Non-Asymptotic Analysis for Two Time-scale TDC with General Smooth Function Approximation

Yue Wang
Department of Electrical Engineering
University at Buffalo
Buffalo, NY, USA
ywang294@buffalo.edu

Shaofeng Zou
Department of Electrical Engineering
University at Buffalo
Buffalo, NY, USA
szou3@buffalo.edu

Yi Zhou
Department of Electrical and Computer Engineering
University of Utah
Salt Lake City, Utah, USA
yi.zhou@utah.edu

Abstract

Temporal-difference learning with gradient correction (TDC) is a two time-scale algorithm for policy evaluation in reinforcement learning. This algorithm was initially proposed with linear function approximation, and was later extended to the one with general smooth function approximation. The asymptotic convergence for the on-policy setting with general smooth function approximation was established in [Bhatnagar et al., 2009], however, the non-asymptotic convergence analysis remains unsolved due to challenges in the non-linear and two-time-scale update structure, non-convex objective function and the projection onto a time-varying tangent plane. In this paper, we develop novel techniques to address the above challenges and explicitly characterize the non-asymptotic error bound for the general off-policy setting with i.i.d. or Markovian samples, and show that it converges as fast as $\mathcal{O}(1/\sqrt{T})$ (up to a factor of $\mathcal{O}(\log T)$). Our approach can be applied to a wide range of value-based reinforcement learning algorithms with general smooth function approximation.

1 Introduction

In reinforcement learning (RL), an agent interacts with a stochastic environment in order to maximize the total reward [Sutton and Barto, 2018]. Towards this goal, it is often needed to evaluate how good a policy performs, and more specifically, to learn its value function. Temporal difference (TD) learning algorithm is one of the most popular policy evaluation approaches. However, when applied with function approximation approach and/or under the off-policy setting, the TD learning algorithm may diverge [Baird, 1995, Tsitsiklis and Van Roy, 1997]. To address this issue, a family of gradient-based TD (GTD) algorithms, e.g., GTD, GTD2, temporal-difference learning with gradient correction (TDC) and Greedy-GQ, were developed for the case with linear function approximation [Maei, 2011, Sutton et al., 2009b, Maei et al., 2010, Sutton et al., 2009a,b]. These algorithms were later extended to the case with general smooth function approximation in [Bhatnagar et al., 2009], where asymptotic convergence guarantee was established for the on-policy setting with i.i.d. samples.

Despite the success of the GTD methods in practice, previous theoretical studies only showed that these algorithms converge asymptotically, and did not suggest how fast these algorithms converge and how the accuracy of the solution depends on various parameters of the algorithms. Not until

recently have the non-asymptotic error bounds for these algorithms been investigated, e.g., [Dalal et al., 2020, Karmakar and Bhatnagar, 2018, Wang and Zou, 2020, Xu et al., 2019, Kaledin et al., 2020, Dalal et al., 2018, Wang et al., 2017], which mainly focus on the case with linear function approximation. These results thus cannot be directly applied to more practical applications with general smooth function approximation, e.g., neural networks, which have greater representation power, do not need to construct feature mapping, and are widely used in practice.

In this paper, we develop a non-asymptotic analysis for the TDC algorithm with general smooth function approximation (which we refer to as non-linear TDC) for both i.i.d. and Markovian samples. Technically, the analysis in this paper is not a straightforward extension of previous studies on those GTD algorithms with linear function approximation. First of all, different from existing studies with linear function approximation whose objective functions are *convex* and the updates are *linear*, the objective function of the non-linear TDC algorithm is *non-convex*, and the two time-scale updates are *non-linear* functions of the parameters. Second, the objective function of the non-linear TDC algorithm, the mean-square projected Bellman error (MSPBE), involves a projection onto a *time-varying* tangent plane which depends on the sample trajectory, whereas for GTD algorithms with linear function approximation, this projection is time-invariant. Third, due to the two time-scale structure of the algorithm and the Markovian noise, novel techniques to deal with the stochastic bias and the tracking error need to be developed.

1.1 Challenges and Contributions

In this section, we summarize the technical challenges and our contributions.

Analysis for two time-scale non-linear updates and non-convex objective. Unlike many existing results on two time-scale stochastic approximation, e.g., [Konda et al., 2004, Gupta et al., 2019, Kaledin et al., 2020] and the studies of linear GTD algorithms in [Xu et al., 2019, Ma et al., 2020, Wang et al., 2017, Dalal et al., 2020], the objective function of the non-linear TDC is non-convex, and its two time-scale updates are non-linear. Therefore, existing studies on linear two time-scale algorithms cannot be directly applied. Moreover, the convergence to global optimum cannot be guaranteed for the non-linear TDC algorithm, and therefore, we study the convergence to stationary points. In this paper, we develop a novel non-asymptotic analysis of the non-linear TDC algorithm, which solves RL problems from a non-convex optimization perspective. We note that our analysis is not a straightforward extension of analyses of non-convex optimization, as the update rule here is two time-scale and the noise is Markovian. The framework we develop in this paper can be applied to analyze a wide range of value-based RL algorithms with general smooth function approximation.

Time-varying projection. For the MSPBE, a projection of the Bellman error onto the parameterized function class is involved. However, unlike linear function approximation, the projection onto a general smooth class of functions usually does not have a closed-form solution. Thus, a projection onto the tangent plane at the current parameter is used instead, which incurs a time-varying projection that depends on the current parameter and thus the sample trajectory. This brings in additional challenges in the bias and variance analysis due to such dependency. We develop a novel approach to decouple such a dependency and characterize the bias by exploiting the uniform ergodicity of the underlying MDP and the smoothness of the parameterized function. The new challenges posed by the time-varying projection and the dependence between the projection and the sample trajectory are not special to the non-linear TDC investigated in this paper, and they exist in a wide range of value-based algorithms with general smooth function approximation, where our techniques can be applied.

A tight tracking error analysis. Due to the two time-scale structure of the update rule, the tracking error, which measures how fast the fast time-scale tracks its own limit, needs to be explicitly bounded. Unlike the studies on two time-scale linear stochastic approximation [Dalal et al., 2020, Kaledin et al., 2020, Konda et al., 2004], where a linear transformation can asymptotically decouple the dependence between the fast and slow time-scale updates, it is non-trivial to construct such a transformation for non-linear updates. To develop a tight bound on the tracking error, we develop a novel technique that bounds the tracking error as a function of the gradient of the MSPBE. This leads to a tighter bound on the tracking error compared to many existing works on two time-scale analysis, e.g., [Wu et al., 2020, Hong et al., 2020]. Although we do not decouple the fast and slow time-scale updates, we still obtain a desired convergence rate of $\mathcal{O}(1/\sqrt{T})$ (up to a factor of $\log T$), which matches with the complexity of stochastic gradient descent for non-convex problems [Ghadimi and Lan, 2013].

1.2 Related Work

TD, Q-learning and SARSA. The asymptotic convergence of TD with linear function approximation was shown in [Tsitsiklis and Van Roy, 1997], and the non-asymptotic analysis of TD was developed in [Srikant and Ying, 2019, Lakshminarayanan and Szepesvari, 2018, Bhandari et al., 2018, Dalal et al., 2018, Sun et al., 2020]. Moreover, [Cai et al., 2019] further studied the non-asymptotic error bound of TD learning with neural function approximation. Q-learning and SARSA are usually used for solving the optimal control problem and were shown to converge asymptotically under some conditions in [Melo et al., 2008, Perkins and Precup, 2003]. Their non-asymptotic error bounds were also studied in [Zou et al., 2019]. The non-asymptotic analysis of Q-learning under the neural function approximation was developed in [Cai et al., 2019, Xu and Gu, 2020]. Note that all these algorithms are one time-scale, while the TDC algorithm we study is a two time-scale algorithm.

GTD methods with linear function approximation. A class of GTD algorithms were proposed to address the divergence issue for off-policy training [Baird, 1995] and arbitrary smooth function approximation [Tsitsiklis and Van Roy, 1997], e.g., GTD, GTD2 and TDC [Maei, 2011, Sutton et al., 2009b, Maei et al., 2010, Sutton et al., 2009a,b]. Recent studies established their non-asymptotic convergence rate, e.g., [Dalal et al., 2018, Wang et al., 2017, Liu et al., 2015, Gupta et al., 2019, Xu et al., 2019, Dalal et al., 2020, Kaledin et al., 2020, Ma et al., 2020, Wang and Zou, 2020, Ma et al., 2021] under i.i.d. and Markovian settings. These studies focus on the case with linear function approximation, and thus the objective functions are convex, and the updates are linear. In this paper, we focus on the non-linear TDC algorithm with general smooth function approximation, where the two time-scale update rule is non-linear, the objective is non-convex, and the projection is time-varying, and thus new techniques are required to develop the non-asymptotic analysis.

Non-linear two time-scale stochastic approximation. There are also studies on asymptotic convergence rate and non-asymptotic analysis for non-linear two time-scale stochastic approximation, e.g., [Mokkadem et al., 2006, Doan, 2021]. Although the non-linear update rule is investigated, it is assumed that the algorithm converges to the global optimum. In this paper, we do not make such an assumption on the global convergence, which may not necessarily hold for the non-linear TDC algorithm, and instead, we study the convergence to stationary points, which is a widely used convergence criterion for non-convex optimization problems. We also note that there is a resent work studying the batch-based non-linear TDC in [Xu and Liang, 2021], where at each update, a batch of samples is used. To achieve a sample complexity of $\mathcal{O}(\epsilon^{-2})$, a batch size of $\mathcal{O}(\epsilon^{-1})$ is required in [Xu and Liang, 2021] to control the bias and variance. We note that by setting the batch size being one in [Xu and Liang, 2021], the desired sample complexity cannot be obtained, and their error bound will be a constant. In this paper, we focus on the non-linear TDC algorithm without using the batch method, where the parameters update in an online and incremental fashion and at each update only one sample is used. Our error analysis is novel and more refined as it does not require a large batch size of $\mathcal{O}(\epsilon^{-1})$ while still achieving the same sample complexity.

2 Preliminaries

2.1 Markov Decision Process

A Markov decision process (MDP) is a tuple $(\mathcal{S}, \mathcal{A}, \mathsf{P}, r, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, $\mathsf{P} = \mathsf{P}(s'|s, a)$ is the transition kernel, $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^+$ is the reward function bounded by $r_{\max}$, and $\gamma \in [0, 1]$ is the discount factor. A stationary policy π maps a state $s \in \mathcal{S}$ to a probability distribution $\pi(\cdot|s)$ over the action space $\mathcal{A}$. At each time-step t, suppose the process is at some state $s_t \in \mathcal{S}$, and an action $a_t \in \mathcal{A}$ is taken. Then the system transits to the next state s_{t+1} following the transition kernel $\mathsf{P}(\cdot|s_t, a_t)$, and the agent receives a reward $r(s_t, a_t, s_{t+1})$.

For a given policy π and any initial state $s \in \mathcal{S}$, we define its value function as $V^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t, S_{t+1})|S_0 = s, \pi\right]$. The goal of policy evaluation is to use the samples generated from the MDP to estimate the value function. The value function satisfies the Bellman equation: $V^\pi(s) = T^\pi V^\pi(s)$ for any $s \in \mathcal{S}$, where the Bellman operator T^π is defined as

$$T^\pi V(s) = \sum_{s' \in \mathcal{S}, a \in \mathcal{A}} \mathsf{P}(s'|s, a)\pi(a|s)r(s, a, s') + \gamma \sum_{s' \in \mathcal{S}, a \in \mathcal{A}} \mathsf{P}(s'|s, a)\pi(a|s)V(s'). \quad (1)$$

Hence the value function V^π is the fixed point of the Bellman operator T^π [Bertsekas, 2011].

2.2 Function Approximation

In practice, the state space $\mathcal{S}$ usually contains a large number of states or is even continuous, which will induce a heavy computational overhead. A popular approach is to approximate the value function using a parameterized class of functions. Consider a parameterized family of functions $\{V_\theta : \mathcal{S} \to \mathbb{R} | \theta \in \mathbb{R}^N\}$, e.g., neural networks. The goal is to find a V_θ with a compact representation in θ to approximate the value function V^π. In this paper, we focus on a general family of smooth functions, which may not be linear in θ.

3 TDC with Non-Linear Function Approximation

In this section, we introduce the TDC algorithm with general smooth function approximation in [Bhatnagar et al., 2009] for the off-policy setting with both i.i.d. samples and Markovian samples, and further characterize the non-asymptotic error bounds.

Consider the the following mean-square projected Bellman error (MSPBE):

$$J(\theta) = \mathbb{E}_{\mu^\pi}\left[\|V_\theta(s) - \mathbf{\Pi}_\theta T^\pi V_\theta(s)\|^2\right], \tag{2}$$

where μ^π is the stationary distribution induced by the policy π, and $\mathbf{\Pi}_\theta$ is the orthogonal projection onto the tangent plane of V_θ at θ: $\left\{\hat{V}_\zeta(s) | \zeta \in \mathbb{R}^N \text{ and } \hat{V}_\zeta(s) = \phi_\theta(s)^\top \zeta\right\}$ and $\phi_\theta(s) = \nabla V_\theta(s)$. Note that the projection is onto the tangent plane instead of $\{V_\theta : \theta \in \mathbb{R}^N\}$ since the projection onto the latter one may not be computationally tractable if V_θ is non-linear.

In [Bhatnagar et al., 2009], the authors proposed a two time-scale TDC algorithm to minimize the MSPBE $J(\theta)$. Specifically, a stochastic gradient descent approach is used with the weight doubling trick (for the double sampling problem) [Sutton et al., 2009a], which yield a two time-scale update rule. We note that the algorithm developed in [Bhatnagar et al., 2009] was for the on-policy setting with i.i.d. samples from the stationary distribution, and the asymptotic convergence of the algorithm to stationary points was established.

In the off-policy setting, the goal is to estimate the value function V^π of the target policy π using the samples from a different behavior policy π_b. In this case, the MSPBE can be written as

$$J(\theta) = \mathbb{E}_{\mu^{\pi_b}}[\|V_\theta(s) - \mathbf{\Pi}_\theta T^\pi V_\theta(s)\|^2], \tag{3}$$

and we use the approach of importance sampling. Following steps similar to those in [Maei, 2011], $J(\theta)$ can be further written as

$$J(\theta) = \mathbb{E}_{\mu^{\pi_b}}[\rho(S,A)\delta_{S,A,S'}(\theta)\phi_\theta(S)]^\top A_\theta^{-1}\mathbb{E}_{\mu^{\pi_b}}[\rho(S,A)\delta_{S,A,S'}(\theta)\phi_\theta(S)], \tag{4}$$

where $\delta_{s,a,s'}(\theta) = r(s,a,s') + \gamma V_\theta(s') - V_\theta(s)$ is the TD error, $\phi_\theta(s) = \nabla V_\theta(s)$ is the character vector, $\rho(s,a) = \frac{\pi(a|s)}{\pi_b(a|s)}$ is the importance sampling ratio for a given sample $O = (s,a,r,s')$ and $A_\theta = \mathbb{E}_{\mu^{\pi_b}}[\phi_\theta(S)\phi_\theta(S)^\top]$.

To compute $\nabla J(\theta)$, we consider its i-th entry, i.e., the partial derivative w.r.t. the i-th entry of θ:

$$-\frac{1}{2}\frac{\partial J(\theta)}{\partial \theta^i}$$

$$= \underbrace{-\mathbb{E}_{\mu^{\pi_b}}\left[\frac{\partial}{\partial \theta^i}(\rho\delta\phi)\right]^\top A_\theta^{-1}\mathbb{E}_{\mu^{\pi_b}}[\rho\delta\phi]}_{(a)} + \underbrace{\frac{1}{2}\left(A_\theta^{-1}\mathbb{E}_{\mu^{\pi_b}}[\rho\delta\phi]\right)^\top \mathbb{E}_{\mu^{\pi_b}}\left[\frac{\partial}{\partial \theta^i}(\phi\phi^\top)\right](A_\theta^{-1}\mathbb{E}_{\mu^{\pi_b}}[\rho\delta\phi])}_{(b)},$$

$$\tag{5}$$

where to simplify notations, we omit the dependence on θ, S, A and S'. To get an unbiased estimate of the terms in (5), several independent samples are needed, but this is not applicable when there is only one sample trajectory. Hence we employ the weight doubling trick [Sutton et al., 2009a]. Define $\omega(\theta) = A_\theta^{-1}\mathbb{E}_{\mu^{\pi_b}}[\rho(S,A)\delta_{S,A,S'}(\theta)\phi_\theta(S)]$, then term (a) can be written as follows:

$$-\mathbb{E}_{\mu^{\pi_b}}\left[\frac{\partial}{\partial \theta^i}(\rho\delta\phi)\right]^\top A_\theta^{-1}\mathbb{E}_{\mu^{\pi_b}}[\rho\delta\phi]$$

4

$$= -\mathbb{E}_{\mu^{\pi_b}}\left[\rho(\gamma(\phi_\theta(S'))_i - (\phi_\theta(S))_i)\phi_\theta(S)\right]^\top \omega(\theta) - \mathbb{E}_{\mu^{\pi_b}}\left[\rho\delta(\nabla^2 V)_i\right]^\top \omega(\theta); \quad (6)$$

and term (b) can be written as follows:

$$(A_\theta^{-1}\mathbb{E}_{\mu^{\pi_b}}\left[\rho\delta\phi\right])^\top \mathbb{E}_{\mu^{\pi_b}}\left[\frac{\partial}{\partial\theta^i}(\phi\phi^\top)\right](A_\theta^{-1}\mathbb{E}_{\mu^{\pi_b}}\left[\rho\delta\phi\right]) = 2\mathbb{E}_{\mu^{\pi_b}}\left[\phi^\top\omega(\theta)(\frac{\partial}{\partial\theta^i}\phi^\top)\omega(\theta)\right]. \quad (7)$$

Hence the gradient can be re-written as

$$-\frac{\nabla J(\theta)}{2} = \mathbb{E}_{\mu^{\pi_b}}\left[\rho(S,A)\delta_{S,A,S'}(\theta)\phi_\theta(S)\right] - h(\theta,\omega(\theta)) - \gamma\mathbb{E}_{\mu^{\pi_b}}\left[\rho(S,A)\phi_\theta(S')\phi_\theta(S)^\top\right]\omega(\theta),$$
$$(8)$$

where $h(\theta,\omega) = \mathbb{E}_{\mu^{\pi_b}}[\left(\rho(S,A)\delta_{S,A,S'}(\theta) - \phi_\theta(S)^\top\omega\right)\nabla^2 V_\theta(S)\omega]$. Thus with this weight doubling trick [Sutton et al., 2009a], a two time-scale stochastic gradient descent algorithm can be constructed. In Algorithm 1, we present the algorithm for the Markovian setting. The algorithm under the i.i.d. setting is slightly different, hence we refer the readers to Algorithm 2 in Appendix B.

Algorithm 1 Non-Linear Off-Policy TDC under the Markovian Setting

---

Input: $T, \alpha, \beta, \pi, \pi_b, \{V_\theta | \theta \in \mathbb{R}^N\}$
Initialization: θ_0, w_0
  1: Choose $W \sim \text{Uniform}(0,1,...,T-1)$
  2: **for** $t = 0,1,...,W-1$ **do**
  3:   Sample $O_t = (s_t, a_t, r_t, s_{t+1})$ following π_b
  4:   $\delta_t(\theta_t) = r(s_t, a_t, s_{t+1}) + \gamma V_{\theta_t}(s_{t+1}) - V_{\theta_t}(s_t)$
  5:   $\rho_t = \frac{\pi(a_t|s_t)}{\pi_b(a_t|s_t)}$
  6:   $h_t(\theta_t, \omega_t) = \left(\rho_t\delta_t(\theta_t) - \phi_{\theta_t}(s_t)^\top\omega_t\right)\nabla^2 V_{\theta_t}(s_t)\omega_t$
  7:   $\omega_{t+1} = \mathbf{\Pi}_{R_\omega}\left(\omega_t + \beta\left(-\phi_{\theta_t}(s_t)\phi_{\theta_t}(s_t)^\top\omega_t + \rho_t\delta_t(\theta_t)\phi_{\theta_t}(s_t)\right)\right)$
  8:   $\theta_{t+1} = \theta_t + \alpha\left(\rho_t\delta_t(\theta_t)\phi_{\theta_t}(s_t) - \gamma\rho_t\phi_{\theta_t}(s_{t+1})\phi_{\theta_t}(s_t)^\top\omega_t - h_t(\theta_t, \omega_t)\right)$
  9: **end for**
Output: θ_W

In Algorithm 1, $\mathbf{\Pi}_{R_\omega}(v) = \arg\min_{\|w\|\le R_\omega}\|v-w\|$ denotes the projection operator, where $R_\omega = \frac{\rho_{\max}C_\phi}{\lambda_v}(r_{\max} + (1+\gamma)C_v)$ (the constants are defined in Section 3.1). As we will show in (44) in the appendix that for any $\theta \in \mathbb{R}^N$, $\omega(\theta)$ is always upper bounded by R_ω, i.e., $\|\omega(\theta)\| \le R_\omega$. The projection step in the algorithm is introduced mainly for the convenience of the analysis. Motivated by the randomized stochastic gradient method in [Ghadimi and Lan, 2013], which is designed to analyze non-convex optimization problems, in this paper, we also consider a randomized version of the non-linear TDC algorithm. Specifically, let W be an independent random variable with a uniform distribution over $\{0,1,...,T-1\}$. We then run the non-linear TDC algorithm for W steps and output θ_W.

3.1 Non-asymptotic Error Bounds

In this section, we present our main results of the non-asymptotic error bounds on the convergence of the off-policy non-linear TDC algorithm. Our results will be based on the following assumptions.

Assumption 1 (Boundedness and Smoothness). *For any $s \in \mathcal{S}$ and any $\theta, \theta' \in \mathbb{R}^N$,*

$$|V_\theta(s)| \le C_v, \qquad\qquad \|\phi_\theta(s)\| \le C_\phi,$$
$$\|\nabla^2 V_\theta(s)\| \le D_v, \qquad\qquad \|\nabla^2 V_\theta(s) - \nabla^2 V_{\theta'}(s)\| \le L_V\|\theta - \theta'\|,$$

where C_ϕ, C_v, D_v and L_V are some positive constants.

From Assumption 1, it follows that for any $\theta, \theta' \in \mathbb{R}^N$, $|V_\theta(s) - V_{\theta'}(s)| \le C_\phi\|\theta - \theta'\|$, and $\|\phi_\theta(s) - \phi_{\theta'}(s)\| \le D_v\|\theta - \theta'\|$. We note that these assumptions are equivalent to the assumptions adopted in the original non-linear TDC asymptotic convergence analysis in [Bhatnagar et al., 2009], and can be easily satisfied by appropriately choosing the function class $\{V_\theta : \theta \in \mathbb{R}^N\}$. For example, in neural networks, these assumptions can be satisfied if the activation function is Lipschitz and smooth [Du et al., 2019, Neyshabur, 2017, Miyato et al., 2018].

