_{k=\tau_t}^{t} \left(\frac{1}{\beta_k} - \frac{1}{\beta_{k-1}} \right) + \frac{1}{\beta_{\tau_t-1}} \right)$$

$$= 4R_{\boldsymbol{\omega}}^2 \frac{1}{\beta_t}$$

$$= 4R_{\boldsymbol{\omega}}^2 (1+t)^{\nu} = \mathcal{O}(t^{\nu}),$$

where the first inequality is due to discarding the last term, and the second inequality is due to $\mathbb{E}\|\mathbf{z}_k\|^2 \leq (R_{\boldsymbol{\omega}} + R_{\boldsymbol{\omega}})^2$.

For term I_2, note that due to Lemma C.7, we actually have

$$\Lambda(O_k, \boldsymbol{\omega}_k, \boldsymbol{\theta}_k) \leq C_1(\tau_t + 1)\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-\tau_t}\| + C_2 m\rho^{\tau_t-1} + C_3\|\boldsymbol{\omega}_k - \boldsymbol{\omega}_{k-\tau_t}\|$$

$$\leq C_1(\tau_t + 1) \sum_{i=k-\tau_t}^{k-1} G_{\boldsymbol{\theta}}\alpha_i + C_2 m\rho^{\tau_t-1} + C_3 \sum_{i=k-\tau_t}^{k-1} U_{\delta}\beta_i$$

$$\leq C_1 G_{\boldsymbol{\theta}}(\tau_t + 1)^2 \alpha_{k-\tau_t} + C_2\alpha_t + C_3 U_{\delta}\tau_t\beta_k,$$

and the summation is

$$I_2 := \sum_{k=\tau_t}^{t} \mathbb{E}\Lambda(O_k, \boldsymbol{\omega}_k, \boldsymbol{\theta}_k)$$

$$\leq C_1 G_{\boldsymbol{\theta}}(\tau_t + 1)^2 \sum_{k=\tau_t}^{t} \alpha_{k-\tau_t} + C_2 \sum_{k=\tau_t}^{t} \alpha_t + C_3 U_{\delta}\tau_t \sum_{k=\tau_t}^{t} \beta_k$$

$$\leq C_1 G_{\boldsymbol{\theta}}(\tau_t + 1)^2 \sum_{k=0}^{t-\tau_t} \alpha_k + C_2(t - \tau_t + 1)\alpha_t + C_3 U_{\delta}\tau_t \sum_{k=0}^{t-\tau_t} \beta_k$$

$$\leq C_1 G_{\boldsymbol{\theta}}(\tau_t + 1)^2 c_{\alpha} \frac{(1 + t - \tau_t)^{1-\sigma}}{1-\sigma} + C_2(t - \tau_t + 1)c_{\alpha}(1+t)^{-\sigma} + C_3 U_{\delta}\tau_t \frac{(1 + t - \tau_t)^{1-\nu}}{1-\nu}$$

$$\leq \left[\frac{C_1 G_{\boldsymbol{\theta}}(\tau_t + 1)^2 c_{\alpha}}{1-\sigma} + C_2 c_{\alpha} + \frac{C_3 U_{\delta}\tau_t}{1-\nu} \right] (1+t)^{1-\nu}$$

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

where the second inequality is due to the monotonicity of α_k and β_k. The $\mathcal{O}(\cdot)$ comes from that $\tau = \mathcal{O}(\log t)$ and $\sum k^{-\nu} = \mathcal{O}(t^{1-\nu})$.

For term I_3 and I_4, we will instead show it can be bounded in a different form. Using Cauchy-Schwartz inequality we have

$$I_3 := \sum_{k=\tau_t}^{t} \frac{\alpha_k}{\beta_k} \sqrt{\mathbb{E}\|\mathbf{z}_k\|^2} \leq \left(\sum_{k=\tau_t}^{t} \frac{\alpha_k^2}{\beta_k^2} \right)^{\frac{1}{2}} \left(\sum_{k=\tau_t}^{t} \mathbb{E}\|\mathbf{z}_k\|^2 \right)^{\frac{1}{2}} \leq \left(\sum_{k=0}^{t-\tau_t} \frac{\alpha_k^2}{\beta_k^2} \right)^{\frac{1}{2}} \left(\sum_{k=\tau_t}^{t} \mathbb{E}\|\mathbf{z}_k\|^2 \right)^{\frac{1}{2}},$$

$$I_4 := \sum_{k=\tau_t}^{t} \sqrt{\mathbb{E}y_k^2} \cdot \sqrt{\mathbb{E}\|\mathbf{z}_k\|^2} \leq \left(\sum_{k=\tau_t}^{t} \mathbb{E}y_k^2 \right)^{\frac{1}{2}} \left(\sum_{k=\tau_t}^{t} \mathbb{E}\|\mathbf{z}_k\|^2 \right)^{\frac{1}{2}} \leq \left(\sum_{k=0}^{t-\tau_t} \mathbb{E}y_k^2 \right)^{\frac{1}{2}} \left(\sum_{k=\tau_t}^{t} \mathbb{E}\|\mathbf{z}_k\|^2 \right)^{\frac{1}{2}}.$$

For term I_5, simply bound it as $\sum_{k=0}^{t-\tau_t} \beta_k \le (1+t)^{1-\nu}/(1-\nu)$.

Collecting the upper bounds of the above five terms, and writing them using $\mathcal{O}(\cdot)$ notation give

$$2\lambda \sum_{k=\tau_t}^{t} \mathbb{E}\|\mathbf{z}_k\|^2 \le 4R_{\boldsymbol{\omega}}^2 (1+t)^{\nu} + 2\left[\frac{C_1 G_{\boldsymbol{\theta}}(\tau_t+1)^2 c_{\alpha}}{1-\sigma} + C_2 c_{\alpha} + \frac{C_3 U_{\delta}\tau_t + C_q}{1-\nu}\right](1+t)^{1-\nu}$$

$$+ 2L_* G_{\boldsymbol{\theta}}\left(\sum_{k=0}^{t-\tau_t} \frac{\alpha_k^2}{\beta_k^2}\right)^{\frac{1}{2}}\left(\sum_{k=\tau_t}^{t} \mathbb{E}\|\mathbf{z}_k\|^2\right)^{\frac{1}{2}}$$

$$+ \left(\sum_{k=0}^{t-\tau_t} \mathbb{E}y_k^2\right)^{\frac{1}{2}}\left(\sum_{k=\tau_t}^{t} \mathbb{E}\|\mathbf{z}_k\|^2\right)^{\frac{1}{2}}. \tag{C.9}$$

