Deterministic Policy Gradient: Convergence Analysis

Abstract

The deterministic policy gradient (DPG) method proposed in Silver et al. [2014] has been demonstrated to exhibit superior performance particularly for applications with multi-dimensional and continuous action spaces. However, it remains unclear whether DPG converges, and if so, how fast it converges and whether it converges as efficiently as other PG methods. In this paper, we provide a theoretical analysis of DPG to answer those questions. We study the single timescale DPG (often the case in practice) in both on-policy and off-policy settings, and show that both algorithms attain an ϵ-accurate stationary policy with a sample complexity of $\mathcal{O}(\epsilon^{-2})$. Moreover, we establish the convergence rate for DPG under Gaussian noise exploration, which is widely adopted in practice to improve the performance of DPG. To our best knowledge, this is the first non-asymptotic convergence characterization for DPG methods.

1 INTRODUCTION

Reinforcement learning (RL) has achieved tremendous success so far in many applications such as playing video games Mnih et al. [2013], bipedal walking Castillo et al. [2019] and online advertising Pednault et al. [2002], to name a few. The central aim of RL is to learn a policy that maximizes an accumulative reward for a task via the interaction with the environment. To this end, one popular method is to directly parameterize the policy and then optimize over the parameter space via (stochastic) gradient descent, which is referred to as the policy gradient (PG) algorithm [Williams, 1992]. More variants of policy gradient have been developed to further improve the performance, including natural policy gradient (NPG) [Kakade, 2002], trust region policy optimization (TRPO) [Schulman et al., 2015], proximal policy optimization (PPO) [Schulman et al., 2017], actor-critic (AC) Konda and Borkar [1999], Konda and Tsitsiklis [2000], Asynchronous Advantage Actor-Critic (A3C) [Mnih et al., 2016], Soft Actor-Critic (SAC) [Haarnoja et al., 2018], etc.

All the aforementioned PG algorithms adopt stochastic policies where the policy is modeled as a probability distribution over the action space. Rather, many RL applications have multi-dimensional **continuous** action spaces, for which **deterministic** policy gradient (DPG) algorithms have been proposed and demonstrated to significantly outperform stochastic PG algorithms in Silver et al. [2014]. Motivated by this, Lillicrap et al. [2016] combined DPG with DQN and proposed Deep Deterministic Policy Gradient (DDPG), which extends DQN in discrete action space to a continuous setting. Later, DDPG has also gained great success in distributional [Barth-Maron et al., 2018] and multi-agent [Lowe et al., 2017] scenarios. Although DPG and its variants exhibit superior performance in practice, the theoretical understanding of its convergence is rather limited. In fact, the only attempt was made in Kumar et al. [2020], which provided the convergence results for a modified zeroth-order DPG algorithm. However, what is used commonly in practice is the DPG algorithm originally proposed in Silver et al. [2014], for which the convergence guarantee remains open.

In fact, the convergence theory of DPG does not follow from that for stochastic PG algorithms due to a few unique features that DPG has. (a) The policy gradient in DPG takes a very different form from that in PG, and admits different compatibility for function approximation and consequently different actors to estimate. There is thus no guarantee that such a designed update rule must have guaranteed convergence as PG. (b) Through determinism in policy, practical implementation of DPG introduces **stochastic noisy** sampling to improve exploration. There is no previous theory on such a mixed deterministic policy update with noisy sampling for exploration. (c) DPG takes alternative simultaneous updates between critic and actor with constant learning rates for both. Previous analysis for stochastic PG typically requires sufficiently fast update for critic so that its track-

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

ing error can be (asymptotically) decoupled from actor's convergence error. Such analysis is not applicable to DPG.

Thus, the goal of this paper is to develop new tools to address the aforementioned challenges and provide the first finite-sample convergence guarantee for DPG algorithms in Silver et al. [2014].

1.1 MAIN CONTRIBUTION

The main contribution of this work lies in establishing the first finite-sample analysis for both on-policy and off-policy DPG algorithms proposed in Silver et al. [2014].

For the on-policy setting, we study DPG-TD, which uses the compatible approximation rule given by Silver et al. [2014] to update actor, and adopts temporal difference (TD) learning with linear function approximation to update critic. We show that DPG-TD finds a stationary point (up to a system error) with a sample complexity of $\mathcal{O}(\epsilon^{-2})$. In addition, we also show that noised DPG-TD (NoiDPG-TD) which uses a mixture of noisy exploration and the deterministic policies can also achieve a sample complexity of $\mathcal{O}(\epsilon^{-2})$. For the off-policy setting, we study DPG-TDC, which uses TD with gradient correction (TDC) to update critic under off-policy data and show that DPG-TDC also achieves the convergence with a sample complexity of $\mathcal{O}(\epsilon^{-2})$.

Our results bring the following insights to the understanding of DPG. (a) Our sample complexity of DPG matches the best known actor-critic (AC) type PG in Xu et al. [2020b]. This implies that although the policy gradient of DPG is more challenging to estimate, the compatible function approximation for DPG in Silver et al. [2014] is as efficient as that for PG, which yields the same complexity for DPG. (b) DPG achieves the same sample complexity for more challenging continuous and possibly unbounded action space whereas the known theorems for PG typically require the bounded action space except for Gaussian policy. (c) The noisy exploration does not cause higher sample complexity. (d) The simultaneous updates for actor and critic without sufficient estimation accuracy for critic still yield convergence without causing more sample complexity.

Technically, our analysis develops the following novel techniques to handle the unique challenges arising due to deterministic policies. (a) We develop a new analysis to bound the estimation error of the Fisher information of deterministic policy arising via the compatibility property, and then further capture how such a metric affects the convergence via its minimum eigenvalue. (b) We develop a new tool to analyze the coupled actor and critic's stochastic approximation processes, due to their simultaneous updates both with constant stepsizes (which are commonly used in practice). Previous analysis of (stochastic) PG-type algorithms including AC algorithms mainly decouples the critic's error from actor's error either by sufficient updates of critic before each actor's update [Wang et al., 2020, Kumar et al., 2019, Qiu et al., 2019, Xu et al., 2020b], or by updating critic much faster than actor via two timescale learning rates [Wu et al., 2020, Xu et al., 2020c]. Our analysis allows the coupling between the two and develops the idea to cancel the critic's coupling error by the actor's overall positive progress to the stationary policy.

1.2 RELATED WORK

Convergence of DPG and its variants: Since proposed and practically justified in Silver et al. [2014], DPG has inspired many variants and gained great success. However, there is almost no theoretical study of DPG with only one exception in Kumar et al. [2020], which provided the convergence guarantee for a zeroth-order DPG rather than the original form of DPG used widely in practice. Our work aims to provide the finite-sample convergence guarantee for DPG in its original practical form [Silver et al., 2014].

Convergence of stochastic PG: The vanilla PG is a fundamental policy-based RL algorithm. Its asymptotic convergence has been established in Williams [1992], Baxter and Bartlett [2001], Sutton et al. [2000], Kakade [2002], Pirotta et al. [2015], Tadić et al. [2017] via modeling PG as stochastic approximation (SA). PG has been shown to find the optimal policy under convex policy function approximation [Bhandari and Russo, 2019] or in some specific applications such as LQR [Fazel et al., 2018, Malik et al., 2019, Tu and Recht, 2019]. Convergence of (N)PG under a more general function approximation has been also provided in Shen et al. [2019], Papini et al. [2017, 2018, 2019], Xu et al. [2019, 2020a], Zhang et al. [2019], Agarwal et al. [2019], Karimi et al. [2019], Wang et al. [2020], Cen et al. [2020].

Convergence of stochastic actor-critic: The AC algorithm was proposed in Konda and Borkar [1999], and since then has aroused wide interest in understanding its convergence. Konda and Tsitsiklis [2000], Konda [2002], Peters and Schaal [2008], Bhatnagar et al. [2008, 2009], Bhatnagar [2010], Castro and Meir [2010], Maei [2018] established the asymptotic convergence for (natural) AC. The non-asymptotic convergence for (N)AC has been also explored recently. Under a double-loop setting, where critic can run sufficiently many iterations before updating actor, the convergence rate for (N)AC has been characterized in Yang et al. [2019], Wang et al. [2020], Kumar et al. [2019], Qiu et al. [2019], Xu et al. [2020b]. Under a two-timescale setting, where critic and actor update simultaneously but undergo different diminishing learning rates, the convergence rate has been provided in Wu et al. [2020], Xu et al. [2020c], Hong et al. [2020], Zhang et al. [2020], Shen et al. [2020]. Under the single timescale setting, where critic and actor are simultaneously updated with constant learning rates, the convergence rate was given in Fu et al. [2021]. Differently

from the above stochastic PG and stochastic AC, our work studies the deterministic PG algorithms. The update structure of our work is the same as that in Fu et al. [2021], but we adopt the more practical TD updates for critic rather than LSTD in Fu et al. [2021], which causes substantial difference in our analysis besides the deterministic policy.

2 PRELIMINARY

2.1 PROBLEM SETUP

We consider the standard RL settings where an agent interacts with a stochastic environment. Such a system is usually modeled as a discrete-time discounted Markov Decision Process (MDP) which is represented by a tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$. Here, $\mathcal{S}$ denotes the state space, $\mathcal{A}$ denotes the action space, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0, 1]$ denotes the transition kernel for the state transitions, e.g., $P(s'|s, a)$ represents the probability that the system takes the next state $s' \in \mathcal{S}$ given the current state s and action a; $r : \mathcal{S} \times \mathcal{A} \mapsto [0, R_{\max}]$ is the reward function mapping the station-action pairs to a bounded subset of $\mathbb{R}$, and $\gamma \in (0, 1)$ is the discount factor.

Here, we consider a deterministic policy μ_θ parameterized by $\theta \in \mathbb{R}^d$, namely, given the current state s, the policy follows a deterministic function mapping to generate an action $a = \mu_\theta(s)$. We also assume that the Markov chains generated by the policies are ergodic throughout this paper.

2.2 ON-POLICY DETERMINISTIC POLICY GRADIENT

We first consider the on-policy case where the interaction with the environment (i.e., the sampling) can follow the instantaneous target policy μ_θ.

The goal is to maximize the expected cumulative reward in the infinite-horizon case given by

$$J(\mu_\theta) = \int_{\mathcal{S}} \nu_{\mu_\theta}(s) r(s, \mu_\theta(s)) ds = \mathbb{E}_{s\sim\nu_{\mu_\theta}} \left[r(s, \mu_\theta(s)) \right], \tag{1}$$

where $\nu_\mu(s') = \int_{\mathcal{S}} \sum_{t=0}^{\infty} \gamma^t p_0(s) p(s \to s', t, \mu) ds$ is the (improper) discounted state visitation distribution and $p(s \to s', t, \mu)$ denotes the density at state s' after t steps from state s under policy μ. In the remaining of this paper, we denote $J(\theta) := J(\mu_\theta)$ and $\nu_\theta := \nu_{\mu_\theta}$ for brevity.

One popular method to optimize the loss function defined in (1) is to use gradient-based algorithms such as stochastic gradient descent (SGD). To this end, the gradient of the loss function has been given by the so-called deterministic policy gradient theorem [Silver et al., 2014] as follows:

$$\nabla J(\theta) = \int_{\mathcal{S}} \nu_\theta(s) \nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s, a)|_{a=\mu_\theta(s)} ds$$
$$= \mathbb{E}_{\nu_\theta} \left[\nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s, a)|_{a=\mu_\theta(s)} \right]. \tag{2}$$

The policy gradient theorem for deterministic policies suggests a way to estimate the gradient via sampling, and then model-free policy gradient algorithms can be developed by following SGD updates for optimizing over policies. The difficulty of estimating the policy gradient $\nabla J(\theta)$ in (2) lies in approximating $\nabla_a Q^{\mu_\theta}(s, a)$. To address this difficulty, the compatible function approximation was established in Silver et al. [2014] which guarantees that $\nabla_a Q^{\mu_\theta}(s, a)$ can be replaced by $\nabla_a Q^w(s, a)$ in the policy gradient. We state such a property below, which is critical for designing the deterministic policy gradient algorithms.

Proposition 1. *(Compatible function approximation [Silver et al., 2014]) A function estimator $Q^w(s, a)$ is said to be compatible with a deterministic policy μ_θ, i.e., $\nabla J(\theta) = \mathbb{E}_{\nu_\theta} \left[\nabla_\theta \mu_\theta(s) \nabla_a Q^w(s, a)|_{a=\mu_\theta(s)} \right]$, if it satisfies the following two conditions:*

1. $\nabla_a Q^w(s, a)|_{a=\mu_\theta(s)} = \nabla_\theta \mu_\theta(s)^T w$;
2. $w = w^*_{\xi_\theta}$ *minimizes the mean square error* $\mathbb{E}_{\nu_\theta} \left[\xi(s; \theta, w)^T \xi(s; \theta, w) \right]$, *where* $\xi(s; \theta, w) = \nabla_a Q^w(s, a)|_{a=\mu_\theta(s)} - \nabla_a Q^{\mu_\theta}(s, a)|_{a=\mu_\theta(s)}$.

Following the compatibility property, the deterministic policy gradient can be rewritten as

$$\nabla J(\theta) = \mathbb{E}_{\nu_\theta} \left[\nabla_\theta \mu_\theta(s) \nabla_\theta \mu_\theta(s)^T w^*_{\xi_\theta} \right], \tag{3}$$

where $w^*_{\xi_\theta}$ can be approximated easily by solving a regression problem.

2.3 OFF-POLICY DETERMINISTIC POLICY GRADIENT

In practice, it is often convenient to estimate the policy gradient via sampling under a behavior policy β, which is different from the target policy μ_θ. Silver et al. [2014] has also provided the deterministic policy gradient theorem for such an off-policy case, which is given by

$$\nabla J_\beta(\theta) = \int_{\mathcal{S}} \nu_\beta(s) \nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s, a)|_{a=\mu_\theta(s)} ds$$
$$= \mathbb{E}_{\nu_\beta} \left[\nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s, a)|_{a=\mu_\theta(s)} \right], \tag{4}$$

where ν_β is the state visitation measure of the policy β. Correspondingly, the compatible form is given by

$$\nabla J_\beta(\theta) = \mathbb{E}_{\nu_\beta} \left[\nabla_\theta \mu_\theta(s) \nabla_\theta \mu_\theta(s)^T w^*_{\beta, \xi_\theta} \right], \tag{5}$$

where $w^*_{\beta, \xi_\theta} = \arg\min_w \mathbb{E}_{\nu_\beta} \left[\xi(s; \theta, w)^T \xi(s; \theta, w) \right]$ and $\xi(s; \theta, w)$ holds the same form as in Proposition 1.

3 ON-POLICY DPG ALGORITHM

In this section, we first describe the on-policy DPG algorithm proposed in Silver et al. [2014] and then provide the finite-sample convergence result for this algorithm.

3.1 ALGORITHM

In Silver et al. [2014], a compatible DPG algorithm using TD critic update was proposed, which we call as DPG-TD and describe in Algorithm 1. This algorithm introduces a critic parameter w to estimate the gradient of Q-function based on the compatibility property. At each iteration, w is updated by TD with a linear function approximator $Q^w(s,a) = \phi(s,a)^T w$ (line 9 of Algorithm 1). The algorithm uses θ as an actor parameter to update the policy (line 13 in Algorithm 1) based on the compatibility property.

Algorithm 1 DPG-TD

1: **Input:** $\alpha_w, \alpha_\theta, w_0, \theta_0$, batch size M.
2: **for** $t = 0, 1, \ldots, T$ **do**
3: **for** $j = 0, 1, \ldots, M-1$ **do**
4: Sample $s_{t,j} \sim d_{\theta_t}$. Generate $a_{t,j} = \mu_{\theta_t}(s_{t,j})$.
5: Sample $s_{t+1,j} \sim P(\cdot|s_{t,j}, a_{t,j})$ and $r_{t,j}$. Generate $a_{t+1,j} = \mu_{\theta_t}(s_{t+1,j})$.
6: Denote $x_{t,j} = (s_{t,j}, a_{t,j})$.
7: $\delta_{t,j} = r_{t,j} + \gamma\phi(x_{t+1,j})^T w_t - \phi(x_{t,j})^T w_t$.
8: **end for**
9: $w_{t+1} = w_t + \frac{\alpha_w}{M}\sum_{j=0}^{M-1} \delta_{t,j}\phi(x_{t,j})$.
10: **for** $j = 0, 1, \ldots, M-1$ **do**
11: Sample $s'_{t,j} \sim \nu_{\theta_t}$.
12: **end for**
13: $\theta_{t+1} = \theta_t + \frac{\alpha_\theta}{M}\sum_{j=0}^{M-1} \nabla_\theta \mu_{\theta_t}(s'_{t,j}) \nabla_\theta \mu_{\theta_t}(s'_{t,j})^T w_t$.
14: **end for**

3.2 TECHNICAL ASSUMPTIONS

Before providing the result, we first introduce technical assumptions, all of which are standard or necessary mild regularity requirements.

Assumption 1. *For any $\theta_1, \theta_2, \theta \in \mathbb{R}^d$, there exist positive constants L_μ, L_ψ and λ_Ψ, such that (1) $\|\mu_{\theta_1}(s) - \mu_{\theta_2}(s)\| \leq L_\mu \|\theta_1 - \theta_2\|, \forall s \in \mathcal{S}$; (2) $\|\nabla_\theta \mu_{\theta_1}(s) - \nabla_\theta \mu_{\theta_2}(s)\| \leq L_\psi \|\theta_1 - \theta_2\|, \forall s \in \mathcal{S}$; (3) the matrix $\Psi_\theta := \mathbb{E}_{\nu_\theta}\left[\nabla_\theta \mu_\theta(s) \nabla_\theta \mu_\theta(s)^T\right]$ (which we call as* **Fisher information of deterministic policy***) is non-singular with the minimal eigenvalue uniformly lower-bounded as $\sigma_{\min}(\Psi_\theta) \geq \lambda_\Psi$.*

The first two statements can be easily satisfied for properly parameterized policy classes such as the linear approximator used in Silver et al. [2014] and the smooth neural network class. The last one ensures that $w^*_{\xi_\theta}$ defined in Proposition 1 is solvable and unique. A similar assumption has been also used in Liu et al. [2020].

Assumption 2. *For any $a_1, a_2 \in \mathcal{A}$, there exist positive constants L_P, L_r, such that (1) the transition kernel satisfies $|P(s'|s,a_1) - P(s'|s,a_2)| \leq L_P \|a_1 - a_2\|, \forall s, s' \in \mathcal{S}$; (2) the reward function satisfies $|r(s,a_1) - r(s,a_2)| \leq L_r \|a_1 - a_2\|, \forall s, s' \in \mathcal{S}$.*

In Assumption 2, the first statement is standard in the theoretical studies of RL [Bertsekas, 1975, Chow and Tsitsiklis, 1991, Dufour and Prieto-Rumeau, 2013, 2015], where the transition kernel is assumed to be Lipschitz continuous with respect to (w.r.t.) both state and action. Shah and Xie [2018] relaxes the Lipschitz continuity to be only w.r.t. state when considering a continuous state space. In this paper, we need the Lipschitz continuity to hold only w.r.t. action, because DPG algorithms are commonly used for continuous action space. The second statement can be easily satisfied for a properly defined reward function.