Assumption 2 (Non-singularity). *For any $\theta \in \mathbb{R}^N$, $\lambda_L(A_\theta) \geq \lambda_v > 0$, where $\lambda_L(A)$ denotes the minimal eigenvalue of the matrix A and λ_v is a positive constant.*

Assumption 3 (Bounded Importance Sampling Ratio). *For any $(s, a) \in \mathcal{S} \times \mathcal{A}$, $\rho(s, a) = \frac{\pi(a|s)}{\pi_b(a|s)} \leq \rho_{\max}$, for some positive constant $\rho_{\max}$.*

The following assumption is only needed for the analysis under the Markovian setting, and is widely used for analyzing the Markovian noise, e.g., [Wang and Zou, 2020, Kaledin et al., 2020, Xu and Liang, 2021, Zou et al., 2019, Srikant and Ying, 2019, Bhandari et al., 2018].

Assumption 4 (Geometric uniform ergodicity). *There exist some constants $m > 0$ and $\kappa \in (0, 1)$ such that $\sup_{s \in \mathcal{S}} d_{TV}(\mathbb{P}(s_t = \cdot | s_0 = s, \pi), \mu^\pi) \leq m\kappa^t$, for any $t > 0$, where d_{TV} denotes the total-variation distance between the probability measures.*

We then present the bounds on the convergence of the TDC algorithm with general smooth function approximation in the following theorem.

Theorem 1. *Consider the following step-sizes: $\alpha = \mathcal{O}\left(\frac{1}{T^a}\right)$, and $\beta = \mathcal{O}\left(\frac{1}{T^b}\right)$, where $\frac{1}{2} \leq a \leq 1$ and $0 < b \leq a$. Then, (1) under the i.i.d. setting, $\|\nabla J(\theta_W)\|^2 = \mathcal{O}\left(\frac{1}{T^{1-a}} + \frac{1}{T^b} + \frac{1}{T^{1-b}}\right)$; and (2) under the Markovian setting, $\|\nabla J(\theta_W)\|^2 = \mathcal{O}\left(\frac{\log T}{T^{1-a}} + \frac{1}{T^{1-b}} + \frac{\log T}{T^b}\right)$.*

Here we only assume the order of the step-sizes in terms of T for simplicity, their exact assumptions on them can be found in Section B.3 and Section C.3. Similarly, we only provide the order of the bounds here, and the explicit bounds can be found in (86) and (133) in the appendix. It can be seen that the rate under the Markovian setting is slower than the one under the i.i.d. setting by a factor of $\log T$, which is essentially the mixing time introduced by the dependence of samples.

Theorem 1 characterizes the dependence between convergence rate and the step-sizes α and β. We also optimize over the step-sizes in the following corollary.

Corollary 1. *Let $a = b = \frac{1}{2}$, i.e., $\alpha, \beta = \mathcal{O}(1/\sqrt{T})$, then (1) under the i.i.d. setting, $\|\nabla J(\theta_W)\|^2 = \mathcal{O}(1/\sqrt{T})$; and (2) under the Markovian setting, $\|\nabla J(\theta_W)\|^2 = \mathcal{O}(\log T/\sqrt{T})$.*

Remark 1. Our result matches with the sample complexity for the batch-based algorithm in [Xu and Liang, 2021]. But their work requires a large batch size of $\mathcal{O}(\epsilon^{-1})$ to control the bias and variance, while ours only needs one sample in each step to update θ and ω and can still obtain the same convergence rate. We note that by setting the batch size being one in [Xu and Liang, 2021], their desired sample complexity cannot be obtained, and their error bound will be a *constant*. To obtain our non-asymptotic bound and sample complexity for the non-linear TDC algorithm, we develop a novel and more refined analysis on the tracking error, which will be discussed in the next section. Moreover, our result matches with the convergence rate of solving general non-convex optimization problems using stochastic gradient descent in [Ghadimi and Lan, 2013]. Compared to their work, our analysis is more challenging due to the two time-scale structure and the gradient bias from the Markovian noise and the tracking error.

Remark 2. Some analyses on two time-scale stochastic approximation bound the tracking error in terms of $\frac{\alpha}{\beta}$, and require $\frac{\alpha}{\beta} \to 0$ in order to drive the tracking error to zero resulting in a convergence rate of $\mathcal{O}\left(\beta + \frac{\alpha}{\beta}\right)$ [Borkar, 2009]. In this paper, we develop a much tighter bound on the tracking error in terms of the slow time-scale parameter $\nabla J(\theta)$. Therefore, the tracking error in our analysis is driven to zero by $\nabla J(\theta) \to 0$ not $\frac{\alpha}{\beta} \to 0$. Similar results that do not need $\frac{\alpha}{\beta} \to 0$ can also be found, e.g., in [Konda et al., 2004, Kaledin et al., 2020]. We would like to point out that the techniques in [Konda et al., 2004, Kaledin et al., 2020] cannot be applied in our analysis due to the non-linear two time-scale updates in this paper.

4 Proof Sketch

In this section, we provide an outline of the proof of Theorem 1 under the Markovian setting, and highlight our major technical contributions. For the complete proof of Theorem 1, we refer the readers to Appendices B.2 and C.2.

Let $O_t = (s_t, a_t, r_t, s_{t+1})$ be the sample observed at time t. Denote the tracking error by $z_t = \omega_t - \omega(\theta_t)$, which characterizes the error between the fast time-scale update and its limit if the slow time-scale update θ_t is kept fixed and only the fast time-scale is being updated. Denote by $G_{t+1}(\theta, \omega) \triangleq \rho_t \delta_t(\theta) \phi_\theta(s_t) - \gamma \rho_t \phi_\theta(s_{t+1}) \phi_\theta(s_t)^\top \omega_t - h_t(\theta, \omega)$. Denote by τ_β the mixing time of the MDP, i.e., $\tau_\beta \triangleq \min\{t : m\kappa^t \leq \beta\}$.

Step 1. In this step, we decompose the error of gradient norm into two parts: the stochastic bias and the tracking error. We first show in Appendix A that $J(\theta)$ is L_J-smooth: for any $\theta_1, \theta_2 \in \mathbb{R}^N$,

$$\|\nabla J(\theta_1) - \nabla J(\theta_2)\| \leq L_J \|\theta_1 - \theta_2\|. \tag{9}$$

We note that the smoothness of $J(\theta)$ is also used in [Xu and Liang, 2021], which, however, is assumed instead of being proved as in this paper. It then follows that

$$\frac{\alpha}{2}\|\nabla J(\theta_t)\|^2 \leq J(\theta_t) - J(\theta_{t+1}) + \underbrace{\alpha\langle \nabla J(\theta_t), -G_{t+1}(\theta_t, \omega(\theta_t)) + G_{t+1}(\theta_t, \omega_t)\rangle}_{(a)}$$

$$+ \alpha \underbrace{\left\langle \nabla J(\theta_t), \frac{\nabla J(\theta_t)}{2} + G_{t+1}(\theta_t, \omega(\theta_t))\right\rangle}_{(b)} + \frac{L_J}{2}\alpha^2\|G_{t+1}(\theta_t, \omega_t)\|^2. \tag{10}$$

This implies that the error bound on the gradient norm is controlled by the tracking error (a) which is introduced by the two time-scale update rule, and the stochastic bias (b) which is due to the time-varying projection and the Markovian sampling.

Step 2. We first bound the tracking error. Re-write the update of ω_t in terms of z_t: $z_{t+1} = z_t + \beta(-A_{\theta_t}(s_t)z_t + b_t(\theta_t)) + \omega(\theta_t) - \omega(\theta_{t+1})$, where $A_{\theta_t}(s_t) = \phi_{\theta_t}(s_t)\phi_{\theta_t}(s_t)^\top$ and $b_t(\theta_t) = -A_{\theta_t}(s_t)\omega(\theta_t) + \rho_t \delta_t(\theta_t)\phi_{\theta_t}(s_t)$. From the Lipschitz continuity of $\omega(\theta)$, it follows that

$$\|z_{t+1}\| \leq (1 + \beta C_\phi^2)\|z_t\| + \beta(b_{\max} + L_\omega C_g),$$
$$\|z_{t+1} - z_t\| \leq \beta C_\phi^2 \|z_t\| + \beta(b_{\max} + L_\omega C_g), \tag{11}$$

which further implies

$$\mathbb{E}\left[\|z_{t+1}\|^2 - \|z_t\|^2\right]$$
$$\leq \underbrace{\mathbb{E}[2z_t^\top(z_{t+1} - z_t + \beta A_{\theta_t} z_t)]}_{(c)} + \mathcal{O}\left(\beta^2 \mathbb{E}[\|z_t\|^2] + \beta^2\right) + \beta\mathbb{E}\left[2z_t^\top(-A_{\theta_t})z_t\right], \tag{12}$$

where the last term in (12) can be further upper bounded by $-2\beta\lambda_v \mathbb{E}[\|z_t\|^2]$.

One challenging part in our analysis is to bound term (c). Equivalently, we decompose the following term into three parts:

$$\mathbb{E}\left[z_t^\top\left(-A_{\theta_t}z_t - \frac{1}{\beta}(z_{t+1} - z_t)\right)\right]$$
$$= \underbrace{\mathbb{E}[z_t^\top(-A_{\theta_t} + A_{\theta_t}(s_t))z_t]}_{(d)} - \underbrace{\mathbb{E}[z_t^\top b_t(\theta_t)]}_{(e)} - \underbrace{\mathbb{E}\left[z_t^\top \frac{\omega(\theta_t) - \omega(\theta_{t+1})}{\beta}\right]}_{(f)}. \tag{13}$$

Consider term (d) in (13). Unlike the case with linear function approximation, where the character function $\nabla V_\theta(s) = \phi(s)$ is independent with θ, here the character function $\phi_\theta(s)$ depends on θ. We use the geometric uniform ergodicity property of the MDP and the Lipschitz continuity of A_θ and $A_\theta(s)$ to decouple the dependence. More specifically, for any fixed θ, $\mathbb{E}[A_\theta(s_t)]$ converges to A_θ as t increases. Let $t = \tau_\beta$, then we have that

$$\mathbb{E}\left[z_{\tau_\beta}^\top(-A_{\theta_{\tau_\beta}} + A_{\theta_{\tau_\beta}}(s_{\tau_\beta}))z_{\tau_\beta}\right]$$
$$= \mathbb{E}\left[z_0^\top(-A_{\theta_0} + A_{\theta_0}(s_{\tau_\beta}))z_0\right] + \mathbb{E}\left[z_0^\top(-A_{\theta_{\tau_\beta}} + A_{\theta_{\tau_\beta}}(s_{\tau_\beta}) + A_{\theta_0} - A_{\theta_0}(s_{\tau_\beta}))z_0\right]$$
$$+ \mathbb{E}\left[(z_{\tau_\beta} - z_0)^\top(-A_{\theta_{\tau_\beta}} + A_{\theta_{\tau_\beta}}(s_{\tau_\beta}))(z_{\tau_\beta} - z_0)\right] + 2\mathbb{E}\left[(z_{\tau_\beta} - z_0)^\top(-A_{\theta_{\tau_\beta}} + A_{\theta_{\tau_\beta}}(s_{\tau_\beta}))z_0\right], \tag{14}$$

which can be further bounded using the mixing time τ_β and the Lipschitz property of A_θ and $A_\theta(s_{\tau_\beta})$. We note that from the update of z_t, we can bound $\|z_{\tau_\beta} - z_0\|$ and $\|z_0\|$ by $\|z_{\tau_\beta}\|$, hence the bound in (14) can be bounded in terms of $\|z_{\tau_\beta}\|$.

Similarly, note that $\mathbb{E}[b_t(\theta)]$ converges to 0 as $t \to \infty$, then we can also bound term (e) in (13):

$$\mathbb{E}[z_{\tau_\beta}^\top b_{\tau_\beta}(\theta_{\tau_\beta})] = \mathbb{E}[(z_{\tau_\beta} - z_0)^\top b_{\tau_\beta}(\theta_{\tau_\beta})] + \mathbb{E}[z_0^\top b_{\tau_\beta}(\theta_0)] + \mathbb{E}[z_0^\top (b_{\tau_\beta}(\theta_{\tau_\beta}) - b_{\tau_\beta}(\theta_0))], \quad (15)$$

which can be similarly bounded in terms of $\|z_{\tau_\beta}\|$.

The challenge of bounding the third term (f) in (13) lies in bounding the difference between $\omega(\theta_t)$ and $\omega(\theta_{t+1})$. One simple approach is to use the Lipschitz continuity of $\omega(\theta)$ and bound $\|\theta_t - \theta_{t+1}\|$ by a constant of order $\mathcal{O}(\alpha)$, but this will lead to a loose bound because the update $G_{t+1}(\theta_t, \omega_t)$ is actually an estimator of the gradient, which will also converge to zero. The key idea in our analysis is to bound term (f) in terms of the gradient of the objective function $\nabla J(\theta)$. Specifically, we first rewrite term $\langle z_t, \omega(\theta_t) - \omega(\theta_{t+1})\rangle = -\langle z_t, \nabla\omega(\hat{\theta}_t)(\theta_{t+1} - \theta_t)\rangle = -\alpha\langle \nabla\omega(\hat{\theta}_t)G_{t+1}(\theta_t, \omega_t)\rangle$, where $\hat{\theta}_t = c\theta_t + (1-c)\theta_{t+1}$ for some $c \in [0,1]$. It can be shown that

$$\mathbb{E}\left[z_{\tau_\beta}^\top \frac{\omega(\theta_{\tau_\beta}) - \omega(\theta_{\tau_\beta+1})}{\beta}\right] = -\frac{\alpha}{\beta}\mathbb{E}[z_{\tau_\beta}^\top \nabla\omega(\hat{\theta}_{\tau_\beta})(G_{\tau_\beta+1}(\theta_{\tau_\beta}, \omega_{\tau_\beta}) - G_{\tau_\beta+1}(\theta_{\tau_\beta}, \omega(\theta_{\tau_\beta})))]$$

$$-\frac{\alpha}{\beta}\mathbb{E}\left[z_{\tau_\beta}^\top \nabla\omega(\hat{\theta}_{\tau_\beta})\left(G_{\tau_\beta+1}(\theta_{\tau_\beta}, \omega(\theta_{\tau_\beta})) + \frac{\nabla J(\theta_{\tau_\beta})}{2}\right)\right] + \frac{\alpha}{\beta}\mathbb{E}\left[z_{\tau_\beta}^\top \nabla\omega(\hat{\theta}_{\tau_\beta})\left(\frac{\nabla J(\theta_{\tau_\beta})}{2}\right)\right].$$
$$(16)$$

The first term in (16) can be bounded in terms of $\|z_{\tau_\beta}\|^2$ using the Lipschitz property of $G_{\tau_\beta+1}(\theta, \omega)$ in ω. The second term can be bounded using the uniform ergodicity of the MDP and the Lipschitz property of $z_0^\top \nabla\omega(\theta)\left(G_{\tau_\beta+1}(\theta, \omega(\theta)) + \frac{\nabla J(\theta)}{2}\right)$ in θ. The third term can be bounded in terms of $\|z_{\tau_\beta}\|^2$ and $\|\nabla J(\theta_{\tau_\beta})\|^2$. Combining all bounds together, we have the bound on term (f) in (13):

$$\left|\mathbb{E}\left[z_{\tau_\beta}^\top \frac{\omega(\theta_{\tau_\beta}) - \omega(\theta_{\tau_\beta+1})}{\beta}\right]\right|$$
$$\leq \mathcal{O}\left(\frac{\alpha}{\beta}\right)\mathbb{E}\left[\|z_{\tau_\beta}\|^2\right] + \mathcal{O}(\alpha\tau_\beta)\mathbb{E}\left[\|z_{\tau_\beta}\|\right] + \mathcal{O}(\alpha\tau_\beta) + \mathcal{O}\left(\frac{\alpha}{8\beta}\right)\mathbb{E}\left[\|\nabla J(\theta_{\tau_\beta})\|^2\right]. \quad (17)$$

We combine all the bounds on terms $(d), (e)$ and (f) and hence get the error bound on (13):

$$\mathbb{E}\left[z_t^\top\left(-A_{\theta_t}z_t - \frac{1}{\beta}(z_{t+1} - z_t)\right)\right] \leq \mathcal{O}\left(\frac{\alpha}{\beta}\right)\mathbb{E}[\|z_t\|^2] + \mathcal{O}(\beta\tau_\beta) + \mathcal{O}\left(\frac{\alpha}{\beta}\right)\mathbb{E}[\|\nabla J(\theta_t)\|^2].$$
$$(18)$$

Plugging the above bound in (12), we have the following recursive bound on the tracking error:

$$\mathbb{E}\left[\|z_{t+1}\|^2\right] \leq \mathcal{O}(1-\beta)\mathbb{E}\left[\|z_t\|^2\right] + \mathcal{O}(\alpha)\mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right] + \mathcal{O}(\beta^2\tau_\beta). \quad (19)$$

Then by recursively applying the inequality in (19) and summing up w.r.t. t from 0 to $T-1$, we obtain the bound on the tracking error $\sum_{t=0}^{T-1}\mathbb{E}[\|z_t\|^2]/T$:

$$\frac{\sum_{t=0}^{T-1}\mathbb{E}[\|z_t\|^2]}{T} \leq \mathcal{O}\left(\frac{1}{T\beta} + \frac{\alpha}{\beta}\frac{\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T} + \beta\tau_\beta\right).$$

Step 3. In this step we bound the stochastic bias term $\mathbb{E}\left[\left\langle \nabla J(\theta_t), \frac{\nabla J(\theta_t)}{2} + G_{t+1}(\theta_t, \omega(\theta_t))\right\rangle\right]$. Similarly, we add and subtract $\nabla J(\theta_0)$ and $G_{\tau_\beta+1}(\theta_0, \omega(\theta_0))$, and obtain that

$$\left\langle \nabla J(\theta_{\tau_\beta}), \frac{\nabla J(\theta_{\tau_\beta})}{2} + G_{\tau_\beta+1}(\theta_{\tau_\beta}, \omega(\theta_{\tau_\beta}))\right\rangle$$

$$= \left\langle \nabla J(\theta_0), \frac{\nabla J(\theta_0)}{2} + G_{\tau_\beta+1}(\theta_0, \omega(\theta_0))\right\rangle + \left(\left\langle \nabla J(\theta_{\tau_\beta}), \frac{\nabla J(\theta_{\tau_\beta})}{2} + G_{\tau_\beta+1}(\theta_{\tau_\beta}, \omega(\theta_{\tau_\beta}))\right\rangle\right.$$

8

$$- \left\langle \nabla J(\theta_0), \frac{\nabla J(\theta_0)}{2} + G_{\tau_\beta+1}(\theta_0, \omega(\theta_0)) \right\rangle \Bigg), \tag{20}$$

which again can be bounded using the geometry uniform ergodicity of the MDP and the Lipschitz continuity of $\left\langle \nabla J(\theta), \frac{\nabla J(\theta)}{2} + G_{\tau_\beta+1}(\theta, \omega(\theta)) \right\rangle$.

Step 4. Plugging in the bounds on the tracking error and the stochastic bias and rearranging the terms, then it follows that $\frac{\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T} \leq U \sqrt{\frac{\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}} + V$, where U and V are some constants depending on the step sizes, and the explicit definitions can be found in (132). By solving the inequality of $\frac{\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}$, we obtain that

$$\frac{\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T} \leq \mathcal{O}\left(\beta\tau_\beta + \frac{1}{T\beta} + \alpha\tau_\beta + \frac{1}{T\alpha} \right).$$

5 Conclusion

In this paper, we extend the on-policy non-linear TDC algorithm to the off-policy setting, and characterize its non-asymptotic error bounds under both the i.i.d. and the Markovian settings. We show that the non-linear TDC algorithm converges as fast as $\mathcal{O}(1/\sqrt{T})$ (up to a factor of $\log T$). The techniques and tools developed in this paper can be used to analyze a wide range of value-based RL algorithms with general smooth function approximation.

Limitations: It is not clear yet whether the stationary points that the TDC converges to are second-order stationary or potentially saddle points.

Negative social impacts: This work is a theoretical investigation of some fundamental RL algorithms, and therefore, the authors do not foresee any negative societal impact.

6 Acknowledgment

References

TW Archibald, KIM McKinnon, and LC Thomas. On the generation of Markov decision processes. *Journal of the Operational Research Society*, 46(3):354–361, 1995.

Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings*, pages 30–37. Elsevier, 1995.

Dimitri P Bertsekas. Dynamic programming and optimal control 3rd edition, volume II. *Belmont, MA: Athena Scientific*, 2011.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Proc. Annual Conference on Learning Theory (CoLT)*, pages 1691–1692. PMLR, 2018.

Shalabh Bhatnagar, Doina Precup, David Silver, Richard S Sutton, Hamid Maei, and Csaba Szepesvári. Convergent temporal-difference learning with arbitrary smooth function approximation. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, volume 22, pages 1204–1212, 2009.

Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

Qi Cai, Zhuoran Yang, Jason D Lee, and Zhaoran Wang. Neural temporal-difference learning converges to global optima. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 11312–11322, 2019.

Gal Dalal, Balázs Szörényi, Gugan Thoppe, and Shie Mannor. Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. *Proceedings of Machine Learning Research*, 75:1–35, 2018.

Gal Dalal, Balazs Szorenyi, and Gugan Thoppe. A tale of two-timescale reinforcement learning with the tightest finite-time bound. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, pages 3701–3708, 2020.

Thinh T Doan. Nonlinear two-time-scale stochastic approximation: Convergence and finite-time performance. In *Learning for Dynamics and Control*, pages 47–47. PMLR, 2021.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *Proc. International Conference on Machine Learning (ICML)*, pages 1675–1685. PMLR, 2019.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

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

Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *arXiv preprint arXiv:2007.05170*, 2020.

Maxim Kaledin, Eric Moulines, Alexey Naumov, Vladislav Tadic, and Hoi-To Wai. Finite time analysis of linear two-timescale stochastic approximation with Markovian noise. In *Proc. Annual Conference on Learning Theory (CoLT)*, pages 2144–2203. PMLR, 2020.

Prasenjit Karmakar and Shalabh Bhatnagar. Two time-scale stochastic approximation with controlled Markov noise and off-policy temporal-difference learning. *Mathematics of Operations Research*, 43(1):130–151, 2018.

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.

Chandrashekar Lakshminarayanan and Csaba Szepesvari. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *Proc. International Conference on Artificial Intelligence and Statistics*, pages 1347–1355, 2018.

Bo Liu, Ji Liu, Mohammad Ghavamzadeh, Sridhar Mahadevan, and Marek Petrik. Finite-sample analysis of proximal gradient TD algorithms. In *Proc. International Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 504–513. Citeseer, 2015.

Shaocong Ma, Yi Zhou, and Shaofeng Zou. Variance-reduced off-policy TDC learning: Non-asymptotic convergence analysis. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 14796–14806, 2020.

Shaocong Ma, Yi Zhou, and Shaofeng Zou. Greedy-GQ with variance reduction: Finite-time analysis and improved complexity. In *Proc. International Conference on Learning Representations (ICLR)*, 2021.

Hamid Reza Maei. Gradient temporal-difference learning algorithms. *Thesis, University of Alberta*, 2011.

Hamid Reza Maei, Csaba Szepesvári, Shalabh Bhatnagar, and Richard S Sutton. Toward off-policy learning control with function approximation. In *Proc. International Conference on Machine Learning (ICML)*, pages 719–726, 2010.

