Deterministic Policy Gradient: Convergence Analysis

Huaqing Xiong1 Tengyu Xu†1 Lin Zhao2 Yingbin Liang1 Wei Zhang3

1Department of Electrical and Computer Engineering, The Ohio State University, Columbus, Ohio, USA
2Department of Electrical and Computer Engineering, National University of Singapore, Singapore, Republic of Singapore
3Department of Mechanical and Energy Engineering, Southern University of Science and Technology (SUSTech), Shenzhen, Guangdong, China

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 $\epsilon$-accurate stationary policy up to a system error with a sample complexity of $O(\epsilon^{-2})$. 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. [2015] 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

*Equal contribution
†Corresponding author

Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022), PMLR 180:2159–2169.
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 \( O(\epsilon^{-2}) \). 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 \( O(\epsilon^{-2}) \). 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 \( O(\epsilon^{-2}) \).

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; Qui 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], 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, Jader 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, Qui 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 \((S, A, P, r, \gamma)\). Here, \(S\) denotes the state space, \(A\) denotes the action space, \(P: S \times A \times S \mapsto [0, 1]\) denotes the transition kernel for the state transitions, \(E\), \(P(s'|s, a)\) represents the probability that the system takes the next state \(s' \in S\) given the current state \(s\) and action \(a; r: S \times A \mapsto [0, R_{\max}\}]\) 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 \(\mu_\theta\) 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 \(\mu_\theta\).

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

\[
J(\mu_\theta) = \int_S \nu_{\mu_\theta}(s)r(s, \mu_\theta(s))ds = \mathbb{E}_{s \sim \nu_{\mu_\theta}} [r(s, \mu_\theta(s))],
\]

(1)

where \(\nu_{\mu}(s') = \int_S \sum_{t=0}^{\infty} \gamma^t p_\theta(s)p(s \rightarrow s', t, \mu)ds\) is the (improper) discounted state visitation distribution and \(p(s \rightarrow s', t, \mu)\) denotes the density at state \(s'\) after \(t\) steps from state \(s\) under policy \(\mu\). 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:

\[
\nabla J(\theta) = \int_S \nu_\theta(s)\nabla_\theta \mu_\theta(s)\nabla_a Q^{\mu_\theta}(s, a)|_{a=\mu_\theta(s)}ds
= \mathbb{E}_{\nu_\theta} [\nabla_\theta \mu_\theta(s)\nabla_a Q^{\mu_\theta}(s, a)|_{a=\mu_\theta(s)}].
\]

(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 \(Q^w(s, a)\) compatible with a deterministic policy \(\mu_\theta\), i.e., \(\nabla J(\theta) = \mathbb{E}_{\nu_\theta} [\nabla_\theta \mu_\theta(s)\nabla_a Q^w(s, a)|_{a=\mu_\theta(s)}]\). 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. \(w = w^*_\xi\) minimizes the mean square error \(\mathbb{E}_{\nu_\theta} [\xi(s; \theta, w)^T \xi(s; \theta, w)]\), 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

\[
\nabla J(\theta) = \mathbb{E}_{\nu_\theta} [\nabla_\theta \mu_\theta(s)\nabla_a Q^w(s, a)^T w^*_\xi];
\]

(3)

where \(w^*_\xi\) 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 \(\beta\), which is different from the target policy \(\mu_\theta\). [Silver et al., 2014] has also provided the deterministic policy gradient theorem for such an off-policy case, which is given by

\[
\nabla J_\beta(\theta) = \int_S \nu_\beta(s)\nabla_\theta \mu_\theta(s)\nabla_a Q^{\mu_\theta}(s, a)|_{a=\mu_\beta(s)}ds
= \mathbb{E}_{\nu_\beta} [\nabla_\theta \mu_\theta(s)\nabla_a Q^{\mu_\theta}(s, a)|_{a=\mu_\beta(s)}];
\]

(4)

where \(\nu_\beta\) is the state visitation measure of the policy \(\beta\). Correspondingly, the compatible form is given by

\[
\nabla J_\beta(\theta) = \mathbb{E}_{\nu_\beta} [\nabla_\theta \mu_\theta(s)\nabla_a Q^w(s, a)^T w^*_\xi];
\]

(5)

where \(w^*_\xi\) = \(\arg\min_w \mathbb{E}_{\nu_\beta} [\xi(s; \theta, w)^T \xi(s; \theta, w)]\) 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 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 \( \theta \) as an actor parameter to update the policy (line 13 in Algorithm 1) based on the compatibility property.