Assumption 3. *For any $a_1, a_2 \in \mathcal{A}$, there exists a positive constant L_Q, such that $\|\nabla_a Q^{\mu_\theta}(s,a_1) - \nabla_a Q^{\mu_\theta}(s,a_2)\| \leq L_Q \|a_1 - a_2\|, \forall \theta \in \mathbb{R}^d, s \in \mathcal{S}$.*

Assumption 3 indicates that the Q-function is smooth over action, which is a standard assumption in deterministic policy related studies [Kumar et al., 2020], and is also known as a principle to mitigate overfitting in the value estimation for actor-critic algorithms [Fujimoto et al., 2018].

Assumption 4. *The feature function $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is uniformly bounded, i.e., $\|\phi(\cdot,\cdot)\| \leq C_\phi$ for some positive constant C_ϕ. In addition, we define $A = \mathbb{E}_{d_\theta}\left[\phi(x)(\gamma\phi(x') - \phi(x))^T\right]$ and $D = \mathbb{E}_{d_\theta}\left[\phi(x)\phi(x)^T\right]$, and assume that A and D are non-singular. We further assume that the absolute value of the eigenvalues of A are uniformly lower bounded, i.e., $|\sigma(A)| \geq \lambda_A$ for some positive constant λ_A.*

Assumption 4 is standard in the studies of TD learning with linear function approximation [Zou et al., 2019, Wu et al., 2020, Xu et al., 2020b,c]. This assumption guarantees the solvability of TD learning with linear function approximation. To be more specific, it ensures that $\mathbb{E}_{d_\theta}[\delta\phi] = 0$ has a unique root, namely w^*_θ, which is also the global optimum of TD learning for a fixed policy μ_θ.

Remark 1. *We will abuse the notations a bit and assume that Assumptions 1 and 4 also hold for the off-policy case, with the expectations taken over the behavior stationary and (improper) visitation distributions d_β and ν_β.*

3.3 CONVERGENCE RESULT

In this subsection, we provide the finite-sample convergence analysis for DPG-TD in Algorithm 1.

Note that the deterministic policy gradient $\nabla J(\theta)$ in (2) has a different and more challenging form to analyze compared with stochastic PG, which requires the development of several new tools. First, we characterize the Lipschitz property for the deterministic policy gradient, which serves as a crucial step in the finite-sample analysis of DPG-TD. The previous study of DPG in Kumar et al. [2020] takes such a property as an assumption. Here, we formally establish such a Lipschitz property with the proof provided in Appendix A, and characterize the dependence of the deterministic policy gradient on the basic parameters of the MDP.

Lemma 1. *Suppose Assumptions 1-3 hold. Then the deterministic policy gradient $\nabla J(\theta)$ defined in* (2) *is Lipschitz continuous with the parameter L_J, i.e., $\forall \theta_1, \theta_2 \in \mathbb{R}^d$,*

$$\|\nabla J(\theta_1) - \nabla J(\theta_2)\| \leq L_J \|\theta_1 - \theta_2\|, \tag{6}$$

where $L_J = \left(\frac{1}{2} L_P L_\mu^2 L_\nu C_\nu + \frac{L_\psi}{1-\gamma}\right)\left(L_r + \frac{\gamma R_{\max} L_P}{1-\gamma}\right) + \frac{L_\mu}{1-\gamma}\left(L_Q L_\mu + \frac{\gamma}{2} L_P^2 R_{\max} L_\mu C_\nu + \frac{\gamma L_P L_r L_\mu}{1-\gamma}\right)$.

In the following, we provide the convergence guarantee for DPG-TD. Our main technical novelty lies in the development of a new framework to analyze the **coupled** actor and critic's stochastic approximation processes, due to their simultaneous updates both with constant stepsizes. Our central idea is to cancel the critic's cumulative coupling error by the overall positive progress of actor's approach to the stationary policy. This is different from the previous analysis of (stochastic) PG-type algorithms which mainly decouples or asymptotically decouples the critic's error from actor's error. We provide a proof sketch in Section 3.4 with the full proof given in Appendix B.

Theorem 1. *Suppose that Assumptions 1-4 hold. Let $\alpha_w \leq \frac{\lambda}{2C_A^2}; M \geq \frac{48\alpha_w C_A^2}{\lambda}; \alpha_\theta \leq \min\left\{\frac{1}{4L_J}, \frac{\lambda\alpha_w}{24L_h L_w}\right\}$. Then the output of DPG-TD in Algorithm 1 satisfies*

$$\min_{t\in[T]} \mathbb{E}\|\nabla J(\theta_t)\|^2 \leq \frac{c_1}{T} + \frac{c_2}{M} + c_3\kappa^2,$$

where $c_1 = \frac{8R_{\max}}{\alpha_\theta(1-\gamma)} + \frac{144L_h^2}{\lambda\alpha_w}\left\|w_0 - w_{\theta_0}^*\right\|^2, c_2 = \left[48\alpha_w^2(C_A^2 C_w^2 + C_b^2) + \frac{48L_w^2 L_\mu^4 C_{w_\xi}^2 \alpha_\theta^2}{\lambda\alpha_w}\right] \cdot \frac{144L_h^2}{\lambda\alpha_w} + 36L_\mu^4 C_{w_\xi}^2, c_3 = 18L_h^2 + \frac{24L_w^2 L_h^2 \alpha_\theta^2}{\lambda\alpha_w}$ *with* $C_A = 2C_\phi^2, C_b = R_{\max} C_\phi, C_w = \frac{R_{\max} C_\phi}{\lambda_A}, C_{w_\xi} = \frac{L_\mu C_Q}{\lambda_\Psi(1-\gamma)}, L_w = \frac{L_J}{\lambda_\Psi} + \frac{L_\mu C_Q}{\lambda_\Psi^2(1-\gamma)}\left(L_\mu^2 L_\nu + \frac{2L_\mu L_\psi}{1-\gamma}\right), L_h = L_\mu^2, C_Q = L_r + L_P \cdot \frac{\gamma R_{\max}}{1-\gamma}, L_\nu = \frac{1}{2} C_\nu L_P L_\mu$, *and L_J defined in Lemma 1, and we define*

$$\kappa := \max_\theta \left\|w_\theta^* - w_{\xi_\theta}^*\right\|. \tag{7}$$

Theorem 1 indicates that the convergence upper bound consists of three parts. The first term captures the convergence rate and vanishes sublinearly with the number of iterations. The second term captures the variance caused by the stochastic sampling and can be controlled by the batch size. The last term captures the system error $\left\|w_\theta^* - w_{\xi_\theta}^*\right\|$ which is uniformly bounded by some constant κ. Such a system error is determined by the expressive power of the function approximation class. In practice, the high capacity of the neural network class can significantly help to reduce such an error and achieves better convergence accuracy.

Theorem 1 also captures how the Fisher information of deterministic policy $\Psi_\theta := \mathbb{E}_{\nu_\theta}\left[\nabla_\theta \mu_\theta(s) \nabla_\theta \mu_\theta(s)^T\right]$ affects the convergence rate via its minimum eigenvalue bound λ_Ψ. Such a metric arises due to the compatible function approximation and captures how well actor estimates the deterministic policy gradient. Clearly, larger λ_Ψ indicates a better system condition (smaller C_{w_ξ} and L_w) and hence a faster convergence.

Based on the convergence rate in Theorem 1, we provide the sample complexity of the algorithm as follows.

Corollary 1. *Suppose that the same assumptions in Theorem 1 hold. Then the output of DPG-TD in Algorithm 1 satisfies $\min_{t\in[T]} \mathbb{E}\|\nabla J(\theta_t)\|^2 \leq \epsilon + c_3\kappa^2$, by using the total number of samples $2MT = \mathcal{O}\left(1/\epsilon^2\right)$.*

Corollary 1 shows that DPG-TD attains an ϵ-accurate stationary point (up to the system error) with a sample complexity of $\mathcal{O}(\epsilon^{-2})$. To our best knowledge, this is the first finite-sample characterization for DPG.

Despite the policy gradient of DPG is more challenging to estimate, the sample complexity of DPG in Corollary 1 matches the best known stochastic PG (with AC scheme) in Xu et al. [2020b]. Furthermore, such a result does not require critic's update in DPG to accurately track the deterministic policy gradient at each step and hence is practically desired, whereas the sample complexity guarantee in Xu et al. [2020b] for stochastic PG requires sufficient accuracy of tracking the policy gradient at each iteration.

DPG with noisy sampling for exploration. In practice, the deterministic policy used in Algorithm 1 usually suffers from the inefficient exploration. To overcome such an issue, Silver et al. [2014] proposed to use a noisy sampling for DPG. To be specific, in lines 4-5 of Algorithm 1, a noisy policy, e.g., $\pi_{\theta_t}(s) = \mu_{\theta_t}(s) + \mathcal{N}(0, \sigma^2)$, is adopted to generate actions $a_{t,j}, a_{t+1,j}$. Correspondingly, the states for critic's updates in lines 4-5 of Algorithm 1 are generated by the stationary distribution $d_{\pi_{\theta_t}}$ associated with the noisy policy π_{θ_t}. The rest of Algorithm 1 is unchanged. We refer to such an algorithm as noisy DPG-TD (NoiDPG-TD).

Following the same techniques for the proof of Theorem 1 and replace the stationary distribution d_{θ_t} by $d_{\pi_{\theta_t}}$, we readily obtain the convergence result for NoiDPG-TD as follows.

Corollary 2. *Suppose that the same assumptions in Theorem 1 hold. Then the output of NoiDPG-TD satisfies* $\min_{t\in[T]}\mathbb{E}\left\|\nabla J(\theta_t)\right\|^2 \leq \epsilon + \mathcal{O}\left(\kappa^2\right)$*, by using the total number of samples* $\mathcal{O}\left(\epsilon^{-2}\right)$.

In Corollary 2, the system error κ is determined by the noisy policy. Corollary 2 indicates that the noisy sampling for exploration does not cause higher sample complexity compared to DPG-TD.

3.4 PROOF SKETCH OF THEOREM 1

In the following, we outline the proof of Theorem 1 to highlight our new approach to analyzing the **coupled** actor and critic's stochastic approximation processes, due to their simultaneous updates both with constant stepsizes. The central idea is to cancel the critic's cumulative tracking error by the actor's overall positive progress to the stationary policy, which is different from the existing analysis of (stochastic) PG-type algorithms that mainly decouples or asymptotically decouples the critic's error from actor's error. Further, we develop a new analysis to bound the estimation error of the Fisher information of deterministic policy arising via the compatibility theorem, and then further capture how such a metric affects the convergence via its minimum eigenvalue.

The main proof consists of three steps. First, we characterize the error propagation of tracking a dynamic critic target (i.e., dynamic tracking error) based on its coupling with actor's update progress. Second, we bound the critic's cumulative tracking error in terms of actor's update progress via the compatibility properties of DPG. Last, we establish the overall convergence by canceling out the cumulative tracking error via the actor's overall positive progress towards the stationary policy.

Step I: Characterizing dynamics of critic's error via coupling with actor.

In the first step, we characterize the propagation of the dynamics of critic's dynamic tracking error based on its coupling with actor's updates. That is, we develop the relationship between $\left\|w_{t+1} - w^*_{\theta_{t+1}}\right\|^2$ and $\left\|w_t - w^*_{\theta_t}\right\|^2$ by their coupling with actor's updates.

Recall that $w^*_{\theta_t}$ is the global optimum of TD given a fixed policy μ_{θ_t}, or is equivalently the unique root of $\bar{g}_{\theta_t}(w_t) := \mathbb{E}_{d_{\theta_t}}\left[\frac{1}{M}\sum_{j=0}^{M-1}\delta_{t,j}\phi(x_{t,j})\right] = 0$. We first give the following bound on the update rule of w_t in Algorithm 1 given by the TD learning property [Tsitsiklis and Van Roy, 1997, Bhandari et al., 2018, Xiong et al., 2020],

$$\mathbb{E}\left\|w_{t+1} - w^*_{\theta_t}\right\|^2 \leq \left(1 - \frac{\alpha_w\lambda}{2}\right)\mathbb{E}\left\|w_t - w^*_{\theta_t}\right\|^2 + \frac{24\alpha_w^2(C_A^2C_w^2 + C_b^2)}{M},$$

where $\alpha_w \leq \frac{\lambda}{2C_A^2}$, $M \geq \frac{48\alpha_w C_A^2}{\lambda}$.

In the previous analysis of (stochastic) AC algorithms, sufficient TD updates of critic result in a controlled small tracking error before updating the actor, which is hence decoupled from the actor's progress. In contrast, DPG-TD takes alternative updates between critic and actor, so that the critic's tracking error is inherent and non-vanishing. Thus, we take a new approach to characterize the moving dynamics of the tracking error and directly couple it with the actor's update as follows,

$$\mathbb{E}\left\|w_{t+1} - w^*_{\theta_{t+1}}\right\|^2 \leq \frac{4L_w^2}{\lambda\alpha_w}\mathbb{E}\left\|\theta_{t+1} - \theta_t\right\|^2 + \left(1 - \frac{\lambda\alpha_w}{4}\right)\mathbb{E}\left\|w_t - w^*_{\theta_t}\right\|^2 + \frac{48\alpha_w^2(C_A^2C_w^2 + C_b^2)}{M}.$$

Clearly, in the above bound, the two tracking errors at times $t+1$ and t have different targets $w^*_{\theta_{t+1}}$ and $w^*_{\theta_t}$ due to actor's one update between critic's two consecutive updates. Hence, actor's update is necessarily coupled into the dynamics of the critic's tracking error.

Step II: Bounding cumulative tracking error via compatibility theorem for DPG.

In this step, we bound the cumulative tracking error based on the dynamics of the tracking error from the last step.

To this end, we first bound the difference between two consecutive actor parameters via DPG's properties. By the update rule of θ_t in Algorithm 1, we have $\theta_{t+1} - \theta_t = \frac{\alpha_\theta}{M}\sum_{j=0}^{M-1}\nabla_\theta\mu_{\theta_t}(s'_{t,j})\nabla_\theta\mu_{\theta_t}(s'_{t,j})^T w_t := \alpha_\theta h_{\theta_t}(w_t,\mathcal{B}_t)$. Since $h_{\theta_t}(w_t,\mathcal{B}_t)$ is not an unbiased estimator of the deterministic policy gradient $\nabla J(\theta_t)$, we characterize such a bias by exploiting the compatibility theorem as well as the property of Fisher information of deterministic policy defined in Assumption 1 and obtain the following bound (see Lemma 8 for the proof)

$$\mathbb{E}\left\|h_{\theta_t}(w_t,\mathcal{B}_t) - \nabla J(\theta_t)\right\|^2 \leq 3L_h^2\mathbb{E}\left\|w_t - w^*_{\theta_t}\right\|^2 + 3L_h^2\kappa^2 + \frac{6L_\mu^4C_{w_\xi}^2}{M}.$$

The above bound then connects the critic's error dynamics from Step I to the policy gradient and yields the following result:

$$\mathbb{E}\left\|w_{t+1} - w^*_{\theta_{t+1}}\right\|^2 \leq \left(1 - \frac{\lambda\alpha_w}{8}\right)\mathbb{E}\left\|w_t - w^*_{\theta_t}\right\|^2 + \frac{48\alpha_w^2(C_A^2C_w^2 + C_b^2)}{M} + \frac{8L_w^2\alpha_\theta^2}{\lambda\alpha_w}\mathbb{E}\left\|\nabla J(\theta_t)\right\|^2 + \frac{8L_w^2\alpha_\theta^2}{\lambda\alpha_w}\left(3L_h^2\kappa^2 + \frac{6L_\mu^4C_{w_\xi}^2}{M}\right),$$

where it requires $\alpha_\theta \leq \frac{\lambda\alpha_w}{\sqrt{96}L_hL_w}$.

Thus, we obtain the cumulative dynamic tracking error as

$$\sum_{t=0}^{T-1} \mathbb{E}\left\|w_t - w_{\theta_t}^*\right\|^2 \leq \frac{8\left\|w_0 - w_{\theta_0}^*\right\|^2}{\lambda\alpha_w} + \left[\frac{48\alpha_w^2(C_A^2C_w^2+C_b^2)}{M} + \frac{8L_w^2\alpha_\theta^2}{\lambda\alpha_w}\left(3L_h^2\kappa^2 + \frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\right] \cdot \frac{8T}{\lambda\alpha_w} + \frac{64L_w^2\alpha_\theta^2}{\lambda^2\alpha_w^2}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla J(\theta_t)\right\|^2.$$

The above bound connects the cumulative dynamic tracking error to the convergence rate of actor's update via policy gradient, i.e., such an error depends on how fast actor's update approaches to the stationary point.

Step III: Overall convergence by canceling tracking error via actor's positive progress.

In this step, we establish the overall convergence to a stationary policy by novel cancellation of the above cumulative tracking error via actor's update progress.

We first bound the cumulative policy gradient by the cumulative tracking error via the relationship between the progress of loss function and the tracking error as follows:

$$\frac{\alpha_\theta}{4}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla J(\theta_t)\right\|^2 \leq \frac{9\alpha_\theta L_h^2}{4}\sum_{t=0}^{T-1}\mathbb{E}\left\|w_t - w_{\theta_t}^*\right\|^2 + \frac{R_{\max}}{1-\gamma} + \frac{3\alpha_\theta}{4}\left(3L_h^2\kappa^2 + \frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\cdot T.$$

The previous analysis of (stochastic) AC typically exploits the fact that the above critic's tracking error can decay sufficiently fast by decoupling it from actor's update, which does not hold here. In contrast, we exploit the connection of the cumulative tracking errors and the cumulative policy gradient that we establish in Step II, and show that such a tracking error can ultimately be canceled by the actor's positive progress towards a stationary point. This also explains why the critic's inaccurate estimation does not affect the overall convergence guarantee. Such an idea is captured as follows:

$$\left(\frac{\alpha_\theta}{4} - \frac{144L_h^2L_w^2\alpha_\theta^3}{\lambda^2\alpha_w^2}\right)\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla J(\theta_t)\right\|^2 \leq \frac{R_{\max}}{1-\gamma} + \frac{18\alpha_\theta L_h^2}{\lambda\alpha_w}\left\|w_0 - w_{\theta_0}^*\right\|^2 + \left[\frac{48\alpha_w^2(C_A^2C_w^2+C_b^2)}{M} + \frac{8L_w^2\alpha_\theta^2}{\lambda\alpha_w}\left(3L_h^2\kappa^2 + \frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\right] \cdot \frac{18\alpha_\theta L_h^2T}{\lambda\alpha_w} + \frac{3\alpha_\theta}{4}\left(3L_h^2\kappa^2 + \frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\cdot T.$$

Finally, by letting $\alpha_\theta \leq \frac{\lambda\alpha_w}{24L_hL_w}$ and rearranging the above terms, we complete the proof.

4 OFF-POLICY DPG ALGORITHM

In this section, we consider the off-policy setting, where the behavior policy is different from the target policy, and provide the convergence guarantee of such an off-policy DPG algorithm.