Now, we first divide both sides by $(1+t-\tau_t)$, and denote

$$Z(t) := \frac{1}{1+t-\tau_t}\sum_{k=\tau_t}^{t} \mathbb{E}\|\mathbf{z}_k\|^2,$$

$$F(t) := \frac{1}{1+t-\tau_t}\sum_{k=0}^{t-\tau_t} \frac{\alpha_k^2}{\beta_k^2} \le \frac{t^{-2(\sigma-\nu)}}{1-2(\sigma-\nu)} = \mathcal{O}(t^{-2(\sigma-\nu)}),$$

$$G(t) := \frac{1}{1+t-\tau_t}\sum_{k=0}^{t-\tau_t} \mathbb{E}[y_k^2] = \mathcal{O}(t^{\nu-1}) + \mathcal{O}(\log t \cdot t^{-\nu}) + \mathcal{O}(t^{-2(\sigma-\nu)}),$$

and the rest as $A(t) = \mathcal{O}(t^{\nu}) + \mathcal{O}(t^{1-\nu})$. $G(t)$'s constants appear at (C.6) in exact form. This simplification leads to

$$2\lambda\left(\sqrt{Z(t)} - \frac{L_* G_{\boldsymbol{\theta}}}{2\lambda}\cdot\sqrt{F(t)} - \frac{1}{4\lambda}\sqrt{G(t)}\right)^2 \le A(t) + 2\lambda\left(\frac{L_* G_{\boldsymbol{\theta}}}{2\lambda}\sqrt{F(t)} + \frac{1}{4\lambda}\sqrt{G(t)}\right)^2,$$

which further gives

$$Z(t) \le A(t)/\lambda + 16F(t) + 16G(t).$$

This is again a similar reasoning as in the end of the proof of Theorem 4.5. We actually show that

$$\frac{1}{1+t-\tau_t}\sum_{k=\tau_t}^{t} \mathbb{E}\|\boldsymbol{\omega}_k - \boldsymbol{\omega}_k^*\|^2 = \mathcal{O}\left(\frac{1}{t^{1-\nu}}\right) + \mathcal{O}\left(\frac{\log t}{t^{\nu}}\right) + \mathcal{O}\left(\frac{1}{t^{2(\sigma-\nu)}}\right).$$

This completes the proof. To obtain the exact constant, please refer to (C.6) and (C.9). $\qquad\square$

C.4 Proof of Corollary 4.9

Proof of Corollary 4.9. By Theorem 4.7, we have

$$\frac{1}{1+t-\tau_t}\sum_{k=\tau_t}^{t} \mathbb{E}\|\boldsymbol{\omega}_k - \boldsymbol{\omega}_k^*\|^2 = \mathcal{O}\left(\frac{1}{t^{1-\nu}}\right) + \mathcal{O}\left(\frac{\log t}{t^{\nu}}\right) + \mathcal{O}\left(\frac{1}{t^{2(\sigma-\nu)}}\right).$$

By Lemma B.3, $\mathcal{E}(t)$ in Theorem 4.5 is of the equivalent order:

$$\mathcal{E}_1(t) = \frac{1}{t}\sum_{k=1}^{t} \mathbb{E}\|\boldsymbol{\omega}_k - \boldsymbol{\omega}_k^*\|^2$$

$$= \mathcal{O}\left(\frac{1}{1+t-\tau_t}\sum_{k=\tau_t}^{t} \mathbb{E}\|\boldsymbol{\omega}_k - \boldsymbol{\omega}_k^*\|^2\right) + \mathcal{O}\left(\frac{\log t}{t}\right)$$

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

$$= \mathcal{O}\left(\frac{1}{t^{1-\nu}}\right) + \mathcal{O}\left(\frac{\log t}{t^{\nu}}\right) + \mathcal{O}\left(\frac{1}{t^{2(\sigma-\nu)}}\right).$$

The same reasoning also applies to

$$\mathcal{E}_2(t) = \frac{1}{t}\sum_{k=1}^{t}\mathbb{E}(\eta_k - r(\boldsymbol{\theta}_k))^2$$

$$= \mathcal{O}\left(\frac{1}{t^{1-\nu}}\right) + \mathcal{O}\left(\frac{\log t}{t^\nu}\right) + \mathcal{O}\left(\frac{1}{t^{2(\sigma-\nu)}}\right).$$

Plugging the above results into Theorem 4.5, and optimizing over the choice of σ and ν (which gives $\sigma = 3/5$ and $\nu = 2/5$), we have

$$\min_{0\leq k\leq t}\mathbb{E}\|\nabla J(\boldsymbol{\theta}_k)\|^2 = \mathcal{O}\left(\frac{1}{t^{1-\sigma}}\right) + \mathcal{O}\left(\frac{\log^2 t}{t^\sigma}\right) + \mathcal{O}\left(\frac{1}{t^{1-\nu}}\right) + \mathcal{O}\left(\frac{\log t}{t^\nu}\right) + \mathcal{O}\left(\frac{1}{t^{2(\sigma-\nu)}}\right) + \mathcal{O}(\epsilon_{\text{app}})$$

$$= \mathcal{O}\left(\frac{1}{t^{1-\sigma}}\right) + \mathcal{O}\left(\frac{\log t}{t^\nu}\right) + \mathcal{O}\left(\frac{1}{t^{2(\sigma-\nu)}}\right) + \mathcal{O}(\epsilon_{\text{app}})$$

$$= \mathcal{O}\left(\frac{\log t}{t^{2/5}}\right) + \mathcal{O}(\epsilon_{\text{app}}).$$

Therefore, in order to obtain an ϵ-approximate(ignoring the approximation error) stationary point of J, namely,

$$\min_{0\leq k\leq T}\mathbb{E}\|\nabla J(\boldsymbol{\theta}_k)\|^2 = \mathcal{O}\left(\frac{\log T}{T^{2/5}}\right) + \mathcal{O}(\epsilon_{\text{app}}) \leq \mathcal{O}(\epsilon_{\text{app}}) + \epsilon,$$

we need to set $T = \widetilde{\mathcal{O}}(\epsilon^{-2.5})$. $\qquad\square$

D   Proof of Technical Lemmas

D.1   Proof of Lemma C.1

Proof of Lemma C.1. The first inequality comes from Lemma 3.2 in Zhang et al. [42].

The second inequality is well known as a partial result of $[-L, L]$-smoothness of non-convex functions.
$\qquad\square$

D.2   Proof of Lemma C.2

Proof of Lemma C.2. Applying the definition of $\Delta h()$ and Cauchy-Schwartz inequality immediately yields the result. $\qquad\square$

D.3   Proof of Lemma C.3

The proof of Lemma C.3 will be built on the following supporting lemmas.