Algorithm 1 DPG-TD

1: Input: \( \alpha, \omega, \theta_0 \), batch size \( M \).
2: for \( t = 0, 1, \ldots, T \) do
3:     for \( j = 0, 1, \ldots, M - 1 \) do
4:         Sample \( s_{t,j} \sim \mu_\theta \). Generate \( a_{t,j} = \mu_\theta(s_{t,j}) \).
5:         Sample \( s_{t+1,j} \sim P(\cdot | s_{t,j}, a_{t,j}) \) and \( r_{t,j} \). Generate \( a_{t+1,j} = \mu_\theta(s_{t+1,j}) \).
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:     \( w_{t+1} = w_t + \frac{\alpha}{M} \sum_{j=0}^{M-1} \delta_{t,j} \phi(x_{t,j}) \).
10: end for
11: Sample \( s'_{t,j} \sim \nu_\theta \).
12: end for
13: \( \theta_{t+1} = \theta_t + \frac{\alpha}{M} \sum_{j=0}^{M-1} \nabla_\theta \mu_\theta(s'_{t,j}) \nabla_\theta \phi_\theta(s_{t,j})^T w_t \).
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 \( L_\mu, L_\nu \) and \( \lambda_\Psi \), such that (1) \( \| \mu_\theta(s) - \mu_\theta(s) \| \leq L_\mu \| \theta_1 - \theta_2 \|, \forall s \in S \); (2) \( \| \nabla_\theta \mu_\theta(s) - \nabla_\theta \mu_\theta(s) \| \leq L_\nu \| \theta_1 - \theta_2 \|, \forall s \in S \); (3) the matrix \( \Psi_\theta := \mathbb{E}_d \left[ (\nabla_\theta \phi_\theta(s) \nabla_\theta \phi_\theta(s))^T \right] \) (which we call Fisher information of deterministic policy) is non-singular with the minimal eigenvalue uniformly lower-bounded as \( \sigma_{\min}(\Psi_\theta) \geq \lambda_\Psi \).

Assumption 1 is standard in the studies of TD learning with linear function approximation. To be more specific, it ensures that \( \mathbb{E}_d [\delta \phi] = 0 \) has a unique root, namely \( \nu_\phi \), which is also the global optimum of TD learning for a fixed policy \( \mu_\theta \).

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

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

Assumption 4 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 \( \phi : S \times A \rightarrow \mathbb{R}^d \) is uniformly bounded, i.e., \( \| \phi(\cdot, \cdot) \| \leq C_\phi \) for some positive constant \( C_\phi \). In addition, we define \( A = \mathbb{E}_d [\phi(x)(\gamma \phi(\cdot)^T - \phi(\cdot))^T] \) and \( D = \mathbb{E}_d [\phi(x)\phi(x)^T] \), 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 \( \lambda_A \).

Remark 1. We will abuse the notations a bit and assume that Assumptions 2, 3 also hold for the off-policy case, with the expectations taken over the behavior stationary (and improper) visitation distributions \( d_\beta \) and \( \nu_\beta \).

3.3 CONVERGENCE RESULT

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

2162
Theorem 1. Suppose that Assumptions 1-4 hold. Let Theorem 1 hold. Then the deterministic policy gradient \( \nabla J(\theta) \) defined in (2) is Lipschitz continuous with the parameter \( L_{\theta}, \theta \in \mathbb{R}^d \),

\[
\| \nabla J(\theta_1) - \nabla J(\theta_2) \| \leq L_{\theta} \| \theta_1 - \theta_2 \| ,
\]

where \( L_{\theta} = \left( \frac{1}{2} L_{P} L_{\mu} + \frac{L_{Q} L_{\mu} + \frac{L_{P} R_{\max} L_{\mu} C_{\nu} + \frac{\gamma R_{\max} L_{P} L_{\nu}}{1 - \gamma}}{1 - \gamma} \right) \).