4.1 ALGORITHM

The design of the algorithm is based on the compatibility property in (5). However, off-policy TD with linear function approximation is known to not necessarily converge [Baird, 1995]. To overcome such a divergence issue, Silver et al. [2014] adopted TD with gradient correction (TDC) to update the critic parameter. Note that since critic in DPG estimates the Q-function rather than the value function, there is no need to use the importance sampling to adjust the sampling distribution [Silver et al., 2014]. We call the compatible DPG using TDC updates as DPG-TDC, with its details given in Algorithm 2.

Algorithm 2 DPG-TDC

1: **Input:** $\alpha_w, \eta, \alpha_\theta, w_0, u_0, \theta_0$, batch size M, behavior policy β.
2: **for** $t = 0, 1, \ldots, T$ **do**
3: **for** $j = 0, 1, \ldots, M-1$ **do**
4: Sample $x_{t,j} := (s_{t,j}, a_{t,j}) \sim d_\beta$.
5: Sample $s_{t+1,j} \sim P(\cdot|s_{t,j}, a_{t,j})$ and $r_{t,j}$. Generate $a_{t+1,j} = \mu_{\theta_t}(s_{t+1,j})$.
6: Denote $\phi_{t,j} = \phi(s_{t,j}, a_{t,j})$.
7: $\delta_{t,j} = r_{t,j} + \gamma\phi_{t+1,j}^T w_t - \phi_{t,j}^T w_t$.
8: **end for**
9: $w_{t+1} = w_t + \frac{\alpha_w}{M}\sum_{j=0}^{M-1}\left[\delta_{t,j}\phi_{t,j} - \gamma\phi_{t+1,j}\phi_{t,j}^T u_t\right]$.
10: $u_{t+1} = u_t + \frac{\eta\alpha_w}{M}\sum_{j=0}^{M-1}\left[\delta_{t,j}\phi_{t,j} - \phi_{t,j}\phi_{t,j}^T u_t\right]$.
11: **for** $j = 0, 1, \ldots, M-1$ **do**
12: Sample $s'_{t,j} \sim \nu_{\theta_t}$.
13: **end for**
14: $\theta_{t+1} = \theta_t + \frac{\alpha_\theta}{M}\sum_{j=0}^{M-1}\nabla_\theta\mu_{\theta_t}(s'_{t,j})\nabla_\theta\mu_{\theta_t}(s'_{t,j})^T w_t$.
15: **end for**

As shown in Algorithm 2, while the state-action pair at time t is sampled by the stationary distribution with the behavior policy β, the action corresponding to state s_{t+1} is still generated by the target policy. In addition to the updates of w_t as critic and θ_t as actor, a gradient correction parameter u_t is also updated in line 10 of Algorithm 2.

4.2 CONVERGENCE RESULT

In this subsection, we characterize the convergence rate and sample complexity for DPG-TDC in Algorithm 2. Similarly to the analysis of DPG-TD, we first show the Lipschitz continuity property for off-policy DPG-TDC.

Lemma 2. *Suppose Assumptions 1-3 hold. Then the deterministic policy gradient $\nabla J_\beta(\theta)$ defined in* (4) *is Lipschitz continuous with the parameter L_{J_β}, i.e., $\forall \theta_1, \theta_2 \in \mathbb{R}^d$,*

$$\|\nabla J_\beta(\theta_1) - \nabla J_\beta(\theta_2)\| \leq L_{J_\beta} \|\theta_1 - \theta_2\|, \qquad (8)$$

where $L_{J_\beta} = \frac{L_\mu}{1-\gamma}\left(L_Q L_\mu + \frac{1}{2}\gamma L_P^2 R_{\max} L_\mu C_\nu + \frac{\gamma L_P L_r L_\mu}{1-\gamma}\right) + \frac{L_\psi}{1-\gamma}\left(L_r + \frac{\gamma R_{\max} L_P}{1-\gamma}\right)$.

Compared with Lemma 1 for the on-policy case, the Lipschitz parameter L_{J_β} in Lemma 2 has the same dependence on $1-\gamma$ as L_J, but does not have the state visitation error term because the behavior policy does not change as actor updates the policy.

To analyze the off-policy algorithm DPG-TDC, one of the main challenges lies in the complication arising due to the extra correction parameter u_t. To overcome this, we treat the update of critic as a lifted linear system with respect to a grouped state $z_t = [w_t^T u_t^T]^T \in \mathbb{R}^{2d}$ and then analyze the key properties of such a system matrix.

We next provide the convergence guarantee for DPG-TDC in the following theorem. The full proof of Theorem 2 can be found in Appendix D.

Theorem 2. *Suppose that Assumptions 1-4 hold. Let $\alpha_w \leq \frac{\lambda'}{2C_G^2}; M \geq \frac{48\alpha_w C_G^2}{\lambda'}; \alpha_\theta \leq \min\left\{\frac{1}{4L_{J_\beta}}, \frac{\lambda'\alpha_w}{24 L_h L_{w'}}\right\}; \eta > \max\left\{0, \sigma_{\min}\left(D^{-1}\cdot\frac{A+A^T}{2}\right)\right\}$ where A, D are defined in Assumption 4. Then the output of DPG-TDC in Algorithm 2 satisfies*

$$\min_{t\in[T]} \mathbb{E}\|\nabla J_\beta(\theta_t)\|^2 \leq \frac{c_4}{T} + \frac{c_5}{M} + c_6\kappa^2,$$

where $c_4 = \frac{8R_{\max}}{\alpha_\theta(1-\gamma)} + \frac{144L_h^2}{\lambda'\alpha_w}\left\|z_0 - z_{\theta_0}^*\right\|^2, c_5 = \left[48\alpha_w^2(C_G^2C_w^2 + C_\ell^2) + \frac{48L_{w'}^2 L_\mu^4 C_{w_\xi}^2 \alpha_\theta^2}{\lambda'\alpha_w}\right]\cdot\frac{144L_h^2}{\lambda'\alpha_w} + 36L_\mu^4 C_{w_\xi}^2, c_6 = 18L_h^2 + \frac{24L_{w'}^2 L_h^2\alpha_\theta^2}{\lambda'\alpha_w}$ *with* $C_G^2 = 5(1+\eta^2)C_\phi^4, C_\ell^2 = (1+\eta^2)R_{\max}^2 C_\phi^2, L_{w'} = \frac{L_{J_\beta}}{\lambda_\Psi} + \frac{2L_\mu^2 L_\psi C_Q}{\lambda_\Psi^2(1-\gamma)^2}$, κ *given by* (7)*, L_{J_β} defined in Lemma 2, and other constants remain the same as those in Theorem 1.*

Theorem 2 can readily imply the sample complexity for the convergence of DPG-TDC as given below.

Corollary 3. *Suppose the conditions in Theorem 2 still hold. Then the output of DPG-TDC in Algorithm 2 satisfies*

$$\min_{t\in[T]} \mathbb{E}\|\nabla J_\beta(\theta_t)\|^2 \leq \epsilon + c_6\kappa^2,$$

by using the total number of samples $2MT = \mathcal{O}\left(1/\epsilon^2\right)$.

In Corollary 3, the system error κ is determined by the off-policy distributions, and thus differs from that of the on-policy DPG-TD algorithm. Overall, Corollary 3 shows that off-policy DPG-TDC achieves the same sample complexity as on-policy DPG-TD in Corollary 1 (up to a different system error). To our best knowledge, there has been no existing study on off-policy *stochastic* AC, where critic uses TDC with general nonconvex policy function approximation. Our techniques can be extended to fill such a gap.

5 CONCLUSION

This paper provides the first finite-sample analysis for DPG algorithms in both on-policy (DPG-TD) and off-policy (DPG-TDC) settings. Up to the system error that necessarily exists for actor-critic algorithms, we show that both DPG-TD and DPG-TDC can find an ϵ-accurate stationary policy with a sample complexity of $\mathcal{O}(\epsilon^{-2})$. Our results and the analysis techniques can lead to several promising extension directions. For example, it would be important to explore whether DPG converges to a globally optimal policy as stochastic PG/NPG. Convergence of more popular algorithms such as DDPG is also interesting to study.

References

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *arXiv preprint arXiv:1908.00261*, 2019.

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

Gabriel Barth-Maron, Matthew W. Hoffman, David Budden, Will Dabney, Dan Horgan, Dhruva TB, Alistair Muldal, Nicolas Heess, and Timothy Lillicrap. Distributed distributional deterministic policy gradients. In *International Conference on Learning Representations (ICLR)*, 2018.

Jonathan Baxter and Peter L Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

Dimitri Bertsekas. Convergence of discretization procedures in dynamic programming. *IEEE Transactions on Automatic Control*, 20(3):415–419, 1975.

Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on Learning Theory (COLT)*, 2018.

S. Bhatnagar, M. Ghavamzadeh, M. Lee, and R. S. Sutton. Incremental natural actor-critic algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 105–112, 2008.

Shalabh Bhatnagar. An actor-critic algorithm with function approximation for discounted cost constrained markov decision processes. *Systems & Control Letters*, 59(12): 760–766, 2010.

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

Guillermo A Castillo, Bowen Weng, Ayonga Hereid, Zheng Wang, and Wei Zhang. Reinforcement learning meets hybrid zero dynamics: A case study for rabbit. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 284–290, 2019.

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

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

CHEE-S Chow and John N Tsitsiklis. An optimal one-way multigrid algorithm for discrete-time stochastic control. *IEEE transactions on automatic control*, 36(8):898–914, 1991.

Francois Dufour and Tomas Prieto-Rumeau. Finite linear programming approximations of constrained discounted markov decision processes. *SIAM Journal on Control and Optimization*, 51(2):1298–1324, 2013.

Francois Dufour and Tomas Prieto-Rumeau. Approximation of average cost markov decision processes using empirical distributions and concentration inequalities. *Stochastics An International Journal of Probability and Stochastic Processes*, 87(2):273–307, 2015.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning (ICML)*, pages 1467–1476, 2018.

Zeyue Fu, Zhuoran Yang, and Zhaoran Wang. Single-timescale actor-critic provably finds globally optimal policy. In *International Conference on Learning Representations (ICLR)*, 2021.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning (ICML)*, pages 1587–1596, 2018.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018.

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.

Sham M Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1531–1538, 2002.

Belhal Karimi, Blazej Miasojedow, Eric Moulines, and Hoi-To Wai. Non-asymptotic analysis of biased stochastic approximation scheme. In *Conference on Learning Theory (COLT)*, 2019.

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

Vijaymohan Konda. *Actor-critic algorithms*. PhD thesis, Massachusetts Institute of Technology, 2002.

Vijaymohan R Konda and Vivek S Borkar. Actor-critic–type learning algorithms for markov decision processes. *SIAM Journal on control and Optimization*, 38(1):94–123, 1999.

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

Harshat Kumar, Dionysios S. Kalogerias, George J. Pappas, and Alejandro Ribeiro. Zeroth-order deterministic policy gradient. *arXiv preprint arXiv:2006.07314*, 2020.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2016.

Yanli Liu, Kaiqing Zhang, Tamer Basar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in neural information processing systems (NeurIPS)*, pages 6379–6390, 2017.

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

Hamid Reza Maei. Convergent actor-critic algorithms under off-policy training and function approximation. *arXiv preprint arXiv:1802.07842*, 2018.

Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter Bartlett, and Martin Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2916–2925, 2019.

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

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1928–1937. PMLR, 20–22 Jun 2016.

Matteo Papini, Matteo Pirotta, and Marcello Restelli. Adaptive batch size for safe policy gradients. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3591–3600, 2017.

Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. In *International Conference on Machine Learning (ICML)*, pages 4026–4035, 2018.

Matteo Papini, Matteo Pirotta, and Marcello Restelli. Smoothing policies and safe policy gradients. *arXiv preprint arXiv:1905.03231*, 2019.

Edwin Pednault, Naoki Abe, and Bianca Zadrozny. Sequential cost-sensitive decision making with reinforcement learning. In *Proceedings of the eighth ACM SIGKDD International conference on Knowledge Discovery and Data Mining*, pages 259–268, 2002.

Jan Peters and Stefan Schaal. Natural actor-critic. *Neurocomputing*, 71(7-9):1180–1190, 2008.

Matteo Pirotta, Marcello Restelli, and Luca Bascetta. Policy gradient in lipschitz Markov decision processes. *Machine Learning*, 100(2-3):255–283, 2015.

Shuang Qiu, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. On the finite-time convergence of actor-critic algorithm. In *Optimization Foundations for Reinforcement Learning Workshop at Advances in Neural Information Processing Systems*, 2019.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning (ICML)*, pages 1889–1897, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Devavrat Shah and Qiaomin Xie. Q-learning with nearest neighbors. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3111–3121, 2018.

Han Shen, Kaiqing Zhang, Mingyi Hong, and Tianyi Chen. Asynchronous advantage actor critic: Non-asymptotic analysis and linear speedup. *arXiv preprint arXiv:2012.15511*, 2020.

Zebang Shen, Alejandro Ribeiro, Hamed Hassani, Hui Qian, and Chao Mi. Hessian aided policy gradient. In *International Conference on Machine Learning (ICML)*, pages 5729–5738, 2019.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International Conference on Machine Learning (ICML)*, pages 387–395, 2014.

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

Vladislav B Tadić, Arnaud Doucet, et al. Asymptotic bias of stochastic gradient search. *The Annals of Applied Probability*, 27(6):3255–3304, 2017.

John N. Tsitsiklis and Benjamin Van Roy. An analysis of temporal-diffference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674 – 690, 1997.

Stephen Tu and Benjamin Recht. The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint. In *Conference on Learning Theory (COLT)*, pages 3036–3083, 2019.

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

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

Yue Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite time analysis of two time-scale actor critic methods. *arXiv preprint arXiv:2005.01350*, 2020.

Huaqing Xiong, Tengyu Xu, Yingbin Liang, and Wei Zhang. Non-asymptotic convergence of adam-type reinforcement learning algorithms under markovian sampling. *arXiv preprint arXiv:2002.06286*, 2020.

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

Pan Xu, Felicia Gao, and Quanquan Gu. Sample efficient policy gradient methods with recursive variance reduction. In *International Conference on Learning Representations (ICLR)*, 2020a.

Tengyu Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for actor-critic algorithms. *arXiv preprint arXiv:2004.12956*, 2020b.

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

Zhuoran Yang, Yongxin Chen, Mingyi Hong, and Zhaoran Wang. On the global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. *arXiv preprint arXiv:1907.06246*, 2019.

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

Shangtong Zhang, Bo Liu, Hengshuai Yao, and Shimon Whiteson. Provably convergent two-timescale off-policy actor-critic with function approximation. In *International Conference on Machine Learning*, pages 11204–11213. PMLR, 2020.

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

Supplementary Materials

A PROOF OF LEMMA 1

A.1 SUPPORTING LEMMAS

We first provide some useful lemmas. The first lemma provides the Lipschitz continuity property of the state visitation measure.

Lemma 3. *Suppose Assumptions 1 and 2 hold. We define the total variation norm between two state visitation distributions respectively corresponding to two policies $\mu_{\theta_1}, \mu_{\theta_2}$ as $\|\nu_{\theta_1}(\cdot) - \nu_{\theta_2}(\cdot)\|_{TV} = \int_{\mathcal{S}} |\nu_{\theta_1}(ds) - \nu_{\theta_2}(ds)|$. Then there exists some constant $L_\nu > 0$, such that*

$$\|\nu_{\theta_1}(\cdot) - \nu_{\theta_2}(\cdot)\|_{TV} \leq L_\nu \|\theta_1 - \theta_2\| .$$

Proof. Since we consider ergodic Markov chains, Theorem 3.1 of Mitrophanov [2005] shows that there exists some constant $C_\nu > 1$, such that

$$\|\nu_{\theta_1}(\cdot) - \nu_{\theta_2}(\cdot)\|_{TV} \leq C_\nu \|P_{\theta_1} - P_{\theta_2}\|_{\text{op}} , \tag{9}$$

where P_θ denotes the state transition kernel corresponding to a policy μ_θ, and the operator norm $\|\cdot\|_{\text{op}}$ is given by $\|P\|_{\text{op}} = \sup_{\|q\|_{TV}=1} \|qP\|_{TV}$. Then we have

$$\begin{aligned}
\|P_{\theta_1} - P_{\theta_2}\|_{\text{op}} &= \sup_{\|q\|_{TV}=1} \left\| \int_{\mathcal{S}} (P_{\theta_1} - P_{\theta_2})(s, \cdot) q(ds) \right\|_{TV} \\
&= \frac{1}{2} \sup_{\|q\|_{TV}=1} \int_{s'} \left| \int_s \left(P_{\theta_1}(s, ds') - P_{\theta_2}(s, ds') \right) q(ds) \right| \\
&\leq \frac{1}{2} \sup_{\|q\|_{TV}=1} \int_{s'} \int_s |P_{\theta_1}(s, ds') - P_{\theta_2}(s, ds')| \, q(ds) \\
&= \frac{1}{2} \sup_{\|q\|_{TV}=1} \int_{s'} \int_s |P(ds'|s, \mu_{\theta_1}(s)) - P(ds'|s, \mu_{\theta_1}(s))| \, q(ds) \\
&\overset{\text{(i)}}{\leq} \frac{1}{2} \sup_{\|q\|_{TV}=1} \int_s L_P \|\mu_{\theta_1}(s) - \mu_{\theta_2}(s)\| \, q(ds) \\
&\overset{\text{(ii)}}{\leq} \frac{1}{2} L_P L_\mu \|\theta_1 - \theta_2\| ,
\end{aligned}$$

where (i) follows form Assumption 2, and (ii) follows from Assumption 1. Then, combining the above bound together with (9) completes the proof. □

Next, we show that the value function of a deterministic policy is Lipschitz continuous.

Lemma 4. *Suppose Assumptions 1 and 2 hold. The value function is Lipschitz continuous w.r.t. the policies. That is, for any $\theta_1, \theta_2 \in \mathbb{R}^d, s \in \mathcal{S}$, we have*

$$\|V^{\mu_{\theta_1}}(s) - V^{\mu_{\theta_2}}(s)\| \leq L_V \|\theta_1 - \theta_2\| ,$$

where $L_V = R_{\max} L_\nu + \frac{L_r L_\mu}{1-\gamma}$.