Francisco S Melo, Sean P Meyn, and M Isabel Ribeiro. An analysis of reinforcement learning with function approximation. In *Proc. International Conference on Machine Learning (ICML)*, pages 664–671. ACM, 2008.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *Proc. International Conference on Learning Representations (ICLR)*, 2018.

Abdelkader Mokkadem, Mariane Pelletier, et al. Convergence rate and averaging of nonlinear two-time-scale stochastic approximation algorithms. *The Annals of Applied Probability*, 16(3): 1671–1702, 2006.

Behnam Neyshabur. Implicit regularization in deep learning. *arXiv preprint arXiv:1709.01953*, 2017.

Theodore J Perkins and Doina Precup. A convergent form of approximate policy iteration. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 1627–1634, 2003.

Rayadurgam Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation and TD learning. In *Proc. Annual Conference on Learning Theory (CoLT)*, pages 2803–2830. PMLR, 2019.

Jun Sun, Gang Wang, Georgios B Giannakis, Qinmin Yang, and Zaiyue Yang. Finite-sample analysis of decentralized temporal-difference learning with linear function approximation. In *Proeedings of the International Workshop on Artificial Intelligence and Statistics*, 2020.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction, Second Edition*. The MIT Press, Cambridge, Massachusetts, 2018.

Richard S Sutton, Hamid R Maei, and Csaba Szepesvári. A convergent $O(n)$ temporal-difference algorithm for off-policy learning with linear function approximation. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 1609–1616, 2009a.

Richard S Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proc. International Conference on Machine Learning (ICML)*, pages 993–1000, 2009b.

John N Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE transactions on automatic control*, 42(5):674–690, 1997.

Yue Wang and Shaofeng Zou. Finite-sample analysis of greedy-GQ with linear function approximation under Markovian noise. In *Proc. International Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 11–20. PMLR, 2020.

Yue Wang, Wei Chen, Yuting Liu, Zhi-Ming Ma, and Tie-Yan Liu. Finite sample analysis of the GTD policy evaluation algorithms in Markov setting. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 5504–5513, 2017.

Yue Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite time analysis of two time-scale actor critic methods. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Pan Xu and Quanquan Gu. A finite-time analysis of Q-learning with neural network function approximation. In *Proc. International Conference on Machine Learning (ICML)*, pages 10555–10565. PMLR, 2020.

Tengyu Xu and Yingbin Liang. Sample complexity bounds for two timescale value-based reinforcement learning algorithms. In *Proc. International Conference on Artifical Intelligence and Statistics (AISTATS)*, pages 811–819. PMLR, 2021.

Tengyu Xu, Shaofeng Zou, and Yingbin Liang. Two time-scale off-policy TD learning: Non-asymptotic analysis over Markovian samples. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 10633–10643, 2019.

Shaofeng Zou, Tengyu Xu, and Yingbin Liang. Finite-sample analysis for SARSA with linear function approximation. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 8665–8675, 2019.

Part

Appendix

Table of Contents

We first introduce some notations. In the following proofs, $\|a\|$ denotes the ℓ_2 norm if a is a vector; and $\|A\|$ denotes the operator norm if A is a matrix.

In Appendix A, we prove the Lipschitz continuity of some important functions, including $\omega(\theta)$, $\nabla\omega(\theta)$ and the gradient $\nabla J(\theta)$ of objective function. In Appendix B, we present the non-asymptotic analysis for the i.i.d. setting. In Appendix C, we present the non-asymptotic analysis for the Markovian setting. In appendix D, we present some numerical experiments.

A Useful Lemmas

A.1 Lipschitz Continuity of $\omega(\theta)$

In this section, we show that $\omega(\theta)$ is Lipschitz in θ.

Lemma 1. *For any $\theta, \theta' \in \mathbb{R}^N$, we have that*

$$\|\omega(\theta) - \omega(\theta')\| \le L_\omega \|\theta - \theta'\|, \tag{21}$$

where $L_\omega = \frac{1}{\lambda_v}\left((1+\gamma)C_\phi^2 + (r_{\max} + (1+\gamma)C_v)D_v\right) + \frac{2C_\phi^2 D_v}{\lambda_v^2}(r_{\max} + (1+\gamma)C_v)$.

Proof. Recall that

$$\begin{aligned}
\omega(\theta) &= \mathbb{E}_{\mu^{\pi_b}}[\phi_\theta(S)\phi_\theta(S)^\top]^{-1}\mathbb{E}_{\mu^{\pi_b}}[\rho(S,A)\delta_{S,A,S'}(\theta)\phi_\theta(S)] \\
&= A_\theta^{-1}\mathbb{E}_{\mu^{\pi_b}}[\rho(S,A)\delta_{S,A,S'}(\theta)\phi_\theta(S)],
\end{aligned} \tag{22}$$

hence we can show the conclusion by showing that A_θ^{-1} and $\mathbb{E}_{\mu^{\pi_b}}[\rho(S,A)\delta_{S,A,S'}(\theta)\phi_\theta(S)]$ are both Lipschitz and bounded.

From Assumption 2, we know that

$$\|A_\theta^{-1}\| \le \frac{1}{\lambda_v}. \tag{23}$$

We also show that

$$
\begin{aligned}
&\|A_\theta^{-1} - A_{\theta'}^{-1}\| \\
&= \|A_\theta^{-1} A_{\theta'} A_{\theta'}^{-1} - A_\theta^{-1} A_\theta A_{\theta'}^{-1}\| \\
&= \|A_\theta^{-1}(A_{\theta'} - A_\theta)A_{\theta'}^{-1}\| \\
&\leq \frac{2C_\phi D_v}{\lambda_v^2}\|\theta - \theta'\|,
\end{aligned}
\tag{24}
$$

which is from the fact that $\|A_\theta - A_{\theta'}\| = \left\|\mathbb{E}_{\mu^{\pi_b}}[\phi_\theta(S)\phi_\theta(S)^\top] - \mathbb{E}_{\mu^{\pi_b}}[\phi_{\theta'}(S)\phi_{\theta'}(S)^\top]\right\| \leq 2C_\phi D_v\|\theta - \theta'\|$.

By Assumption 1 and the boundedness of the reward function, it can be shown that for any $\theta \in \mathbb{R}^N$ and any $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$,

$$
|\delta_{s,a,s'}(\theta)| = |r(s, a, s') + \gamma V_\theta(s') - V_\theta(s)| \leq r_{\max} + (1 + \gamma)C_v.
\tag{25}
$$

We then show that $\delta_{s,a,s'}(\theta)$ is Lipschitz, i.e., for any $\theta, \theta' \in \mathbb{R}^N$ and any $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$,

$$
\begin{aligned}
&|\delta_{s,a,s'}(\theta) - \delta_{s,a,s'}(\theta')| \\
&= |\gamma V_\theta(s') - V_\theta(s) - \gamma V_{\theta'}(s') - V_{\theta'}(s)| \\
&\leq (\gamma + 1)C_\phi\|\theta - \theta'\|.
\end{aligned}
\tag{26}
$$

Hence, the function $\|\mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)\delta_{S,A,S'}(\theta)\phi_\theta(S)]\|$ is Lipschitz:

$$
\begin{aligned}
&\|\mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)\delta_{S,A,S'}(\theta)\phi_\theta(S)] - \mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)\delta_{S,A,S'}(\theta')\phi_{\theta'}(S)]\| \\
&= \|\mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)\delta_{S,A,S'}(\theta)\phi_\theta(S)] - \mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)\delta_{S,A,S'}(\theta')\phi_\theta(S)] \\
&\quad + \mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)\delta_{S,A,S'}(\theta')\phi_\theta(S)] - \mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)\delta_{S,A,S'}(\theta')\phi_{\theta'}(S)]\| \\
&\leq \|\mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)\delta_{S,A,S'}(\theta)\phi_\theta(S)] - \mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)\delta_{S,A,S'}(\theta')\phi_\theta(S)]\| \\
&\quad + \|\mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)\delta_{S,A,S'}(\theta')\phi_\theta(S)] - \mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)\delta_{S,A,S'}(\theta')\phi_{\theta'}(S)]\| \\
&\leq \mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)|\delta_{S,A,S'}(\theta) - \delta_{S,A,S'}(\theta')|\|\phi_\theta(S)\|] \\
&\quad + \mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)|\delta_{S,A,S'}(\theta')|\|\phi_\theta(S) - \phi_{\theta'}(S)\|] \\
&\overset{(a)}{\leq} (1 + \gamma)C_\phi^2\|\theta - \theta'\| + (r_{\max} + (1 + \gamma)C_v)D_v\|\theta - \theta'\| \\
&= \left((1 + \gamma)C_\phi^2 + (r_{\max} + (1 + \gamma)C_v)D_v\right)\|\theta - \theta'\|,
\end{aligned}
\tag{27}
$$

where (a) is from (26) and the fact that $\mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)] = 1$. Also $\|\mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)\delta_{S,A,S'}(\theta)\phi_\theta(S)]\|$ can be upper bounded as follows:

$$
\|\mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)\delta_{S,A,S'}(\theta)\phi_\theta(S)]\| \leq C_\phi(r_{\max} + (1 + \gamma)C_v).
\tag{28}
$$

Combining (23), (24), (28) and (27), we show that $\omega(\cdot)$ is Lipschitz in θ:

$$
\begin{aligned}
&\|\omega(\theta) - \omega(\theta')\| \\
&\leq \left(\frac{1}{\lambda_v}\left((1 + \gamma)C_\phi^2 + (r_{\max} + (1 + \gamma)C_v)D_v\right) + \frac{2C_\phi^2 D_v}{\lambda_v^2}(r_{\max} + (1 + \gamma)C_v)\right)\|\theta - \theta'\| \\
&\triangleq L_\omega\|\theta - \theta'\|,
\end{aligned}
\tag{29}
$$

where $L_\omega = \frac{1}{\lambda_v}\left((1 + \gamma)C_\phi^2 + (r_{\max} + (1 + \gamma)C_v)D_v\right) + \frac{2C_\phi^2 D_v}{\lambda_v^2}(r_{\max} + (1 + \gamma)C_v)$. $\qquad \square$

A.2 Lipschitz Continuity of $\nabla\omega(\theta)$

In this section, we show that $\nabla\omega(\theta)$ is Lipschitz.

Lemma 2. *For any $\theta, \theta' \in \mathbb{R}^N$, it follows that*

$$
\|\nabla\omega(\theta) - \nabla\omega(\theta')\| \leq D_\omega\|\theta - \theta'\|,
\tag{30}
$$

where

$$D_\omega = \left(\frac{(C_\phi L_v + 2D_v^2 + D_v C_\phi)}{\lambda_v^2} + \frac{8C_\phi^2 D_v^2}{\lambda_v^3} \right) C_\phi (r_{\max} + C_v + \gamma C_v)$$

$$+ \frac{4C_\phi D_v}{\lambda_v^2} \left(C_\phi^2 (1 + \gamma) + D_v (r_{\max} + (1 + \gamma)C_v) \right)$$

$$+ \frac{3C_\phi D_v (1 + \gamma) + L_v (r_{\max} + (1 + \gamma)C_v)}{\lambda_v}. \tag{31}$$

Proof. Recall the definition of $\omega(\theta) = A_\theta^{-1} \mathbb{E}_{\mu^{\pi_b}} [\rho(S, A)\delta_{S,A,S'}(\theta)\phi_\theta(S)]$, hence we have

$$\nabla\omega(\theta) = -A_\theta^{-1}(\nabla A_\theta)A_\theta^{-1}\mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)\delta_{S,A,S'}(\theta)\phi_\theta(S)]$$

$$+ A_\theta^{-1}\mathbb{E}_{\mu^{\pi_b}}[\nabla\rho(S, A)\delta_{S,A,S'}(\theta)\phi_\theta(S)], \tag{32}$$

where the tensor ∇A_θ can be equivalently viewed as an operator: $\mathbb{R}^N \to \mathbb{R}^{N \times N}$, i.e., $\nabla A_\theta(w) = \nabla(A_\theta w)$ for any $w \in \mathbb{R}^N$.

We show that the operator norm of ∇A_θ is bounded as follows:

$$\|\nabla A_\theta\| = \sup_{\|w\|=1} \|\nabla A_\theta(w)\|$$

$$= \sup_{\|w\|=1} \|\nabla(A_\theta w)\|$$

$$= \sup_{\|w\|=1} \|\nabla\mathbb{E}_{\mu^{\pi_b}}[\phi_\theta(S)\phi_\theta(S)^\top w]\|$$

$$= \sup_{\|w\|=1} \|\mathbb{E}_{\mu^{\pi_b}}[(\phi_\theta(S)^\top w)\nabla\phi_\theta(S)] + \mathbb{E}_{\mu^{\pi_b}}[\phi_\theta(S)(\nabla\phi_\theta(S)^\top w)^\top]\|$$

$$\leq \sup_{\|w\|=1} 2C_\phi D_v\|w\|$$

$$= 2C_\phi D_v. \tag{33}$$

The Lipschitz continuous of ∇A_θ can be shown as follows:

$$\|\nabla A_\theta - \nabla A_{\theta'}\|$$

$$= \sup_{\|w\|=1} \|\nabla(A_\theta w) - \nabla(A_{\theta'} w)\|$$

$$= \sup_{\|w\|=1} \|\mathbb{E}_{\mu^{\pi_b}}[\nabla\phi_\theta(S)(\phi_\theta(S)^\top w) + (\nabla\phi_\theta(S)^\top w)\phi_\theta(S)^\top - \nabla\phi_{\theta'}(S)(\phi_{\theta'}(S)^\top w)$$

$$- (\nabla\phi_{\theta'}(S)^\top w)\phi_{\theta'}(S)^\top]\|$$

$$\leq \sup_{\|w\|=1} (C_\phi L_v + 2D_v^2 + D_v C_\phi)\|\theta - \theta'\|\|w\|$$

$$= (C_\phi L_v + 2D_v^2 + D_v C_\phi)\|\theta - \theta'\|. \tag{34}$$

Then we conclude that the operator norm of $-A_\theta^{-1}(\nabla A_\theta)$ is upper bounded by $\frac{2C_\phi D_v}{\lambda_v}$, and is Lipschitz with constant $\frac{(C_\phi L_v + 2D_v^2 + D_v C_\phi)}{\lambda_v} + \frac{4C_\phi^2 D_v^2}{\lambda_v^2}$. It can be further seen that $-A_\theta^{-1}(\nabla A_\theta)A_\theta^{-1}$ is upper bounded by $\frac{2C_\phi D_v}{\lambda_v^2}$, and Lipschitz with constant $\frac{(C_\phi L_v + 2D_v^2 + D_v C_\phi)}{\lambda_v^2} + \frac{8C_\phi^2 D_v^2}{\lambda_v^3}$.

Recall that we have shown in (28) that

$$\|\mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)\delta_{S,A,S'}(\theta)\phi_\theta(S)] - \mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)\delta_{S,A,S'}(\theta')\phi_{\theta'}(S)]\|$$

$$\leq \left((1 + \gamma)C_\phi^2 + (r_{\max} + (1 + \gamma)C_v)D_v \right) \|\theta - \theta'\|, \tag{35}$$

and it is upper bounded by $C_\phi(r_{\max} + (1 + \gamma)C_V)$. Hence we have that $-A_\theta^{-1}(\nabla A_\theta)A_\theta^{-1}\mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)\delta_{S,A,S'}(\theta)\phi_\theta(S)]$ can be upper bounded by $(r_{\max} + (1 + \gamma)C_V)\frac{2C_\phi^2 D_v}{\lambda_v^2}$, and it is Lipschitz with constant $\left(\frac{(C_\phi L_v + 2D_v^2 + D_v C_\phi)}{\lambda_v^2} + \frac{8C_\phi^2 D_v^2}{\lambda_v^3} \right) C_\phi(r_{\max} + C_v + \gamma C_v) + \frac{2C_\phi D_v}{\lambda_v^2} \left((1 + \gamma)C_\phi^2 + (r_{\max} + (1 + \gamma)C_v)D_v \right) \triangleq L_A$.

For the second term of (32), we also show it is Lipschitz as follows. First note that $\nabla \delta_{s,a,s'}(\theta)\phi_\theta(s) = \nabla \delta_{s,a,s'}(\theta)\phi_\theta(s)^\top + \delta_{s,a,s'}(\theta)\nabla\phi_\theta(s)$, hence we know $\mathbb{E}_{\mu^{\pi_b}}[\nabla\rho(S,A)\delta_{S,A,S'}(\theta)\phi_\theta(S)]$ can be upper bounded by $C_\phi^2(1+\gamma) + D_v(r_{\max} + (1+\gamma)C_v)$, and is Lipschitz with constant $3C_\phi D_v(1+\gamma) + L_v(r_{\max} + (1+\gamma)C_v)$. Finally we conclude that the second term in (32) $A_\theta^{-1}\mathbb{E}_{\mu^{\pi_b}}[\nabla\rho(S,A)\delta_{S,A,S'}(\theta)\phi_\theta(S)]$ is Lipschitz with constant $\frac{2C_\phi D_v}{\lambda_v^2}\left(C_\phi^2(1+\gamma) + D_v(r_{\max} + (1+\gamma)C_v)\right) + \frac{3C_\phi D_v(1+\gamma)+L_v(r_{\max}+(1+\gamma)C_v)}{\lambda_v} \triangleq L_A'$.

Hence $\nabla\omega(\theta)$ is Lipschitz with constant $L_A + L_A' \triangleq D_\omega$, where

$$
\begin{aligned}
D_\omega = & \left(\frac{(C_\phi L_v + 2D_v^2 + D_v C_\phi)}{\lambda_v^2} + \frac{8C_\phi^2 D_v^2}{\lambda_v^3}\right) C_\phi(r_{\max} + C_v + \gamma C_v) \\
& + \frac{4C_\phi D_v}{\lambda_v^2}\left(C_\phi^2(1+\gamma) + D_v(r_{\max} + (1+\gamma)C_v)\right) \\
& + \frac{3C_\phi D_v(1+\gamma) + L_v(r_{\max} + (1+\gamma)C_v)}{\lambda_v}.
\end{aligned}
\tag{36}
$$

$\square$

A.3 Smoothness of $J(\theta)$

In the following lemma, we show that the objective function $J(\theta)$ is L_J-smooth. We note that the smoothness of $J(\theta)$ is assumed in [Xu and Liang, 2021] instead of being proved as in this paper.

Lemma 3. $J(\theta)$ is L_J-smooth, i.e., for any $\theta, \theta' \in \mathbb{R}^N$,

$$
\|\nabla J(\theta) - \nabla J(\theta')\| \le L_J\|\theta - \theta'\|,
\tag{37}
$$

where

$$
\begin{aligned}
L_J = & 2\left((1+\gamma)C_\phi^2 + (r_{\max} + (1+\gamma)C_v)D_v\right) + 2\gamma\left(C_\phi^2 L_\omega + 2D_v\frac{C_\phi^2}{\lambda_v}(r_{\max} + (1+\gamma)C_v)\right) \\
& + 2\bigg((D_v R_\omega + C_\phi L_\omega + (1+\gamma)C_\phi)D_v R_\omega \\
& + (R_\omega L_V + D_v L_\omega)((r_{\max} + (1+\gamma)C_v) + C_\phi R_\omega)\bigg).
\end{aligned}
\tag{38}
$$

Proof. Before we prove the main statement, we first drive some boundedness and Lipschitz properties. Recall that

$$
\begin{aligned}
-\frac{\nabla J(\theta)}{2} = \mathbb{E}_{\mu^{\pi_b}}\bigg[&\left(\rho(S,A)\delta_{S,A,S'}(\theta)\phi_\theta(S) - \gamma\rho(S,A)\phi_\theta(S')\phi_\theta(S)^\top\omega(\theta)\right. \\
& \left. - h_{S,A,S'}(\theta, \omega(\theta))\right)\bigg],
\end{aligned}
\tag{39}
$$

$$
\omega(\theta) = \mathbb{E}_{\mu^{\pi_b}}[\phi_\theta(S)\phi_\theta(S)^\top]^{-1}\mathbb{E}_{\mu^{\pi_b}}[\rho(S,A)\delta_{S,A,S'}(\theta)\phi_\theta(S)],
\tag{40}
$$

$$
h_{s,a,s'}(\theta, \omega(\theta)) = (\rho(s,a)\delta_{s,a,s'}(\theta) - \phi_\theta(s)^\top\omega(\theta))\nabla^2 V_\theta(s)\omega(\theta).
\tag{41}
$$

We have shown in Lemma 1 that for any $\theta \in \mathbb{R}^N$ and any $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$,

$$
|\delta_{s,a,s'}(\theta)| = |r(s,a,s') + \gamma V_\theta(s') - V_\theta(s)| \le r_{\max} + (1+\gamma)C_v;
\tag{42}
$$

and that

$$
\begin{aligned}
& \|\mathbb{E}_{\mu^{\pi_b}}[\rho(S,A)\delta_{S,A,S'}(\theta)\phi_\theta(S)] - \mathbb{E}_{\mu^{\pi_b}}[\rho(S,A)\delta_{S,A,S'}(\theta')\phi_{\theta'}(S)]\| \\
& \le \left((1+\gamma)C_\phi^2 + (r_{\max} + (1+\gamma)C_v)D_v\right)\|\theta - \theta'\|.
\end{aligned}
\tag{43}
$$

Also it is easy to see from the definition that

$$
\|\omega(\theta)\| \le \frac{C_\phi}{\lambda_v}(r_{\max} + (1+\gamma)C_v) \triangleq R_\omega.
\tag{44}
$$

Hence the Lipschitz continuity of $\mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)\phi_\theta(S')\phi_\theta(S)^\top]\omega(\theta)$ can be shown as follows

$$
\begin{aligned}
&\left\|\mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)\phi_\theta(S')\phi_\theta(S)^\top]\omega(\theta) - \mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)\phi_{\theta'}(S')\phi_{\theta'}(S)^\top]\omega(\theta')\right\| \\
&\leq \left\|\mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)\phi_\theta(S')\phi_\theta(S)^\top]\omega(\theta) - \mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)\phi_\theta(S')\phi_\theta(S)^\top]\omega(\theta')\right\| \\
&\quad + \left\|\mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)\phi_\theta(S')\phi_\theta(S)^\top]\omega(\theta') - \mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)\phi_{\theta'}(S')\phi_{\theta'}(S)^\top]\omega(\theta')\right\| \\
&\overset{(a)}{\leq} C_\phi^2 L_\omega\|\theta - \theta'\| + 2C_\phi D_v R_\omega\|\theta - \theta'\| \\
&= \left(C_\phi^2 L_\omega + 2D_v \frac{C_\phi^2}{\lambda_v}(r_{\max} + (1 + \gamma)C_v)\right)\|\theta - \theta'\|,
\end{aligned}
\tag{45}
$$

where (a) is due to the fact that $\omega(\theta)$ is Lipschitz in (21) and the fact that

$$
\left\|\mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)\phi_\theta(S')\phi_\theta(S)^\top] - \mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)\phi_{\theta'}(S')\phi_{\theta'}(S)^\top]\right\| \leq 2C_\phi D_v\|\theta - \theta'\|.
\tag{46}
$$