Lemma D.1. For any $t \geq 0$,
$$\left|\Gamma(O_t, \boldsymbol{\theta}_t) - \Gamma(O_t, \boldsymbol{\theta}_{t-\tau})\right| \leq G_{\boldsymbol{\theta}}(U_\delta L_l + 2L_*B + 3L_J)\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-\tau}\|.$$

Lemma D.2. For any $t \geq 0$,
$$\left|\mathbb{E}[\Gamma(O_t, \boldsymbol{\theta}_{t-\tau}) - \Gamma(\widetilde{O}_t, \boldsymbol{\theta}_{t-\tau})]\right| \leq 2U_\delta BG_{\boldsymbol{\theta}}|\mathcal{A}|L\sum_{i=t-\tau}^{t}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{t-\tau}\|.$$

Lemma D.3. For any $t \geq 0$,
$$\left|\mathbb{E}[\Gamma(\widetilde{O}_t, \boldsymbol{\theta}_{t-\tau}) - \Gamma(O'_t, \boldsymbol{\theta}_{t-\tau})]\right| \leq 4U_\delta BG_{\boldsymbol{\theta}}m\rho^{\tau-1}.$$

Proof of Lemma C.3. First note that
$$\delta = \left|r(s,a) - J(\boldsymbol{\theta}) + \boldsymbol{\phi}^\top(s')\boldsymbol{\omega} - \boldsymbol{\phi}^\top(s)\boldsymbol{\omega}\right|$$
$$\leq \left|r(s,a)\right| + \left|J(\boldsymbol{\theta})\right| + \left|\boldsymbol{\phi}^\top(s')\boldsymbol{\omega}\right| + \left|\boldsymbol{\phi}^\top(s)\boldsymbol{\omega}\right|$$
$$= 2U_r + 2R_{\boldsymbol{\omega}}$$
$$=: U_\delta,$$

which immediately implies

$$\big\|\delta\nabla\log\pi_{\boldsymbol{\theta}}(a|s)\big\| \le |\delta| \cdot \big\|\nabla\log\pi_{\boldsymbol{\theta}}(a|s)\big\| \le U_\delta \cdot B,$$

where the last inequality is due to Assumption 4.3. We decompose the Markovian bias as

$$\mathbb{E}[\Gamma(O_t, \boldsymbol{\theta}_t)] = \mathbb{E}[\Gamma(O_t, \boldsymbol{\theta}_t) - \Gamma(O_t, \boldsymbol{\theta}_{t-\tau})] + \mathbb{E}[\Gamma(O_t, \boldsymbol{\theta}_{t-\tau}) - \Gamma(\widetilde{O}_t, \boldsymbol{\theta}_{t-\tau})]$$
$$+ \mathbb{E}[\Gamma(\widetilde{O}_t, \boldsymbol{\theta}_{t-\tau}) - \Gamma(O'_t, \boldsymbol{\theta}_{t-\tau})] + \mathbb{E}[\Gamma(O'_t, \boldsymbol{\theta}_{t-\tau})],$$

where $\widetilde{O}_t$ is from the auxiliary Markovian chain and O'_t is from the stationary distribution which actually satisfy $\mathbb{E}[\Gamma(O'_t, \boldsymbol{\theta}_{t-\tau})] = 0$. By collecting the corresponding bounds from Lemmas D.1, D.2 and D.3, we have that

$$\mathbb{E}[\Gamma(O_t, \boldsymbol{\theta}_t)] \ge -G_{\boldsymbol{\theta}}(U_\delta L_l + 2L_* B + 3L_J)\mathbb{E}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-\tau}\| - 2U_\delta B G_{\boldsymbol{\theta}}|\mathcal{A}|L\sum_{i=t-\tau}^{t}\mathbb{E}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{t-\tau}\|$$

$$- 4U_\delta B G_{\boldsymbol{\theta}} m\rho^{\tau-1}$$

$$\ge -G_{\boldsymbol{\theta}}(U_\delta L_l + 2L_* B + 3L_J)\sum_{i=t-\tau+1}^{t}\mathbb{E}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|$$

$$- 2U_\delta B G_{\boldsymbol{\theta}}|\mathcal{A}|L\sum_{i=t-\tau+1}^{t}\sum_{j=t-\tau+1}^{i}\mathbb{E}\|\boldsymbol{\theta}_j - \boldsymbol{\theta}_{j-1}\| - 4U_\delta B G_{\boldsymbol{\theta}} m\rho^{\tau-1}$$

$$\ge -G_{\boldsymbol{\theta}}(U_\delta L_l + 2L_* B + 3L_J)\sum_{i=t-\tau+1}^{t}\mathbb{E}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|$$

$$- 2U_\delta B G_{\boldsymbol{\theta}}|\mathcal{A}|L\tau\sum_{j=t-\tau+1}^{t}\mathbb{E}\|\boldsymbol{\theta}_j - \boldsymbol{\theta}_{j-1}\| - 4U_\delta B G_{\boldsymbol{\theta}} m\rho^{\tau-1}$$

$$\ge -G_{\boldsymbol{\theta}}\left(D_1(\tau+1)\sum_{k=t-\tau+1}^{t}\mathbb{E}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}\| + D_2 m\rho^{\tau-1}\right),$$

where $D_1 := \max\{(U_\delta L_l + 2L_* B + 3L_J), 2U_\delta B|\mathcal{A}|L\}$ and $D_2 := 4U_\delta B$, which completes the proof. $\qquad\square$

D.4 Proof of Lemma C.4

Proof of Lemma C.4. By definition, we have

$$J(\boldsymbol{\theta}_1) - J(\boldsymbol{\theta}_2) = \mathbb{E}[r(s^{(1)}, a^{(1)}) - r(s^{(2)}, a^{(2)})],$$

where $s^{(i)} \sim \mu_{\boldsymbol{\theta}_i}, a^{(i)} \sim \pi_{\boldsymbol{\theta}_i}$. Therefore, it holds that

$$J(\boldsymbol{\theta}_1) - J(\boldsymbol{\theta}_2) = \mathbb{E}[r(s^{(1)}, a^{(1)}) - r(s^{(2)}, a^{(2)})]$$
$$\le 2U_r d_{TV}(\mu_{\boldsymbol{\theta}_1} \otimes \pi_{\boldsymbol{\theta}_1}, \mu_{\boldsymbol{\theta}_2} \otimes \pi_{\boldsymbol{\theta}_2})$$
$$\le 2U_r|\mathcal{A}|L\left(1 + \lceil\log_\rho m^{-1}\rceil + \frac{1}{1-\rho}\right)\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$$
$$= C_J\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.$$

$$\square$$

D.5 Proof of Lemma C.5

The proof of this lemma depends on several auxiliary lemmas as follows.