Proof. By definition, we have $V^{\mu_\theta}(s_0) = \int_{\mathcal{S}} r(s, \mu_\theta(s)) \nu_{\mu_\theta}^{s_0}(ds)$, where $\nu_{\mu_\theta}^{s_0}(\cdot)$ is the discounted state visitation measure

given the initial state, i.e., $\nu_{\mu_\theta}^{s_0}(s) = \int_{\mathcal{S}} \sum_{t=0}^{\infty} \gamma^t p(s_0 \to s, t, \mu_\theta) ds$. We then derive

$$
\begin{aligned}
&|V^{\mu_{\theta_1}}(s_0) - V^{\mu_{\theta_2}}(s_0)| \\
&= \left| \int_{\mathcal{S}} r(s, \mu_{\theta_1}(s)) \nu_{\mu_{\theta_1}}^{s_0}(ds) - \int_{\mathcal{S}} r(s, \mu_{\theta_2}(s)) \nu_{\mu_{\theta_2}}^{s_0}(ds) \right| \\
&\leq \left| \int_{\mathcal{S}} r(s, \mu_{\theta_1}(s)) \nu_{\mu_{\theta_1}}^{s_0}(ds) - \int_{\mathcal{S}} r(s, \mu_{\theta_1}(s)) \nu_{\mu_{\theta_2}}^{s_0}(ds) \right| \\
&\quad + \left| \int_{\mathcal{S}} r(s, \mu_{\theta_1}(s)) \nu_{\mu_{\theta_2}}^{s_0}(ds) - \int_{\mathcal{S}} r(s, \mu_{\theta_2}(s)) \nu_{\mu_{\theta_2}}^{s_0}(ds) \right| \\
&\leq \int_{\mathcal{S}} |r(s, \mu_{\theta_1}(s))| \cdot \left| \nu_{\mu_{\theta_1}}^{s_0}(ds) - \nu_{\mu_{\theta_2}}^{s_0}(ds) \right| + \int_{\mathcal{S}} |r(s, \mu_{\theta_1}(s)) - r(s, \mu_{\theta_2}(s))| \, \nu_{\mu_{\theta_2}}^{s_0}(ds) \\
&\overset{\text{(i)}}{\leq} R_{\max} \left\| \nu_{\mu_{\theta_1}}^{s_0}(\cdot) - \nu_{\mu_{\theta_2}}^{s_0}(\cdot) \right\|_{TV} + L_r \int_{\mathcal{S}} \|\mu_{\theta_1}(s) - \mu_{\theta_2}(s)\| \, \nu_{\mu_{\theta_2}}^{s_0}(ds) \\
&\overset{\text{(ii)}}{\leq} R_{\max} L_\nu \|\theta_1 - \theta_2\| + L_r L_\mu \|\theta_1 - \theta_2\| \int_{\mathcal{S}} \nu_{\mu_{\theta_2}}^{s_0}(ds) \\
&= \left(R_{\max} L_\nu + \frac{L_r L_\mu}{1-\gamma} \right) \|\theta_1 - \theta_2\| ,
\end{aligned}
$$

where (i) follows from Assumption 2, and (ii) follows from Lemma 3 and Assumption 1.

□

The next lemma establishes the boundedness and Lipschitz continuity property for the gradient of Q-function.

Lemma 5. *Suppose Assumptions 1-3 hold. The gradient of Q-function w.r.t. action is uniformly bounded. That is, for any $(s, a) \in \mathcal{S} \times \mathcal{A}, \theta \in \mathbb{R}^d$,*

$$\|\nabla_a Q^{\mu_\theta}(s, a)\| \leq C_Q,$$

where $C_Q = L_r + L_P \cdot \frac{\gamma R_{\max}}{1-\gamma}$. Furthermore, $\nabla_a Q^{\mu_\theta}(s, a_\theta)$ is Lipschitz continuous w.r.t. θ, that is, for any $\theta_1, \theta_2 \in \mathbb{R}^d$, we have

$$\|\nabla_a Q^{\mu_{\theta_1}}(s, a_{\theta_1}) - \nabla_a Q^{\mu_{\theta_2}}(s, a_{\theta_2})\| \leq L'_Q \|\theta_1 - \theta_2\| ,$$

where $L'_Q = L_Q L_\mu + \gamma L_P L_V$.

Proof. For the boundedness property, we have

$$
\begin{aligned}
\|\nabla_a Q^{\mu_\theta}(s, a)\| &= \left\| \nabla_a \int_{\mathcal{S}} \left(r(s, a) + \gamma P(s'|s, a) V^{\mu_\theta}(s') \right) ds' \right\| \\
&\leq \|\nabla_a r(s, a)\| + \gamma \int_{\mathcal{S}} \|\nabla_a P(s'|s, a)\| \cdot |V^{\mu_\theta}(s')| \, ds' \\
&\leq L_r + L_P \cdot \frac{\gamma R_{\max}}{1-\gamma},
\end{aligned}
$$

where the last inequality follows from Assumptions 1, 2 and the fact that $|V^{\mu_\theta}(s')| \leq \frac{R_{\max}}{1-\gamma}$.

We next show the Lipschitz property as follows.

$$
\begin{aligned}
&\|\nabla_a Q^{\mu_{\theta_1}}(s, a_{\theta_1}) - \nabla_a Q^{\mu_{\theta_2}}(s, a_{\theta_2})\| \\
&\quad \leq \|\nabla_a Q^{\mu_{\theta_1}}(s, a_{\theta_1}) - \nabla_a Q^{\mu_{\theta_1}}(s, a_{\theta_2})\| + \|\nabla_a Q^{\mu_{\theta_1}}(s, a_{\theta_2}) - \nabla_a Q^{\mu_{\theta_2}}(s, a_{\theta_2})\| \\
&\quad \overset{\text{(i)}}{\leq} L_Q \|a_{\theta_1} - a_{\theta_2}\| + \|\nabla_a Q^{\mu_{\theta_1}}(s, a_{\theta_2}) - \nabla_a Q^{\mu_{\theta_2}}(s, a_{\theta_2})\| \\
&\quad = L_Q \|\mu_{\theta_1}(s) - \mu_{\theta_2}(s)\| + \|\nabla_a Q^{\mu_{\theta_1}}(s, a_{\theta_2}) - \nabla_a Q^{\mu_{\theta_2}}(s, a_{\theta_2})\| \\
&\quad \overset{\text{(ii)}}{\leq} L_Q L_\mu \|\theta_1 - \theta_2\| + \left\| \int_{\mathcal{S}} \gamma \nabla_a P(s'|s, a) \left(V^{\mu_{\theta_1}}(s') - V^{\mu_{\theta_2}}(s') \right) ds' \right\|
\end{aligned}
$$

$$
\begin{aligned}
&\leq L_Q L_\mu \|\theta_1 - \theta_2\| + \gamma \int_{\mathcal{S}} \|\nabla_a P(s'|s,a)\| \cdot |V^{\mu_{\theta_1}}(s') - V^{\mu_{\theta_2}}(s')| \, ds' \\
&\overset{\text{(iii)}}{\leq} (L_Q L_\mu + \gamma L_P L_V) \|\theta_1 - \theta_2\| ,
\end{aligned}
$$

where (i) follows from Assumption 3, (ii) follows from Assumption 1 and (iii) follows from Assumption 2 and Lemma 4.

□

A.2 PROOF OF LEMMA 1

To simplify the notation, we define $\psi_\theta(s) := \nabla_\theta \mu_\theta(s)$, $a_\theta = \mu_\theta(s)$ and $\nabla_a Q^{\mu_\theta}(s, a_\theta) = \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)}$ in the following proof.

We start from the form of the off-policy deterministic policy gradient given in (2), and have

$$
\begin{aligned}
&\|\nabla J(\theta_1) - \nabla J(\theta_2)\| \\
&= \left\| \int_{\mathcal{S}} \psi_{\theta_1}(s) \nabla_a Q^{\mu_{\theta_1}}(s, a_{\theta_1}) \nu_{\theta_1}(ds) - \int_{\mathcal{S}} \psi_{\theta_2}(s) \nabla_a Q^{\mu_{\theta_2}}(s, a_{\theta_2}) \nu_{\theta_2}(ds) \right\| \\
&= \Bigg\| \int_{\mathcal{S}} \psi_{\theta_1}(s) \nabla_a Q^{\mu_{\theta_1}}(s, a_{\theta_1}) \nu_{\theta_1}(ds) - \int_{\mathcal{S}} \psi_{\theta_1}(s) \nabla_a Q^{\mu_{\theta_1}}(s, a_{\theta_1}) \nu_{\theta_2}(ds) \\
&\quad + \int_{\mathcal{S}} \psi_{\theta_1}(s) \nabla_a Q^{\mu_{\theta_1}}(s, a_{\theta_1}) \nu_{\theta_2}(ds) - \int_{\mathcal{S}} \psi_{\theta_1}(s) \nabla_a Q^{\mu_{\theta_2}}(s, a_{\theta_2}) \nu_{\theta_2}(ds) \\
&\quad + \int_{\mathcal{S}} \psi_{\theta_1}(s) \nabla_a Q^{\mu_{\theta_2}}(s, a_{\theta_2}) \nu_{\theta_2}(ds) - \int_{\mathcal{S}} \psi_{\theta_2}(s) \nabla_a Q^{\mu_{\theta_2}}(s, a_{\theta_2}) \nu_{\theta_2}(ds) \Bigg\| \\
&\leq \left\| \int_{\mathcal{S}} \psi_{\theta_1}(s) \nabla_a Q^{\mu_{\theta_1}}(s, a_{\theta_1}) \nu_{\theta_1}(ds) - \int_{\mathcal{S}} \psi_{\theta_1}(s) \nabla_a Q^{\mu_{\theta_1}}(s, a_{\theta_1}) \nu_{\theta_2}(ds) \right\| \\
&\quad + \left\| \int_{\mathcal{S}} \psi_{\theta_1}(s) \nabla_a Q^{\mu_{\theta_1}}(s, a_{\theta_1}) \nu_{\theta_2}(ds) - \int_{\mathcal{S}} \psi_{\theta_1}(s) \nabla_a Q^{\mu_{\theta_2}}(s, a_{\theta_2}) \nu_{\theta_2}(ds) \right\| \\
&\quad + \left\| \int_{\mathcal{S}} \psi_{\theta_1}(s) \nabla_a Q^{\mu_{\theta_2}}(s, a_{\theta_2}) \nu_{\theta_2}(ds) - \int_{\mathcal{S}} \psi_{\theta_2}(s) \nabla_a Q^{\mu_{\theta_2}}(s, a_{\theta_2}) \nu_{\theta_2}(ds) \right\| \\
&\leq \int_{\mathcal{S}} \|\psi_{\theta_1}(s)\| \cdot \|\nabla_a Q^{\mu_{\theta_1}}(s, a_{\theta_1})\| \, |\nu_{\theta_1}(ds) - \nu_{\theta_2}(ds)| \\
&\quad + \int_{\mathcal{S}} \|\psi_{\theta_1}(s)\| \cdot \|\nabla_a Q^{\mu_{\theta_1}}(s, a_{\theta_1}) - \nabla_a Q^{\mu_{\theta_2}}(s, a_{\theta_2})\| \, \nu_{\theta_2}(ds) \\
&\quad + \int_{\mathcal{S}} \|\psi_{\theta_1}(s) - \psi_{\theta_2}(s)\| \cdot \|\nabla_a Q^{\mu_{\theta_2}}(s, a_{\theta_2})\| \, \nu_{\theta_2}(ds) \\
&\overset{\text{(i)}}{\leq} L_\mu C_Q \|\nu_{\theta_1}(\cdot) - \nu_{\theta_2}(\cdot)\|_{TV} + L_\mu \int_{\mathcal{S}} \|\nabla_a Q^{\mu_{\theta_1}}(s, a_{\theta_1}) - \nabla_a Q^{\mu_{\theta_2}}(s, a_{\theta_2})\| \, \nu_{\theta_2}(ds) \\
&\quad + C_Q \int_{\mathcal{S}} \|\psi_{\theta_1}(s) - \psi_{\theta_2}(s)\| \, \nu_{\theta_2}(ds) \\
&\overset{\text{(ii)}}{\leq} L_\mu C_Q \|\nu_{\theta_1}(\cdot) - \nu_{\theta_2}(\cdot)\|_{TV} + L_\mu L'_Q \|\theta_1 - \theta_2\| \int_{\mathcal{S}} \nu_{\theta_2}(ds) + C_Q L_\psi \|\theta_1 - \theta_2\| \int_{\mathcal{S}} \nu_{\theta_2}(ds) \\
&\overset{\text{(iii)}}{=} L_\mu C_Q \|\nu_{\theta_1}(\cdot) - \nu_{\theta_2}(\cdot)\|_{TV} + \frac{L_\mu L'_Q}{1-\gamma} \|\theta_1 - \theta_2\| + \frac{C_Q L_\psi}{1-\gamma} \|\theta_1 - \theta_2\| \\
&\overset{\text{(iv)}}{\leq} \left(L_\mu C_Q L_\nu + \frac{L_\mu L'_Q}{1-\gamma} + \frac{C_Q L_\psi}{1-\gamma} \right) \|\theta_1 - \theta_2\| \\
&:= L_J \|\theta_1 - \theta_2\| ,
\end{aligned}
$$

where (i) follows because $\|\psi_\theta(s)\| \leq L_\mu$ as indicated by Assumption 1 and $\|\nabla_a Q^{\mu_\theta}(s,a)\| \leq C_Q$ by Lemma 5, (ii) follows from Assumption 1 and Lemma 5, (iii) follows because $\int_{\mathcal{S}} \nu_\theta(ds) = \frac{1}{1-\gamma}$, and (iv) follows from Lemma 3.

B PROOF OF THEOREM 1 AND THEOREM 1

B.1 SUPPORTING LEMMAS

In the following, we provide a few supporting lemmas that are used in the main proof of Theorem 1. The first lemma characterizes the properties of mini-batch sampling.

Lemma 6. *The following two properties hold.*

1. *Let $\hat{Y}, \bar{Y} \in \mathbb{R}^{d_1 \times d_2}$ be matrices satisfying $\left\|\hat{Y}\right\|_F \leq C_Y, \left\|\bar{Y}\right\|_F \leq C_Y$. If $\hat{Y}$ is an unbiased estimator of $\bar{Y}$ and $\{\hat{Y}_j\}_j$ are i.i.d. estimators, then we have*

$$\mathbb{E}\left\|\frac{1}{M}\sum_{j=0}^{M-1}\hat{Y}_j - \bar{Y}\right\|_F^2 \leq \frac{4C_Y^2}{M}.$$

2. *Let $\hat{y}, \bar{y} \in \mathbb{R}^d$ be vectors satisfying $\|\hat{y}\| \leq C_y, \|\bar{y}\| \leq C_y$. If $\hat{y}$ is an unbiased estimator of $\bar{y}$ and $\{y_j\}_j$ are i.i.d. estimators, then we have*

$$\mathbb{E}\left\|\frac{1}{M}\sum_{j=0}^{M-1}\hat{y}_j - \bar{y}\right\|^2 \leq \frac{4C_y^2}{M}.$$

Proof. We first prove the first statement of the matrix case as follows.

$$\begin{aligned}
\mathbb{E}\left\|\frac{1}{M}\sum_{j=0}^{M-1}\hat{Y}_j - \bar{Y}\right\|_F^2 &= \frac{1}{M^2}\sum_{c}^{d_2}\sum_{i=0}^{M-1}\sum_{j=0}^{M-1}\mathbb{E}\langle \hat{Y}_i^c - \bar{Y}^c, Y_j^c - \bar{Y}^c\rangle \\
&= \frac{1}{M^2}\sum_{j=0}^{M-1}\mathbb{E}\left\|\hat{Y}_j - \bar{Y}\right\|_F^2 + \frac{1}{M^2}\sum_{c}^{d_2}\sum_{i\neq j}\mathbb{E}\langle \hat{Y}_i^c - \bar{Y}^c, Y_j^c - \bar{Y}^c\rangle \\
&= \frac{1}{M^2}\sum_{j=0}^{M-1}\mathbb{E}\left\|\hat{Y}_j - \bar{Y}\right\|_F^2 \\
&\leq \frac{2}{M^2}\sum_{j=0}^{M-1}\left(\mathbb{E}\left\|\hat{Y}_j\right\|_F^2 + \mathbb{E}\left\|\bar{Y}\right\|_F^2\right) \\
&\leq \frac{4C_Y^2}{M},
\end{aligned}$$

where $\hat{Y}_j^c$ is the c-th column of $\hat{Y}_j$.

We next prove the second statement of the vector case as follows.

$$\begin{aligned}
\mathbb{E}\left\|\frac{1}{M}\sum_{j=0}^{M-1}\hat{y}_j - \bar{y}\right\|^2 &= \frac{1}{M^2}\sum_{i=0}^{M-1}\sum_{j=0}^{M-1}\mathbb{E}\langle \hat{y}_i - \bar{y}, y_j - \bar{y}\rangle \\
&= \frac{1}{M^2}\sum_{j=0}^{M-1}\mathbb{E}\left\|\hat{y}_j - \bar{y}\right\|^2 + \frac{1}{M^2}\sum_{i\neq j}\mathbb{E}\langle \hat{y}_i - \bar{y}, y_j - \bar{y}\rangle \\
&= \frac{1}{M^2}\sum_{j=0}^{M-1}\mathbb{E}\left\|\hat{y}_j - \bar{y}\right\|^2 \\
&\leq \frac{2}{M^2}\sum_{j=0}^{M-1}\left(\mathbb{E}\left\|\hat{y}_j\right\|^2 + \mathbb{E}\left\|\bar{y}\right\|^2\right) \\
&\leq \frac{4C_y^2}{M}.
\end{aligned}$$

□

Next, we provide some important properties of $w^*_{\xi_\theta}$.

Lemma 7. *Let $w^*_{\xi_\theta}$ be defined in Proposition 1. Suppose Assumptions 1-3 hold. Then we have*

$$\left\|w^*_{\xi_\theta}\right\| \leq C_{w_\xi},$$

where $C_{w_\xi} = \frac{L_\mu C_Q}{\lambda_\Psi(1-\gamma)}$. Furthermore, for any θ_1, θ_2, we have

$$\left\|w^*_{\xi_{\theta_1}} - w^*_{\xi_{\theta_2}}\right\| \leq L_w \left\|\theta_1 - \theta_2\right\|,$$

where $L_w = \frac{L_J}{\lambda_\Psi} + \frac{L_\mu C_Q}{\lambda_\Psi^2(1-\gamma)}\left(L_\mu^2 L_\nu + \frac{2L_\mu L_\psi}{1-\gamma}\right)$.

Proof. We first show the boundedness of $\|\nabla J(\theta)\|$.

$$\begin{aligned}
\|\nabla J(\theta)\| &= \left\|\int_{\mathcal{S}} \nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} \nu_\theta(ds)\right\| \\
&\leq \int_{\mathcal{S}} \|\nabla_\theta \mu_\theta(s)\| \left\|\nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)}\right\| \nu_\theta(ds) \\
&\overset{\text{(i)}}{\leq} L_\mu C_Q \int_{\mathcal{S}} \nu_\theta(ds) = \frac{L_\mu C_Q}{(1-\gamma)},
\end{aligned}$$

where (i) follows from Assumption 1 and Lemma 5.