We then show that the function $h_{s,a,s'}(\theta, \omega(\theta))$ is Lipschitz in θ as follows. We first note that for any $s \in \mathcal{S}$ and $\theta, \theta' \in \mathbb{R}^N$,

$$
\begin{aligned}
&\|\phi_\theta(s)^\top\omega(\theta) - \phi_{\theta'}(s)^\top\omega(\theta')\| \\
&\leq \|\phi_\theta(s)^\top\omega(\theta) - \phi_{\theta'}(s)^\top\omega(\theta)\| + \|\phi_{\theta'}(s)^\top\omega(\theta) - \phi_{\theta'}(s)^\top\omega(\theta')\| \\
&\leq (D_v R_\omega + C_\phi L_\omega)\|\theta - \theta'\|.
\end{aligned}
\tag{47}
$$

This implies that for any $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ and $\theta, \theta' \in \mathbb{R}^N$,

$$
\begin{aligned}
&\|\rho(s, a)\delta_{s,a,s'}(\theta) - \phi_\theta(s)^\top\omega(\theta) - \rho(s, a)\delta_{s,a,s'}(\theta') + \phi_{\theta'}(s)^\top\omega(\theta')\| \\
&\leq (D_v R_\omega + C_\phi L_\omega + (1 + \gamma)C_\phi\rho(s, a))\|\theta - \theta'\|.
\end{aligned}
\tag{48}
$$

We also show the following function is Lipschitz:

$$
\begin{aligned}
&\|\nabla^2 V_\theta(s)\omega(\theta) - \nabla^2 V_{\theta'}(s)\omega(\theta')\| \\
&\leq \|\nabla^2 V_\theta(s)\omega(\theta) - \nabla^2 V_{\theta'}(s)\omega(\theta)\| + \|\nabla^2 V_{\theta'}(s)\omega(\theta) - \nabla^2 V_{\theta'}(s)\omega(\theta')\| \\
&\leq R_\omega L_V\|\theta - \theta'\| + D_v L_\omega\|\theta - \theta'\| \\
&= (R_\omega L_V + D_v L_\omega)\|\theta - \theta'\|.
\end{aligned}
\tag{49}
$$

Combining (48) and (49), it can be shown that $h_{s,a,s'}(\theta, \omega(\theta))$ is Lipschitz in θ as follows

$$
\begin{aligned}
&\|h_{s,a,s'}(\theta, \omega(\theta)) - h_{s,a,s'}(\theta', \omega(\theta'))\| \\
&= \|\left(\rho(s, a)\delta_{s,a,s'}(\theta) - \phi_\theta(s)^\top\omega(\theta)\right)\nabla^2 V_\theta(s)\omega(\theta) \\
&\quad - \left(\rho(s, a)\delta_{s,a,s'}(\theta') - \phi_{\theta'}(s)^\top\omega(\theta')\right)\nabla^2 V_{\theta'}(s)\omega(\theta')\| \\
&\leq ((D_v R_\omega + C_\phi L_\omega + (1 + \gamma)C_\phi\rho(s, a))D_v R_\omega)\|\theta - \theta'\| \\
&\quad + (R_\omega L_V + D_v L_\omega)(\rho(s, a)(r_{\max} + (1 + \gamma)C_v) + C_\phi R_\omega)\|\theta - \theta'\|.
\end{aligned}
\tag{50}
$$

From the results in (43), (45) and (50), it follows that

$$
\begin{aligned}
&\|\nabla J(\theta) - \nabla J(\theta')\| \\
&\leq 2\left\|\mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)\delta_{S,A,S'}(\theta)\phi_\theta(S) - \rho(S, A)\delta_{S,A,S'}(\theta')\phi_{\theta'}(S)]\right\| \\
&\quad + 2\gamma\left\|\mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)\phi_\theta(S')\phi_\theta(S)^\top\omega(\theta) - \rho(S, A)\phi_{\theta'}(S')\phi_{\theta'}(S)^\top\omega(\theta')]\right\| \\
&\quad + 2\left\|\mathbb{E}_{\mu^{\pi_b}}[h_{S,A,S'}(\theta, \omega(\theta)) - h_{S,A,S'}(\theta', \omega(\theta'))]\right\| \\
&\leq 2\left((1 + \gamma)C_\phi^2 + (r_{\max} + (1 + \gamma)C_v)D_v\right)\|\theta - \theta'\| \\
&\quad + 2\gamma\left(C_\phi^2 L_\omega + 2D_v\frac{C_\phi^2}{\lambda_v}(r_{\max} + (1 + \gamma)C_v)\right)\|\theta - \theta'\| \\
&\quad + 2\mathbb{E}_{\mu^{\pi_b}}[((D_v R_\omega + C_\phi L_\omega + (1 + \gamma)C_\phi\rho(S, A))D_v R_\omega)]\|\theta - \theta'\| \\
&\quad + 2\mathbb{E}_{\mu^{\pi_b}}[(R_\omega L_V + D_v L_\omega)(\rho(S, A)(r_{\max} + (1 + \gamma)C_v) + C_\phi R_\omega)]\|\theta - \theta'\|
\end{aligned}
$$

17

$$\overset{(a)}{\leq} 2\left((1+\gamma)C_\phi^2 + (r_{\max} + (1+\gamma)C_v)D_v\right)\|\theta - \theta'\|$$

$$+ 2\gamma\left(C_\phi^2 L_\omega + 2D_v\frac{C_\phi^2}{\lambda_v}(r_{\max} + (1+\gamma)C_v)\right)\|\theta - \theta'\|$$

$$+ 2\big((D_v R_\omega + C_\phi L_\omega + (1+\gamma)C_\phi)D_v R_\omega$$

$$+ (R_\omega L_V + D_v L_\omega)((r_{\max} + (1+\gamma)C_v) + C_\phi R_\omega)\big)\|\theta - \theta'\|$$

$$\triangleq L_J\|\theta - \theta'\|, \tag{51}$$

where (a) is due to the fact that $\mathbb{E}_{\mu^{\pi_b}}[\rho(S, A)] = 1$, and

$$L_J = 2\left((1+\gamma)C_\phi^2 + (r_{\max} + (1+\gamma)C_v)D_v\right) + 2\gamma\left(C_\phi^2 L_\omega + 2D_v\frac{C_\phi^2}{\lambda_v}(r_{\max} + (1+\gamma)C_v)\right)$$

$$+ 2\big((D_v R_\omega + C_\phi L_\omega + (1+\gamma)C_\phi)D_v R_\omega$$

$$+ (R_\omega L_V + D_v L_\omega)((r_{\max} + (1+\gamma)C_v) + C_\phi R_\omega)\big). \tag{52}$$

This completes the proof. $\qquad\square$

B Non-asymptotic Analysis under the i.i.d. Setting

First we introduce the off-policy TDC learning with non-linear function approximation algorithm under the i.i.d. setting in Algorithm 2. We then bound the tracking error in Appendix B.1, and prove the Theorem 1 under the i.i.d. setting in Appendix B.2.

Algorithm 2 Non-Linear Off-Policy TDC under the i.i.d. Setting
---
Input: $T, \alpha, \beta, \pi, \pi_b, \{V_\theta | \theta \in \mathbb{R}^N\}$
Initialization: θ_0, ω_0
1: Choose $W \sim \text{Uniform}(0, 1, ..., T-1)$
2: **for** $t = 0, 1, ..., W-1$ **do**
3:     Sample $O_t = (s_t, a_t, r_t, s'_t)$ according to μ^{π_b}
4:     $\rho_t = \frac{\pi(a_t|s_t)}{\pi_b(a_t|s_t)}$
5:     $\delta_t(\theta_t) = r(s_t, a_t, s'_t) + \gamma V_{\theta_t}(s'_t) - V_{\theta_t}(s_t)$
6:     $h_t(\theta_t, \omega_t) = \left(\rho_t\delta_t(\theta_t) - \phi_{\theta_t}(s_t)^\top\omega_t\right)\nabla^2 V_{\theta_t}(s_t)\omega_t$
7:     $\omega_{t+1} = \mathbf{\Pi}_{R_\omega}\left(\omega_t + \beta\left(-\phi_{\theta_t}(s_t)\phi_{\theta_t}(s_t)^\top\omega_t + \rho_t\delta_t(\theta_t)\phi_{\theta_t}(s_t)\right)\right)$
8:     $\theta_{t+1} = \theta_t + \alpha\left(\rho_t\delta_t(\theta_t)\phi_{\theta_t}(s_t) - \gamma\rho_t\phi_{\theta_t}(s'_t)\phi_{\theta_t}(s_t)^\top\omega_t - h_t(\theta_t, \omega_t)\right)$
9: **end for**
Output: θ_W

We note that under the i.i.d. setting, it is assumed that at each time step t, a sample $O_t = (s_t, a_t, r_t, s'_t)$ is available, where $s_t \sim \mu^{\pi_b}(\cdot)$, $a_t \sim \pi_b(\cdot|s_t)$ and $s'_t \sim \mathsf{P}(\cdot|s_t, a_t)$.

B.1 Tracking Error Analysis under the i.i.d. Setting

Denote the tracking error by $z_t = \omega_t - \omega(\theta_t)$. Then by the update of ω_t, the update of z_t can be written as

$$z_{t+1} = \omega_{t+1} - \omega(\theta_{t+1})$$

$$= \omega_t + \beta\left(-\phi_{\theta_t}(s_t)\phi_{\theta_t}(s_t)^\top\omega_t + \rho_t\delta_t(\theta_t)\phi_{\theta_t}(s_t)\right) - \omega(\theta_{t+1})$$

$$= z_t + \omega(\theta_t) - \omega(\theta_{t+1}) + \beta\left(-\phi_{\theta_t}(s_t)\phi_{\theta_t}(s_t)^\top(z_t + \omega(\theta_t)) + \rho_t\delta_t(\theta_t)\phi_{\theta_t}(s_t)\right)$$

$$= z_t + \omega(\theta_t) - \omega(\theta_{t+1}) + \beta\left(-A_{\theta_t}(s_t)z_t - A_{\theta_t}(s_t)\omega(\theta_t) + \rho_t\delta_t(\theta_t)\phi_{\theta_t}(s_t)\right), \tag{53}$$

where $A_{\theta_t}(s_t) = \phi_{\theta_t}(s_t)\phi_{\theta_t}(s_t)^\top$. It then follows that

$$\|z_{t+1}\|^2$$

$$\begin{aligned}
&= \|z_t + \omega(\theta_t) - \omega(\theta_{t+1}) + \beta\left(-A_{\theta_t}(s_t)z_t - A_{\theta_t}(s_t)\omega(\theta_t) + \rho_t\delta_t(\theta_t)\phi_{\theta_t}(s_t)\right)\|^2 \\
&= \|z_t\|^2 + \|\omega(\theta_t) - \omega(\theta_{t+1}) + \beta\left(-A_{\theta_t}(s_t)z_t - A_{\theta_t}(s_t)\omega(\theta_t) + \rho_t\delta_t(\theta_t)\phi_{\theta_t}(s_t)\right)\|^2 \\
&\quad + 2\langle z_t, \omega(\theta_t) - \omega(\theta_{t+1})\rangle - 2\beta\langle z_t, A_{\theta_t}(s_t)z_t\rangle + 2\beta\langle z_t, -A_{\theta_t}(s_t)\omega(\theta_t) + \rho_t\delta_t(\theta_t)\phi_{\theta_t}(s_t)\rangle \\
&\leq \|z_t\|^2 + \underbrace{2\beta^2\|\left(-A_{\theta_t}(s_t)z_t - A_{\theta_t}(s_t)\omega(\theta_t) + \rho_t\delta_t(\theta_t)\phi_{\theta_t}(s_t)\right)\|^2}_{(a)} \\
&\quad + \underbrace{2\|\omega(\theta_t) - \omega(\theta_{t+1})\|^2}_{(b)} + \underbrace{2\langle z_t, \omega(\theta_t) - \omega(\theta_{t+1})\rangle}_{(c)} \underbrace{-2\beta\langle z_t, A_{\theta_t}(s_t)z_t\rangle}_{(d)} \\
&\quad + 2\beta\langle z_t, -A_{\theta_t}(s_t)\omega(\theta_t) + \rho_t\delta_t(\theta_t)\phi_{\theta_t}(s_t)\rangle.
\end{aligned} \tag{54}$$

We then provide the bounds of the terms in (54) one by one. Their proofs can be found in Appendices B.1.1 to B.1.4.

Term (a) can be bounded as follows:

$$2\beta^2\|\left(-A_{\theta_t}(s_t)z_t - A_{\theta_t}(s_t)\omega(\theta_t) + \rho_t\delta_t(\theta_t)\phi_{\theta_t}(s_t)\right)\|^2 \leq 4\beta^2 C_\phi^2\|z_t\|^2 + 4\beta^2 C_{g1}, \tag{55}$$

where $C_{g1} = \left(\frac{C_\phi^3}{\lambda_v}(r_{\max} + (1+\gamma)C_v) + \rho_{\max}C_\phi(r_{\max} + (1+\gamma)C_v)\right)^2$.

Term (b) can be bounded as follows:

$$2\|\omega(\theta_t) - \omega(\theta_{t+1})\|^2 \leq 4\alpha^2 L_\omega^2 L_g^2\|z_t\|^2 + 4\alpha^2 C_g^2 L_\omega^2, \tag{56}$$

where $C_g = \rho_{\max}C_\phi(r_{\max} + (1+\gamma)C_v) + \gamma\rho_{\max}R_\omega C_\phi^2 + D_v R_\omega(R_\omega C_\phi + \rho_{\max}(r_{\max} + C_v + \gamma C_v))$.

Term (c) can be bounded as follows:

$$2\langle z_t, \omega(\theta_t) - \omega(\theta_{t+1})\rangle$$
$$\leq 2(\alpha L_\omega L_g + \tfrac{1}{2}\alpha L_\omega + 4\alpha^2 C_g L_g D_\omega)\|z_t\|^2 + \frac{\alpha L_\omega}{4}\|\nabla J(\theta_t)\|^2 + \frac{\alpha^2 C_g^3 D_\omega}{L_g} + 2\alpha\eta_G(\theta_t, z_t, O_t), \tag{57}$$

where $\eta_G(\theta_t, z_t, O_t) = -\left\langle z_t, \nabla\omega(\theta_t)\left(G_{t+1}(\theta_t, \omega(\theta_t)) + \frac{\nabla J(\theta_t)}{2}\right)\right\rangle$.

Term (d) can be bounded as follows:

$$-2\beta\langle z_t, A_{\theta_t}(s_t)z_t\rangle \leq -2\beta\lambda_v\|z_t\|^2 + 2\beta\langle z_t, (A_{\theta_t} - A_{\theta_t}(s_t))z_t\rangle, \tag{58}$$

where $A_\theta = \mathbb{E}_{\mu^{\pi_b}}\left[\phi_\theta(S)\phi_\theta(S)^\top\right]$ is the expectation of $A_\theta(S)$.

By plugging all the bounds from (55), (56), (57) and (58) in (54), it follows that

$$\begin{aligned}
&\|z_{t+1}\|^2 \\
&\leq (1 + 4\beta^2 C_\phi^2 + 4\alpha^2 L_\omega^2 L_g^2 + 2\alpha L_w L_g + \alpha L_w + 8\alpha^2 C_g L_g D_\omega - 2\beta\lambda_v)\|z_t\|^2 \\
&\quad + \frac{1}{4}\alpha L_\omega\|\nabla J(\theta_t)\|^2 + 4\beta^2 C_{g1} + 4\alpha^2 C_g^2 L_\omega^2 + \frac{\alpha^2 C_g^3 D_\omega}{L_g} + 2\alpha\eta_G(\theta_t, z_t, O_t) \\
&\quad + 2\beta\langle z_t, (A_{\theta_t} - A_{\theta_t}(s_t))z_t\rangle + 2\beta\langle z_t, -A_{\theta_t}(s_t)\omega(\theta_t) + \rho_t\delta_t(\theta_t)\phi_{\theta_t}(s_t)\rangle \\
&\triangleq (1-q)\|z_t\|^2 + \frac{\alpha L_\omega}{4}\|\nabla J(\theta_t)\|^2 + 4\beta^2 C_{g1} + 4\alpha^2 C_g^2 L_\omega^2 + \frac{\alpha^2 C_g^3 D_\omega}{L_g} + 2\alpha\eta_G(\theta_t, z_t, O_t) \\
&\quad + 2\beta\langle z_t, (A_{\theta_t} - A_{\theta_t}(s_t))z_t\rangle + 2\beta\langle z_t, -A_{\theta_t}(s_t)\omega(\theta_t) + \rho_t\delta_t(\theta_t)\phi_{\theta_t}(s_t)\rangle, \tag{59}
\end{aligned}$$

where $q = 2\beta\lambda_v - 4\beta^2 C_\phi^2 - 4\alpha^2 L_\omega^2 L_g^2 - 2\alpha L_w L_g - \alpha L_w - 8\alpha^2 C_g L_g D_\omega$. Note that $q = \mathcal{O}(\beta - \beta^2 - \alpha - \alpha^2) = \mathcal{O}(\beta)$, hence we can choose α and β such that $q > 0$.

Note that under the i.i.d. setting,

$$\begin{aligned}
\mathbb{E}\left[\eta_G(\theta_t, z_t, O_t)\right] &= \mathbb{E}\left[\mathbb{E}\left[\eta_G(\theta_t, z_t, O_t)|\mathcal{F}_t\right]\right] \\
&= \mathbb{E}\left[-\left\langle z_t, \nabla\omega(\theta_t)\mathbb{E}\left[\left(G_{t+1}(\theta_t, \omega(\theta_t)) + \frac{\nabla J(\theta_t)}{2}\right)\middle|\mathcal{F}_t\right]\right\rangle\right]
\end{aligned}$$

19

$$= 0, \tag{60}$$

which is due to the fact that $\mathbb{E}_{\mu^{\pi_b}}[G_{t+1}(\theta, \omega(\theta))] = -\frac{\nabla J(\theta)}{2}$ when θ is fixed, and $\mathcal{F}_t$ is the σ-field generated by the randomness until θ_t and ω_t. Similarly, it can also be shown that

$$\mathbb{E}[\langle z_t, (A_{\theta_t} - A_{\theta_t}(s_t))z_t \rangle] = 0 \tag{61}$$

$$\mathbb{E}[\langle z_t, -A_{\theta_t}(s_t)\omega(\theta_t) + \rho_t \delta_t(\theta_t)\phi_{\theta_t}(s_t) \rangle] = 0. \tag{62}$$

Hence the tracking error in (59) can be further bounded as

$$\mathbb{E}[\|z_{t+1}\|^2] \le (1-q)\mathbb{E}\left[\|z_t\|^2\right] + \frac{\alpha L_\omega}{4}\mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right] + 4\beta^2 C_{g1} + 4\alpha^2 C_g^2 L_\omega^2 + \frac{\alpha^2 C_g^3 D_\omega}{L_g}. \tag{63}$$

Recursively applying the inequality in (63), it follows that

$$\mathbb{E}\left[\|z_t\|^2\right] \le (1-q)^t \|z_0\|^2 + \frac{\alpha L_\omega}{4}\sum_{i=0}^t (1-q)^{t-i}\mathbb{E}\left[\|\nabla J(\theta_i)\|^2\right]$$

$$+ \frac{1}{q}\left(4\beta^2 C_{g1} + 4\alpha^2 C_g^2 L_\omega^2 + \frac{\alpha^2 C_g^3 D_\omega}{L_g}\right), \tag{64}$$

and summing up w.r.t. t from 0 to $T-1$, it follows that

$$\frac{\sum_{t=0}^{T-1} \mathbb{E}\left[\|z_t\|^2\right]}{T} \le \frac{\sum_{t=0}^{T-1}(1-q)^t}{T}\|z_0\|^2 + \frac{\alpha L_\omega}{4T}\sum_{t=0}^{T-1}\sum_{i=0}^t (1-q)^{t-i}\mathbb{E}\left[\|\nabla J(\theta_i)\|^2\right]$$

$$+ \frac{1}{q}\left(4\beta^2 C_{g1} + 4\alpha^2 C_g^2 L_\omega^2 + \frac{\alpha^2 C_g^3 D_\omega}{L_g}\right)$$

$$\overset{(a)}{\le} \frac{\|z_0\|^2}{Tq} + \frac{\alpha L_\omega}{4q}\frac{\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right]}{T}$$

$$+ \frac{1}{q}\left(4\beta^2 C_{g1} + 4\alpha^2 C_g^2 L_\omega^2 + \frac{\alpha^2 C_g^3 D_\omega}{L_g}\right)$$

$$= \mathcal{O}\left(\frac{1}{T\beta} + \frac{\alpha}{\beta}\frac{\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right]}{T} + \beta\right), \tag{65}$$

where (a) is due to the double-sum trick, i.e., for any $x_i \ge 0$, $\sum_{t=0}^{T-1}\sum_{i=0}^t (1-q)^{t-i}x_i \le \sum_{t=0}^{T-1}(1-q)^t\sum_{t=0}^{T-1}x_t \le \frac{1}{q}\sum_{t=0}^{T-1}x_t$, and the last step is because $q = \mathcal{O}(\beta)$.

B.1.1 Bound on Term (a)

In this section we provide the detailed proof of the bound on term (a) in (55).

It can be shown that

$$\|\left(-A_{\theta_t}(s_t)z_t - A_{\theta_t}(s_t)\omega(\theta_t) + \rho_t \delta_t(\theta_t)\phi_{\theta_t}(s_t)\right)\|^2$$

$$\le 2\| - A_{\theta_t}(s_t)z_t\|^2 + 2\| - A_{\theta_t}(s_t)\omega(\theta_t) + \rho_t \delta_t(\theta_t)\phi_{\theta_t}(s_t)\|^2$$

$$\overset{(a)}{\le} 2C_\phi^2\|z_t\|^2 + 2\left(\frac{C_\phi^3}{\lambda_v}(r_{\max} + (1+\gamma)C_v) + \rho_{\max}C_\phi(r_{\max} + (1+\gamma)C_v)\right)^2, \tag{66}$$

where (a) is from the fact that $\|A_\theta(s)\| = \|\phi_\theta(s)\phi_\theta(s)^\top\| \le C_\phi^2$ and the bounds in (42) and (44).

B.1.2 Bound on Term (b)

In this section we provide the detailed proof of the bound on term (b) in (56).

We first show that $G_{t+1}(\theta, \omega)$ is Lipschitz in ω for any fixed θ. Specifically, for any $\theta, \omega_1, \omega_2 \in \mathbb{R}^N$, it follows that

$$
\begin{aligned}
&\|G_{t+1}(\theta, \omega_1) - G_{t+1}(\theta, \omega_2)\| \\
&= \|\rho_t \delta_t(\theta) \phi_\theta(s_t) - \gamma \rho_t \phi_\theta(s_t') \phi_\theta(s_t)^\top \omega_1 - h_t(\theta, \omega_1) - \rho_t \delta_t(\theta) \phi_\theta(s_t) + \gamma \rho_t \phi_\theta(s_t') \phi_\theta(s_t)^\top \omega_2 \\
&\quad + h_t(\theta, \omega_2)\| \\
&\leq \|h_t(\theta, \omega_1) - h_t(\theta, \omega_2)\| + \|\gamma \rho_t \phi_\theta(s_t') \phi_\theta(s_t)^\top \omega_1 - \gamma \rho_t \phi_\theta(s_t') \phi_\theta(s_t)^\top \omega_2\| \\
&\overset{(a)}{\leq} \left(C_\phi D_v R_\omega + D_v(C_\phi R_\omega + \rho_{\max}(r_{\max} + C_v + \gamma C_v)) + \gamma \rho_{\max} C_\phi^2\right) \|\omega_1 - \omega_2\| \\
&\triangleq L_g \|\omega_1 - \omega_2\|,
\end{aligned}
\tag{67}
$$

where $L_g = D_v(2C_\phi R_\omega + \rho_{\max}(r_{\max} + C_v + \gamma C_v)) + \gamma \rho_{\max} C_\phi^2$, and (a) is from the Lipschitz continuous of $h_t(\theta, \cdot)$, i.e.,

$$
\|h_t(\theta, \omega_1) - h_t(\theta, \omega_2)\| \leq \rho_{\max}(r_{\max} + (1+\gamma)C_v)D_v\|\omega_1 - \omega_2\| + 2C_\phi D_v R_\omega \|\omega_1 - \omega_2\|.
\tag{68}
$$

We note that to show (67), we use the bound on ω_t, which is guaranteed by the projection step. And this is the only step in our proof where the projection is used.