Lemma D.4. For any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, eta, O = (s, a, s')$, we have

$$\big|\Xi(O, \eta, \boldsymbol{\theta}_1) - \Xi(O, \eta, \boldsymbol{\theta}_2)\big| \le 4U_r C_J\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.$$

Lemma D.5. For any $\boldsymbol{\theta}, \eta_1, \eta_2, O$, we have
$$\left|\Xi(O, \eta_1, \boldsymbol{\theta}) - \Xi(O, \eta_2, \boldsymbol{\theta})\right| \leq 2U_r|\eta_1 - \eta_2|.$$

Lemma D.6. Consider original tuples $O_t = (s_t, a_t, s_{t+1})$ and the auxiliary tuples $\widetilde{O}_t = (\widetilde{s}_t, \widetilde{a}_t, \widetilde{s}_{t+1})$. Conditioned on $s_{t-\tau+1}$ and $\boldsymbol{\theta}_{t-\tau}$, we have
$$\left|\mathbb{E}[\Xi(O_t, \eta_{t-\tau}, \boldsymbol{\theta}_{t-\tau}) - \Xi(\widetilde{O}_t, \eta_{t-\tau}, \boldsymbol{\theta}_{t-\tau})]\right| \leq 2U_r^2|\mathcal{A}|L \sum_{i=t-\tau}^{t} \mathbb{E}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{t-\tau}\|.$$

Lemma D.7. Conditioned on $s_{t-\tau+1}$ and $\boldsymbol{\theta}_{t-\tau}$, we have
$$\mathbb{E}[\Xi(\widetilde{O}_t, \eta_{t-\tau}, \boldsymbol{\theta}_{t-\tau})] \leq 4U_r^2 m \rho^{\tau-1}.$$

Proof. By the Lemma D.4, D.5, D.6 and D.7, we can collect the corresponding term and get the bound

$$\mathbb{E}[\Xi(O_t, \eta_t, \boldsymbol{\theta}_t)] = \mathbb{E}[\Xi(O_t, \eta_t, \boldsymbol{\theta}_t) - \Xi(O_t, \eta_t, \boldsymbol{\theta}_{t-\tau})] + \mathbb{E}[\Xi(O_t, \eta_t, \boldsymbol{\theta}_{t-\tau}) - \Xi(O_t, \eta_{t-\tau}, \boldsymbol{\theta}_{t-\tau})]$$
$$+ \mathbb{E}[\Xi(O_t, \eta_{t-\tau}, \boldsymbol{\theta}_{t-\tau}) - \Xi(\widetilde{O}_t, \eta_{t-\tau}, \boldsymbol{\theta}_{t-\tau})] + \mathbb{E}[\Xi(\widetilde{O}_t, \eta_{t-\tau}, \boldsymbol{\theta}_{t-\tau})]$$
$$\leq 4U_r C_J \|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-\tau}\| + 2U_r|\eta_t - \eta_{t-\tau}| + 2U_r^2|\mathcal{A}|L \sum_{i=t-\tau}^{t} \mathbb{E}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{t-\tau}\| + 4U_r^2 m \rho^{\tau-1}.$$

$\square$

D.6 Proof of Lemma C.6

Proof of Lemma C.6. For the first inequality, apply the property of norm and the Cauchy-Schwartz inequality:

$$\|g(O, \boldsymbol{\omega}, \boldsymbol{\theta})\| = \|(r(s, a) - J(\boldsymbol{\theta}) + \boldsymbol{\phi}^\top(s')\boldsymbol{\omega} - \boldsymbol{\phi}^\top(s)\boldsymbol{\omega})\boldsymbol{\phi}(s)\|$$
$$\leq |r(s, a)| + \|J(\boldsymbol{\theta})\| + |\boldsymbol{\phi}^\top(s')\boldsymbol{\omega}| \cdot \|\boldsymbol{\phi}^\top(s)\| + |\boldsymbol{\phi}^\top(s)\boldsymbol{\omega}| \cdot \|\boldsymbol{\phi}^\top(s)\|$$
$$= U_r + U_r + R_{\boldsymbol{\omega}} + R_{\boldsymbol{\omega}} \leq 2U_r + 2R_{\boldsymbol{\omega}}.$$

For the second inequality, we can directly apply Cauchy-Schwartz inequality and obtain the result. For the third inequality, apply Cauchy-Schwartz inequality as we have

$$\left|\Lambda(O, \boldsymbol{\omega}, \boldsymbol{\theta})\right| = \left|\langle \boldsymbol{\omega} - \boldsymbol{\omega}^*, g(O, \boldsymbol{\omega}, \boldsymbol{\theta}) - \bar{g}(\boldsymbol{\omega}, \boldsymbol{\theta})\rangle\right|$$
$$\leq \|\boldsymbol{\omega} - \boldsymbol{\omega}^*\| \cdot \|g(O, \boldsymbol{\omega}, \boldsymbol{\theta}) - \bar{g}(\boldsymbol{\omega}, \boldsymbol{\theta})\|$$
$$\leq 2R_{\boldsymbol{\omega}} \cdot 2U_\delta \leq 2U_\delta^2,$$

which completes the proof. $\square$

D.7 Proof of Lemma C.7

This Lemma is actually a combination of several auxiliary lemmas listed here:

Lemma D.8. For any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\omega}$ and tuple $O = (s, a, s')$,
$$\left|\Lambda(O, \boldsymbol{\omega}, \boldsymbol{\theta}_1) - \Lambda(O, \boldsymbol{\omega}, \boldsymbol{\theta}_2)\right| \leq K_1\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|,$$
where $K_1 = 2U_\delta^2|\mathcal{A}|L(1 + \lceil\log_\rho m^{-1}\rceil + 1/(1-\rho)) + 2U_\delta L_*$.

Lemma D.9. For any $\boldsymbol{\theta}, \boldsymbol{\omega}_1, \boldsymbol{\omega}_2$ and tuple $O = (s, a, s')$,
$$\left|\Lambda(O, \boldsymbol{\omega}_1, \boldsymbol{\theta}) - \Lambda(O, \boldsymbol{\omega}_2, \boldsymbol{\theta})\right| \leq 6U_\delta\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\|.$$

Lemma D.10. Consider original tuples $O_t = (s_t, a_t, s_{t+1})$ and the auxiliary tuples $\widetilde{O}_t = (\widetilde{s}_t, \widetilde{a}_t, \widetilde{s}_{t+1})$. Conditioned on $s_{t-\tau+1}$ and $\boldsymbol{\theta}_{t-\tau}$, we have

$$\mathbb{E}[\Lambda(O_t, \boldsymbol{\omega}_{t-\tau}, \boldsymbol{\theta}_{t-\tau}) - \Lambda(\widetilde{O}_t, \boldsymbol{\omega}_{t-\tau}, \boldsymbol{\theta}_{t-\tau})] \leq U_\delta^2|\mathcal{A}|L \sum_{i=t-\tau}^{t} \mathbb{E}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{t-\tau}\| \qquad \text{(D.1)}$$

Lemma D.11. Conditioned on $s_{t-\tau+1}$ and $\boldsymbol{\theta}_{t-\tau}$,
$$\mathbb{E}[\Lambda(\widetilde{O}_t, \boldsymbol{\omega}_{t-\tau}, \boldsymbol{\theta}_{t-\tau})] \le 2U_\delta^2 m \rho^{\tau-1}.$$