Recall we define $\Psi_\theta = \mathbb{E}_{\nu_{\mu_\theta}}\left[\nabla_\theta \mu_\theta(s) \nabla_\theta \mu_\theta(s)^T\right]$. Assumption 1 implies that Ψ_θ is non-singular. Then by definition, we have

$$\left\|w^*_{\xi_\theta}\right\| = \left\|\Psi_\theta^{-1} \nabla J(\theta)\right\| \leq \frac{1}{\lambda_\Psi} \|\nabla J(\theta)\| \leq \frac{L_\mu C_Q}{\lambda_\Psi(1-\gamma)}.$$

Next, we show the Lipschitz continuity property.

$$\begin{aligned}
&\left\|w^*_{\xi_{\theta_1}} - w^*_{\xi_{\theta_2}}\right\| \\
&= \left\|\Psi_{\theta_1}^{-1} \nabla J(\theta_1) - \Psi_{\theta_2}^{-1} \nabla J(\theta_2)\right\| \\
&= \left\|\Psi_{\theta_1}^{-1} \nabla J(\theta_1) - \Psi_{\theta_1}^{-1} \nabla J(\theta_2) + \Psi_{\theta_1}^{-1} \nabla J(\theta_2) - \Psi_{\theta_2}^{-1} \nabla J(\theta_2)\right\| \\
&\leq \left\|\Psi_{\theta_1}^{-1} (\nabla J(\theta_1) - \nabla J(\theta_2))\right\| + \left\|\left(\Psi_{\theta_1}^{-1} - \Psi_{\theta_2}^{-1}\right) \nabla J(\theta_2)\right\| \\
&\overset{\text{(i)}}{\leq} \frac{L_J}{\lambda_\Psi} \|\theta_1 - \theta_2\| + \left\|\left(\Psi_{\theta_1}^{-1} - \Psi_{\theta_2}^{-1}\right) \nabla J(\theta_2)\right\| \\
&= \frac{L_J}{\lambda_\Psi} \|\theta_1 - \theta_2\| + \left\|\left(\Psi_{\theta_1}^{-1} \Psi_{\theta_2} \Psi_{\theta_2}^{-1} - \Psi_{\theta_1}^{-1} \Psi_{\theta_1} \Psi_{\theta_2}^{-1}\right) \nabla J(\theta_2)\right\| \\
&= \frac{L_J}{\lambda_\Psi} \|\theta_1 - \theta_2\| + \left\|\Psi_{\theta_1}^{-1} \left(\Psi_{\theta_2} - \Psi_{\theta_1}\right) \Psi_{\theta_2}^{-1} \nabla J(\theta_2)\right\| \\
&\leq \frac{L_J}{\lambda_\Psi} \|\theta_1 - \theta_2\| + \frac{1}{\lambda_\Psi^2} \|\Psi_{\theta_2} - \Psi_{\theta_1}\| \|\nabla J(\theta_2)\| \\
&\leq \frac{L_J}{\lambda_\Psi} \|\theta_1 - \theta_2\| + \frac{L_\mu C_Q}{\lambda_\Psi^2(1-\gamma)} \|\Psi_{\theta_2} - \Psi_{\theta_1}\|,
\end{aligned}$$

where (i) follows from Lemma 1 and Assumption 1.

Observe that

$$
\begin{aligned}
&\|\Psi_{\theta_2} - \Psi_{\theta_1}\| \\
&= \left\| \int_{\mathcal{S}} \nabla_\theta \mu_{\theta_2}(s) \nabla_\theta \mu_{\theta_2}(s)^T \nu_{\theta_2}(ds) - \int_{\mathcal{S}} \nabla_\theta \mu_{\theta_1}(s) \nabla_\theta \mu_{\theta_1}(s)^T \nu_{\theta_1}(ds) \right\| \\
&\leq \left\| \int_{\mathcal{S}} \nabla_\theta \mu_{\theta_2}(s) \nabla_\theta \mu_{\theta_2}(s)^T \nu_{\theta_2}(ds) - \int_{\mathcal{S}} \nabla_\theta \mu_{\theta_2}(s) \nabla_\theta \mu_{\theta_2}(s)^T \nu_{\theta_1}(ds) \right\| \\
&\quad + \left\| \int_{\mathcal{S}} \nabla_\theta \mu_{\theta_2}(s) \nabla_\theta \mu_{\theta_2}(s)^T \nu_{\theta_1}(ds) - \int_{\mathcal{S}} \nabla_\theta \mu_{\theta_2}(s) \nabla_\theta \mu_{\theta_1}(s)^T \nu_{\theta_1}(ds) \right\| \\
&\quad + \left\| \int_{\mathcal{S}} \nabla_\theta \mu_{\theta_2}(s) \nabla_\theta \mu_{\theta_1}(s)^T \nu_{\theta_1}(ds) - \int_{\mathcal{S}} \nabla_\theta \mu_{\theta_1}(s) \nabla_\theta \mu_{\theta_1}(s)^T \nu_{\theta_1}(ds) \right\| \\
&\overset{\text{(i)}}{\leq} L_\mu^2 \left\| \nu_{\theta_1}(\cdot) - \nu_{\theta_2}(\cdot) \right\|_{TV} + 2L_\mu \int_{\mathcal{S}} \left\| \nabla_\theta \mu_{\theta_2}(s) - \nabla_\theta \mu_{\theta_1}(s) \right\| \nu_{\theta_1}(ds) \\
&\overset{\text{(ii)}}{\leq} L_\mu^2 \left\| \nu_{\theta_1}(\cdot) - \nu_{\theta_2}(\cdot) \right\|_{TV} + \frac{2L_\mu L_\psi}{1-\gamma} \|\theta_1 - \theta_2\| \\
&\overset{\text{(iii)}}{\leq} \left(L_\mu^2 L_\nu + \frac{2L_\mu L_\psi}{1-\gamma} \right) \|\theta_1 - \theta_2\| ,
\end{aligned}
$$

where both (i) and (ii) follow from Assumption 1, and (iii) follows from Lemma 3.

Thus, we have

$$
\begin{aligned}
&\left\| w^*_{\xi_{\theta_1}} - w^*_{\xi_{\theta_2}} \right\| \\
&\leq \frac{L_J}{\lambda_\Psi} \|\theta_1 - \theta_2\| + \frac{L_\mu C_Q}{\lambda_\Psi^2 (1-\gamma)} \|\Psi_{\theta_2} - \Psi_{\theta_1}\| \\
&\leq \left[\frac{L_J}{\lambda_\Psi} + \frac{L_\mu C_Q}{\lambda_\Psi^2 (1-\gamma)} \left(L_\mu^2 L_\nu + \frac{2L_\mu L_\psi}{1-\gamma} \right) \right] \|\theta_1 - \theta_2\| .
\end{aligned}
$$

□

The next lemma provides an important bound for the difference between the gradient estimators and the true gradient.

Lemma 8. *Suppose Assumptions 1-3. Then we have*

$$
\mathbb{E} \left\| h_{\theta_t}(w_t, \mathcal{B}_t) - \nabla J(\theta_t) \right\|^2 \leq 3L_h^2 \mathbb{E} \left\| w_t - w^*_{\theta_t} \right\|^2 + 3L_h^2 \kappa^2 + \frac{6L_\mu^4 C_{w_\xi}^2}{M},
$$

where $L_h = L_\mu^2$ and κ is defined in (7)*.*

Proof. By definition, we have

$$
\begin{aligned}
&\mathbb{E} \left\| h_{\theta_t}(w_t, \mathcal{B}_t) - \nabla J(\theta_t) \right\|^2 \\
&= \mathbb{E} \left\| h_{\theta_t}(w_t, \mathcal{B}_t) - h_{\theta_t}(w^*_{\theta_t}, \mathcal{B}_t) + h_{\theta_t}(w^*_{\theta_t}, \mathcal{B}_t) - h_{\theta_t}(w^*_{\xi_{\theta_t}}, \mathcal{B}_t) + h_{\theta_t}(w^*_{\xi_{\theta_t}}, \mathcal{B}_t) - \nabla J(\theta_t) \right\|^2 \\
&\leq 3\mathbb{E} \left\| h_{\theta_t}(w_t, \mathcal{B}_t) - h_{\theta_t}(w^*_{\theta_t}, \mathcal{B}_t) \right\|^2 + 3\mathbb{E} \left\| h_{\theta_t}(w^*_{\theta_t}, \mathcal{B}_t) - h_{\theta_t}(w^*_{\xi_{\theta_t}}, \mathcal{B}_t) \right\|^2 \\
&\quad + 3\mathbb{E} \left\| h_{\theta_t}(w^*_{\xi_{\theta_t}}, \mathcal{B}_t) - \nabla J(\theta_t) \right\|^2 \\
&\overset{\text{(i)}}{\leq} 3L_h^2 \mathbb{E} \left\| w_t - w^*_{\theta_t} \right\|^2 + 3L_h^2 \mathbb{E} \left\| w^*_{\theta_t} - w^*_{\xi_{\theta_t}} \right\|^2 + 3\mathbb{E} \left\| h_{\theta_t}(w^*_{\xi_{\theta_t}}, \mathcal{B}_t) - \nabla J(\theta_t) \right\|^2 \\
&\overset{\text{(ii)}}{\leq} 3L_h^2 \mathbb{E} \left\| w_t - w^*_{\theta_t} \right\|^2 + 3L_h^2 \kappa^2 + 3\mathbb{E} \left\| h_{\theta_t}(w^*_{\xi_{\theta_t}}, \mathcal{B}_t) - \nabla J(\theta_t) \right\|^2 \\
&\overset{\text{(iii)}}{\leq} 3L_h^2 \mathbb{E} \left\| w_t - w^*_{\theta_t} \right\|^2 + 3L_h^2 \kappa^2 + \frac{6L_\mu^4 C_{w_\xi}^2}{M},
\end{aligned}
$$

where (i) follows because for any $w_1, w_2, \theta \in \mathbb{R}^d$, we have

$$\|h_\theta(w_1, \mathcal{B}_t) - h_\theta(w_2, \mathcal{B}_t)\| = \left\| \frac{1}{M} \sum_{j=0}^{M-1} \nabla_\theta \mu_{\theta_t}(s'_{t,j}) \nabla_\theta \mu_{\theta_t}(s'_{t,j})^T (w_1 - w_2) \right\| \\ \leq L_\mu^2 \|w_1 - w_2\| := L_h \|w_1 - w_2\|,$$

(ii) follows from (7), and (iii) holds due to the fact that

$$\begin{aligned}
&\mathbb{E} \left\| h_{\theta_t}(w^*_{\xi_{\theta_t}}, \mathcal{B}_t) - \nabla J(\theta_t) \right\|^2 \\
&= \mathbb{E} \left\| \frac{1}{M} \sum_{j=0}^{M-1} \nabla_\theta \mu_{\theta_t}(s'_{t,j}) \nabla_\theta \mu_{\theta_t}(s'_{t,j})^T w^*_{\xi_{\theta_t}} - \nabla J(\theta_t) \right\|^2 \\
&= \frac{1}{M^2} \sum_{i=0}^{M-1} \sum_{j=0}^{M-1} \mathbb{E} \langle \nabla_\theta \mu_{\theta_t}(s'_{t,i}) \nabla_\theta \mu_{\theta_t}(s'_{t,i})^T w^*_{\xi_{\theta_t}} - \nabla J(\theta_t), \\
&\quad \nabla_\theta \mu_{\theta_t}(s'_{t,j}) \nabla_\theta \mu_{\theta_t}(s'_{t,j})^T w^*_{\xi_{\theta_t}} - \nabla J(\theta_t) \rangle \\
&= \frac{1}{M^2} \sum_{j=0}^{M-1} \mathbb{E} \left\| \nabla_\theta \mu_{\theta_t}(s'_{t,j}) \nabla_\theta \mu_{\theta_t}(s'_{t,j})^T w^*_{\xi_{\theta_t}} - \nabla J(\theta_t) \right\|^2 \\
&\overset{\text{(i)}}{\leq} \frac{1}{M^2} \sum_{j=0}^{M-1} 2 L_\mu^4 C^2_{w_\xi} = \frac{2 L_\mu^4 C^2_{w_\xi}}{M},
\end{aligned}$$

where (i) follows from Assumption 1, Lemma 6 and Lemma 7. □

B.2 PROOF OF THEOREM 1

We use the following notations for the clarity of the presentation:

$$\begin{aligned}
g_{\theta_t}(w_t, \mathcal{B}_t) &= \frac{1}{M} \sum_{j=0}^{M-1} \delta_{t,j} \phi(x_{t,j}) = \frac{1}{M} \sum_{j=0}^{M-1} (A_{t,j} w_t + b_{t,j}) := \hat{A}_t w_t + \hat{b}_t; \\
\bar{g}_{\theta_t}(w_t) &= \mathbb{E}_{d_{\theta_t}} [\delta_t \phi(x_t)] = \bar{A}_t w_t + \bar{b}_t; \\
\bar{g}_{\theta_t}(w^*_{\theta_t}) &= \bar{A}_t w^*_{\theta_t} + \bar{b}_t = 0; \\
h_{\theta_t}(w_t, \mathcal{B}_t) &= \frac{1}{M} \sum_{j=0}^{M-1} \nabla_\theta \mu_{\theta_t}(s'_{t,j}) \nabla_\theta \mu_{\theta_t}(s'_{t,j})^T w_t.
\end{aligned}$$

In this proof, we develop a new approach to analyzing the coupled actor and critic's stochastic approximation processes, due to their simultaneous updates both with constant stepsizes. The central idea is to cancel the critic's cumulative tracking error by the actor's overall positive progress to the stationary policy, which is different from the existing analysis of (stochastic) PG-type algorithms that mainly decouples or asymptotically decouples the critic's error from actor's error. Further, we develop a new analysis to bound the estimation error of the Fisher information of deterministic policy arising via the compatibility theorem, and then further capture how such a metric affects the convergence via its minimum eigenvalue.

The main proof consists of three steps.

Step I: Characterizing dynamics of critic's error via coupling with actor.

In the first step, we characterize the propagation of the dynamics of critic's dynamic tracking error based on its coupling with actor's updates. That is, we develop the relationship between $\left\| w_{t+1} - w^*_{\theta_{t+1}} \right\|^2$ and $\left\| w_t - w^*_{\theta_t} \right\|^2$ by their coupling with actor's updates.

We first use the dynamics of the critic to obtain

$$
\begin{aligned}
&\left\|w_{t+1}-w_{\theta_t}^*\right\|^2 \\
&= \left\|w_t+\alpha_w g_{\theta_t}(w_t,\mathcal{B}_t)-w_{\theta_t}^*\right\|^2 \\
&= \left\|w_t-w_{\theta_t}^*\right\|^2 + 2\alpha_w\langle w_t-w_{\theta_t}^*, g_{\theta_t}(w_t,\mathcal{B}_t)\rangle + \alpha_w^2\left\|g_{\theta_t}(w_t,\mathcal{B}_t)\right\|^2 \\
&= \left\|w_t-w_{\theta_t}^*\right\|^2 + 2\alpha_w\langle w_t-w_{\theta_t}^*, \bar{g}_{\theta_t}(w_t)\rangle + 2\alpha_w\langle w_t-w_{\theta_t}^*, g_{\theta_t}(w_t,\mathcal{B}_t)-\bar{g}_{\theta_t}(w_t)\rangle \\
&\quad + \alpha_w^2\left\|g_{\theta_t}(w_t,\mathcal{B}_t)\right\|^2 \\
&= \left\|w_t-w_{\theta_t}^*\right\|^2 + 2\alpha_w(w_t-w_{\theta_t}^*)^T\bar{A}_t(w_t-w_{\theta_t}^*) + 2\alpha_w\langle w_t-w_{\theta_t}^*, g_{\theta_t}(w_t,\mathcal{B}_t)-\bar{g}_{\theta_t}(w_t)\rangle \\
&\quad + \alpha_w^2\left\|g_{\theta_t}(w_t,\mathcal{B}_t)\right\|^2 \\
&\overset{\text{(i)}}{\leq} (1-2\alpha_w\lambda)\left\|w_t-w_{\theta_t}^*\right\|^2 + 2\alpha_w\langle w_t-w_{\theta_t}^*, g_{\theta_t}(w_t,\mathcal{B}_t)-\bar{g}_{\theta_t}(w_t)\rangle + \alpha_w^2\left\|g_{\theta_t}(w_t,\mathcal{B}_t)\right\|^2 \\
&\leq (1-2\alpha_w\lambda)\left\|w_t-w_{\theta_t}^*\right\|^2 + 2\alpha_w\langle w_t-w_{\theta_t}^*, g_{\theta_t}(w_t,\mathcal{B}_t)-\bar{g}_{\theta_t}(w_t)\rangle \\
&\quad + 2\alpha_w^2\left\|g_{\theta_t}(w_t,\mathcal{B}_t)-\bar{g}_{\theta_t}(w_t)\right\|^2 + 2\alpha_w^2\left\|\bar{g}_{\theta_t}(w_t)\right\|^2 \\
&= (1-2\alpha_w\lambda)\left\|w_t-w_{\theta_t}^*\right\|^2 + 2\alpha_w\langle w_t-w_{\theta_t}^*, g_{\theta_t}(w_t,\mathcal{B}_t)-\bar{g}_{\theta_t}(w_t)\rangle \\
&\quad + 2\alpha_w^2\left\|g_{\theta_t}(w_t,\mathcal{B}_t)-\bar{g}_{\theta_t}(w_t)\right\|^2 + 2\alpha_w^2\left\|\bar{g}_{\theta_t}(w_t)-\bar{g}_{\theta_t}(w_{\theta_t}^*)\right\|^2 \\
&= (1-2\alpha_w\lambda)\left\|w_t-w_{\theta_t}^*\right\|^2 + 2\alpha_w\langle w_t-w_{\theta_t}^*, g_{\theta_t}(w_t,\mathcal{B}_t)-\bar{g}_{\theta_t}(w_t)\rangle \\
&\quad + 2\alpha_w^2\left\|g_{\theta_t}(w_t,\mathcal{B}_t)-\bar{g}_{\theta_t}(w_t)\right\|^2 + 2\alpha_w^2\left\|\bar{A}_t(w_t-w_{\theta_t}^*)\right\|^2 \\
&\leq (1-2\alpha_w\lambda)\left\|w_t-w_{\theta_t}^*\right\|^2 + 2\alpha_w\langle w_t-w_{\theta_t}^*, g_{\theta_t}(w_t,\mathcal{B}_t)-\bar{g}_{\theta_t}(w_t)\rangle \\
&\quad + 2\alpha_w^2\left\|g_{\theta_t}(w_t,\mathcal{B}_t)-\bar{g}_{\theta_t}(w_t)\right\|^2 + 2\alpha_w^2\left\|\bar{A}_t\right\|^2\left\|(w_t-w_{\theta_t}^*)\right\|^2 \\
&\overset{\text{(ii)}}{\leq} (1-2\alpha_w\lambda+2\alpha_w^2C_A^2)\left\|w_t-w_{\theta_t}^*\right\|^2 + 2\alpha_w\langle w_t-w_{\theta_t}^*, g_{\theta_t}(w_t,\mathcal{B}_t)-\bar{g}_{\theta_t}(w_t)\rangle \\
&\quad + 2\alpha_w^2\left\|g_{\theta_t}(w_t,\mathcal{B}_t)-\bar{g}_{\theta_t}(w_t)\right\|^2,
\end{aligned}
$$

where (i) follows from the property $(w_t-w_{\theta_t}^*)^T\bar{A}_t(w_t-w_{\theta_t}^*) \leq -\lambda\left\|w_t-w_{\theta_t}^*\right\|^2$ with some constant $\lambda>0$ for any policy, which has been proved in Tsitsiklis and Van Roy [1997], Bhandari et al. [2018], Tu and Recht [2019], Xiong et al. [2020], and (ii) follows because $\|A\|^2 \leq 2(1+\gamma^2)C_\phi^4 \leq 4C_\phi^4 := C_A^2$.