Then it follows that

$$
\begin{aligned}
\|\theta_{t+1} - \theta_t\| &= \alpha \|G_{t+1}(\theta_t, \omega_t)\| \\
&\leq \alpha \|G_{t+1}(\theta_t, \omega_t) - G_{t+1}(\theta_t, \omega(\theta_t)) + G_{t+1}(\theta_t, \omega(\theta_t))\| \\
&\leq \alpha L_g \|z_t\| + \alpha \|G_{t+1}(\theta_t, \omega(\theta_t))\| \\
&\leq \alpha L_g \|z_t\| + \alpha C_g,
\end{aligned}
\tag{69}
$$

where $C_g = \rho_{\max} C_\phi(r_{\max} + (1+\gamma)C_v) + \gamma \rho_{\max} R_\omega C_\phi^2 + D_v R_\omega(R_\omega C_\phi + \rho_{\max}(r_{\max} + C_v + \gamma C_v))$, and the last step in (69) can be shown as follows

$$
\begin{aligned}
&\|G_{t+1}(\theta_t, \omega(\theta_t))\| \\
&= \|\rho_t \delta_t(\theta) \phi_\theta(s_t) - \gamma \rho_t \phi_\theta(s_t') \phi_\theta(s_t)^\top \omega(\theta) - h_t(\theta, \omega(\theta))\| \\
&\leq \rho_{\max} C_\phi(r_{\max} + (1+\gamma)C_v) + \gamma \rho_{\max} R_\omega C_\phi^2 + D_v R_\omega(R_\omega C_\phi + \rho_{\max}(r_{\max} + C_v + \gamma C_v)).
\end{aligned}
\tag{70}
$$

Using (21) and (69), it follows that

$$
\|\omega(\theta_t) - \omega(\theta_{t+1})\| \leq L_\omega \|\theta_{t+1} - \theta_t\| \leq \alpha L_\omega L_g \|z_t\| + \alpha C_g L_\omega,
\tag{71}
$$

and

$$
\|\omega(\theta_t) - \omega(\theta_{t+1})\|^2 \leq 2\alpha^2 L_\omega^2 L_g^2 \|z_t\|^2 + 2\alpha^2 C_g^2 L_\omega^2.
\tag{72}
$$

This completes the proof for term (b).

B.1.3 Bound on Term (c)

In this section we provide the detailed proof of the bound on term (c) in (57).

Consider the inner product $\langle z_t, \omega(\theta_t) - \omega(\theta_{t+1}) \rangle$. By the Mean-Value Theorem, it follows that

$$
\langle z_t, \omega(\theta_t) \rangle - \langle z_t, \omega(\theta_{t+1}) \rangle = \langle z_t, \omega(\theta_t) - \omega(\theta_{t+1}) \rangle = \langle z_t, \nabla\omega(\hat{\theta}_t)(\theta_t - \theta_{t+1}) \rangle,
\tag{73}
$$

where $\hat{\theta}_t = c\theta_t + (1-c)\theta_{t+1}$ for some $c \in [0, 1]$. Thus, it follows that

$$
\begin{aligned}
&\langle z_t, \omega(\theta_t) - \omega(\theta_{t+1}) \rangle \\
&= \langle z_t, \nabla\omega(\hat{\theta}_t)(\theta_t - \theta_{t+1}) \rangle \\
&= -\alpha \langle z_t, \nabla\omega(\hat{\theta}_t) G_{t+1}(\theta_t, \omega_t) \rangle \\
&= -\alpha \left\langle z_t, \nabla\omega(\hat{\theta}_t) \left(G_{t+1}(\theta_t, \omega_t) - G_{t+1}(\theta_t, \omega(\theta_t)) + G_{t+1}(\theta_t, \omega(\theta_t)) + \frac{\nabla J(\theta_t)}{2} \right) \right\rangle
\end{aligned}
$$

21

$$+ \alpha \left\langle z_t, \nabla\omega(\hat{\theta}_t)\frac{\nabla J(\theta_t)}{2} \right\rangle$$

$$= -\alpha \left\langle z_t, \nabla\omega(\hat{\theta}_t)\left(G_{t+1}(\theta_t, \omega_t) - G_{t+1}(\theta_t, \omega(\theta_t))\right) \right\rangle + \alpha \left\langle z_t, \nabla\omega(\hat{\theta}_t)\frac{\nabla J(\theta_t)}{2} \right\rangle$$

$$- \alpha \left\langle z_t, \nabla\omega(\hat{\theta}_t)\left(G_{t+1}(\theta_t, \omega(\theta_t)) + \frac{\nabla J(\theta_t)}{2}\right) \right\rangle$$

$$\overset{(a)}{\leq} \alpha L_\omega L_g \|z_t\|^2 + \alpha L_\omega \|z_t\| \left\|\frac{\nabla J(\theta_t)}{2}\right\| - \alpha \left\langle z_t, \nabla\omega(\theta_t)\left(G_{t+1}(\theta_t, \omega(\theta_t)) + \frac{\nabla J(\theta_t)}{2}\right) \right\rangle$$

$$+ \alpha \left\langle z_t, (\nabla\omega(\theta_t) - \nabla\omega(\hat{\theta}_t))\left(G_{t+1}(\theta_t, \omega(\theta_t)) + \frac{\nabla J(\theta_t)}{2}\right) \right\rangle$$

$$\leq \alpha L_\omega L_g \|z_t\|^2 + \frac{1}{2}\alpha L_\omega \|z_t\|^2 + \frac{\alpha L_\omega}{8}\|\nabla J(\theta_t)\|^2 + \alpha\eta_G(\theta_t, z_t, O_t)$$

$$+ \alpha \|z_t\|\|\nabla\omega(\theta_t) - \nabla\omega(\hat{\theta}_t)\| \left\|G_{t+1}(\theta_t, \omega(\theta_t)) + \frac{\nabla J(\theta_t)}{2}\right\|$$

$$\overset{(b)}{\leq} \alpha L_\omega L_g \|z_t\|^2 + \frac{1}{2}\alpha L_\omega \|z_t\|^2 + \frac{\alpha L_\omega}{8}\|\nabla J(\theta_t)\|^2 + \alpha\eta_G(\theta_t, z_t, O_t) + 2\alpha C_g D_\omega \|z_t\|\|\theta_t - \hat{\theta}_t\|$$

$$\overset{(c)}{\leq} \alpha L_\omega L_g \|z_t\|^2 + \frac{1}{2}\alpha L_\omega \|z_t\|^2 + \frac{\alpha L_\omega}{8}\|\nabla J(\theta_t)\|^2 + \alpha\eta_G(\theta_t, z_t, O_t)$$

$$+ 2\alpha C_g D_\omega \|z_t\|\|\theta_t - \theta_{t+1}\|$$

$$\overset{(d)}{\leq} \alpha L_\omega L_g \|z_t\|^2 + \frac{1}{2}\alpha L_\omega \|z_t\|^2 + \frac{\alpha L_\omega}{8}\|\nabla J(\theta_t)\|^2 + \alpha\eta_G(\theta_t, z_t, O_t)$$

$$+ 2\alpha C_g D_\omega \|z_t\|(\alpha L_g \|z_t\| + \alpha C_g)$$

$$\overset{(e)}{\leq} \alpha L_\omega L_g \|z_t\|^2 + \frac{1}{2}\alpha L_\omega \|z_t\|^2 + \frac{\alpha L_\omega}{8}\|\nabla J(\theta_t)\|^2 + \alpha\eta_G(\theta_t, z_t, O_t)$$

$$+ 2\alpha^2 C_g D_\omega \left(2L_g\|z_t\|^2 + \frac{C_g^2}{4L_g}\right)$$

$$\leq (\alpha L_\omega L_g + \frac{1}{2}\alpha L_\omega + 4\alpha^2 C_g L_g D_\omega)\|z_t\|^2 + \frac{\alpha L_\omega}{8}\|\nabla J(\theta_t)\|^2 + \frac{\alpha^2 C_g^3 D_\omega}{2L_g} + \alpha\eta_G(\theta_t, z_t, O_t),$$
$$\tag{74}$$

where $\eta_G(\theta_t, z_t, O_t) = -\left\langle z_t, \nabla\omega(\theta_t)\left(G_{t+1}(\theta_t, \omega(\theta_t)) + \frac{\nabla J(\theta_t)}{2}\right)\right\rangle$, (a) is from the Lipschitz continuity of $G_{t+1}(\theta, \cdot)$ proved in (67), (b) is from the Lipschitz continuity of $\nabla\omega(\theta)$, which is shown in (30), (c) is from the fact that $\|\theta_t - \hat{\theta}_t\| = (1-c)\|\theta_t - \theta_{t+1}\| \leq \|\theta_t - \theta_{t+1}\|$, (d) is from the bound of $\|\theta_t - \theta_{t+1}\|$ in (69), and (e) is from the fact that $C_g\|z_t\| \leq L_g\|z_t\|^2 + \frac{C_g^2}{4L_g}$.

This completes the proof.

B.1.4 Bound on Term (d)

In this section we provide the detailed proof of the bound on term (d) in (58).

It can be shown that

$$-2\beta\langle z_t, A_{\theta_t}(s_t)z_t\rangle = -2\beta\langle z_t, A_{\theta_t}z_t\rangle + 2\beta\langle z_t, (A_{\theta_t} - A_{\theta_t}(s_t))z_t\rangle$$
$$\leq -2\beta\lambda_v\|z_t\|^2 + 2\beta\langle z_t, (A_{\theta_t} - A_{\theta_t}(s_t))z_t\rangle, \tag{75}$$

where the inequality is due to the fact that $\langle z_t, A_{\theta_t}z_t\rangle = z_t^\top A_{\theta_t} z_t \geq \lambda_L(A_{\theta_t})\|z_t\|^2 \geq \lambda_v\|z_t\|^2$.

B.2 Proof under the i.i.d. Setting

In this section we provide the proof of Theorem 1 under the i.i.d. setting.

From Lemma 3, we know that the objective function $J(\theta)$ is L_J-smooth, hence it follows that

$$J(\theta_{t+1}) \leq J(\theta_t) + \langle \nabla J(\theta_t), \theta_{t+1} - \theta_t \rangle + \frac{L_J}{2}\|\theta_{t+1} - \theta_t\|^2$$

$$= J(\theta_t) + \alpha \langle \nabla J(\theta_t), G_{t+1}(\theta_t, \omega_t) \rangle + \frac{L_J}{2}\alpha^2\|G_{t+1}(\theta_t, \omega_t)\|^2$$

$$= J(\theta_t) - \alpha \left\langle \nabla J(\theta_t), -G_{t+1}(\theta_t, \omega_t) - \frac{\nabla J(\theta_t)}{2} + G_{t+1}(\theta_t, \omega(\theta_t)) - G_{t+1}(\theta_t, \omega(\theta_t)) \right\rangle$$

$$\quad - \frac{\alpha}{2}\|\nabla J(\theta_t)\|^2 + \frac{L_J}{2}\alpha^2\|G_{t+1}(\theta_t, \omega_t)\|^2$$

$$= J(\theta_t) - \alpha \langle \nabla J(\theta_t), -G_{t+1}(\theta_t, \omega_t) + G_{t+1}(\theta_t, \omega(\theta_t)) \rangle$$

$$\quad + \alpha \left\langle \nabla J(\theta_t), \frac{\nabla J(\theta_t)}{2} + G_{t+1}(\theta_t, \omega(\theta_t)) \right\rangle - \frac{\alpha}{2}\|\nabla J(\theta_t)\|^2 + \frac{L_J}{2}\alpha^2\|G_{t+1}(\theta_t, \omega_t)\|^2$$

$$\overset{(a)}{\leq} J(\theta_t) + \alpha L_g\|\nabla J(\theta_t)\|\|\omega(\theta_t) - \omega_t\| + \alpha \left\langle \nabla J(\theta_t), \frac{\nabla J(\theta_t)}{2} + G_{t+1}(\theta_t, \omega(\theta_t)) \right\rangle$$

$$\quad - \frac{\alpha}{2}\|\nabla J(\theta_t)\|^2 + \frac{L_J}{2}\alpha^2\|G_{t+1}(\theta_t, \omega_t)\|^2$$

$$\overset{(b)}{\leq} J(\theta_t) + \alpha L_g\|\nabla J(\theta_t)\|\|z_t\| + \alpha \left\langle \nabla J(\theta_t), \frac{\nabla J(\theta_t)}{2} + G_{t+1}(\theta_t, \omega(\theta_t)) \right\rangle$$

$$\quad - \frac{\alpha}{2}\|\nabla J(\theta_t)\|^2 + \frac{L_J}{2}\alpha^2 \left(2L_g^2\|z_t\|^2 + 2C_g^2\right), \tag{76}$$

where (a) is from (67) and (b) is because $\|\theta_{t+1} - \theta_t\| = \alpha\|G_{t+1}(\theta_t, \omega_t)\| \leq \alpha L_g\|z_t\| + \alpha C_g$, whose detailed proof is provided in (69). Thus by re-arranging the terms, taking expectation and summing up w.r.t. t from 0 to $T-1$, it follows that

$$\frac{\alpha}{2}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla J(\theta_t)\|^2]$$

$$\leq -\mathbb{E}[J(\theta_T)] + J(\theta_0) + \alpha L_g\sqrt{\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla J(\theta_t)\|^2]}\sqrt{\sum_{t=0}^{T-1}\mathbb{E}[\|z_t\|^2]} + \alpha^2 L_J L_g^2\sum_{t=0}^{T-1}\mathbb{E}[\|z_t\|^2]$$

$$+ \alpha^2 C_g^2 L_J T, \tag{77}$$

which is due to the fact that under the i.i.d. setting,

$$\mathbb{E}\left[\left\langle \nabla J(\theta_t), \frac{\nabla J(\theta_t)}{2} + G_{t+1}(\theta_t, \omega(\theta_t)) \right\rangle\right]$$

$$= \mathbb{E}\left[\left\langle \nabla J(\theta_t), \mathbb{E}\left[\frac{\nabla J(\theta_t)}{2} + G_{t+1}(\theta_t, \omega(\theta_t))\Big|\mathcal{F}_t\right] \right\rangle\right] = 0, \tag{78}$$

and the Cauchy's inequality

$$\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla J(\theta_t)\|\|z_t\|] \leq \sqrt{\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla J(\theta_t)\|^2]}\sqrt{\sum_{t=0}^{T-1}\mathbb{E}[\|z_t\|^2]}. \tag{79}$$

Thus dividing both sides by $\frac{\alpha T}{2}$, it follows that

$$\frac{\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}$$

$$\leq \frac{2(J(\theta_0) - J^*)}{T\alpha} + 2L_g\sqrt{\frac{\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}}\sqrt{\frac{\sum_{t=0}^{T-1}\mathbb{E}[\|z_t\|^2]}{T}}$$

$$+ 2\alpha L_J L_g^2\frac{\sum_{t=0}^{T-1}\mathbb{E}[\|z_t\|^2]}{T} + 2\alpha C_g^2 L_J, \tag{80}$$

23

where $J^* \triangleq \min_\theta J(\theta)$.

Recall the tracking error in (65):

$$\frac{\sum_{t=0}^{T-1} \mathbb{E}\left[\|z_t\|^2\right]}{T}$$

$$\leq \frac{\|z_0\|^2}{Tq} + \frac{\alpha L_\omega}{4q} \frac{\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right]}{T} + \frac{1}{q}\left(4\beta^2 C_{g1} + 4\alpha^2 C_g^2 L_\omega^2 + \frac{\alpha^2 C_g^3 D_\omega}{L_g}\right). \qquad (81)$$

We then plug in the tracking error and obtain that

$$\frac{\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}$$

$$\leq \frac{2(J(\theta_0) - J^*)}{T\alpha} + 2\alpha C_g^2 L_J + 2L_g \sqrt{\frac{\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}}$$

$$\times \sqrt{\frac{\|z_0\|^2}{Tq} + \alpha L_\omega \frac{1}{4q} \frac{\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right]}{T} + \frac{1}{q}\left(4\beta^2 C_{g1} + 4\alpha^2 C_g^2 L_\omega^2 + \frac{\alpha^2 C_g^3 D_\omega}{L_g}\right)}$$

$$+ 2\alpha L_J L_g^2 \left(\frac{\|z_0\|^2}{Tq} + \alpha L_\omega \frac{1}{4q} \frac{\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right]}{T}\right.$$

$$\left. + \frac{1}{q}\left(4\beta^2 C_{g1} + 4\alpha^2 C_g^2 L_\omega^2 + \frac{\alpha^2 C_g^3 D_\omega}{L_g}\right)\right)$$

$$\leq \frac{2(J(\theta_0) - J^*)}{T\alpha} + 2\alpha C_g^2 L_J + L_g \sqrt{\frac{\alpha L_\omega}{q} \frac{\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right]}{T}}$$

$$+ 2L_g \sqrt{\frac{\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}} \sqrt{\frac{\|z_0\|^2}{Tq} + \frac{1}{q}\left(4\beta^2 C_{g1} + 4\alpha^2 C_g^2 L_\omega^2 + \frac{\alpha^2 C_g^3 D_\omega}{L_g}\right)}$$

$$+ 2\alpha L_J L_g^2 \left(\frac{\|z_0\|^2}{Tq} + \alpha L_\omega \frac{1}{4q} \frac{\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right]}{T}\right.$$

$$\left. + \frac{1}{q}\left(4\beta^2 C_{g1} + 4\alpha^2 C_g^2 L_\omega^2 + \frac{\alpha^2 C_g^3 D_\omega}{L_g}\right)\right), \qquad (82)$$

where the last step is from the fact that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for any $x, y \geq 0$. Re-arranging the terms, it follows that

$$\left(1 - L_g \sqrt{\frac{\alpha L_\omega}{q}} - \frac{\alpha^2 L_J L_g^2 L_\omega}{2q}\right) \frac{\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}$$

$$\leq \frac{2(J(\theta_0) - J^*)}{T\alpha} + 2\alpha C_g^2 L_J + 2\alpha L_J L_g^2 \left(\frac{\|z_0\|^2}{Tq} + \frac{1}{q}\left(4\beta^2 C_{g1} + 4\alpha^2 C_g^2 L_\omega^2 + \frac{\alpha^2 C_g^3 D_\omega}{L_g}\right)\right)$$

$$+ 2L_g \sqrt{\frac{\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}} \sqrt{\frac{\|z_0\|^2}{Tq} + \frac{1}{q}\left(4\beta^2 C_{g1} + 4\alpha^2 C_g^2 L_\omega^2 + \frac{\alpha^2 C_g^3 D_\omega}{L_g}\right)}. \qquad (83)$$

Note that $\left(L_g \sqrt{\frac{\alpha L_\omega}{q}} + \frac{\alpha^2 L_J L_g^2 L_\omega}{2q}\right) = \mathcal{O}\left(\sqrt{\frac{\alpha}{\beta}} + \frac{\alpha^2}{\beta}\right)$, hence we can choose α and β such that $\left(1 - L_g \sqrt{\frac{\alpha L_\omega}{q}} - \frac{\alpha^2 L_J L_g^2 L_\omega}{2q}\right) \geq \frac{1}{2}$. Thus (83) implies that

$$\frac{\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}$$

$$\leq \frac{4(J(\theta_0) - J^*)}{T\alpha} + 4\alpha C_g^2 L_J + 4\alpha L_J L_g^2 \left(\frac{\|z_0\|^2}{Tq} + \frac{1}{q}\left(4\beta^2 C_{g1} + 4\alpha^2 C_g^2 L_\omega^2 + \frac{\alpha^2 C_g^3 D_\omega}{L_g}\right)\right)$$

$$+ 4L_g \sqrt{\frac{\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}} \sqrt{\frac{\|z_0\|^2}{Tq} + \frac{1}{q}\left(4\beta^2 C_{g1} + 4\alpha^2 C_g^2 L_\omega^2 + \frac{\alpha^2 C_g^3 D_\omega}{L_g}\right)}. \quad (84)$$

Denote $U = \frac{4(J(\theta_0) - J^*)}{T\alpha} + 4\alpha C_g^2 L_J + 4\alpha L_J L_g^2 \left(\frac{\|z_0\|^2}{Tq} + \frac{1}{q}\left(4\beta^2 C_{g1} + 4\alpha^2 C_g^2 L_\omega^2 + \frac{\alpha^2 C_g^3 D_\omega}{L_g}\right)\right)$,

and $V = 4L_g \sqrt{\frac{\|z_0\|^2}{Tq} + \frac{1}{q}\left(4\beta^2 C_{g1} + 4\alpha^2 C_g^2 L_\omega^2 + \frac{\alpha^2 C_g^3 D_\omega}{L_g}\right)}$. Then it follows that

$$\frac{\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T} \leq V \sqrt{\frac{\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}} + U, \quad (85)$$

which further implies that

$$\frac{\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}$$

$$\leq V^2 + 2U$$

$$= 16L_g^2 \left(\frac{\|z_0\|^2}{Tq} + \frac{1}{q}\left(4\beta^2 C_{g1} + 4\alpha^2 C_g^2 L_\omega^2 + \frac{\alpha^2 C_g^3 D_\omega}{L_g}\right)\right) + \frac{8(J(\theta_0) - J^*)}{T\alpha}$$

$$\qquad + 8\alpha C_g^2 L_J + 8\alpha L_J L_g^2 \left(\frac{\|z_0\|^2}{Tq} + \frac{1}{q}\left(4\beta^2 C_{g1} + 4\alpha^2 C_g^2 L_\omega^2 + \frac{\alpha^2 C_g^3 D_\omega}{L_g}\right)\right)$$

$$= (16L_g^2 + 8\alpha L_J L_g^2)\left(\frac{\|z_0\|^2}{Tq} + \frac{1}{q}\left(4\beta^2 C_{g1} + 4\alpha^2 C_g^2 L_\omega^2 + \frac{\alpha^2 C_g^3 D_\omega}{L_g}\right)\right)$$

$$\qquad + \frac{8(J(\theta_0) - J^*)}{T\alpha} + 8\alpha C_g^2 L_J$$

$$= \mathcal{O}\left(\frac{1}{T\beta} + \beta + \frac{1}{T\alpha}\right)$$

$$= \mathcal{O}\left(\frac{1}{T^{1-a}} + \frac{1}{T^b} + \frac{1}{T^{1-b}}\right). \quad (86)$$

This completes the proof.