Proof of Lemma C.7. By the Lemma D.8, D.9, D.10 and D.11, we can collect the corresponding term and get the bound
$$
\mathbb{E}[\Lambda(O_t, \boldsymbol{\omega}_t, \boldsymbol{\theta}_t)] = \mathbb{E}[\Lambda(O_t, \boldsymbol{\omega}_t, \boldsymbol{\theta}_t) - \Lambda(O_t, \boldsymbol{\omega}_t, \boldsymbol{\theta}_{t-\tau})] + \mathbb{E}[\Lambda(O_t, \boldsymbol{\omega}_t, \boldsymbol{\theta}_{t-\tau}) - \Lambda(O_t, \boldsymbol{\omega}_{t-\tau}, \boldsymbol{\theta}_{t-\tau})]
$$
$$
+ \mathbb{E}[\Lambda(O_t, \boldsymbol{\omega}_{t-\tau}, \boldsymbol{\theta}_{t-\tau}) - \Lambda(\widetilde{O}_t, \boldsymbol{\omega}_{t-\tau}, \boldsymbol{\theta}_{t-\tau})] + \mathbb{E}[\Lambda(\widetilde{O}_t, \boldsymbol{\omega}_{t-\tau}, \boldsymbol{\theta}_{t-\tau})]
$$
$$
\le C_1(\tau+1)\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-\tau}\| + C_2 m \rho^{\tau-1} + C_3\|\boldsymbol{\omega}_t - \boldsymbol{\omega}_{t-\tau}\|,
$$
where $C_1 = 2U_\delta^2 |\mathcal{A}| L(1 + \lceil \log_\rho m^{-1} \rceil + 1/(1-\rho)) + 2U_\delta L_*, C_2 = 2U_\delta^2, C_3 = 4U_\delta.$ $\qquad\square$

E Proof of Auxiliary Lemmas

E.1 Proof of Lemma D.1

Proof of Lemma D.1. Let $\delta(O_t, \boldsymbol{\theta}) := r(s_t, a_t) + (\phi(s_{t+1}) - \phi(s_t))^\top \boldsymbol{\omega}^* - r(\boldsymbol{\theta})$ and it can be shown that $\delta(O_t, \boldsymbol{\theta}_1) - \delta(O_t, \boldsymbol{\theta}_2) = (\phi(s_{t+1}) - \phi(s_t))^\top (\boldsymbol{\omega}_1^* - \boldsymbol{\omega}_2^*) - (r(\boldsymbol{\theta}_1) - r(\boldsymbol{\theta}_2)).$
$$
\big\|h(O_t, \boldsymbol{\theta}_t) - h(O_t, \boldsymbol{\theta}_{t-\tau})\big\| = \big\|\delta(O_t, \boldsymbol{\theta}_t)\nabla \log \pi_{\boldsymbol{\theta}_t}(a_t|s_t) - \delta(O_t, \boldsymbol{\theta}_{t-\tau})\nabla \log \pi_{\boldsymbol{\theta}_{t-\tau}}(a_t|s_t)\big\|
$$
$$
\le \big\|\delta(O_t, \boldsymbol{\theta}_t)\nabla \log \pi_{\boldsymbol{\theta}_t}(a_t|s_t) - \delta(O_t, \boldsymbol{\theta}_t)\nabla \log \pi_{\boldsymbol{\theta}_{t-\tau}}(a_t|s_t)\big\|
$$
$$
+ \big\|\delta(O_t, \boldsymbol{\theta}_t)\nabla \log \pi_{\boldsymbol{\theta}_{t-\tau}}(a_t|s_t) - \delta(O_t, \boldsymbol{\theta}_{t-\tau})\nabla \log \pi_{\boldsymbol{\theta}_{t-\tau}}(a_t|s_t)\big\|
$$
$$
\le U_\delta L_l\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-\tau}\| + 2L_* B\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-\tau}\|.
$$
By triangle inequality, we have
$$
\big|\Gamma(O_t, \boldsymbol{\theta}_t) - \Gamma(O_t, \boldsymbol{\theta}_{t-\tau})\big| \le G_{\boldsymbol{\theta}}\big\|h(O_t, \boldsymbol{\theta}_t) - h(O_t, \boldsymbol{\theta}_{t-\tau})\big\| + 3G_{\boldsymbol{\theta}}\big\|\nabla J(\boldsymbol{\theta}_t) - \nabla J(\boldsymbol{\theta}_{t-\tau})\big\|
$$
$$
\le G_{\boldsymbol{\theta}}(U_\delta L_l + 2L_* B + 3L_J)\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-\tau}\|.
$$
$\qquad\square$

E.2 Proof of Lemma D.2

Proof of Lemma D.2. By the definition of in (C.1),
$$
\mathbb{E}\big[\Gamma(O_t, \boldsymbol{\theta}_{t-\tau}) - \Gamma(\widetilde{O}_t, \boldsymbol{\theta}_{t-\tau})\big] = \mathbb{E}\big[\langle \nabla J(\boldsymbol{\theta}_{t-\tau}), h(O_t, \boldsymbol{\theta}_{t-\tau}) - h(\widetilde{O}_t, \boldsymbol{\theta}_{t-\tau})\rangle\big]
$$
$$
= \mathbb{E}\Big[\langle \nabla J(\boldsymbol{\theta}_{t-\tau}), h(O_t, \boldsymbol{\theta}_{t-\tau})\rangle - \langle \nabla J(\boldsymbol{\theta}_{t-\tau}), h(\widetilde{O}_t, \boldsymbol{\theta}_{t-\tau})\rangle\Big]
$$
$$
\le 4U_\delta B G_{\boldsymbol{\theta}} d_{TV}\big(\mathbb{P}(O_t = \cdot|s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau}), \mathbb{P}(\widetilde{O}_t = \cdot|s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau})\big),
$$
$$(E.1)$$
where the inequality is by the definition of total variation. By Lemma B.2 we have
$$
d_{TV}\big(\mathbb{P}(O_t \in \cdot|s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau}), \mathbb{P}(\widetilde{O}_t \in \cdot|s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau})\big)
$$
$$
= d_{TV}\big(\mathbb{P}((s_t, a_t) \in \cdot|s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau}), \mathbb{P}((\widetilde{s}_t, \widetilde{a}_t) \in \cdot|s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau})\big)
$$
$$
\le d_{TV}\big(\mathbb{P}(s_t \in \cdot|s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau}), \mathbb{P}(\widetilde{s}_t \in \cdot|s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau})\big) + \frac{1}{2}|\mathcal{A}|L\mathbb{E}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-\tau}\|
$$
$$
\le d_{TV}\big(\mathbb{P}(O_{t-1} \in \cdot|s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau}), \mathbb{P}(\widetilde{O}_{t-1} \in \cdot|s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau})\big) + \frac{1}{2}|\mathcal{A}|L\mathbb{E}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-\tau}\|.
$$
Repeat the inequality above over t to $t - \tau + 1$ we have
$$
d_{TV}\big(\mathbb{P}(O_t \in \cdot|s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau}), \mathbb{P}(\widetilde{O}_t \in \cdot|s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau})\big) \le \frac{1}{2}|\mathcal{A}|L \sum_{i=t-\tau}^{t} \mathbb{E}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{t-\tau}\|. \quad (E.2)
$$
Plugging (E.2) into (E.1) we get
$$
\mathbb{E}\big[\Gamma(O_t, \boldsymbol{\theta}_{t-\tau}) - \Gamma(\widetilde{O}_t, \boldsymbol{\theta}_{t-\tau})\big] \le 2U_\delta B G_{\boldsymbol{\theta}}|\mathcal{A}|L \sum_{i=t-\tau}^{t} \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{t-\tau}\|.
$$
$\qquad\square$