Taking the expectation on both sides yields

$$
\begin{aligned}
&\mathbb{E}\left\|w_{t+1}-w_{\theta_t}^*\right\|^2 \\
&\leq (1-2\alpha_w\lambda+2\alpha_w^2C_A^2)\mathbb{E}\left\|w_t-w_{\theta_t}^*\right\|^2 + 2\alpha_w\mathbb{E}\langle w_t-w_{\theta_t}^*, g_{\theta_t}(w_t,\mathcal{B}_t)-\bar{g}_{\theta_t}(w_t)\rangle \\
&\quad + 2\alpha_w^2\mathbb{E}\left\|g_{\theta_t}(w_t,\mathcal{B}_t)-\bar{g}_{\theta_t}(w_t)\right\|^2 \\
&= (1-2\alpha_w\lambda+2\alpha_w^2C_A^2)\mathbb{E}\left\|w_t-w_{\theta_t}^*\right\|^2 + 2\alpha_w^2\mathbb{E}\left\|g_{\theta_t}(w_t,\mathcal{B}_t)-\bar{g}_{\theta_t}(w_t)\right\|^2. \qquad (10)
\end{aligned}
$$

Observe that

$$
\begin{aligned}
&\mathbb{E}\left\|g_{\theta_t}(w_t,\mathcal{B}_t)-\bar{g}_{\theta_t}(w_t)\right\|^2 \\
&= \mathbb{E}\left\|\hat{A}_tw_t+\hat{b}_t-\bar{A}_tw_t-\bar{b}_t\right\|^2 \\
&\overset{\text{(i)}}{\leq} 3\mathbb{E}\left\|(\hat{A}_t-\bar{A}_t)(w_t-w_{\theta_t}^*)\right\|^2 + 3\mathbb{E}\left\|(\hat{A}_t-\bar{A}_t)w_{\theta_t}^*\right\|^2 + 3\mathbb{E}\left\|\hat{b}_t-\bar{b}_t\right\|^2 \\
&\leq 3\mathbb{E}\left\|\hat{A}_t-\bar{A}_t\right\|_F^2\left\|w_t-w_{\theta_t}^*\right\|^2 + 3\mathbb{E}\left\|\hat{A}_t-\bar{A}_t\right\|_F^2\left\|w_{\theta_t}^*\right\|^2 + 3\mathbb{E}\left\|\hat{b}_t-\bar{b}_t\right\|^2 \\
&\overset{\text{(ii)}}{\leq} \frac{12C_A^2}{M}\mathbb{E}\left\|w_t-w_{\theta_t}^*\right\|^2 + \frac{12(C_A^2\mathbb{E}\left\|w_{\theta_t}^*\right\|^2+C_b^2)}{M} \\
&\overset{\text{(iii)}}{\leq} \frac{12C_A^2}{M}\mathbb{E}\left\|w_t-w_{\theta_t}^*\right\|^2 + \frac{12(C_A^2C_w^2+C_b^2)}{M},
\end{aligned}
$$

where (i) follows because $(x+y+z)^2 \leq 3x^2+3y^2+3z^2$, (ii) follows from Lemma 6 and $C_b := R_{\max}C_\phi \geq \|b\|$, and (iii) follows because $\left\|w_{\theta_t}^*\right\|^2 = \left\|\bar{A}_t^{-1}\bar{b}_t\right\|^2 \leq C_b/\lambda_A = R_{\max}C_\phi/\lambda_A := C_w$ by Assumption 4.

Substituting the above bound into (10), we have

$$
\begin{aligned}
&\mathbb{E}\left\|w_{t+1}-w_{\theta_t}^*\right\|^2 \\
&\quad\leq (1-2\alpha_w\lambda+2\alpha_w^2C_A^2)\mathbb{E}\left\|w_t-w_{\theta_t}^*\right\|^2 + 2\alpha_w^2\mathbb{E}\left\|g_{\theta_t}(w_t,\mathcal{B}_t)-\bar{g}_{\theta_t}(w_t)\right\|^2 \\
&\quad\leq \left(1-2\alpha_w\lambda+2\alpha_w^2C_A^2+\frac{24\alpha_w^2C_A^2}{M}\right)\mathbb{E}\left\|w_t-w_{\theta_t}^*\right\|^2 + \frac{24\alpha_w^2(C_A^2C_w^2+C_b^2)}{M}.
\end{aligned}
$$

Since $\alpha_w \leq \frac{\lambda}{2C_A^2}; M \geq \frac{48\alpha_w C_A^2}{\lambda}$, we further obtain

$$
\begin{aligned}
&\mathbb{E}\left\|w_{t+1}-w_{\theta_t}^*\right\|^2 \\
&\quad\leq \left(1-2\alpha_w\lambda+2\alpha_w^2C_A^2+\frac{24\alpha_w^2C_A^2}{M}\right)\mathbb{E}\left\|w_t-w_{\theta_t}^*\right\|^2 + \frac{24\alpha_w^2(C_A^2C_w^2+C_b^2)}{M} \\
&\quad\leq \left(1-\frac{\alpha_w\lambda}{2}\right)\mathbb{E}\left\|w_t-w_{\theta_t}^*\right\|^2 + \frac{24\alpha_w^2(C_A^2C_w^2+C_b^2)}{M}.
\end{aligned} \tag{11}
$$

Next, we use Young's inequality, and obtain

$$
\begin{aligned}
&\mathbb{E}\left\|w_{t+1}-w_{\theta_{t+1}}^*\right\|^2 \\
&\quad\leq \left(1+\frac{1}{2(2/\lambda\alpha_w-1)}\right)\mathbb{E}\left\|w_{t+1}-w_{\theta_t}^*\right\|^2 + (1+2(2/\lambda\alpha_w-1))\,\mathbb{E}\left\|w_{\theta_t}^*-w_{\theta_{t+1}}^*\right\|^2 \\
&\quad\overset{\text{(i)}}{\leq} \left(1-\frac{\lambda\alpha_w}{4}\right)\mathbb{E}\left\|w_t-w_{\theta_t}^*\right\|^2 + \frac{4-\lambda\alpha_w}{4-2\lambda\alpha_w}\cdot\frac{24\alpha_w^2(C_A^2C_w^2+C_b^2)}{M} + \frac{4}{\lambda\alpha_w}\mathbb{E}\left\|w_{\theta_t}^*-w_{\theta_{t+1}}^*\right\|^2 \\
&\quad\overset{\text{(ii)}}{\leq} \left(1-\frac{\lambda\alpha_w}{4}\right)\mathbb{E}\left\|w_t-w_{\theta_t}^*\right\|^2 + \frac{4-\lambda\alpha_w}{4-2\lambda\alpha_w}\cdot\frac{24\alpha_w^2(C_A^2C_w^2+C_b^2)}{M} + \frac{4L_w^2}{\lambda\alpha_w}\mathbb{E}\left\|\theta_{t+1}-\theta_t\right\|^2,
\end{aligned} \tag{12}
$$

where (i) follows from the bound derived in (11), and (ii) follows from Lemma 7.

Step II: Bounding cumulative tracking error via compatibility theorem for DPG.

In this step, we bound the cumulative tracking error based on the dynamics of the tracking error from the last step. To this end, we need to first bound the difference between two consecutive actor parameters.

Observe that $\theta_{t+1}-\theta_t = \frac{1}{M}\sum_{j=0}^{M-1}\nabla_\theta\mu_{\theta_t}(s'_{t,j})\nabla_\theta\mu_{\theta_t}(s'_{t,j})^T w_t := h_{\theta_t}(w_t,\mathcal{B}_t)$ and $\mathbb{E}\left\|h_{\theta_t}(w_t,\mathcal{B}_t)\right\|^2 \leq 2\mathbb{E}\left\|\nabla J(\theta_t)\right\|^2 + 2\mathbb{E}\left\|h_{\theta_t}(w_t,\mathcal{B}_t)-\nabla J(\theta_t)\right\|^2$. We proceed to bound (12) as follows

$$
\begin{aligned}
&\mathbb{E}\left\|w_{t+1}-w_{\theta_{t+1}}^*\right\|^2 \\
&\quad\leq \left(1-\frac{\lambda\alpha_w}{4}\right)\mathbb{E}\left\|w_t-w_{\theta_t}^*\right\|^2 + \frac{4-\lambda\alpha_w}{4-2\lambda\alpha_w}\cdot\frac{24\alpha_w^2(C_A^2C_w^2+C_b^2)}{M} + \frac{4L_w^2}{\lambda\alpha_w}\mathbb{E}\left\|\theta_{t+1}-\theta_t\right\|^2 \\
&\quad\leq \left(1-\frac{\lambda\alpha_w}{4}\right)\mathbb{E}\left\|w_t-w_{\theta_t}^*\right\|^2 + \frac{48\alpha_w^2(C_A^2C_w^2+C_b^2)}{M} + \frac{8L_w^2\alpha_\theta^2}{\lambda\alpha_w}\mathbb{E}\left\|\nabla J(\theta_t)\right\|^2 \\
&\qquad + \frac{8L_w^2\alpha_\theta^2}{\lambda\alpha_w}\mathbb{E}\left\|h_{\theta_t}(w_t,\mathcal{B}_t)-\nabla J(\theta_t)\right\|^2 \\
&\quad\overset{\text{(i)}}{\leq} \left(1-\frac{\lambda\alpha_w}{4}+\frac{24L_h^2L_w^2\alpha_\theta^2}{\lambda\alpha_w}\right)\mathbb{E}\left\|w_t-w_{\theta_t}^*\right\|^2 + \frac{48\alpha_w^2(C_A^2C_w^2+C_b^2)}{M} + \frac{8L_w^2\alpha_\theta^2}{\lambda\alpha_w}\mathbb{E}\left\|\nabla J(\theta_t)\right\|^2 \\
&\qquad + \frac{8L_w^2\alpha_\theta^2}{\lambda\alpha_w}\left(3L_h^2\kappa^2+\frac{6L_\mu^4C_{w_\xi}^2}{M}\right) \\
&\quad\overset{\text{(ii)}}{\leq} \left(1-\frac{\lambda\alpha_w}{8}\right)\mathbb{E}\left\|w_t-w_{\theta_t}^*\right\|^2 + \frac{8L_w^2\alpha_\theta^2}{\lambda\alpha_w}\mathbb{E}\left\|\nabla J(\theta_t)\right\|^2 + \frac{48\alpha_w^2(C_A^2C_w^2+C_b^2)}{M}
\end{aligned}
$$

$$+ \frac{8L_w^2\alpha_\theta^2}{\lambda\alpha_w}\left(3L_h^2\kappa^2 + \frac{6L_\mu^4 C_{w_\xi}^2}{M}\right), \tag{13}$$

where (i) follows from Lemma 8, and (ii) follows because $\alpha_\theta \leq \frac{\lambda\alpha_w}{\sqrt{96}L_h L_w}$.

We further take the summation over all iterations on both sides of (13) and have

$$\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}\left\|w_t - w_{\theta_t}^*\right\|^2 \\
&\leq \sum_{t=0}^{T-1}\left(1-\frac{\lambda\alpha_w}{8}\right)^t\left\|w_0 - w_{\theta_0}^*\right\|^2 + \frac{8L_w^2\alpha_\theta^2}{\lambda\alpha_w}\sum_{t=0}^{T-1}\sum_{i=0}^{t-1}\left(1-\frac{\lambda\alpha_w}{8}\right)^{t-1-i}\mathbb{E}\left\|\nabla J(\theta_t)\right\|^2 \\
&\quad + \left[\frac{48\alpha_w^2(C_A^2C_w^2 + C_b^2)}{M} + \frac{8L_w^2\alpha_\theta^2}{\lambda\alpha_w}\left(3L_h^2\kappa^2 + \frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\right]\sum_{t=0}^{T-1}\sum_{i=0}^{t-1}\left(1-\frac{\lambda\alpha_w}{8}\right)^{t-1-i} \\
&\leq \frac{8\left\|w_0 - w_{\theta_0}^*\right\|^2}{\lambda\alpha_w} + \left[\frac{48\alpha_w^2(C_A^2C_w^2 + C_b^2)}{M} + \frac{8L_w^2\alpha_\theta^2}{\lambda\alpha_w}\left(3L_h^2\kappa^2 + \frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\right]\cdot\frac{8T}{\lambda\alpha_w} \\
&\quad + \frac{64L_w^2\alpha_\theta^2}{\lambda^2\alpha_w^2}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla J(\theta_t)\right\|^2.
\end{aligned} \tag{14}$$

Step III: Overall convergence by canceling tracking error via actor's positive progress.

In this step, we establish the overall convergence to a stationary policy by novel cancellation of the above cumulative tracking error via actor's update progress.

Based on Lemma 1, we have

$$\begin{aligned}
&\mathbb{E}[J(\theta_{t+1})] - \mathbb{E}[J(\theta_t)] \\
&\geq \mathbb{E}\langle\nabla J(\theta_t), \theta_{t+1} - \theta_t\rangle - \frac{L_J}{2}\mathbb{E}\left\|\theta_{t+1} - \theta_t\right\|^2 \\
&= \alpha_\theta\mathbb{E}\langle\nabla J(\theta_t), h_{\theta_t}(w_t, \mathcal{B}_t)\rangle - \frac{L_J\alpha_\theta^2}{2}\mathbb{E}\left\|h_{\theta_t}(w_t, \mathcal{B}_t)\right\|^2 \\
&= \alpha_\theta\mathbb{E}\left\|\nabla J(\theta_t)\right\|^2 + \alpha_\theta\mathbb{E}\langle\nabla J(\theta_t), h_{\theta_t}(w_t, \mathcal{B}_t) - \nabla J(\theta_t)\rangle - \frac{L_J\alpha_\theta^2}{2}\mathbb{E}\left\|h_{\theta_t}(w_t, \mathcal{B}_t)\right\|^2 \\
&\overset{\text{(i)}}{\geq} \frac{\alpha_\theta}{2}\mathbb{E}\left\|\nabla J(\theta_t)\right\|^2 - \frac{\alpha_\theta}{2}\mathbb{E}\left\|h_{\theta_t}(w_t, \mathcal{B}_t) - \nabla J(\theta_t)\right\|^2 \\
&\quad - \frac{L_J\alpha_\theta^2}{2}\mathbb{E}\left\|h_{\theta_t}(w_t, \mathcal{B}_t) - \nabla J(\theta_t) + \nabla J(\theta_t)\right\|^2 \\
&\geq \left(\frac{\alpha_\theta}{2} - L_J\alpha_\theta^2\right)\mathbb{E}\left\|\nabla J(\theta_t)\right\|^2 - \left(\frac{\alpha_\theta}{2} + L_J\alpha_\theta^2\right)\mathbb{E}\left\|h_{\theta_t}(w_t, \mathcal{B}_t) - \nabla J(\theta_t)\right\|^2 \\
&\overset{\text{(ii)}}{\geq} \left(\frac{\alpha_\theta}{2} - L_J\alpha_\theta^2\right)\mathbb{E}\left\|\nabla J(\theta_t)\right\|^2 - \left(\frac{\alpha_\theta}{2} + L_J\alpha_\theta^2\right)\left(3L_h^2\mathbb{E}\left\|w_t - w_{\theta_t}^*\right\|^2 + 3L_h^2\kappa^2 + \frac{6L_\mu^4C_{w_\xi}^2}{M}\right) \\
&\overset{\text{(iii)}}{\geq} \frac{\alpha_\theta}{4}\mathbb{E}\left\|\nabla J(\theta_t)\right\|^2 - \frac{3\alpha_\theta}{4}\left(3L_h^2\mathbb{E}\left\|w_t - w_{\theta_t}^*\right\|^2 + 3L_h^2\kappa^2 + \frac{6L_\mu^4C_{w_\xi}^2}{M}\right),
\end{aligned} \tag{15}$$

where (i) follows because $x^Ty \geq -\frac{1}{2}x^2 - \frac{1}{2}y^2$, (ii) follows from Lemma 8, and (iii) follows from the condition $\alpha_\theta \leq \frac{1}{4L_J}$.

We next take the summation over all iterations on both sides of the above bound and obtain

$$\begin{aligned}
&\frac{\alpha_\theta}{4}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla J(\theta_t)\right\|^2 \\
&\leq \mathbb{E}[J(\theta_{T+1})] - \mathbb{E}[J(\theta_0)] + \frac{3\alpha_\theta}{4}\left(3L_h^2\kappa^2 + \frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\cdot T + \frac{9\alpha_\theta L_h^2}{4}\sum_{t=0}^{T-1}\mathbb{E}\left\|w_t - w_{\theta_t}^*\right\|^2
\end{aligned}$$

$$\leq \frac{R_{\max}}{1-\gamma} + \frac{3\alpha_\theta}{4}\left(3L_h^2\kappa^2 + \frac{6L_\mu^4 C_{w_\xi}^2}{M}\right)\cdot T + \frac{9\alpha_\theta L_h^2}{4}\sum_{t=0}^{T-1}\mathbb{E}\left\|w_t - w_{\theta_t}^*\right\|^2. \tag{16}$$

Substituting the cumulative tracking error bound derived in (14) into (16) yields

$$\begin{aligned}
&\frac{\alpha_\theta}{8}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla J(\theta_t)\right\|^2 \\
&\overset{\text{(i)}}{\leq}\left(\frac{\alpha_\theta}{4} - \frac{144L_h^2L_w^2\alpha_\theta^3}{\lambda^2\alpha_w^2}\right)\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla J(\theta_t)\right\|^2 \\
&\leq \frac{R_{\max}}{1-\gamma} + \frac{3\alpha_\theta}{4}\left(3L_h^2\kappa^2 + \frac{6L_\mu^4 C_{w_\xi}^2}{M}\right)\cdot T + \frac{18\alpha_\theta L_h^2}{\lambda\alpha_w}\left\|w_0 - w_{\theta_0}^*\right\|^2 \\
&\quad + \left[\frac{48\alpha_w^2(C_A^2C_w^2 + C_b^2)}{M} + \frac{8L_w^2\alpha_\theta^2}{\lambda\alpha_w}\left(3L_h^2\kappa^2 + \frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\right]\cdot\frac{18\alpha_\theta L_h^2 T}{\lambda\alpha_w},
\end{aligned}$$

where (i) follows from the condition $\alpha_\theta \leq \frac{\lambda\alpha_w}{24L_hL_w}$.

Finally, we have

$$\begin{aligned}
\min_{t\in[T]}\mathbb{E}\left\|\nabla J(\theta_t)\right\|^2 &\leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla J(\theta_t)\right\|^2 \\
&\leq \left(\frac{8R_{\max}}{\alpha_\theta(1-\gamma)} + \frac{144L_h^2}{\lambda\alpha_w}\left\|w_0 - w_{\theta_0}^*\right\|^2\right)\cdot\frac{1}{T} + 6\left(3L_h^2\kappa^2 + \frac{6L_\mu^4C_{w_\xi}^2}{M}\right) \\
&\quad + \left[\frac{48\alpha_w^2(C_A^2C_w^2 + C_b^2)}{M} + \frac{8L_w^2\alpha_\theta^2}{\lambda\alpha_w}\left(3L_h^2\kappa^2 + \frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\right]\cdot\frac{144L_h^2}{\lambda\alpha_w} \\
&= \frac{c_1}{T} + \frac{c_2}{M} + c_3\kappa^2,
\end{aligned}$$

where

$$c_1 = \frac{8R_{\max}}{\alpha_\theta(1-\gamma)} + \frac{144L_h^2}{\lambda\alpha_w}\left\|w_0 - w_{\theta_0}^*\right\|^2, \tag{17}$$

$$c_2 = 36L_\mu^4C_{w_\xi}^2 + \left[48\alpha_w^2(C_A^2C_w^2 + C_b^2) + \frac{48L_w^2L_\mu^4C_{w_\xi}^2\alpha_\theta^2}{\lambda\alpha_w}\right]\cdot\frac{144L_h^2}{\lambda\alpha_w}, \tag{18}$$

$$c_3 = 18L_h^2 + \frac{24L_w^2L_h^2\alpha_\theta^2}{\lambda\alpha_w}. \tag{19}$$

B.3 PROOF OF COROLLARY 1

Following from the upper bound in Theorem 1, we let $\frac{c_1}{T} \leq \frac{\epsilon}{2}$ and $\frac{c_2}{M} \leq \frac{\epsilon}{2}$ to achieve the ϵ-accuracy. Then we obtain $T \geq \frac{2c_1}{\epsilon}$ and $M \geq \frac{2c_2}{\epsilon}$. Further, since we generate M samples in the update steps of both critic and actor in Algorithm 1, the total number of samples we use is thus $2MT = \frac{8c_1c_2}{\epsilon^2}$.