B.3 Choice of Step-sizes

As the proof is complicated and we have made several assumptions on the step-sizes, in this section we summarize all the assumptions we made on the step-sizes. This would help the readers to have a more clear understanding of the choice of α and β.

In the proof under the i.i.d. setting, we made two assumptions on step-sizes. In (59), we assume
$$q = 2\beta\lambda_v - 4\beta^2 C_\phi^2 - 4\alpha^2 L_\omega^2 L_g^2 - 2\alpha L_w L_g - \alpha L_w - 8\alpha^2 C_g L_g D_\omega > 0; \quad (87)$$
And in (83), we moreover assume

$$\left(1 - L_g \sqrt{\frac{\alpha L_\omega}{q}} - \frac{\alpha^2 L_J L_g^2 L_\omega}{2q}\right) \geq \frac{1}{2}. \quad (88)$$

Note that the first one can be satisfied if $\beta \leq \min\left\{1, \frac{\lambda_v}{4C_\phi^2}\right\}$ and $\frac{\alpha}{\beta} \leq \frac{\lambda_v}{4L_\omega^2 L_g^2 + 2L_w L_g + L_w + 8C_g L_g D_\omega}$. As for assumption (88), we only need to find α and β such that

$$L_g \sqrt{\frac{\alpha L_\omega}{q}} \leq \frac{1}{4},$$

$$\frac{\alpha^2 L_J L_g^2 L_\omega}{2q} \leq \frac{1}{4}. \quad (89)$$

Note that these two conditions are satisfied if condition (87) is satisfied.

Hence to meet all the requirements on the step-sizes, we can set $\beta \leq \min\left\{1, \frac{\lambda_v}{4C_\phi^2}\right\}$ and $\frac{\alpha}{\beta} \leq \min\left\{1, \frac{\lambda_v}{4L_\omega^2 L_g^2 + 2L_w L_g + L_w + 8C_g L_g D_\omega}\right\}$.

C Non-asymptotic Analysis under the Markovian Setting

In this section we provide the proof of Theorem 1 under that Markovian setting. In Appendix C.1 we develop the finite-time analysis of the tracking error and in Appendix C.2 we prove Theorem 1.

C.1 Tracking Error Analysis under the Markovian Setting

We first define the mixing time $\tau_\beta = \inf\{t : m\kappa^t \leq \beta\}$ (Assumption 4). It can be shown that for any bounded function $\|f(O_t)\| \leq C_f$, for any $t \geq \tau_\beta$, $\|\mathbb{E}[f(O_t)] - \mathbb{E}_{O\sim\mu^{\pi_b}}[f(O)]\| \leq C_f\beta$ and $\tau_\beta = \mathcal{O}(-\log\beta)$. We note that $\beta\tau_\beta \to 0$ as $\beta \to 0$, and we assume that $\beta\tau_\beta C_\phi^2 \leq \frac{1}{4}$.

From (53), the update of the tracking error z_t can be written as

$$z_{t+1} = z_t + \beta(-A_{\theta_t}(s_t)z_t + b_t(\theta_t)) + \omega(\theta_t) - \omega(\theta_{t+1}), \tag{90}$$

where $A_{\theta_t}(s_t) = \phi_{\theta_t}(s_t)\phi_{\theta_t}(s_t)^\top$ and $b_t(\theta_t) = -A_{\theta_t}(s_t)\omega(\theta_t) + \rho_t\delta_t(\theta_t)\phi_{\theta_t}(s_t)$. Note that for any $\theta \in \mathbb{R}^N$ and any sample $O_t = (s_t, a_t, r_t, s_{t+1}) \in \mathcal{S} \times \mathcal{A} \times \mathbb{R} \times \mathcal{S}$, $\|b_t(\theta_t)\| \leq C_\phi^2 R_\omega + \rho_{\max}C_\phi(r_{\max} + C_v + \gamma C_v) \triangleq b_{\max}$.

Then it can be shown that

$$\begin{aligned}
&\mathbb{E}\left[\|z_{t+1}\|^2 - \|z_t\|^2\right] \\
&= \mathbb{E}\left[2z_t^\top(z_{t+1} - z_t) + \|z_{t+1} - z_t\|^2\right] \\
&= \mathbb{E}\left[2z_t^\top(z_{t+1} - z_t + \beta A_{\theta_t}z_t)\right] + \mathbb{E}\left[\|z_{t+1} - z_t\|^2\right] + \beta\mathbb{E}\left[2z_t^\top(-A_{\theta_t})z_t\right] \\
&\leq \underbrace{\mathbb{E}\left[\|z_{t+1} - z_t\|^2\right]}_{(a)} + \underbrace{\mathbb{E}\left[2z_t^\top(z_{t+1} - z_t + \beta A_{\theta_t}z_t)\right]}_{(b)} - 2\beta\lambda_v\mathbb{E}\left[\|z_t\|^2\right],
\end{aligned} \tag{91}$$

where the last inequality is due to the fact that $\lambda_L(A_{\theta_t}) \geq \lambda_v$. We first provide the bounds on terms (a) and (b) as follows, and their detailed proof can be found in Appendices C.1.1 and C.1.2.

Term (a) can be bounded as follows:

For any $t \geq 0$, we have that

$$\|z_{t+1} - z_t\|^2 \leq 2\beta^2 C_\phi^4\|z_t\|^2 + 2\beta^2(b_{\max} + L_\omega C_g)^2. \tag{92}$$

Term (b) can be bounded as follows:

For any $t \geq \tau_\beta$, we have that

$$\begin{aligned}
&\left|\mathbb{E}\left[z_t^\top\left(-A_{\theta_t}z_t - \frac{1}{\beta}(z_{t+1} - z_t)\right)\right]\right| \\
&\leq (R_1 + R_3 + P_1 + P_2 + P_3)\mathbb{E}\left[\|z_t\|^2\right] + (Q_1 + Q_2 + Q_3 + P_1 + P_2 + P_3) \\
&\quad + \frac{\alpha}{8\beta}L_\omega\mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right],
\end{aligned} \tag{93}$$

where the definition of P_i, Q_i and R_i, $i = 1, 2, 3$, can be found in (114), (117) and (120).

From (91), it can be shown that for any $t \geq \tau_\beta$,

$$\begin{aligned}
&\mathbb{E}\left[\|z_{t+1}\|^2 - \|z_t\|^2\right] \\
&\leq 2\beta(R_1 + R_3 + P_1 + P_2 + P_3)\mathbb{E}\left[\|z_t\|^2\right] + 2\beta(Q_1 + Q_2 + Q_3 + P_1 + P_2 + P_3) \\
&\quad + \frac{\alpha}{4}L_\omega\mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right] + 2\beta^2 C_\phi^4\mathbb{E}\left[\|z_t\|^2\right] + 2\beta^2(b_{\max} + L_\omega C_g)^2 - 2\beta\lambda_v\mathbb{E}\left[\|z_t\|^2\right]. \tag{94}
\end{aligned}$$

Thus by re-arranging the terms we obtain that

$$\begin{aligned}
&\mathbb{E}\left[\|z_{t+1}\|^2\right] \\
&\leq (1 - 2\beta\lambda_v + 2\beta(R_1 + R_3 + P_1 + P_2 + P_3) + 2\beta^2 C_\phi^4)\mathbb{E}\left[\|z_t\|^2\right] + \frac{\alpha}{4}L_\omega\mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right]
\end{aligned}$$

$$+ 2\beta(Q_1 + Q_2 + Q_3 + P_1 + P_2 + P_3) + 2\beta^2(b_{\max} + L_\omega C_g)^2$$

$$\triangleq (1-q)\mathbb{E}\left[\|z_t\|^2\right] + \frac{\alpha}{4}L_\omega \mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right] + p, \tag{95}$$

where $q = 2\beta\lambda_v - 2\beta(R_1 + R_3 + P_1 + P_2 + P_3) - 2\beta^2 C_\phi^4 = \mathcal{O}(\beta)$ and $p = 2\beta(Q_1 + Q_2 + Q_3 + P_1 + P_2 + P_3) + 2\beta^2(b_{\max} + L_\omega C_g)^2 = \mathcal{O}(\beta^2\tau_\beta)$. Then by recursively using the previous inequality, it follows that for any $t \geq \tau_\beta$,

$$\mathbb{E}[\|z_t\|^2] \leq (1-q)^{t-\tau_\beta}\mathbb{E}\left[\|z_{\tau_\beta}\|^2\right] + \frac{\alpha L_\omega}{4}\sum_{j=0}^{t}(1-q)^{t-j}\mathbb{E}[\|\nabla J(\theta_j)\|^2] + \frac{p}{q}, \tag{96}$$

and hence

$$\frac{\sum_{t=0}^{T-1}\mathbb{E}[\|z_t\|^2]}{T}$$

$$= \frac{\sum_{t=\tau_\beta}^{T-1}\mathbb{E}[\|z_t\|^2]}{T} + \frac{\sum_{t=0}^{\tau_\beta-1}\mathbb{E}[\|z_t\|^2]}{T}$$

$$\leq \frac{\mathbb{E}\left[\|z_{\tau_\beta}\|^2\right]}{Tq} + \frac{\tau_\beta\left(2\|z_0\| + 2\beta\tau_\beta(b_{\max} + L_\omega C_g)\right)^2}{T} + \frac{\alpha L_\omega \sum_{t=0}^{T-1}\mathbb{E}[\|\nabla J(\theta_t)\|^2]}{4q} + \frac{p}{q}$$

$$\leq (2\|z_0\| + 2\beta\tau_\beta(b_{\max} + L_\omega C_g))^2\left(\frac{1}{Tq} + \frac{\tau_\beta}{T}\right) + \frac{\alpha L_\omega \sum_{t=0}^{T-1}\mathbb{E}[\|\nabla J(\theta_t)\|^2]}{4q} + \frac{p}{q}$$

$$= \mathcal{O}\left(\frac{1}{T\beta} + \frac{\alpha}{\beta}\frac{\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T} + \beta\tau_\beta\right), \tag{97}$$

where the last step is because $q = \mathcal{O}(\beta)$ and $p = \mathcal{O}(\beta^2\tau_\beta)$.

C.1.1 Bound on Term (a)

In this section we provide the detailed proof of the bound on term (a) in (91).

We first note that from the update of z_t in (90), term $\|z_{t+1} - z_t\|$ can be bounded as follows

$$\begin{aligned}\|z_{t+1} - z_t\| &\leq \|\beta(-A_{\theta_t}(s_t)z_t + b_t(\theta_t))\| + \|\omega(\theta_t) - \omega(\theta_{t+1})\| \\ &\leq \beta C_\phi^2\|z_t\| + \beta b_{\max} + L_\omega\|\theta_t - \theta_{t+1}\| \\ &\overset{(a)}{\leq} \beta C_\phi^2\|z_t\| + \beta b_{\max} + \alpha L_\omega C_g \\ &\leq \beta C_\phi^2\|z_t\| + \beta(b_{\max} + L_\omega C_g),\end{aligned} \tag{98}$$

where (a) is due to the fact $\|G_{t+1}(\theta_t, \omega_t)\| \leq C_g$ for any $t \geq 0$, and where the last inequality is from the fact that $\alpha \leq \beta$. Hence term (a) can be bounded as follows

$$\|z_{t+1} - z_t\|^2 \leq 2\beta^2 C_\phi^4\|z_t\|^2 + 2\beta^2(b_{\max} + L_\omega C_g)^2. \tag{99}$$

This completes the proof.

C.1.2 Bound on Term (b)

In this section we provide the detailed proof of the bound on term (b) in (91).

From (98), it follows that

$$\begin{aligned}\|z_{t+1}\| &\leq (1 + \beta C_\phi^2)\|z_t\| + \beta b_{\max} + \alpha L_\omega C_g \\ &\leq (1 + \beta C_\phi^2)\|z_t\| + \beta(b_{\max} + L_\omega C_g).\end{aligned} \tag{100}$$

By applying (100) recursively, it follows that

$$\|z_t\| \leq (1 + \beta C_\phi^2)^t\|z_0\| + \beta(b_{\max} + L_\omega C_g)\frac{(1 + \beta C_\phi^2)^t - 1}{\beta C_\phi^2}$$

$$= (1 + \beta C_\phi^2)^t \|z_0\| + (b_{\max} + L_\omega C_g) \frac{(1 + \beta C_\phi^2)^t - 1}{C_\phi^2}. \tag{101}$$

We first show the following lemma which bounds the update $\|z_t - z_{t-\tau_\beta}\|$ by $\|z_t\|$.

Lemma 4. *For any $t \geq \tau_\beta$ and $t \geq j \geq t - \tau_\beta$, we have that*

$$\|z_j\| \leq 2\|z_{t-\tau_\beta}\| + 2\beta\tau_\beta(b_{\max} + L_\omega C_g); \tag{102}$$

$$\|z_t - z_{t-\tau_\beta}\| \leq 2\beta\tau_\beta C_\phi^2 \|z_{t-\tau_\beta}\| + 2\beta\tau_\beta(b_{\max} + L_\omega C_g), \tag{103}$$

$$\|z_t - z_{t-\tau_\beta}\| \leq 4\beta\tau_\beta C_\phi^2 \|z_t\| + 4\beta\tau_\beta(b_{\max} + L_\omega C_g). \tag{104}$$

Proof. From (100), it follows that

$$\|z_{t+1}\| \leq (1 + \beta C_\phi^2)\|z_t\| + \beta(b_{\max} + L_\omega C_g). \tag{105}$$

First note that $\beta C_\phi^2 \tau_\beta \leq \frac{1}{4}$ and hence $\beta C_\phi^2 \leq \frac{1}{4\tau_\beta} \leq \frac{\log 2}{\tau_\beta - 1}$. This implies that

$$(1 + \beta C_\phi^2)^{\tau_\beta} \leq 1 + 2\tau_\beta \beta C_\phi^2, \tag{106}$$

which is because $(1 + x)^k \leq 1 + 2kx$ for $x \leq \frac{\log 2}{k-1}$.

Applying inequality (105) recursively, it follows that

$$\|z_j\| \leq (1 + \beta C_\phi^2)^{j-t+\tau_\beta} \|z_{t-\tau_\beta}\| + (b_{\max} + L_\omega C_g) \frac{(1 + \beta C_\phi^2)^{\tau_\beta} - 1}{C_\phi^2}$$

$$\leq (1 + \beta C_\phi^2)^{\tau_\beta} \|z_{t-\tau_\beta}\| + (b_{\max} + L_\omega C_g) \frac{(1 + \beta C_\phi^2)^{\tau_\beta} - 1}{C_\phi^2}$$

$$\overset{(a)}{\leq} (1 + 2\tau_\beta \beta C_\phi^2)\|z_{t-\tau_\beta}\| + 2\beta\tau_\beta(b_{\max} + L_\omega C_g)$$

$$\overset{(b)}{\leq} 2\|z_{t-\tau_\beta}\| + 2\beta\tau_\beta(b_{\max} + L_\omega C_g), \tag{107}$$

where (a) is from (106), and (b) is from the fact that $\beta\tau_\beta C_\phi^2 \leq \frac{1}{4}$.

To prove (103) and (104), first note that

$$\|z_t - z_{t-\tau_\beta}\| \leq \sum_{j=t-\tau_\beta}^{t-1} \|z_{j+1} - z_j\|$$

$$\overset{(a)}{\leq} \sum_{j=t-\tau_\beta}^{t-1} \beta C_\phi^2 \|z_j\| + \beta\tau_\beta(b_{\max} + L_\omega C_g)$$

$$\overset{(b)}{\leq} \sum_{j=t-\tau_\beta}^{t-1} \beta C_\phi^2 (2\|z_{t-\tau_\beta}\| + 2\beta\tau_\beta(b_{\max} + L_\omega C_g)) + \beta\tau_\beta(b_{\max} + L_\omega C_g)$$

$$\leq \beta\tau_\beta C_\phi^2 (2\|z_{t-\tau_\beta}\| + 2\beta\tau_\beta(b_{\max} + L_\omega C_g)) + \beta\tau_\beta(b_{\max} + L_\omega C_g)$$

$$= 2\beta\tau_\beta C_\phi^2 \|z_{t-\tau_\beta}\| + (2\beta^2\tau_\beta^2 C_\phi^2 + \beta\tau_\beta)(b_{\max} + L_\omega C_g)$$

$$\overset{(c)}{\leq} 2\beta\tau_\beta C_\phi^2 \|z_{t-\tau_\beta}\| + 2\beta\tau_\beta(b_{\max} + L_\omega C_g), \tag{108}$$

where (a) is from (98), (b) is from (107) and (c) is due to the fact that $\beta\tau_\beta C_\phi^2 \leq \frac{1}{4}$. Moreover, it can be further shown that

$$\|z_t - z_{t-\tau_\beta}\| \leq 2\beta\tau_\beta C_\phi^2 (\|z_t\| + \|z_t - z_{t-\tau_\beta}\|) + 2\beta\tau_\beta(b_{\max} + L_\omega C_g)$$

$$\leq 2\beta\tau_\beta C_\phi^2 \|z_t\| + \frac{1}{2}\|z_t - z_{t-\tau_\beta}\| + 2\beta\tau_\beta(b_{\max} + L_\omega C_g), \tag{109}$$

where the last step is because $\beta\tau_\beta C_\phi^2 \leq \frac{1}{4}$. Hence

$$\|z_t - z_{t-\tau_\beta}\| \leq 4\beta\tau_\beta C_\phi^2 \|z_t\| + 4\beta\tau_\beta(b_{\max} + L_\omega C_g). \tag{110}$$

$\square$

The bound on term (b) in (91) is straightforward from the following lemma.

Lemma 5. *For any $t \geq \tau_\beta$, it follows that*

$$\left| \mathbb{E}\left[z_t^\top \left(-A_{\theta_t} z_t - \frac{1}{\beta}(z_{t+1} - z_t) \right) \right] \right|$$

$$\leq (R_1 + R_3 + P_1 + P_2 + P_3)\mathbb{E}\left[\|z_t\|^2 \right] + (Q_1 + Q_2 + Q_3 + P_1 + P_2 + P_3)$$

$$+ \frac{\alpha}{8\beta} L_\omega \mathbb{E}\left[\|\nabla J(\theta_t)\|^2 \right], \tag{111}$$

where the definition of P_i, Q_i and R_i, $i = 1, 2, 3$, can be found in (114), (117) and (120).

Proof. We only prove the case $t = \tau_\beta$ here. The proof for the general case with $t > \tau_\beta$ is similar, and thus is omitted here. First note that

$$\mathbb{E}\left[z_{\tau_\beta}^\top \left(-A_{\theta_{\tau_\beta}} z_{\tau_\beta} - \frac{1}{\beta}\left(z_{\tau_\beta+1} - z_{\tau_\beta} \right) \right) \right]$$

$$= \mathbb{E}\left[z_{\tau_\beta}^\top \left(-A_{\theta_{\tau_\beta}} + A_{\theta_{\tau_\beta}}\left(s_{\tau_\beta} \right) \right) z_{\tau_\beta} \right] - \mathbb{E}\left[z_{\tau_\beta}^\top b_{\tau_\beta} \right] - \mathbb{E}\left[z_{\tau_\beta}^\top \frac{\omega\left(\theta_{\tau_\beta}\right) - \omega\left(\theta_{\tau_\beta+1}\right)}{\beta} \right]. \tag{112}$$

We then bound the terms in (112) one by one. First, it can be shown that

$$\left| \mathbb{E}\left[z_{\tau_\beta}^\top \left(-A_{\theta_{\tau_\beta}} + A_{\theta_{\tau_\beta}}\left(s_{\tau_\beta} \right) \right) z_{\tau_\beta} \right] \right|$$

$$\leq \left| \mathbb{E}\left[z_0^\top \left(-A_{\theta_{\tau_\beta}} + A_{\theta_{\tau_\beta}}\left(s_{\tau_\beta} \right) \right) z_0 \right] \right| + \left| \mathbb{E}\left[\left(z_{\tau_\beta} - z_0 \right)^\top \left(-A_{\theta_{\tau_\beta}} + A_{\theta_{\tau_\beta}}\left(s_{\tau_\beta} \right) \right) \left(z_{\tau_\beta} - z_0 \right) \right] \right|$$

$$+ 2\left| \mathbb{E}\left[\left(z_{\tau_\beta} - z_0 \right)^\top \left(-A_{\theta_{\tau_\beta}} + A_{\theta_{\tau_\beta}}\left(s_{\tau_\beta} \right) \right) z_0 \right] \right|$$

$$\leq \|z_0\|^2 \left\| \mathbb{E}\left[-A_{\theta_{\tau_\beta}} + A_{\theta_{\tau_\beta}}\left(s_{\tau_\beta} \right) \right] \right\| + 2C_\phi^2 \mathbb{E}\left[\|z_{\tau_\beta} - z_0\|^2 \right] + 4\|z_0\| C_\phi^2 \mathbb{E}\left[\|z_{\tau_\beta} - z_0\| \right]$$

$$\leq \|z_0\|^2 \left\| \mathbb{E}\left[-A_{\theta_0} + A_{\theta_0}\left(s_{\tau_\beta} \right) \right] \right\| + \|z_0\|^2 \left\| \mathbb{E}\left[-A_{\theta_0} + A_{\theta_{\tau_\beta}} \right] \right\|$$

$$+ \|z_0\|^2 \left\| \mathbb{E}\left[-A_{\theta_{\tau_\beta}}\left(s_{\tau_\beta} \right) + A_{\theta_0}\left(s_{\tau_\beta} \right) \right] \right\| + 2C_\phi^2 \mathbb{E}\left[\|z_{\tau_\beta} - z_0\|^2 \right] + 4\|z_0\| C_\phi^2 \mathbb{E}\left[\|z_{\tau_\beta} - z_0\| \right]$$

$$\overset{(a)}{\leq} \left(\beta C_\phi^2 + 4C_\phi D_v C_g \alpha \tau_\beta \right) \|z_0\|^2 + 2C_\phi^2 \mathbb{E}\left[\|z_{\tau_\beta} - z_0\|^2 \right] + 4\|z_0\| C_\phi^2 \mathbb{E}\left[\|z_{\tau_\beta} - z_0\| \right], \tag{113}$$

where (a) is due to the facts that $\left\| \mathbb{E}\left[-A_{\theta_0} + A_{\theta_0}(s_{\tau_\beta}) \right] \right\| \leq C_\phi^2 \beta$ from the uniform ergodicity of the MDP, both A_θ and $A_\theta(s_{\tau_\beta})$ are Lipschitz with constant $2C_\phi D_v$, and $\|\theta_0 - \theta_{\tau_\beta}\| \leq \sum_{j=0}^{\tau_\beta-1} \|\theta_{j+1} - \theta_j\| \leq \alpha\tau_\beta C_g$.