E.3 Proof of Lemma D.3

Proof of Lemma D.3.

$$\mathbb{E}\big[\Gamma(\widetilde{O}_t, \boldsymbol{\theta}_{t-\tau}) - \Gamma(O'_t, \boldsymbol{\theta}_{t-\tau})\big] \leq 4U_\delta B G_{\boldsymbol{\theta}} d_{TV}\big(\mathbb{P}(\widetilde{O}_t = \cdot | s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau}), \mu_{\boldsymbol{\theta}_{t-\tau}} \otimes \pi_{\boldsymbol{\theta}_{t-\tau}} \otimes \mathcal{P}\big)$$
$$\leq 4U_\delta B G_{\boldsymbol{\theta}} m \rho^{\tau-1}.$$

The first inequality is by the definition of total variation norm and the second inequality is shown in Lemma D.11. $\qquad\square$

E.4 Proof of Lemma D.4

Proof of Lemma D.4. By the definition of $\Xi(O, \eta, \boldsymbol{\theta})$ in (C.5), we have

$$\begin{aligned}
\big|\Xi(O, \eta, \boldsymbol{\theta}_1) - \Xi(O, \eta, \boldsymbol{\theta}_2)\big| &= \big|(\eta - \eta_1^*)(r - \eta_1^*) - (\eta - \eta_2^*)(r - \eta_2^*)\big| \\
&\leq \big|(\eta - \eta_1^*)(r - \eta_1^*) - (\eta - \eta_1^*)(r - \eta_2^*)\big| \\
&\quad + \big|(\eta - \eta_1^*)(r - \eta_2^*) - (\eta - \eta_2^*)(r - \eta_2^*)\big| \\
&\leq 4U_r |\eta_1^* - \eta_2^*| \\
&= 4U_r \big|J(\boldsymbol{\theta}_1) - J(\boldsymbol{\theta}_2)\big| \\
&\leq 4U_r C_J \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.
\end{aligned}$$

$\qquad\square$

E.5 Proof of Lemma D.5

Proof of Lemma D.5. By definition,

$$\begin{aligned}
\big|\Xi(O, \eta_1, \boldsymbol{\theta}) - \Xi(O, \eta_2, \boldsymbol{\theta})\big| &= \big|(\eta_1 - \eta^*)(r - \eta^*) - (\eta_2 - \eta^*)(r - \eta^*)\big| \\
&\leq 2U_r |\eta_1 - \eta_2|.
\end{aligned}$$

$\qquad\square$

E.6 Proof of Lemma D.6

Proof of Lemma D.6. By the Cauchy-Schwartz inequality and the definition of total variation norm, we have

$$\mathbb{E}\big[\Xi(O_t, \eta_{t-\tau}, \boldsymbol{\theta}_{t-\tau}) - \Xi(\widetilde{O}_t, \eta_{t-\tau}, \boldsymbol{\theta}_{t-\tau})\big] = (\eta_{t-\tau} - \eta_{t-\tau}^*)\mathbb{E}[r(s_t, a_t) - r(\widetilde{s}_t, \widetilde{a}_t)].$$

Since

$$\mathbb{E}[r(s_t, a_t) - r(\widetilde{s}_t, \widetilde{a}_t)] \leq 2U_r d_{TV}\big(\mathbb{P}(O_t \in \cdot | s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau}), \mathbb{P}(\widetilde{O}_t \in \cdot | s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau})\big),$$

the total variation between O_t and $\widetilde{O}_t$ has appeared in (E.2), in the proof of Lemma D.2, which is

$$d_{TV}\big(\mathbb{P}(O_t \in \cdot | s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau}), \mathbb{P}(\widetilde{O}_t \in \cdot | s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau})\big) \leq \frac{1}{2}|\mathcal{A}|L \sum_{i=t-\tau}^{t} \mathbb{E}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{t-\tau}\|.$$

Plugging this bound, we have

$$\big|\mathbb{E}[\Xi(O_t, \eta_{t-\tau}, \boldsymbol{\theta}_{t-\tau}) - \Xi(\widetilde{O}_t, \eta_{t-\tau}, \boldsymbol{\theta}_{t-\tau})]\big| \leq 2U_r^2 |\mathcal{A}|L \sum_{i=t-\tau}^{t} \mathbb{E}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{t-\tau}\|.$$

$\qquad\square$

E.7 Proof of Lemma D.7

Proof of Lemma D.7. We first note that according to the definition,
$$\mathbb{E}[\eta(O'_t, \eta_{t-\tau}, \boldsymbol{\theta}_{t-\tau})|\boldsymbol{\theta}_{t-\tau}] = 0,$$
where $O'_t = (s'_t, a'_t, s'_{t+1})$ is the tuple generated by $s'_t \sim \mu_{\boldsymbol{\theta}_{t-\tau}}, a'_t \sim \pi_{\boldsymbol{\theta}_{t-\tau}}, s'_{t+1} \sim \mathcal{P}$. By the ergodicity in Assumption 4.2, it holds that
$$d_{TV}\big(\mathbb{P}(\widetilde{s}_t = \cdot|s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau}), \mu_{\boldsymbol{\theta}_{t-\tau}}\big) \leq m\rho^{\tau-1}.$$
It can be shown that
$$
\begin{aligned}
\mathbb{E}[\Xi(\widetilde{O}_t, \eta_{t-\tau}, \boldsymbol{\theta}_{t-\tau})] &= \mathbb{E}\big[\Xi\big(\widetilde{O}_t, \eta_{t-\tau}, \boldsymbol{\theta}_{t-\tau}\big) - \Xi(O'_t, \eta_{t-\tau}, \boldsymbol{\theta}_{t-\tau})\big] \\
&= \mathbb{E}\big[(\eta_{t-\tau} - \eta^*_{t-\tau})\big(r(\widetilde{s}_t, \widetilde{a}_t) - r(s', a')\big)\big] \\
&\leq 4U_r^2 d_{TV}\big(\mathbb{P}\big(\widetilde{O}_t = \cdot|s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau}\big), \mu_{\boldsymbol{\theta}_{t-\tau}} \otimes \pi_{\boldsymbol{\theta}_{t-\tau}} \otimes \mathcal{P}\big) \\
&\leq 4U_r^2 m\rho^{\tau-1}.
\end{aligned}
$$
The argument used here also appears in the proof of Lemma D.11 and explained in detail there. □