C PROOF OF LEMMA 2

We use the notations $\psi_\theta(s) := \nabla_\theta\mu_\theta(s)$, $a_\theta = \mu_\theta(s)$ and $\nabla_a Q^{\mu_\theta}(s, a_\theta) = \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)}$ in the following proof.

We start from the form of the deterministic policy gradient given in (4), and have

$$
\begin{aligned}
&\|\nabla J_\beta(\theta_1) - \nabla J_\beta(\theta_2)\| \\
&= \left\| \int_{\mathcal{S}} \psi_{\theta_1}(s) \nabla_a Q^{\mu_{\theta_1}}(s, a_{\theta_1}) \nu_\beta(ds) - \int_{\mathcal{S}} \psi_{\theta_2}(s) \nabla_a Q^{\mu_{\theta_2}}(s, a_{\theta_2}) \nu_\beta(ds) \right\| \\
&= \left\| \int_{\mathcal{S}} \psi_{\theta_1}(s) \nabla_a Q^{\mu_{\theta_1}}(s, a_{\theta_1}) \nu_\beta(ds) - \int_{\mathcal{S}} \psi_{\theta_1}(s) \nabla_a Q^{\mu_{\theta_2}}(s, a_{\theta_2}) \nu_\beta(ds) \right. \\
&\quad \left. + \int_{\mathcal{S}} \psi_{\theta_1}(s) \nabla_a Q^{\mu_{\theta_2}}(s, a_{\theta_2}) \nu_\beta(ds) - \int_{\mathcal{S}} \psi_{\theta_2}(s) \nabla_a Q^{\mu_{\theta_2}}(s, a_{\theta_2}) \nu_\beta(ds) \right\| \\
&\leq \left\| \int_{\mathcal{S}} \psi_{\theta_1}(s) \nabla_a Q^{\mu_{\theta_1}}(s, a_{\theta_1}) \nu_\beta(ds) - \int_{\mathcal{S}} \psi_{\theta_1}(s) \nabla_a Q^{\mu_{\theta_2}}(s, a_{\theta_2}) \nu_\beta(ds) \right\| \\
&\quad + \left\| \int_{\mathcal{S}} \psi_{\theta_1}(s) \nabla_a Q^{\mu_{\theta_2}}(s, a_{\theta_2}) \nu_\beta(ds) - \int_{\mathcal{S}} \psi_{\theta_2}(s) \nabla_a Q^{\mu_{\theta_2}}(s, a_{\theta_2}) \nu_\beta(ds) \right\| \\
&\leq \int_{\mathcal{S}} \|\psi_{\theta_1}(s)\| \cdot \|\nabla_a Q^{\mu_{\theta_1}}(s, a_{\theta_1}) - \nabla_a Q^{\mu_{\theta_2}}(s, a_{\theta_2})\| \, \nu_\beta(ds) \\
&\quad + \int_{\mathcal{S}} \|\psi_{\theta_1}(s) - \psi_{\theta_2}(s)\| \cdot \|\nabla_a Q^{\mu_{\theta_2}}(s, a_{\theta_2})\| \, \nu_\beta(ds) \\
&\overset{\text{(i)}}{\leq} L_\mu \int_{\mathcal{S}} \|\nabla_a Q^{\mu_{\theta_1}}(s, a_{\theta_1}) - \nabla_a Q^{\mu_{\theta_2}}(s, a_{\theta_2})\| \, \nu_\beta(ds) + C_Q \int_{\mathcal{S}} \|\psi_{\theta_1}(s) - \psi_{\theta_2}(s)\| \, \nu_\beta(ds) \\
&\overset{\text{(ii)}}{\leq} L_\mu L'_Q \|\theta_1 - \theta_2\| \int_{\mathcal{S}} \nu_\beta(ds) + C_Q L_\psi \|\theta_1 - \theta_2\| \int_{\mathcal{S}} \nu_\beta(ds) \\
&\overset{\text{(iii)}}{=} \left(\frac{L_\mu L'_Q}{1-\gamma} + \frac{C_Q L_\psi}{1-\gamma} \right) \|\theta_1 - \theta_2\| \\
&:= L_{J_\beta} \|\theta_1 - \theta_2\|,
\end{aligned}
$$

where (i) follows because $\|\psi_\theta(s)\| \leq L_\mu$ as indicated by Assumption 1 and $\|\nabla_a Q^{\mu_\theta}(s, a)\| \leq C_Q$ by Lemma 5, (ii) follows from Assumption 1 and Lemma 5, and (iii) follows because $\int_{\mathcal{S}} \nu_\beta(ds) = \frac{1}{1-\gamma}$.

D PROOF OF THEOREM 2 AND COROLLARY 3

D.1 SUPPORTING LEMMAS

The following lemma provides the important properties of w^*_{β,ξ_θ}.

Lemma 9. *Let w^*_{β,ξ_θ} be defined in* (5)*. Suppose Assumptions 1-3 hold. Then we have*

$$
\left\| w^*_{\beta,\xi_\theta} \right\| \leq C_{w_\xi},
$$

where $C_{w_\xi} = \frac{L_\mu C_Q}{\lambda_\Psi (1-\gamma)}$. Furthermore, for any θ_1, θ_2, we have

$$
\left\| w^*_{\beta,\xi_{\theta_1}} - w^*_{\beta,\xi_{\theta_2}} \right\| \leq L_{w'} \|\theta_1 - \theta_2\|,
$$

where $L_{w'} = \frac{L_{J_\beta}}{\lambda_\Psi} + \frac{2L_\mu^2 L_\psi C_Q}{\lambda_\Psi^2 (1-\gamma)^2}$.

Proof. We first show the boundedness of $\|\nabla J_\beta(\theta)\|$.

$$
\begin{aligned}
\|\nabla J_\beta(\theta)\| &= \left\| \int_{\mathcal{S}} \nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s, a)|_{a=\mu_\theta(s)} \nu_\beta(ds) \right\| \\
&\leq \int_{\mathcal{S}} \|\nabla_\theta \mu_\theta(s)\| \left\| \nabla_a Q^{\mu_\theta}(s, a)|_{a=\mu_\theta(s)} \right\| \nu_\beta(ds) \\
&\overset{\text{(i)}}{\leq} L_\mu C_Q \int_{\mathcal{S}} \nu_\beta(ds) = \frac{L_\mu C_Q}{(1-\gamma)},
\end{aligned}
$$

where (i) follows from Assumption 1 and Lemma 5.

We define $\Psi_{\beta,\theta} = \mathbb{E}_{\nu_{\mu_\beta}}\left[\nabla_\theta \mu_\theta(s) \nabla_\theta \mu_\theta(s)^T\right]$. Assumption 1 implies that $\Psi_{\beta,\theta}$ is non-singular. Then by definition, we have

$$\left\|w^*_{\beta,\xi_\theta}\right\| = \left\|\Psi^{-1}_{\beta,\theta}\nabla J_\beta(\theta)\right\| \leq \frac{1}{\lambda_\Psi}\left\|\nabla J_\beta(\theta)\right\| \leq \frac{L_\mu C_Q}{\lambda_\Psi(1-\gamma)}.$$

Next, we show the Lipschitz continuity property.

$$\begin{aligned}
&\left\|w^*_{\xi_{\theta_1}} - w^*_{\xi_{\theta_2}}\right\| \\
&= \left\|\Psi^{-1}_{\beta,\theta_1}\nabla J_\beta(\theta_1) - \Psi^{-1}_{\beta,\theta_2}\nabla J_\beta(\theta_2)\right\| \\
&= \left\|\Psi^{-1}_{\beta,\theta_1}\nabla J_\beta(\theta_1) - \Psi^{-1}_{\beta,\theta_1}\nabla J_\beta(\theta_2) + \Psi^{-1}_{\beta,\theta_1}\nabla J_\beta(\theta_2) - \Psi^{-1}_{\beta,\theta_2}\nabla J_\beta(\theta_2)\right\| \\
&\leq \left\|\Psi^{-1}_{\beta,\theta_1}(\nabla J_\beta(\theta_1) - \nabla J_\beta(\theta_2))\right\| + \left\|\left(\Psi^{-1}_{\beta,\theta_1} - \Psi^{-1}_{\beta,\theta_2}\right)\nabla J_\beta(\theta_2)\right\| \\
&\overset{\text{(i)}}{\leq} \frac{L_J}{\lambda_\Psi}\left\|\theta_1 - \theta_2\right\| + \left\|\left(\Psi^{-1}_{\beta,\theta_1} - \Psi^{-1}_{\beta,\theta_2}\right)\nabla J_\beta(\theta_2)\right\| \\
&= \frac{L_J}{\lambda_\Psi}\left\|\theta_1 - \theta_2\right\| + \left\|\left(\Psi^{-1}_{\beta,\theta_1}\Psi_{\beta,\theta_2}\Psi^{-1}_{\beta,\theta_2} - \Psi^{-1}_{\beta,\theta_1}\Psi_{\beta,\theta_1}\Psi^{-1}_{\beta,\theta_2}\right)\nabla J_\beta(\theta_2)\right\| \\
&= \frac{L_J}{\lambda_\Psi}\left\|\theta_1 - \theta_2\right\| + \left\|\Psi^{-1}_{\beta,\theta_1}\left(\Psi_{\beta,\theta_2} - \Psi_{\beta,\theta_1}\right)\Psi^{-1}_{\beta,\theta_2}\nabla J_\beta(\theta_2)\right\| \\
&\leq \frac{L_J}{\lambda_\Psi}\left\|\theta_1 - \theta_2\right\| + \frac{1}{\lambda^2_\Psi}\left\|\Psi_{\beta,\theta_2} - \Psi_{\beta,\theta_1}\right\|\left\|\nabla J_\beta(\theta_2)\right\| \\
&\leq \frac{L_J}{\lambda_\Psi}\left\|\theta_1 - \theta_2\right\| + \frac{L_\mu C_Q}{\lambda^2_\Psi(1-\gamma)}\left\|\Psi_{\beta,\theta_2} - \Psi_{\beta,\theta_1}\right\|,
\end{aligned}$$

where (i) follows from Lemma 1 and Assumption 1.

We further derive the following bound.

$$\begin{aligned}
&\left\|\Psi_{\beta,\theta_2} - \Psi_{\beta,\theta_1}\right\| \\
&= \left\|\int_{\mathcal{S}}\nabla_\theta\mu_{\theta_2}(s)\nabla_\theta\mu_{\theta_2}(s)^T\nu_\beta(ds) - \int_{\mathcal{S}}\nabla_\theta\mu_{\theta_1}(s)\nabla_\theta\mu_{\theta_1}(s)^T\nu_\beta(ds)\right\| \\
&\leq \left\|\int_{\mathcal{S}}\nabla_\theta\mu_{\theta_2}(s)\nabla_\theta\mu_{\theta_2}(s)^T\nu_\beta(ds) - \int_{\mathcal{S}}\nabla_\theta\mu_{\theta_2}(s)\nabla_\theta\mu_{\theta_1}(s)^T\nu_\beta(ds)\right\| \\
&\quad + \left\|\int_{\mathcal{S}}\nabla_\theta\mu_{\theta_2}(s)\nabla_\theta\mu_{\theta_1}(s)^T\nu_\beta(ds) - \int_{\mathcal{S}}\nabla_\theta\mu_{\theta_1}(s)\nabla_\theta\mu_{\theta_1}(s)^T\nu_\beta(ds)\right\| \\
&\overset{\text{(i)}}{\leq} 2L_\mu\int_{\mathcal{S}}\left\|\nabla_\theta\mu_{\theta_2}(s) - \nabla_\theta\mu_{\theta_1}(s)\right\|\nu_\beta(ds) \\
&\overset{\text{(ii)}}{\leq} \frac{2L_\mu L_\psi}{1-\gamma}\left\|\theta_1 - \theta_2\right\|,
\end{aligned}$$

where both (i) and (ii) follow from Assumption 1.

Thus, we have

$$\begin{aligned}
&\left\|w^*_{\beta,\xi_{\theta_1}} - w^*_{\beta,\xi_{\theta_2}}\right\| \\
&\leq \frac{L_{J_\beta}}{\lambda_\Psi}\left\|\theta_1 - \theta_2\right\| + \frac{L_\mu C_Q}{\lambda^2_\Psi(1-\gamma)}\left\|\Psi_{\beta,\theta_2} - \Psi_{\beta,\theta_1}\right\| \\
&\leq \left(\frac{L_{J_\beta}}{\lambda_\Psi} + \frac{2L^2_\mu L_\psi C_Q}{\lambda^2_\Psi(1-\gamma)^2}\right)\left\|\theta_1 - \theta_2\right\|.
\end{aligned}$$

□

D.2 PROOF OF THEOREM 2

The main difference here from the proof of Theorem 1 lies in the fact that we apply TDC to update critic in Algorithm 2 due to the off-policy sampling, which introduces an extra correction parameter u_t. Thus, we introduce a grouped vector $z_t = [w_t^T u_t^T]^T \in \mathbb{R}^{2d}$ and rewrite the dynamics of critic as a lifted linear system:

$$\begin{aligned} z_{t+1} &= z_t + \alpha_w \begin{bmatrix} \hat{A}_t & \hat{C}_t \\ \eta\hat{A}_t & \eta\hat{D}_t \end{bmatrix} z_t + \alpha_w \begin{bmatrix} \hat{b}_t \\ \eta\hat{b}_t \end{bmatrix} \\ &:= z_t + \alpha_w \left[\hat{G}_t z_t + \hat{\ell}_t\right] \\ &:= z_t + \alpha_w g_{\theta_t}(z_t, \mathcal{B}_t). \end{aligned}$$

TDC algorithm is designed to find the fixed point $w^*_{\beta,\theta}$ satisfying $\bar{A}w^*_{\beta,\theta} + \bar{b} = 0$, where $\bar{A} = \mathbb{E}_{d_\beta}\left[\hat{A}\right], \bar{b} = \mathbb{E}_{d_\beta}\left[\hat{b}\right]$. Correspondingly, if we let $z^*_\theta = [{w^*_{\beta,\theta}}^T \mathbf{0}^T]^T$, then we have

$$\bar{g}_\theta(z^*_\theta) = \bar{G}z^*_\theta + \bar{\ell} = 0,$$

where $\bar{G} = \mathbb{E}_{d_\beta}\left[\hat{G}\right], \bar{\ell} = \mathbb{E}_{d_\beta}\left[\hat{\ell}\right]$. Based on the above lifted linear system, we proceed our proof as follows.

The main proof consists of three steps.

Step I: Characterizing dynamics of critic's error via coupling with actor.

In the following, we first characterize the relationship between $\left\|z_{t+1} - z^*_{\theta_{t+1}}\right\|^2$ and $\left\|z_t - z^*_{\theta_t}\right\|^2$.

We first use the dynamics of the above linear system to obtain

$$\begin{aligned} &\left\|z_{t+1} - z^*_{\theta_t}\right\|^2 \\ &= \left\|z_t + \alpha_w g_{\theta_t}(z_t, \mathcal{B}_t) - z^*_{\theta_t}\right\|^2 \\ &= \left\|z_t - z^*_{\theta_t}\right\|^2 + 2\alpha_w \langle z_t - z^*_{\theta_t}, g_{\theta_t}(z_t, \mathcal{B}_t)\rangle + \alpha_w^2 \left\|g_{\theta_t}(z_t, \mathcal{B}_t)\right\|^2 \\ &= \left\|z_t - z^*_{\theta_t}\right\|^2 + 2\alpha_w \langle z_t - z^*_{\theta_t}, \bar{g}_{\theta_t}(z_t)\rangle + 2\alpha_w \langle z_t - z^*_{\theta_t}, g_{\theta_t}(z_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(z_t)\rangle \\ &\quad + \alpha_w^2 \left\|g_{\theta_t}(z_t, \mathcal{B}_t)\right\|^2 \\ &\overset{\text{(i)}}{\leq} (1 - 2\alpha_w\lambda') \left\|z_t - z^*_{\theta_t}\right\|^2 + 2\alpha_w \langle z_t - z^*_{\theta_t}, g_{\theta_t}(z_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(z_t)\rangle + \alpha_w^2 \left\|g_{\theta_t}(z_t, \mathcal{B}_t)\right\|^2 \\ &\leq (1 - 2\alpha_w\lambda') \left\|z_t - z^*_{\theta_t}\right\|^2 + 2\alpha_w \langle z_t - z^*_{\theta_t}, g_{\theta_t}(z_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(z_t)\rangle \\ &\quad + 2\alpha_w^2 \left\|g_{\theta_t}(z_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(z_t)\right\|^2 + 2\alpha_w^2 \left\|\bar{g}_{\theta_t}(z_t)\right\|^2 \\ &= (1 - 2\alpha_w\lambda') \left\|z_t - z^*_{\theta_t}\right\|^2 + 2\alpha_w \langle z_t - z^*_{\theta_t}, g_{\theta_t}(z_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(z_t)\rangle \\ &\quad + 2\alpha_w^2 \left\|g_{\theta_t}(z_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(z_t)\right\|^2 + 2\alpha_w^2 \left\|\bar{g}_{\theta_t}(z_t) - \bar{g}_{\theta_t}(z^*_{\theta_t})\right\|^2 \\ &= (1 - 2\alpha_w\lambda') \left\|z_t - z^*_{\theta_t}\right\|^2 + 2\alpha_w \langle z_t - z^*_{\theta_t}, g_{\theta_t}(z_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(z_t)\rangle \\ &\quad + 2\alpha_w^2 \left\|g_{\theta_t}(z_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(z_t)\right\|^2 + 2\alpha_w^2 \left\|\bar{G}_t(z_t - z^*_{\theta_t})\right\|^2 \\ &\leq (1 - 2\alpha_w\lambda') \left\|z_t - z^*_{\theta_t}\right\|^2 + 2\alpha_w \langle z_t - z^*_{\theta_t}, g_{\theta_t}(z_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(z_t)\rangle \\ &\quad + 2\alpha_w^2 \left\|g_{\theta_t}(z_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(z_t)\right\|^2 + 2\alpha_w^2 \left\|\bar{G}_t\right\|_F^2 \left\|(z_t - z^*_{\theta_t})\right\|^2 \\ &\overset{\text{(ii)}}{\leq} (1 - 2\alpha_w\lambda' + 2\alpha_w^2 C_G^2) \left\|z_t - z^*_{\theta_t}\right\|^2 + 2\alpha_w \langle z_t - z^*_{\theta_t}, g_{\theta_t}(z_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(z_t)\rangle \\ &\quad + 2\alpha_w^2 \left\|g_{\theta_t}(z_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(z_t)\right\|^2, \end{aligned}$$

where (i) follows from the property $\langle z_t - z^*_{\theta_t}, \bar{g}_{\theta_t}(z_t)\rangle \leq -\lambda' \left\|z_t - z^*_{\theta_t}\right\|^2$ with some constant $\lambda' > 0$ for any policy which has been proved in Theorem 3 of Maei [2011] as long as $\eta > \max\left\{0, \sigma_{\min}\left(D^{-1} \cdot \frac{A+A^T}{2}\right)\right\}$, and (ii) follows from $\left\|\bar{G}\right\|_F^2 = (1+\eta^2)\left\|\bar{A}\right\|_F^2 + \left\|\bar{C}\right\|_F^2 + \eta^2\left\|\bar{D}\right\|_F^2 \leq (1+\eta^2)C_A^2 + C_C^2 + \eta^2 C_D^2 \leq 5(1+\eta^2)C_\phi^4 := C_G^2$.