We then plug in the results from Lemma 4, and hence we have that

$$\left| \mathbb{E}\left[z_{\tau_\beta}^\top \left(-A_{\theta_{\tau_\beta}} + A_{\theta_{\tau_\beta}}\left(s_{\tau_\beta} \right) \right) z_{\tau_\beta} \right] \right|$$

$$\leq \left(\beta C_\phi^2 + 4C_\phi D_v C_g \alpha \tau_\beta \right) \|z_0\|^2 + 2C_\phi^2 \mathbb{E}\left[\|z_{\tau_\beta} - z_0\|^2 \right] + 4\|z_0\| C_\phi^2 \mathbb{E}\left[\|z_{\tau_\beta} - z_0\| \right]$$

$$\overset{(a)}{\leq} \left(\beta C_\phi^2 + 4C_\phi D_v C_g \alpha \tau_\beta \right) \left(2(1 + 4\beta\tau_\beta C_\phi^2)^2 \mathbb{E}\left[\|z_{\tau_\beta}\|^2 \right] + 32\beta^2 \tau_\beta^2 (b_{\max} + L_\omega C_g)^2 \right)$$

$$+ 2C_\phi^2 \left(32\beta^2 \tau_\beta^2 C_\phi^4 \mathbb{E}\left[\|z_{\tau_\beta}\|^2 \right] + 32\beta^2 \tau_\beta^2 (b_{\max} + L_\omega C_g)^2 \right)$$

$$+ 4C_\phi^2 \left(4\beta\tau_\beta C_\phi^2 (1 + 4\beta\tau_\beta C_\phi^2)\mathbb{E}\left[\|z_{\tau_\beta}\|^2 \right] + 4\beta\tau_\beta (b_{\max} + L_\omega C_g)(1 + 8\beta\tau_\beta C_\phi^2)\mathbb{E}\left[\|z_{\tau_\beta}\| \right] \right)$$

$$+ 64C_\phi^2 \beta^2 \tau_\beta^2 (b_{\max} + L_\omega C_g)^2$$

$$\triangleq R_1 \mathbb{E}\left[\|z_{\tau_\beta}\|^2 \right] + P_1 \mathbb{E}\left[\|z_{\tau_\beta}\| \right] + Q_1, \tag{114}$$

where (a) is from (104) and the fact that

$$\|z_0\| \leq \left\| z_{\tau_\beta} - z_0 \right\| + \left\| z_{\tau_\beta} \right\| \leq (1 + 4\beta\tau_\beta C_\phi^2)\left\| z_{\tau_\beta} \right\| + 4\beta\tau_\beta (b_{\max} + L_\omega C_g); \tag{115}$$

and $R_1 = 2(1 + 4\beta\tau_\beta C_\phi^2)^2 \left(\beta C_\phi^2 + 4C_\phi D_v C_g \alpha\tau_\beta\right) + 64\beta^2\tau_\beta^2 C_\phi^6 + 16\beta\tau_\beta C_\phi^4(1 + 4\beta\tau_\beta C_\phi^2) = \mathcal{O}(\beta\tau_\beta)$, $P_1 = 16C_\phi^2\beta\tau_\beta(b_{\max} + L_\omega C_g)(1 + 8\beta\tau_\beta C_\phi^2) = \mathcal{O}(\beta\tau_\beta)$ and $Q_1 = \left(\beta C_\phi^2 + 4C_\phi D_v C_g \alpha\tau_\beta\right)32\beta^2\tau_\beta^2(b_{\max} + L_\omega C_g)^2 + 64C_\phi^2\beta^2\tau_\beta^2(b_{\max} + L_\omega C_g)^2 + 64C_\phi^2\beta^2\tau_\beta^2(b_{\max} + L_\omega C_g)^2 = \mathcal{O}(\beta^2\tau^2)$.

Similarly, the second term in (112) can be bounded as follows

$$
\begin{aligned}
\left|\mathbb{E}\left[z_{\tau_\beta}^\top b_{\tau_\beta}(\theta_{\tau_\beta})\right]\right| &\leq \left|\mathbb{E}\left[(z_{\tau_\beta} - z_0)^\top b_{\tau_\beta}(\theta_{\tau_\beta})\right]\right| + \left|\mathbb{E}\left[z_0^\top b_{\tau_\beta}(\theta_0)\right]\right| \\
&\quad + \|\mathbb{E}\left[z_0^\top(b_{\tau_\beta}(\theta_{\tau_\beta}) - b_{\tau_\beta}(\theta_0))\right]\| \\
&\leq b_{\max}\mathbb{E}\left[\|z_{\tau_\beta} - z_0\|\right] + \beta b_{\max}\|z_0\| + \alpha\tau_\beta C_g L_b\|z_0\|,
\end{aligned}
\tag{116}
$$

where $L_b = 2C_\phi D_v R_\omega + L_\omega C_\phi^2 + \rho_{\max}((1 + \gamma)C_\phi^2 + D_v(r_{\max} + (1 + \gamma)C_v))$ is the Lipschitz constant of $b_t(\theta)$. Again applying Lemma 4 implies that

$$
\begin{aligned}
&\left|\mathbb{E}\left[z_{\tau_\beta}^\top b_{\tau_\beta}(\theta_{\tau_\beta})\right]\right| \\
&\leq b_{\max}\mathbb{E}\left[\|z_{\tau_\beta} - z_0\|\right] + \beta b_{\max}\|z_0\| + \alpha\tau_\beta C_g L_b\|z_0\| \\
&\leq b_{\max}\left(4\beta\tau_\beta C_\phi^2 \mathbb{E}\left[\|z_{\tau_\beta}\|\right] + 4\beta\tau_\beta(b_{\max} + L_\omega C_g)\right) \\
&\quad + (\beta b_{\max} + \alpha\tau_\beta C_g L_b)\left((1 + 4\beta\tau_\beta C_\phi^2)\mathbb{E}\left[\|z_{\tau_\beta}\|\right] + 4\beta\tau_\beta(b_{\max} + L_\omega C_g)\right) \\
&\triangleq P_2\mathbb{E}\left[\|z_{\tau_\beta}\|\right] + Q_2,
\end{aligned}
\tag{117}
$$

where $P_2 = 4\beta\tau_\beta b_{\max}C_\phi^2 + (\beta b_{\max} + \alpha\tau_\beta C_g L_b)\left(1 + 4\beta\tau_\beta C_\phi^2\right) = \mathcal{O}(\beta\tau_\beta)$ and $Q_2 = 4\beta\tau_\beta(b_{\max} + L_\omega C_g)(b_{\max} + \beta b_{\max} + \alpha\tau_\beta C_g L_b) = \mathcal{O}(\beta\tau_\beta)$.

We then bound the last term in (112) as follows

$$
\begin{aligned}
&\left|\mathbb{E}\left[z_{\tau_\beta}^\top \frac{\omega(\theta_{\tau_\beta}) - \omega(\theta_{\tau_\beta+1})}{\beta}\right]\right| \\
&\stackrel{(a)}{=} \left|\frac{1}{\beta}\mathbb{E}[z_{\tau_\beta}^\top \nabla\omega(\hat\theta_{\tau_\beta})(\theta_{\tau_\beta+1} - \theta_{\tau_\beta})]\right| \\
&= \left|\frac{\alpha}{\beta}\mathbb{E}[z_{\tau_\beta}^\top \nabla\omega(\hat\theta_{\tau_\beta})G_{\tau_\beta+1}(\theta_{\tau_\beta}, \omega_{\tau_\beta})]\right| \\
&= \left|\frac{\alpha}{\beta}\mathbb{E}\left[z_{\tau_\beta}^\top \nabla\omega(\hat\theta_{\tau_\beta})\left(G_{\tau_\beta+1}(\theta_{\tau_\beta}, \omega_{\tau_\beta}) - G_{\tau_\beta+1}(\theta_{\tau_\beta}, \omega(\theta_{\tau_\beta})) + G_{\tau_\beta+1}(\theta_{\tau_\beta}, \omega(\theta_{\tau_\beta}))\right.\right.\right. \\
&\quad\left.\left.\left. + \frac{\nabla J(\theta_{\tau_\beta})}{2} - \frac{\nabla J(\theta_{\tau_\beta})}{2}\right)\right]\right| \\
&= \left|\frac{\alpha}{\beta}\mathbb{E}\left[z_{\tau_\beta}^\top \nabla\omega(\hat\theta_{\tau_\beta})(G_{\tau_\beta+1}(\theta_{\tau_\beta}, \omega_{\tau_\beta}) - G_{\tau_\beta+1}(\theta_{\tau_\beta}, \omega(\theta_{\tau_\beta})))\right]\right| \\
&\quad + \left|\frac{\alpha}{\beta}\mathbb{E}\left[z_{\tau_\beta}^\top \nabla\omega(\hat\theta_{\tau_\beta})\left(G_{\tau_\beta+1}(\theta_{\tau_\beta}, \omega(\theta_{\tau_\beta})) + \frac{\nabla J(\theta_{\tau_\beta})}{2}\right)\right]\right| \\
&\quad + \left|\frac{\alpha}{\beta}\mathbb{E}\left[z_{\tau_\beta}^\top \nabla\omega(\hat\theta_{\tau_\beta})\left(-\frac{\nabla J(\theta_{\tau_\beta})}{2}\right)\right]\right| \\
&\stackrel{(b)}{\leq} \frac{\alpha}{\beta}L_\omega L_g \mathbb{E}\left[\|z_{\tau_\beta}\|^2\right] + \frac{\alpha}{2\beta}L_\omega \mathbb{E}\left[\|z_{\tau_\beta}\|^2\right] + \frac{\alpha}{8\beta}L_\omega \mathbb{E}\left[\|\nabla J(\theta_{\tau_\beta})\|^2\right] \\
&\quad + \frac{\alpha}{\beta}\left|\mathbb{E}\left[z_{\tau_\beta}^\top \nabla\omega(\theta_{\tau_\beta})\left(G_{\tau_\beta+1}(\theta_{\tau_\beta}, \omega(\theta_{\tau_\beta})) + \frac{\nabla J(\theta_{\tau_\beta})}{2}\right)\right]\right| \\
&\quad + \frac{\alpha}{\beta}\left|\mathbb{E}\left[z_{\tau_\beta}^\top(\nabla\omega(\hat\theta_{\tau_\beta}) - \nabla\omega(\theta_{\tau_\beta}))\left(G_{\tau_\beta+1}(\theta_{\tau_\beta}, \omega(\theta_{\tau_\beta})) + \frac{\nabla J(\theta_{\tau_\beta})}{2}\right)\right]\right|
\end{aligned}
$$

$$
\begin{aligned}
\leq{}& \frac{\alpha}{\beta}L_\omega L_g \mathbb{E}\left[\|z_{\tau_\beta}\|^2\right] + \frac{\alpha}{2\beta}L_\omega \mathbb{E}\left[\|z_{\tau_\beta}\|^2\right] + \frac{\alpha}{8\beta}L_\omega \mathbb{E}\left[\|\nabla J(\theta_{\tau_\beta})\|^2\right] \\
&+ \frac{\alpha}{\beta}\left|\mathbb{E}\left[z_0^\top \nabla\omega(\theta_{\tau_\beta})\left(G_{\tau_\beta+1}(\theta_{\tau_\beta},\omega(\theta_{\tau_\beta})) + \frac{\nabla J(\theta_{\tau_\beta})}{2}\right)\right]\right| \\
&+ \frac{\alpha}{\beta}\left|\mathbb{E}\left[(z_{\tau_\beta} - z_0)^\top \nabla\omega(\theta_{\tau_\beta})\left(G_{\tau_\beta+1}(\theta_{\tau_\beta},\omega(\theta_{\tau_\beta})) + \frac{\nabla J(\theta_{\tau_\beta})}{2}\right)\right]\right| \\
&+ \frac{2\alpha}{\beta}C_g D_\omega \mathbb{E}\left[\|z_{\tau_\beta}\|\,\|\theta_{\tau_\beta} - \theta_{\tau_\beta+1}\|\right] \\
\leq{}& \frac{\alpha}{\beta}L_\omega L_g \mathbb{E}\left[\|z_{\tau_\beta}\|^2\right] + \frac{\alpha}{2\beta}L_\omega \mathbb{E}\left[\|z_{\tau_\beta}\|^2\right] + \frac{\alpha}{8\beta}L_\omega \mathbb{E}\left[\|\nabla J(\theta_{\tau_\beta})\|^2\right] \\
&+ \frac{\alpha}{\beta}\left|\mathbb{E}\left[z_0^\top \nabla\omega(\theta_0)\left(G_{\tau_\beta+1}(\theta_0,\omega(\theta_0)) + \frac{\nabla J(\theta_0)}{2}\right)\right]\right| \\
&+ \frac{\alpha}{\beta}\left|\mathbb{E}\left[z_0^\top\left(\nabla\omega(\theta_{\tau_\beta})\left(G_{\tau_\beta+1}(\theta_{\tau_\beta},\omega(\theta_{\tau_\beta})) + \frac{\nabla J(\theta_{\tau_\beta})}{2}\right)\right.\right.\right. \\
&\qquad\left.\left.\left. - \nabla\omega(\theta_0)\left(G_{\tau_\beta+1}(\theta_0,\omega(\theta_0)) + \frac{\nabla J(\theta_0)}{2}\right)\right)\right]\right| \\
&+ \frac{\alpha}{\beta}\left|\mathbb{E}\left[(z_{\tau_\beta} - z_0)^\top \nabla\omega(\theta_{\tau_\beta})\left(G_{\tau_\beta+1}(\theta_{\tau_\beta},\omega(\theta_{\tau_\beta}) + \frac{\nabla J(\theta_{\tau_\beta})}{2}\right)\right]\right| \\
&+ \frac{2\alpha}{\beta}C_g D_\omega \mathbb{E}\left[\|z_{\tau_\beta}\|\,\|\theta_{\tau_\beta} - \theta_{\tau_\beta+1}\|\right] \\
\leq{}& \frac{\alpha}{\beta}L_\omega L_g \mathbb{E}\left[\|z_{\tau_\beta}\|^2\right] + \frac{\alpha}{8\beta}L_\omega \mathbb{E}\left[\|\nabla J(\theta_{\tau_\beta})\|^2\right] \\
&+ \frac{\alpha}{\beta}\|z_0\|L_\omega\left\|\mathbb{E}\left[G_{\tau_\beta+1}(\theta_0,\omega(\theta_0)) + \frac{\nabla J(\theta_{\tau_\beta})}{2}\right]\right\| \\
&+ \frac{\alpha}{2\beta}L_\omega \mathbb{E}\left[\|z_{\tau_\beta}\|^2\right] + \frac{\alpha}{\beta}\|z_0\|L_k\mathbb{E}\left[\|\theta_{\tau_\beta} - \theta_0\|\right] + \frac{2\alpha}{\beta}L_\omega C_g\mathbb{E}\left[\|z_{\tau_\beta} - z_0\|\right] \\
&+ \frac{2\alpha}{\beta}C_g D_\omega \mathbb{E}\left[\|z_{\tau_\beta}\|\,\|\theta_{\tau_\beta} - \theta_{\tau_\beta+1}\|\right] \\
\overset{(c)}{\leq}{}& \frac{\alpha}{\beta}L_\omega L_g \mathbb{E}\left[\|z_{\tau_\beta}\|^2\right] + \frac{\alpha}{2\beta}L_\omega \mathbb{E}\left[\|z_{\tau_\beta}\|^2\right] + \frac{\alpha}{8\beta}L_\omega \mathbb{E}\left[\|\nabla J(\theta_{\tau_\beta})\|^2\right] + \frac{\alpha}{\beta}\|z_0\|L_\omega C_g\beta \\
&+ \frac{\alpha^2}{\beta}\tau_\beta L_k\|z_0\|C_g + \frac{2\alpha}{\beta}L_\omega C_g\mathbb{E}\left[\|z_{\tau_\beta} - z_0\|\right] + \frac{2\alpha^2}{\beta}C_g^2 D_\omega \mathbb{E}\left[\|z_{\tau_\beta}\|\right] \\
={}& \left(\frac{\alpha}{\beta}L_\omega L_g + \frac{\alpha}{2\beta}L_\omega\right)\mathbb{E}\left[\|z_{\tau_\beta}\|^2\right] + \frac{2\alpha^2}{\beta}C_g^2 D_\omega \mathbb{E}\left[\|z_{\tau_\beta}\|\right] + \frac{\alpha}{8\beta}L_\omega \mathbb{E}\left[\|\nabla J(\theta_{\tau_\beta})\|^2\right] \\
&+ \left(\alpha L_\omega C_g + \frac{\alpha^2}{\beta}\tau_\beta L_k C_g\right)\|z_0\| + \frac{2\alpha}{\beta}L_\omega C_g\mathbb{E}\left[\|z_{\tau_\beta} - z_0\|\right],
\end{aligned}
\tag{118}
$$

where (a) is from the Mean-Value theorem and $\hat{\theta}_{\tau_\beta} = c\theta_{\tau_\beta} + (1-c)\theta_{\tau_\beta+1}$ for some $c \in [0,1]$, (b) is from Lemmas 1 and 2, (c) is due to the fact that $\left\|\mathbb{E}\left[G_{t+1}(\theta_0,\omega(\theta_0)) + \frac{\nabla J(\theta_0)}{2}\right]\right\| \leq C_g\beta$ for any $t \geq \tau_\beta$ and $\|\theta_{\tau_\beta} - \theta_0\| \leq \alpha\tau_\beta C_g$, and $L_k = 2C_g D_\omega + \left(L_J + \frac{L_g'}{2}\right)L_\omega$ is the Lipschitz constant of $\nabla\omega(\theta)\left(G_{t+1}(\theta,\omega(\theta)) + \frac{\nabla J(\theta)}{2}\right)$, and L_g' is the Lipschitz constant of $G_{t+1}(\theta,\omega(\theta))$.

Our next step is to rewrite the bound in (118) using $\|z_{\tau_\beta}\|$. Note that from Lemma 4, we have that

$$
\|z_0\| \leq \|z_{\tau_\beta} - z_0\| + \|z_{\tau_\beta}\| \leq \left(1 + 4\beta\tau_\beta C_\phi^2\right)\|z_{\tau_\beta}\| + 4\beta\tau_\beta(b_{\max} + L_\omega C_g).
\tag{119}
$$

Plugging in (118), it follows that

$$
\left|\mathbb{E}\left[z_{\tau_\beta}^\top \frac{\omega(\theta_{\tau_\beta}) - \omega(\theta_{\tau_\beta+1})}{\beta}\right]\right|
$$

31

$$
\leq \left(\frac{\alpha}{\beta}L_\omega L_g + \frac{\alpha}{2\beta}L_\omega\right)\mathbb{E}\left[\left\|z_{\tau_\beta}\right\|^2\right] + \frac{2\alpha^2}{\beta}C_g^2 D_\omega \mathbb{E}\left[\left\|z_{\tau_\beta}\right\|\right] + \frac{\alpha}{8\beta}L_\omega \mathbb{E}\left[\left\|\nabla J(\theta_{\tau_\beta})\right\|^2\right]
$$

$$
+ \left(\alpha L_\omega C_g + \frac{\alpha^2}{\beta}\tau_\beta L_k C_g\right)\|z_0\| + \frac{2\alpha}{\beta}L_\omega C_g \mathbb{E}\left[\left\|z_{\tau_\beta} - z_0\right\|\right]
$$

$$
\leq \left(\frac{\alpha}{\beta}L_\omega L_g + \frac{\alpha}{2\beta}L_\omega\right)\mathbb{E}\left[\left\|z_{\tau_\beta}\right\|^2\right] + \frac{2\alpha^2}{\beta}C_g^2 D_\omega \mathbb{E}\left[\left\|z_{\tau_\beta}\right\|\right] + \frac{\alpha}{8\beta}L_\omega \mathbb{E}\left[\left\|\nabla J(\theta_{\tau_\beta})\right\|^2\right]
$$

$$
+ \left(\alpha L_\omega C_g + \frac{\alpha^2}{\beta}\tau_\beta L_k C_g\right)\left((1 + 4\beta\tau_\beta C_\phi^2)\mathbb{E}\left[\left\|z_{\tau_\beta}\right\|\right] + 4\beta\tau_\beta(b_{\max} + L_\omega C_g)\right)
$$

$$
+ \frac{2\alpha}{\beta}L_\omega C_g\left(\mathbb{E}\left[4\beta\tau_\beta C_\phi^2\left\|z_{\tau_\beta}\right\|\right] + 4\beta\tau_\beta(b_{\max} + L_\omega C_g)\right)
$$

$$
= \left(\frac{\alpha}{\beta}L_\omega L_g + \frac{\alpha}{2\beta}L_\omega\right)\mathbb{E}\left[\left\|z_{\tau_\beta}\right\|^2\right]
$$

$$
+ \left(\frac{2\alpha^2}{\beta}C_g^2 D_\omega + \left(\alpha L_\omega C_g + \frac{\alpha^2}{\beta}\tau_\beta L_k C_g\right)(1 + 4\beta\tau_\beta C_\phi^2) + 8\alpha\tau_\beta L_\omega C_g C_\phi^2\right)\mathbb{E}\left[\left\|z_{\tau_\beta}\right\|\right]
$$

$$
+ \frac{\alpha}{8\beta}L_\omega \mathbb{E}\left[\left\|\nabla J(\theta_{\tau_\beta})\right\|^2\right] + \left(\alpha L_\omega C_g + \frac{\alpha^2}{\beta}\tau_\beta L_k C_g\right)(4\beta\tau_\beta(b_{\max} + L_\omega C_g))
$$

$$
+ 8\alpha\tau_\beta L_\omega C_g(b_{\max} + L_\omega C_g)
$$

$$
\triangleq R_3 \mathbb{E}\left[\left\|z_{\tau_\beta}\right\|^2\right] + P_3 \mathbb{E}\left[\left\|z_{\tau_\beta}\right\|\right] + Q_3 + \frac{\alpha}{8\beta}L_\omega \mathbb{E}\left[\left\|\nabla J(\theta_{\tau_\beta})\right\|^2\right], \tag{120}
$$

where $R_3 = \left(\frac{\alpha}{\beta}L_\omega L_g + \frac{\alpha}{2\beta}L_\omega\right) = \mathcal{O}\left(\frac{\alpha}{\beta}\right)$, $P_3 = \left(\frac{2\alpha^2}{\beta}C_g^2 D_\omega + \left(\alpha L_\omega C_g + \frac{\alpha^2}{\beta}\tau_\beta L_k C_g\right)(1 + 4\beta\tau_\beta C_\phi^2) + 8\alpha\tau_\beta L_\omega C_g C_\phi^2\right) = \mathcal{O}(\alpha\tau_\beta)$ and $Q_3 = \left(\alpha L_\omega C_g + \frac{\alpha^2}{\beta}\tau_\beta L_k C_g\right)(4\beta\tau_\beta(b_{\max} + L_\omega C_g)) + 8\alpha\tau_\beta L_\omega C_g(b_{\max} + L_\omega C_g) = \mathcal{O}(\alpha\tau_\beta)$.