E.8 Proof of Lemma D.8

Proof of Lemma D.8.
$$
\begin{aligned}
\big|\Lambda(O, \boldsymbol{\omega}, \boldsymbol{\theta}_1) - \Lambda(O, \boldsymbol{\omega}, \boldsymbol{\theta}_2)\big| &= \Big|\langle\boldsymbol{\omega} - \boldsymbol{\omega}_1^*, g(O, \boldsymbol{\omega}) - \bar{g}(\boldsymbol{\theta}_1, \boldsymbol{\omega})\rangle - \langle\boldsymbol{\omega} - \boldsymbol{\omega}_2^*, g(O, \boldsymbol{\omega}) - \bar{g}(\boldsymbol{\theta}_2, \boldsymbol{\omega})\rangle\Big| \\
&\leq \underbrace{\Big|\langle\boldsymbol{\omega} - \boldsymbol{\omega}_1^*, g(O, \boldsymbol{\omega}) - \bar{g}(\boldsymbol{\theta}_1, \boldsymbol{\omega})\rangle - \langle\boldsymbol{\omega} - \boldsymbol{\omega}_1^*, g(O, \boldsymbol{\omega}) - \bar{g}(\boldsymbol{\theta}_2, \boldsymbol{\omega})\rangle\Big|}_{I_1} \\
&\quad + \underbrace{\Big|\langle\boldsymbol{\omega} - \boldsymbol{\omega}_1^*, g(O, \boldsymbol{\omega}) - \bar{g}(\boldsymbol{\theta}_2, \boldsymbol{\omega})\rangle - \langle\boldsymbol{\omega} - \boldsymbol{\omega}_2^*, g(O, \boldsymbol{\omega}) - \bar{g}(\boldsymbol{\theta}_2, \boldsymbol{\omega})\rangle\Big|}_{I_2}.
\end{aligned}
$$
For the term I_2, we simply use the Cauchy-Schwartz inequality to get $2U_\delta\|\boldsymbol{\omega}_1^* - \boldsymbol{\omega}_2^*\|$.
For the term I_1, it can be bounded as:
$$
\begin{aligned}
&\Big|\langle\boldsymbol{\omega} - \boldsymbol{\omega}_1^*, g(O, \boldsymbol{\omega}) - \bar{g}(\boldsymbol{\theta}_1, \boldsymbol{\omega})\rangle - \langle\boldsymbol{\omega} - \boldsymbol{\omega}_1^*, g(O, \boldsymbol{\omega}) - \bar{g}(\boldsymbol{\theta}_2, \boldsymbol{\omega})\rangle\Big| \\
&= \Big|\langle\boldsymbol{\omega} - \boldsymbol{\omega}_1^*, \bar{g}(\boldsymbol{\theta}_1, \boldsymbol{\omega}) - \bar{g}(\boldsymbol{\theta}_2, \boldsymbol{\omega})\rangle\Big| \\
&\leq 2R_{\boldsymbol{\omega}}\big\|\bar{g}(\boldsymbol{\theta}_1, \boldsymbol{\omega}) - \bar{g}(\boldsymbol{\theta}_2, \boldsymbol{\omega})\big\| \\
&\leq 2R_{\boldsymbol{\omega}} \cdot 2U_\delta \cdot d_{TV}(\mu_{\boldsymbol{\theta}_1} \otimes \pi_{\boldsymbol{\theta}_1} \otimes \mathcal{P}, \mu_{\boldsymbol{\theta}_2} \otimes \pi_{\boldsymbol{\theta}_2} \otimes \mathcal{P}) \\
&\leq 2U_\delta^2 d_{TV}(\mu_{\boldsymbol{\theta}_1} \otimes \pi_{\boldsymbol{\theta}_1} \otimes \mathcal{P}, \mu_{\boldsymbol{\theta}_2} \otimes \pi_{\boldsymbol{\theta}_2} \otimes \mathcal{P}),
\end{aligned}
$$
where the first inequality is due to Cauchy-Schwartz; the second inequality is by the definition of total variation norm; the third inequality is due to the fact $U_\delta \geq 2R_{\boldsymbol{\omega}}$. Therefore, we have
$$
\begin{aligned}
\big|\Lambda(\boldsymbol{\theta}_1, \boldsymbol{\omega}, O) - \Lambda(\boldsymbol{\theta}_2, \boldsymbol{\omega}, O)\big| &\leq 2U_\delta^2 d_{TV}(\mu_{\boldsymbol{\theta}_1} \otimes \pi_{\boldsymbol{\theta}_1} \otimes \mathcal{P}, \mu_{\boldsymbol{\theta}_2} \otimes \pi_{\boldsymbol{\theta}_2} \otimes \mathcal{P}) + 2U_\delta\|\boldsymbol{\omega}_1^* - \boldsymbol{\omega}_2^*\| \\
&\leq 2U_\delta^2|\mathcal{A}|L\Big(1 + \lceil\log_\rho m^{-1}\rceil + \frac{1}{1-\rho}\Big)\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| + 2U_\delta L_*\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \\
&= K_1\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|,
\end{aligned}
$$
where the second inequality is due to Lemma B.1 and Proposition 4.4. □

E.9 Proof of Lemma D.9

Proof of Lemma D.9. By definition,
$$
\begin{aligned}
\big|\Lambda(O, \boldsymbol{\omega}_1, \boldsymbol{\theta}) - \Lambda(O, \boldsymbol{\omega}_2, \boldsymbol{\theta})\big| &= \Big|\langle\boldsymbol{\omega}_1 - \boldsymbol{\omega}^*, g(O, \boldsymbol{\omega}_1) - \bar{g}(\boldsymbol{\omega}_1, \boldsymbol{\theta})\rangle - \langle\boldsymbol{\omega}_2 - \boldsymbol{\omega}^*, g(O, \boldsymbol{\omega}_2) - \bar{g}(\boldsymbol{\omega}_2, \boldsymbol{\theta})\rangle\Big| \\
&\leq \Big|\langle\boldsymbol{\omega}_1 - \boldsymbol{\omega}^*, g(O, \boldsymbol{\omega}_1) - \bar{g}(\boldsymbol{\omega}_1, \boldsymbol{\theta})\rangle - \langle\boldsymbol{\omega}_1 - \boldsymbol{\omega}^*, g(O, \boldsymbol{\omega}_2) - \bar{g}(\boldsymbol{\omega}_2, \boldsymbol{\theta})\rangle\Big| \\
&\quad + \Big|\langle\boldsymbol{\omega}_1 - \boldsymbol{\omega}^*, g(O, \boldsymbol{\omega}_2) - \bar{g}(\boldsymbol{\omega}_2, \boldsymbol{\theta})\rangle - \langle\boldsymbol{\omega}_2 - \boldsymbol{\omega}^*, g(O, \boldsymbol{\omega}_2) - \bar{g}(\boldsymbol{\omega}_2, \boldsymbol{\theta})\rangle\Big| \\
&\leq 2R_{\boldsymbol{\omega}}\big\|(g(O, \boldsymbol{\omega}_1) - g(O, \boldsymbol{\omega}_2)) - (\bar{g}(\boldsymbol{\omega}_1, \boldsymbol{\theta}) - \bar{g}(\boldsymbol{\omega}_2, \boldsymbol{\theta}))\big\| + 2U_\delta\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\|.
\end{aligned}
$$