Taking the expectation on both sides of the above bound yields

$$
\begin{aligned}
&\mathbb{E}\left\|z_{t+1}-z_{\theta_t}^*\right\|^2\\
&\leq (1-2\alpha_w\lambda'+2\alpha_w^2C_G^2)\mathbb{E}\left\|z_t-z_{\theta_t}^*\right\|^2+2\alpha_w\mathbb{E}\langle z_t-z_{\theta_t}^*, g_{\theta_t}(z_t,\mathcal{B}_t)-\bar{g}_{\theta_t}(z_t)\rangle\\
&\quad+2\alpha_w^2\mathbb{E}\left\|g_{\theta_t}(z_t,\mathcal{B}_t)-\bar{g}_{\theta_t}(z_t)\right\|^2\\
&=(1-2\alpha_w\lambda'+2\alpha_w^2C_G^2)\mathbb{E}\left\|z_t-z_{\theta_t}^*\right\|^2+2\alpha_w^2\mathbb{E}\left\|g_{\theta_t}(z_t,\mathcal{B}_t)-\bar{g}_{\theta_t}(z_t)\right\|^2\\
&=(1-2\alpha_w\lambda'+2\alpha_w^2C_G^2)\mathbb{E}\left\|z_t-z_{\theta_t}^*\right\|^2+2\alpha_w^2\mathbb{E}\left\|\hat{G}_tz_t+\hat{\ell}_t-\bar{G}_tz_t-\bar{\ell}_t\right\|^2\\
&\leq (1-2\alpha_w\lambda'+2\alpha_w^2C_G^2)\mathbb{E}\left\|z_t-z_{\theta_t}^*\right\|^2\\
&\quad+6\alpha_w^2\left(\mathbb{E}\left\|(\hat{G}_t-\bar{G}_t)(z_t-z_{\theta_t}^*)\right\|^2+\mathbb{E}\left\|(\hat{G}_t-\bar{G}_t)z_{\theta_t}^*\right\|^2+\mathbb{E}\left\|\hat{\ell}_t-\bar{\ell}_t\right\|^2\right)\\
&\leq (1-2\alpha_w\lambda'+2\alpha_w^2C_G^2)\mathbb{E}\left\|z_t-z_{\theta_t}^*\right\|^2\\
&\quad+6\alpha_w^2\left(\mathbb{E}\left\|\hat{G}_t-\bar{G}_t\right\|_F^2\left\|z_t-z_{\theta_t}^*\right\|^2+\mathbb{E}\left\|\hat{G}_t-\bar{G}_t\right\|_F^2\left\|z_{\theta_t}^*\right\|^2+\mathbb{E}\left\|\hat{\ell}_t-\bar{\ell}_t\right\|^2\right)\\
&\overset{\text{(i)}}{\leq} (1-2\alpha_w\lambda'+2\alpha_w^2C_G^2)\mathbb{E}\left\|z_t-z_{\theta_t}^*\right\|^2+6\alpha_w^2\left(\frac{4C_G^2}{M}\mathbb{E}\left\|z_t-z_{\theta_t}^*\right\|^2+\frac{4(C_G^2\mathbb{E}\left\|z_{\theta_t}^*\right\|^2+C_\ell^2)}{M}\right)\\
&\overset{\text{(ii)}}{=}\left(1-2\alpha_w\lambda'+2\alpha_w^2C_G^2+\frac{24\alpha_w^2C_G^2}{M}\right)\mathbb{E}\left\|z_t-z_{\theta_t}^*\right\|^2+\frac{24\alpha_w^2\left(C_G^2\mathbb{E}\left\|w_{\beta,\theta_t}^*\right\|^2+C_\ell^2\right)}{M}\\
&\overset{\text{(iii)}}{\leq}\left(1-2\alpha_w\lambda'+2\alpha_w^2C_G^2+\frac{24\alpha_w^2C_G^2}{M}\right)\mathbb{E}\left\|z_t-z_{\theta_t}^*\right\|^2+\frac{24\alpha_w^2(C_G^2C_w^2+C_\ell^2)}{M}\\
&\overset{\text{(iv)}}{\leq}\left(1-\frac{\alpha_w\lambda'}{2}\right)\mathbb{E}\left\|z_t-z_{\theta_t}^*\right\|^2+\frac{24\alpha_w^2(C_G^2C_w^2+C_\ell^2)}{M},
\end{aligned}
\tag{20}
$$

where (i) follows from Lemma 6, (ii) follows from $\left\|z_{\theta_t}^*\right\|^2=\left\|w_{\beta,\theta_t}^*\right\|^2$, (iii) follows because $\left\|w_{\beta,\theta_t}^*\right\|^2=\left\|\bar{A}_t^{-1}\bar{b}_t\right\|^2\leq C_b/\lambda_A=R_{\max}C_\phi/\lambda_A:=C_w$ by Assumption 4, and (iv) follows from the conditions $\alpha_w\leq\frac{\lambda'}{2C_G^2}$ and $M\geq\frac{48\alpha_wC_G^2}{\lambda'}$.

We further derive that

$$
\begin{aligned}
&\mathbb{E}\left\|z_{t+1}-z_{\theta_{t+1}}^*\right\|^2\\
&\overset{\text{(i)}}{\leq}\left(1+\frac{1}{2(2/\lambda'\alpha_w-1)}\right)\mathbb{E}\left\|z_{t+1}-z_{\theta_t}^*\right\|^2+(1+2(2/\lambda'\alpha_w-1))\,\mathbb{E}\left\|z_{\theta_t}^*-z_{\theta_{t+1}}^*\right\|^2\\
&\overset{\text{(ii)}}{\leq}\left(1-\frac{\lambda'\alpha_w}{4}\right)\mathbb{E}\left\|z_t-z_{\theta_t}^*\right\|^2+\frac{48\alpha_w^2(C_G^2C_w^2+C_\ell^2)}{M}+\frac{4}{\lambda'\alpha_w}\mathbb{E}\left\|z_{\theta_t}^*-z_{\theta_{t+1}}^*\right\|^2\\
&\overset{\text{(iii)}}{=}\left(1-\frac{\lambda'\alpha_w}{4}\right)\mathbb{E}\left\|z_t-z_{\theta_t}^*\right\|^2+\frac{48\alpha_w^2(C_G^2C_w^2+C_\ell^2)}{M}+\frac{4}{\lambda'\alpha_w}\mathbb{E}\left\|w_{\beta,\theta_t}^*-w_{\beta,\theta_{t+1}}^*\right\|^2\\
&\overset{\text{(iv)}}{\leq}\left(1-\frac{\lambda'\alpha_w}{4}\right)\mathbb{E}\left\|z_t-z_{\theta_t}^*\right\|^2+\frac{48\alpha_w^2(C_G^2C_w^2+C_\ell^2)}{M}+\frac{4L_{w'}^2}{\lambda'\alpha_w}\mathbb{E}\left\|\theta_{t+1}-\theta_t\right\|^2,
\end{aligned}
\tag{21}
$$

where (i) follows from Young's inequality, (ii) follows from the bound derived in (20), (iii) follows because $\left\|z_{\theta_t}^*-z_{\theta_{t+1}}^*\right\|^2=\left\|w_{\beta,\theta_t}^*-w_{\beta,\theta_{t+1}}^*\right\|^2$, and (iv) follows from Lemma 9.

Step II: Bounding cumulative tracking error via compatibility theorem for DPG.

Recall that we define $h_{\theta_t}(w_t,\mathcal{B}_t):=\frac{1}{M}\sum_{j=0}^{M-1}\nabla_\theta\mu_{\theta_t}(s'_{t,j})\nabla_\theta\mu_{\theta_t}(s'_{t,j})^Tw_t$. We continue with (21) and have

$$
\mathbb{E}\left\|z_{t+1}-z_{\theta_{t+1}}^*\right\|^2
$$

$$
\begin{aligned}
&\leq \left(1-\frac{\lambda'\alpha_w}{4}\right)\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2 + \frac{48\alpha_w^2(C_G^2C_w^2+C_\ell^2)}{M} + \frac{4L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\mathbb{E}\left\|h_{\theta_t}(w_t,\mathcal{B}_t)\right\|^2 \\
&\leq \left(1-\frac{\lambda'\alpha_w}{4}\right)\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2 + \frac{48\alpha_w^2(C_G^2C_w^2+C_\ell^2)}{M} + \frac{8L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\mathbb{E}\left\|\nabla J_\beta(\theta_t)\right\|^2 \\
&\quad + \frac{8L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\mathbb{E}\left\|h_{\theta_t}(w_t,\mathcal{B}_t) - \nabla J_\beta(\theta_t)\right\|^2 \\
&\overset{\text{(i)}}{\leq} \left(1-\frac{\lambda'\alpha_w}{4}\right)\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2 + \frac{48\alpha_w^2(C_G^2C_w^2+C_\ell^2)}{M} + \frac{8L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\mathbb{E}\left\|\nabla J_\beta(\theta_t)\right\|^2 \\
&\quad + \frac{8L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\left(3L_h^2\mathbb{E}\left\|w_t - w^*_{\theta_t}\right\|^2 + 3L_h^2\kappa^2 + \frac{6L_\mu^4C_{w_\xi}^2}{M}\right) \\
&\overset{\text{(ii)}}{\leq} \left(1-\frac{\lambda'\alpha_w}{4}\right)\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2 + \frac{48\alpha_w^2(C_G^2C_w^2+C_\ell^2)}{M} + \frac{8L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\mathbb{E}\left\|\nabla J_\beta(\theta_t)\right\|^2 \\
&\quad + \frac{8L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\left(3L_h^2\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2 + 3L_h^2\kappa^2 + \frac{6L_\mu^4C_{w_\xi}^2}{M}\right) \\
&= \left(1-\frac{\lambda'\alpha_w}{4} + \frac{24L_h^2L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\right)\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2 + \frac{48\alpha_w^2(C_G^2C_w^2+C_\ell^2)}{M} \\
&\quad + \frac{8L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\mathbb{E}\left\|\nabla J_\beta(\theta_t)\right\|^2 + \frac{8L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\left(3L_h^2\kappa^2 + \frac{6L_\mu^4C_{w_\xi}^2}{M}\right) \\
&\overset{\text{(iii)}}{\leq} \left(1-\frac{\lambda'\alpha_w}{8}\right)\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2 + \frac{48\alpha_w^2(C_G^2C_w^2+C_\ell^2)}{M} + \frac{8L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\mathbb{E}\left\|\nabla J_\beta(\theta_t)\right\|^2 \\
&\quad + \frac{8L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\left(3L_h^2\kappa^2 + \frac{6L_\mu^4C_{w_\xi}^2}{M}\right),
\end{aligned} \tag{22}
$$

where (i) follows from Lemma 8, (ii) follows since $\left\|w_t - w^*_{\theta_t}\right\|^2 \leq \left\|z_t - z^*_{\theta_t}\right\|^2$, and (iii) follows because $\alpha_\theta \leq \frac{\lambda'\alpha_w}{\sqrt{96}L_hL_{w'}}$. Then, taking the summation over all iterations on both sides of (22) yields

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2 \\
&\leq \sum_{t=0}^{T-1}\left(1-\frac{\lambda'\alpha_w}{8}\right)^t\left\|z_0 - z^*_{\theta_0}\right\|^2 + \frac{8L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\sum_{t=0}^{T-1}\sum_{i=0}^{t-1}\left(1-\frac{\lambda'\alpha_w}{8}\right)^{t-1-i}\mathbb{E}\left\|\nabla J_\beta(\theta_t)\right\|^2 \\
&\quad + \left[\frac{48\alpha_w^2(C_G^2C_w^2+C_\ell^2)}{M} + \frac{8L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\left(3L_h^2\kappa^2 + \frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\right]\sum_{t=0}^{T-1}\sum_{i=0}^{t-1}\left(1-\frac{\lambda'\alpha_w}{8}\right)^{t-1-i} \\
&\leq \frac{8\left\|z_0 - z^*_{\theta_0}\right\|^2}{\lambda'\alpha_w} + \left[\frac{48\alpha_w^2(C_G^2C_w^2+C_\ell^2)}{M} + \frac{8L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\left(3L_h^2\kappa^2 + \frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\right]\cdot\frac{8T}{\lambda'\alpha_w} \\
&\quad + \frac{64L_{w'}^2\alpha_\theta^2}{\lambda'^2\alpha_w^2}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla J_\beta(\theta_t)\right\|^2.
\end{aligned} \tag{23}
$$

Step III: Overall convergence by canceling tracking error via actor's positive progress.

Similarly to the on-policy case, we use the Lipschitz continuity property to obtain (see (15))

$$
\begin{aligned}
&\mathbb{E}[J_\beta(\theta_{t+1})] - \mathbb{E}[J_\beta(\theta_t)] \\
&\quad \geq \frac{\alpha_\theta}{4}\mathbb{E}\left\|\nabla J_\beta(\theta_t)\right\|^2 - \frac{3\alpha_\theta}{4}\left(3L_h^2\mathbb{E}\left\|w_t - w^*_{\beta,\theta_t}\right\|^2 + 3L_h^2\kappa^2 + \frac{6L_\mu^4C_{w_\xi}^2}{M}\right),
\end{aligned}
$$

where we use the condition $\alpha_\theta \leq \frac{1}{4L_{J_\beta}}$.

Further, we take the summation over all iterations on both sides of the above bound and have

$$
\begin{aligned}
&\frac{\alpha_\theta}{4}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla J_\beta(\theta_t)\right\|^2\\
&\leq \mathbb{E}[J_\beta(\theta_{T+1})]-\mathbb{E}[J_\beta(\theta_0)]+\frac{3\alpha_\theta}{4}\left(3L_h^2\kappa^2+\frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\cdot T+\frac{9\alpha_\theta L_h^2}{4}\sum_{t=0}^{T-1}\mathbb{E}\left\|w_t-w_{\beta,\theta_t}^*\right\|^2\\
&\leq \frac{R_{\max}}{1-\gamma}+\frac{3\alpha_\theta}{4}\left(3L_h^2\kappa^2+\frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\cdot T+\frac{9\alpha_\theta L_h^2}{4}\sum_{t=0}^{T-1}\mathbb{E}\left\|w_t-w_{\beta,\theta_t}^*\right\|^2\\
&\leq \frac{R_{\max}}{1-\gamma}+\frac{3\alpha_\theta}{4}\left(3L_h^2\kappa^2+\frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\cdot T+\frac{9\alpha_\theta L_h^2}{4}\sum_{t=0}^{T-1}\mathbb{E}\left\|z_t-z_{\theta_t}^*\right\|^2.
\end{aligned}
$$

Then, we substitute the cumulative error from (23) into the above bound and have

$$
\begin{aligned}
&\frac{\alpha_\theta}{8}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla J_\beta(\theta_t)\right\|^2\\
&\overset{\text{(i)}}{\leq}\left(\frac{\alpha_\theta}{4}-\frac{144L_h^2L_{w'}^2\alpha_\theta^3}{\lambda'^2\alpha_w^2}\right)\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla J_\beta(\theta_t)\right\|^2\\
&\leq \frac{R_{\max}}{1-\gamma}+\frac{3\alpha_\theta}{4}\left(3L_h^2\kappa^2+\frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\cdot T+\frac{18\alpha_\theta L_h^2}{\lambda'\alpha_w}\left\|z_0-z_{\theta_0}^*\right\|^2\\
&\quad+\left[\frac{48\alpha_w^2(C_G^2C_w^2+C_\ell^2)}{M}+\frac{8L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\left(3L_h^2\kappa^2+\frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\right]\cdot\frac{18\alpha_\theta L_h^2T}{\lambda'\alpha_w},
\end{aligned}
$$

where (i) follows from the condition $\alpha_\theta \leq \frac{\lambda'\alpha_w}{24L_hL_{w'}}$.

Finally, we obtain

$$
\begin{aligned}
\min_{t\in[T]}\mathbb{E}\left\|\nabla J_\beta(\theta_t)\right\|^2 &\leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla J_\beta(\theta_t)\right\|^2\\
&\leq \left(\frac{8R_{\max}}{\alpha_\theta(1-\gamma)}+\frac{144L_h^2}{\lambda\alpha_w}\left\|z_0-z_{\theta_0}^*\right\|^2\right)\cdot\frac{1}{T}+6\left(3L_h^2\kappa^2+\frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\\
&\quad+\left[\frac{48\alpha_w^2(C_G^2C_w^2+C_\ell^2)}{M}+\frac{8L_{w'}^2\alpha_\theta^2}{\lambda\alpha_w}\left(3L_h^2\kappa^2+\frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\right]\cdot\frac{144L_h^2}{\lambda\alpha_w}\\
&=\frac{c_4}{T}+\frac{c_5}{M}+c_6\kappa^2,
\end{aligned}
$$

where

$$
c_4=\frac{8R_{\max}}{\alpha_\theta(1-\gamma)}+\frac{144L_h^2}{\lambda'\alpha_w}\left\|z_0-z_{\theta_0}^*\right\|^2, \tag{24}
$$

$$
c_5=36L_\mu^4C_{w_\xi}^2+\left[48\alpha_w^2(C_G^2C_w^2+C_\ell^2)+\frac{48L_{w'}^2L_\mu^4C_{w_\xi}^2\alpha_\theta^2}{\lambda'\alpha_w}\right]\cdot\frac{144L_h^2}{\lambda'\alpha_w}, \tag{25}
$$

$$
c_6=18L_h^2+\frac{24L_{w'}^2L_h^2\alpha_\theta^2}{\lambda'\alpha_w}. \tag{26}
$$

D.3 PROOF OF COROLLARY 3

Following from the upper bound in Theorem 2, we let $\frac{c_4}{T}\leq\frac{\epsilon}{2}$ and $\frac{c_5}{M}\leq\frac{\epsilon}{2}$ to achieve the target ϵ-accuracy. Then we obtain $T\geq\frac{2c_4}{\epsilon}$ and $M\geq\frac{2c_5}{\epsilon}$. Further, since we generate M samples in the update steps of both critic and actor in Algorithm 2, the total number of samples we use is thus $2MT=\frac{8c_4c_5}{\epsilon^2}$.