Then we combine all three bounds in (114), (117) and (120), and it follows that

$$
\left|\mathbb{E}\left[z_{\tau_\beta}^\top\left(-A_{\theta_{\tau_\beta}}z_{\tau_\beta} - \frac{1}{\beta}\left(z_{\tau_\beta+1} - z_{\tau_\beta}\right)\right)\right]\right|
$$

$$
\leq (R_1 + R_3)\mathbb{E}\left[\left\|z_{\tau_\beta}\right\|^2\right] + (P_1 + P_2 + P_3)\mathbb{E}\left[\left\|z_{\tau_\beta}\right\|\right] + (Q_1 + Q_2 + Q_3)
$$

$$
+ \frac{\alpha}{8\beta}L_\omega \mathbb{E}\left[\left\|\nabla J(\theta_{\tau_\beta})\right\|^2\right], \tag{121}
$$

Finally due to the fact that $x \leq x^2 + 1, \forall x \in \mathbb{R}$, it follows that

$$
\left|\mathbb{E}\left[z_{\tau_\beta}^\top\left(-A_{\theta_{\tau_\beta}}z_{\tau_\beta} - \frac{1}{\beta}\left(z_{\tau_\beta+1} - z_{\tau_\beta}\right)\right)\right]\right|
$$

$$
\leq (R_1 + R_3 + P_1 + P_2 + P_3)\mathbb{E}\left[\left\|z_{\tau_\beta}\right\|^2\right] + (Q_1 + Q_2 + Q_3 + P_1 + P_2 + P_3)
$$

$$
+ \frac{\alpha}{8\beta}L_\omega \mathbb{E}\left[\left\|\nabla J(\theta_{\tau_\beta})\right\|^2\right]. \tag{122}
$$

This completes the proof. □

C.2 Proof under the Markovian Setting

In this section, we prove Theorem 1 under the Markovian setting.

From the L_J-smoothness of $J(\theta)$, it follows that

$$
J(\theta_{t+1}) \leq J(\theta_t) + \langle\nabla J(\theta_t), \theta_{t+1} - \theta_t\rangle + \frac{L_J}{2}\|\theta_{t+1} - \theta_t\|^2
$$

$$
= J(\theta_t) + \alpha\langle\nabla J(\theta_t), G_{t+1}(\theta_t, \omega_t)\rangle + \frac{L_J}{2}\alpha^2\|G_{t+1}(\theta_t, \omega_t)\|^2
$$

$$= J(\theta_t) - \alpha \left\langle \nabla J(\theta_t), -G_{t+1}(\theta_t, \omega_t) - \frac{\nabla J(\theta_t)}{2} + G_{t+1}(\theta_t, \omega(\theta_t)) - G_{t+1}(\theta_t, \omega(\theta_t)) \right\rangle$$

$$- \frac{\alpha}{2} \|\nabla J(\theta_t)\|^2 + \frac{L_J}{2} \alpha^2 \|G_{t+1}(\theta_t, \omega_t)\|^2$$

$$= J(\theta_t) - \alpha \left\langle \nabla J(\theta_t), -G_{t+1}(\theta_t, \omega_t) + G_{t+1}(\theta_t, \omega(\theta_t)) \right\rangle$$

$$+ \alpha \left\langle \nabla J(\theta_t), \frac{\nabla J(\theta_t)}{2} + G_{t+1}(\theta_t, \omega(\theta_t)) \right\rangle - \frac{\alpha}{2} \|\nabla J(\theta_t)\|^2 + \frac{L_J}{2} \alpha^2 \|G_{t+1}(\theta_t, \omega_t)\|^2$$

$$\leq J(\theta_t) + \alpha L_g \|\nabla J(\theta_t)\| \|\omega(\theta_t) - \omega_t\| + \alpha \left\langle \nabla J(\theta_t), \frac{\nabla J(\theta_t)}{2} + G_{t+1}(\theta_t, \omega(\theta_t)) \right\rangle$$

$$- \frac{\alpha}{2} \|\nabla J(\theta_t)\|^2 + \frac{L_J}{2} \alpha^2 \|G_{t+1}(\theta_t, \omega_t)\|^2$$

$$\overset{(a)}{\leq} J(\theta_t) + \alpha L_g \|\nabla J(\theta_t)\| \|z_t\| + \alpha \left\langle \nabla J(\theta_t), \frac{\nabla J(\theta_t)}{2} + G_{t+1}(\theta_t, \omega(\theta_t)) \right\rangle$$

$$- \frac{\alpha}{2} \|\nabla J(\theta_t)\|^2 + \frac{L_J}{2} \alpha^2 C_g^2, \tag{123}$$

where (a) is from the fact that $\|\theta_{t+1} - \theta_t\| \leq \alpha C_g$. Thus by re-arranging the terms, taking expectation and summing up w.r.t. t from 0 to $T-1$, it follows that

$$\frac{\alpha}{2} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]$$

$$\leq -\mathbb{E}[J(\theta_T)] + J(\theta_0) + \alpha L_g \sqrt{\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]} \sqrt{\sum_{t=0}^{T-1} \mathbb{E}[\|z_t\|^2]} + \sum_{t=0}^{T-1} \alpha \mathbb{E}[\zeta_G(\theta_t, O_t)]$$

$$+ L_J \alpha^2 T C_g^2, \tag{124}$$

where $\zeta_G(\theta_t, O_t) = \left\langle \nabla J(\theta_t), \frac{\nabla J(\theta_t)}{2} + G_{t+1}(\theta_t, \omega(\theta_t)) \right\rangle$. We then bound ζ_G in the following lemma.

Lemma 6. *For any $t \geq \tau_\beta$,*

$$\mathbb{E}[\zeta_G(\theta_t, O_t)] \leq 2C_g^2 \beta + 2\alpha \tau_\beta L_\zeta C_g. \tag{125}$$

Proof. We only need to consider the case $t = \tau_\beta$, the proof for general case of $t \geq \tau_\beta$ is similar, and thus is omitted here. We first have that

$$\zeta_G(\theta_{\tau_\beta}, O_{\tau_\beta}) = \left\langle \nabla J(\theta_{\tau_\beta}), \frac{\nabla J(\theta_{\tau_\beta})}{2} + G_{\tau_\beta+1}(\theta_{\tau_\beta}, \omega(\theta_{\tau_\beta})) \right\rangle$$

$$= \left\langle \nabla J(\theta_0), \frac{\nabla J(\theta_0)}{2} + G_{\tau_\beta+1}(\theta_0, \omega(\theta_0)) \right\rangle$$

$$+ \left\langle \nabla J(\theta_{\tau_\beta}), \frac{\nabla J(\theta_{\tau_\beta})}{2} + G_{\tau_\beta+1}(\theta_{\tau_\beta}, \omega(\theta_{\tau_\beta})) \right\rangle$$

$$- \left\langle \nabla J(\theta_0), \frac{\nabla J(\theta_0)}{2} + G_{\tau_\beta+1}(\theta_0, \omega(\theta_0)) \right\rangle$$

$$\leq \left\langle \nabla J(\theta_0), \frac{\nabla J(\theta_0)}{2} + G_{\tau_\beta+1}(\theta_0, \omega(\theta_0)) \right\rangle + 2L_\zeta \|\theta_{\tau_\beta} - \theta_0\|$$

$$\leq \left\langle \nabla J(\theta_0), \frac{\nabla J(\theta_0)}{2} + G_{\tau_\beta+1}(\theta_0, \omega(\theta_0)) \right\rangle + 2\alpha \tau_\beta L_\zeta C_g, \tag{126}$$

where $L_\zeta = 2C_g(L_g' + \frac{3L_J}{2})$ is the Lipschitz constant of $\zeta_G(\theta, O_t)$.

Then it follows that

$$\mathbb{E}[\zeta_G(\theta_{\tau_\beta}, O_{\tau_\beta})]$$

33

$$= \mathbb{E}\left[\left\langle \nabla J(\theta_0), \frac{\nabla J(\theta_0)}{2} + G_{\tau_\beta+1}(\theta_0, \omega(\theta_0)) \right\rangle\right] + 2\alpha L_\zeta C_g \tau_\beta$$

$$\le 2C_g^2\beta + 2\alpha\tau_\beta L_\zeta C_g, \tag{127}$$

where the last step follows from the uniform ergodicity of the MDP (Assumption 4). $\qquad\square$

Plugging the bound in (124), it follows that

$$\frac{\alpha}{2} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]$$

$$\le J(\theta_0) - J^* + \alpha L_g \sqrt{\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]} \sqrt{\sum_{t=0}^{T-1} \mathbb{E}[\|z_t\|^2]}$$

$$+ \alpha^2 C_g^2 L_J T + \alpha \left(T(2C_g^2\beta + 2\alpha\tau_\beta L_\zeta C_g) + 4\tau_\beta C_g^2\right), \tag{128}$$

and thus

$$\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]$$

$$\le \frac{2(J(\theta_0) - J^*)}{\alpha} + 2L_g \sqrt{\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]} \sqrt{\sum_{t=0}^{T-1} \mathbb{E}[\|z_t\|^2]} + 2\alpha C_g^2 L_J T$$

$$+ 2\left(T(2C_g^2\beta + 2\alpha\tau_\beta L_\zeta C_g) + 4\tau_\beta C_g^2\right). \tag{129}$$

This further implies that

$$\frac{\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}$$

$$\le \frac{2(J(\theta_0) - J^*)}{\alpha T} + 2L_g \sqrt{\frac{\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}} \sqrt{\frac{\sum_{t=0}^{T-1} \mathbb{E}[\|z_t\|^2]}{T}} + 2\alpha C_g^2 L_J$$

$$+ 2\left((2C_g^2\beta + 2\alpha\tau_\beta L_\zeta C_g) + 4C_g^2 \frac{\tau_\beta}{T}\right). \tag{130}$$

We plug in the tracking error (97), and it follows that

$$\frac{\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}$$

$$\le \frac{2(J(\theta_0) - J^*)}{\alpha T} + 2\alpha C_g^2 L_J + 2\left((2C_g^2\beta + 2\alpha\tau_\beta L_\zeta C_g) + 4C_g^2 \frac{\tau_\beta}{T}\right)$$

$$+ 2L_g \sqrt{\frac{\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}}$$

$$\cdot \sqrt{(2\|z_0\| + 2\beta\tau_\beta(b_{\max} + L_\omega C_g))^2 \left(\frac{1}{Tq} + \frac{\tau_\beta}{T}\right) + \frac{\alpha L_\omega}{4q} \frac{\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T} + \frac{p}{q}}$$

$$\le \frac{2(J(\theta_0) - J^*)}{\alpha T} + 2\alpha C_g^2 L_J + 2\left((2C_g^2\beta + 2\alpha\tau_\beta L_\zeta C_g) + 4C_g^2 \frac{\tau_\beta}{T}\right)$$

$$+ 2L_g \sqrt{\frac{\alpha L_\omega}{4q} \frac{\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}}$$

$$+ 2L_g \sqrt{\frac{\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}} \sqrt{(2\|z_0\| + 2\beta\tau_\beta(b_{\max} + L_\omega C_g))^2 \left(\frac{1}{Tq} + \frac{\tau_\beta}{T}\right) + \frac{p}{q}}. \tag{131}$$

Note that $2L_g\sqrt{\frac{\alpha L_\omega}{4q}} = \mathcal{O}\left(\sqrt{\frac{\alpha}{\beta}}\right)$, hence we can choose α and β such that $2L_g\sqrt{\frac{\alpha L_\omega}{4q}} \le \frac{1}{2}$. Hence it follows that

$$\frac{\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T} \le \frac{4(J(\theta_0) - J^*)}{\alpha T} + 4\alpha C_g^2 L_J + 4\left((2C_g^2\beta + 2\alpha\tau_\beta L_\zeta C_g) + 4C_g^2 \frac{\tau_\beta}{T}\right)$$

$$+ 4L_g \sqrt{\frac{\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}}$$

$$\cdot \sqrt{(2\|z_0\| + 2\beta\tau_\beta(b_{\max} + L_\omega C_g))^2 \left(\frac{1}{Tq} + \frac{\tau_\beta}{T}\right) + \frac{p}{q}}$$

$$\triangleq U \sqrt{\frac{\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}} + V, \tag{132}$$

where $U = 4L_g \sqrt{(2\|z_0\| + 2\beta\tau_\beta(b_{\max} + L_\omega C_g))^2 \left(\frac{1}{Tq} + \frac{\tau_\beta}{T}\right) + \frac{p}{q}} = \mathcal{O}\left(\sqrt{\beta\tau_\beta + \frac{1}{T\beta}}\right)$ and $V = \frac{4(J(\theta_0) - J^*)}{\alpha T} + 4\alpha C_g^2 L_J + 4\left((2C_g^2\beta + 2\alpha\tau_\beta L_\zeta C_g) + 4C_g^2 \frac{\tau_\beta}{T}\right) = \mathcal{O}\left(\frac{1}{T\alpha} + \alpha\tau_\beta + \beta\right)$. Thus it can be shown that

$$\frac{\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T} \leq \left(\frac{U + \sqrt{U^2 + 4V}}{2}\right)^2$$

$$\leq U^2 + 2V$$

$$= 16L_g^2 \left((2\|z_0\| + 2\beta\tau_\beta(b_{\max} + L_\omega C_g))^2 \left(\frac{1}{Tq} + \frac{\tau_\beta}{T}\right) + \frac{p}{q}\right)$$

$$+ \frac{8(J(\theta_0) - J^*)}{\alpha T} + 8\alpha C_g^2 L_J + 8\left((2C_g^2\beta + 2\alpha\tau_\beta L_\zeta C_g) + 4C_g^2 \frac{\tau_\beta}{T}\right)$$

$$= \mathcal{O}\left(\beta\tau_\beta + \frac{1}{T\beta} + \alpha\tau_\beta + \frac{1}{T\alpha}\right). \tag{133}$$

This completes the proof.

C.3 Choice of Step-sizes

In the proof under the Markovian setting, we first assume $\beta\tau_\beta C_\phi^2 \leq \frac{1}{4}$. The last assumption on the step-sizes is $\frac{\alpha}{q} \leq \frac{1}{4L_g^2 L_\omega}$, where $q = 2\beta\lambda_v - 2\beta(R_1 + R_3 + P_1 + P_2 + P_3) - 2\beta^2 C_\phi^4 = \mathcal{O}(\beta)$. Note that this assumption can be satisfied by controlling $\frac{\alpha}{\beta}$ similar to Section B.3, which we omit here. Hence we set $\beta < \min\left\{1, \frac{1}{4\tau_\beta C_\phi^2}\right\}$, and $\frac{\alpha}{q} \leq \left\{1, \frac{1}{4L_g^2 L_\omega}\right\}$.

D Experiments

In this section, we provide some numerical experiments on two RL examples: the Garnet problem [Archibald et al., 1995] and the "spiral" counter example in [Tsitsiklis and Van Roy, 1997].

D.1 Garnet Problem

The first experiment is on the Garnet problem [Archibald et al., 1995], which can be characterized by $\mathcal{G}(|\mathcal{S}|, |\mathcal{A}|, b, N)$. Here b is a branching parameter specifying how many next states are possible for each state-action pair, and these b states are chosen uniformly at random. The transition probabilities are generated by sampling uniformly and randomly between 0 and 1. The parameter N is the dimension of θ to be updated. In our experiments, we generate a reward matrix uniformly and randomly between 0 and 1. For every state s we randomly generate one feature function $k(s) \in [0, 1]$ using as the input. In both experiments, we use a five-layer neural network with (1,2,2,3,1) neurons in each layer as the function approximator. And for the activation function, we use the Sigmoid function, i.e., $f(x) = \frac{1}{1 + e^{-x}}$. We set all the weights and bias of the neurons as the parameter $\theta \in \mathbb{R}^{23}$.

We consider two sets of parameters: $\mathcal{G}(5, 2, 5, 23)$ and $\mathcal{G}(3, 2, 3, 23)$. We set the step-size $\alpha = 0.01$ and $\beta = 0.05$, and also the discount factor $\gamma = 0.95$. In Figures 1 and 2, we plot the squared gradient norm v.s. the number of samples using 40 Garnet MDP trajectories, i.e., at each time t, we plot $\|\nabla J(\theta_t)\|^2$. The upper and lower envelopes of the curves correspond to the 95 and 5 percentiles of the 40 curves, respectively. We also plot the estimated variance of the stochastic update along the

iterations in Figures 1(b) and 2(b). Specifically, we first run the algorithm to get a sequence of θ_t and ω_t. Then we generate 500 different trajectories $O^i = (O_1^i, O_2^i, ..., O_t^i, ...)$ where $i = 1, ..., 500$, and use them to estimate the variance $\|G_{t+1}^i(\theta_t, \omega_t) - \nabla J(\theta_t)\|^2$ and plot $\frac{\sum_{i=1}^{500} \|G_{t+1}^i(\theta_t, \omega_t) - \nabla J(\theta_t)\|^2}{500}$ at each time t.

It can be seen from the figures that both gradient norm $\|\nabla J(\theta_t)\|$ and the estimated variance converge to zero.

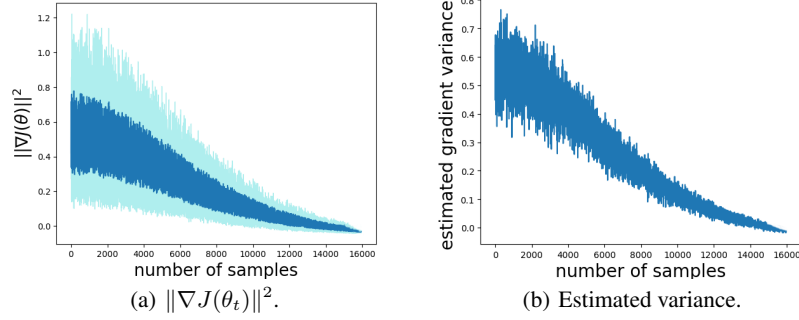(a) $\|\nabla J(\theta_t)\|^2$. (b) Estimated variance.

Figure 1: Garnet problem 1: $\mathcal{G}(5, 2, 5, 23)$.

D.2 Spiral Counter Example

In our second experiment, we consider the spiral counter example proposed in [Tsitsiklis and Van Roy, 1997], which is often used to show the TD algorithm may diverge with nonlinear function approximation. The problem setting is given in Figure 3. There are three states and each state can transit to the next one with probability $\frac{1}{2}$ or stay at the current state with probability $\frac{1}{2}$. The reward is always zero with the discount factor $\gamma = 0.9$. Similar to [Bhatnagar et al., 2009], we consider the value function approximation:

$$V_\theta(s) = (a(s)\cos(k\theta) + b(s)\sin(k\theta))e^{\epsilon\theta}, \tag{134}$$

where in Figure 4, $a = [0.94, -0.43, 0.18]$ and $b = [0.21, -0.52, 0.76]$; and in Figure 5, $a = [0.21, -0.33, 0.29]$ and $b = [0.68, 0.41, 0.82]$. We let $k = 0.866$ and $\epsilon = 0.1$. The step-size are chosen as $\alpha = 0.01$ and $\beta = 0.05$. In Figures 4(a) and 5(a), we plot the squared gradient norm v.s. the number of samples using 40 MDP trajectories. The upper and lower envelopes of the curves correspond to the 95 and 5 percentiles of the 40 curves. Similarly, we also plot the estimated variance $\|G_{t+1}(\theta_t, \omega_t) - \nabla J(\theta_t)\|^2$ of the stochastic update along the iterations using 50 samples at each time step. More specifically, we first run the algorithm to get a sequence of θ_t and ω_t. Then we generate
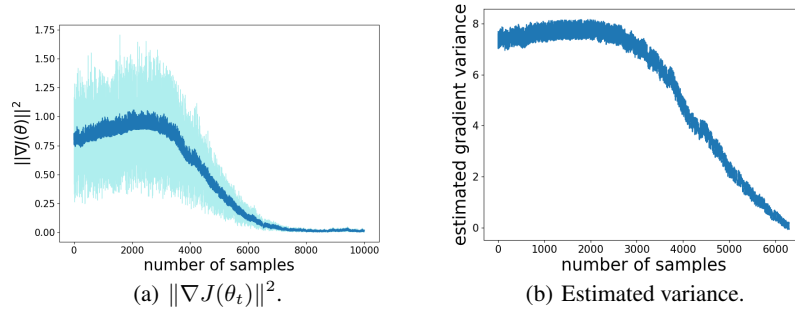

(a) $\|\nabla J(\theta_t)\|^2$. (b) Estimated variance.

Figure 2: Garnet problem 2: $\mathcal{G}(3, 2, 3, 23)$.

50 different trajectories $O^i = (O_1^i, O_2^i, ..., O_t^i, ...)$ where $i = 1, ..., 50$, and use them to estimate the variance $\|G_{t+1}^i(\theta_t, \omega_t) - \nabla J(\theta_t)\|^2$ and plot $\frac{\sum_{i=1}^{50} \|G_{t+1}^i(\theta_t, \omega_t) - \nabla J(\theta_t)\|^2}{50}$ at each time t.

It can be seen that in both experiments, the gradient norm $\|\nabla J(\theta_t)\|$ converges to 0, i.e., the algorithm converges to a stationary point. The estimated variance also decreases to zero.
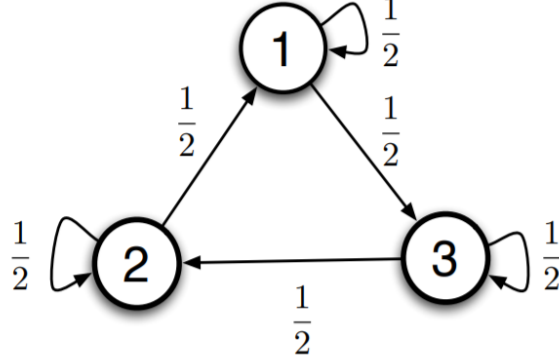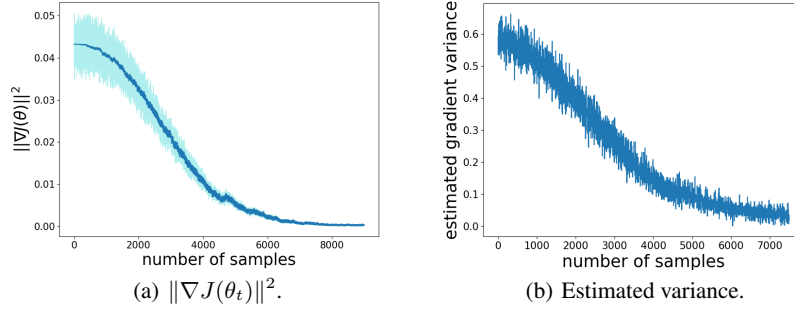

Figure 3: Spiral counter example.

(a) $\|\nabla J(\theta_t)\|^2$.

(b) Estimated variance.

Figure 4: Spiral counter example 1:
$a = [0.94, -0.43, 0.18], b = [0.21, -0.52, 0.76]$.

(a) $\|\nabla J(\theta_t)\|^2$.
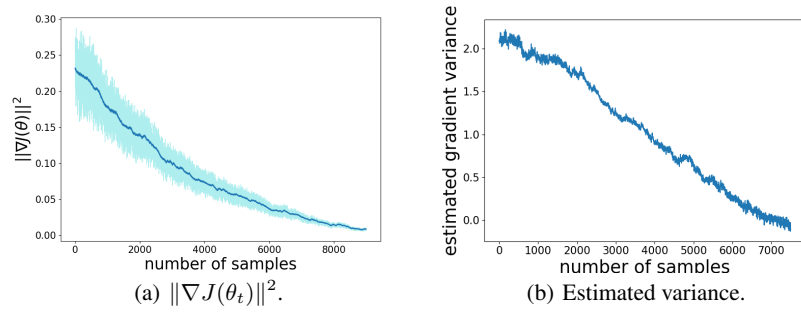

(b) Estimated variance.

Figure 5: Spiral counter example 2:
$a = [0.21, -0.33, 0.29]$, $b = [0.68, 0.41, 0.82]$.