Note that we have $\|g(O, \boldsymbol{\omega}_1, \boldsymbol{\theta}) - g(O, \boldsymbol{\omega}_2, \boldsymbol{\theta})\| = |(\boldsymbol{\phi}(s') - \boldsymbol{\phi}(s))^\top(\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2)| \leq 2\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\|$ and similarly $\|\bar{g}(\boldsymbol{\omega}_1, \boldsymbol{\theta}) - \bar{g}(\boldsymbol{\omega}_2, \boldsymbol{\theta})\| \leq |\mathbb{E}[(\boldsymbol{\phi}(s') - \boldsymbol{\phi}(s))^\top(\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2)]| \leq 2\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\|$. Therefore,

$$
\begin{aligned}
\big|\Lambda(O, \boldsymbol{\omega}_1, \boldsymbol{\theta}) - \Lambda(O, \boldsymbol{\omega}_2, \boldsymbol{\theta})\big| &\leq 2R_{\boldsymbol{\omega}}\Big\|\big(g(O, \boldsymbol{\omega}_1) - g(O, \boldsymbol{\omega}_2)\big) - \big(\bar{g}(\boldsymbol{\omega}_1, \boldsymbol{\theta}) - \bar{g}(\boldsymbol{\omega}_2, \boldsymbol{\theta})\big)\Big\| + 2U_\delta\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\| \\
&\leq 2R_{\boldsymbol{\omega}} \cdot 4\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\| + 2U_\delta\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\| \\
&\leq 6U_\delta\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\|.
\end{aligned}
$$

$\square$

E.10 Proof of Lemma D.10

Proof of Lemma D.10. By the Cauchy-Schwartz inequality and the definition of total variation norm, we have

$$
\begin{aligned}
\mathbb{E}[\Lambda(O_t, \boldsymbol{\omega}_{t-\tau}, \boldsymbol{\theta}_{t-\tau}) - \Lambda(\widetilde{O}_t, \boldsymbol{\omega}_{t-\tau}, \boldsymbol{\theta}_{t-\tau})] &= \mathbb{E}\big[\langle\boldsymbol{\omega}_{t-\tau} - \boldsymbol{\omega}_{t-\tau}^*, g(O_t, \boldsymbol{\omega}_{t-\tau}) - g(\widetilde{O}_t, \boldsymbol{\omega}_{t-\tau})\rangle\big] \\
&\leq 2U_\delta^2 d_{TV}\big(\mathbb{P}(O_t \in \cdot|s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau}), \mathbb{P}(\widetilde{O}_t \in \cdot|s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau})\big).
\end{aligned}
$$
(E.3)

The total variation between O_t and $\widetilde{O}_t$ has appeared in (E.2), in the proof of Lemma D.2, which is

$$
d_{TV}\big(\mathbb{P}(O_t \in \cdot|s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau}), \mathbb{P}(\widetilde{O}_t \in \cdot|s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau})\big) \leq \frac{1}{2}|\mathcal{A}|L \sum_{i=t-\tau}^{t} \mathbb{E}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{t-\tau}\|.
$$

Plugging this bound into (E.3), we have

$$
\mathbb{E}\big|\Lambda(O_t, \boldsymbol{\omega}_{t-\tau}, \boldsymbol{\theta}_{t-\tau}) - \Lambda(\widetilde{O}_t, \boldsymbol{\omega}_{t-\tau}, \boldsymbol{\theta}_{t-\tau})\big| \leq U_\delta^2|\mathcal{A}|L \sum_{i=t-\tau}^{t} \mathbb{E}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{t-\tau}\|.
$$

$\square$

E.11 Proof of Lemma D.11

Proof of Lemma D.11. We first note that according to the definition in Section C.3,

$$
\mathbb{E}[\Lambda(O_t', \boldsymbol{\omega}_{t-\tau}, \boldsymbol{\theta}_{t-\tau})|s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau}] = 0,
$$

where $O_t' = (s_t', a_t', s_{t+1}')$ is the tuple generated by $s_t' \sim \mu_{\boldsymbol{\theta}_{t-\tau}}, a_t' \sim \pi_{\boldsymbol{\theta}_{t-\tau}}, s_{t+1}' \sim \mathcal{P}$. By the ergodicity in Assumption 4.2, it holds that

$$
d_{TV}\big(\mathbb{P}(\widetilde{s}_t = \cdot|s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau}), \mu_{\boldsymbol{\theta}_{t-\tau}}\big) \leq m\rho^{\tau-1}.
$$

It can be shown that

$$
\begin{aligned}
\mathbb{E}[\Lambda(\widetilde{O}_t, \boldsymbol{\omega}_{t-\tau}, \boldsymbol{\theta}_{t-\tau})] &= \mathbb{E}[\Lambda(\widetilde{O}_t, \boldsymbol{\omega}_{t-\tau}, \boldsymbol{\theta}_{t-\tau}) - \Lambda(O_t', \boldsymbol{\omega}_{t-\tau}, \boldsymbol{\theta}_{t-\tau})] \\
&= \mathbb{E}\langle\boldsymbol{\omega}_{t-\tau} - \boldsymbol{\omega}_{t-\tau}^*, g(\widetilde{O}_t, \boldsymbol{\omega}_{t-\tau}) - g(O_t', \boldsymbol{\omega}_{t-\tau})\rangle \\
&\leq 4R_{\boldsymbol{\omega}}U_\delta d_{TV}\big(\mathbb{P}(\widetilde{O}_t = \cdot|s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau}), \mu_{\boldsymbol{\theta}_{t-\tau}} \otimes \pi_{\boldsymbol{\theta}_{t-\tau}} \otimes \mathcal{P}\big) \\
&\leq 2U_\delta^2 d_{TV}\big(\mathbb{P}(\widetilde{s}_t = \cdot|s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau}), \mu_{\boldsymbol{\theta}_{t-\tau}}\big) \\
&\leq 2U_\delta^2 m\rho^{\tau-1}.
\end{aligned}
$$

The third inequality holds because $2R_{\boldsymbol{\omega}} < U_\delta$ and

$$
\begin{aligned}
d_{TV}\big(\mathbb{P}(\widetilde{O}_t = \cdot|s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau}), \mu_{\boldsymbol{\theta}_{t-\tau}} \otimes \pi_{\boldsymbol{\theta}_{t-\tau}} \otimes \mathcal{P}\big) &= d_{TV}\big(\mathbb{P}((\widetilde{s}_t, \widetilde{a}_t) = \cdot|s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau}), \mu_{\boldsymbol{\theta}_{t-\tau}} \otimes \pi_{\boldsymbol{\theta}_{t-\tau}}\big) \\
&= d_{TV}\big(\mathbb{P}(\widetilde{s}_t = \cdot|s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau}), \mu_{\boldsymbol{\theta}_{t-\tau}}\big).
\end{aligned}
$$

This can be shown following the same procedure in (B.1), because $\mathbb{P}(\widetilde{O}_t = \cdot|s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau}) = \mathbb{P}(\widetilde{s}_t = \cdot|s_{t-\tau+1}, \boldsymbol{\theta}_{t-\tau}) \otimes \pi_{\boldsymbol{\theta}_{t-\tau}} \otimes \mathcal{P}$. $\square$