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 3.3 with the full proof given in ?? of the appendix.

Theorem 1. Suppose that Assumptions 1-4 hold. Let \( \alpha_{\omega} \leq \frac{\lambda_{\psi}}{2C_{\phi}} \), \( M \geq \frac{8\alpha_{w} C_{2}^{2}}{\lambda_{\psi}}, \alpha_{\omega} \leq \min \left\{ \frac{1}{4L_{R} C_{\beta}}, \frac{1}{24L_{P},L_{\mu}} \right\} \). Then the output of DPG-TD in Algorithm 4 satisfies

\[
\min_{\theta \in \mathcal{H}} \mathbb{E} \left\| \nabla J(\theta) \right\|^2 \leq \frac{c_{1}}{T} + \frac{c_{2}}{M} + c_{3} \kappa^{2},
\]

where \( c_{1} = \frac{8R_{\max} L_{P} L_{\mu} C_{\nu} + \frac{144L_{P}^{2} L_{Q} L_{\mu} C_{\nu}}{\lambda_{\psi}}}{\lambda_{\psi}} \left\| w_{0} - w_{0}^{*} \right\|^{2}, c_{2} = \left[ 48\alpha_{w}^{2} (C_{A}^{2} C_{\phi}^{2} + C_{A}^{2} C_{\phi}^{2}) + \frac{48L_{P}^{2} L_{Q} L_{\mu} C_{\nu}}{\lambda_{\psi}} \right] \cdot \frac{144L_{P}^{2}}{\lambda_{\psi}} + 36L_{P}^{2} C_{\beta}^{2} C_{\phi}^{2}, c_{3} = 18L_{P}^{2} + \frac{24L_{P}^{2} L_{R} L_{\mu}^{2}}{\lambda_{\psi}} \right\| w_{0} - w_{0}^{*} \right\|^{2}, C_{A} = 2C_{2}^{2}, C_{B} = R_{\max} C_{\phi}, C_{\omega} = \frac{R_{\max} C_{\phi}}{\lambda_{\psi}}, C_{C_{\phi}} = \frac{R_{\max} C_{\phi}}{\lambda_{\psi}}(1 - \gamma), L_{w} = \frac{L_{P} L_{Q}}{\lambda_{\psi}} + \frac{L_{P} L_{Q}}{\lambda_{\psi}}(1 - \gamma), L_{h} = \frac{L_{P} L_{Q}}{\lambda_{\psi}} + \frac{L_{P} L_{Q}}{\lambda_{\psi}}(1 - \gamma) \right\| C_{\phi}^{2} L_{P}, L_{w} = \frac{L_{P} C_{\phi}}{\lambda_{\psi}}(1 - \gamma), L_{r} = L_{P} + \frac{R_{\max} L_{P} L_{\mu}}{1 - \gamma} \right\| L_{P}, L_{r} = \frac{1}{2} C_{2} L_{P} L_{\mu}. \)

Corollary 1. Suppose that the same assumptions in Theorem 4 hold. Then the output of DPG-TD in Algorithm 4 satisfies \( \min_{\theta \in \mathcal{H}} \mathbb{E} \left\| \nabla J(\theta) \right\|^2 \leq \epsilon + c_{3} \kappa^{2} \), by using the total number of samples \( 2MT = O \left( 1/\epsilon^{2} \right) \).

Corollary 1 shows that DPG-TD attains an \( \epsilon \)-accurate stationary point (up to the system error) with a sample complexity of \( O(\epsilon^{-2}) \). 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 guaranteed 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 4 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 4 a noisy policy, e.g., \( \pi_{\theta}(s) = \mu_{\theta}(s) + \mathcal{N}(0, \sigma^2) \), is adopted to generate actions \( a_{t,:}, a_{t+1,:} \). Correspondingly, the states for critic’s updates in lines 4-5 of Algorithm 4 are generated by the stationary distribution \( d_{\pi_{\theta}} \) associated with the noisy policy \( \pi_{\theta} \). The rest of Algorithm 4 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 $d_\theta$ by $d_{\mu_\theta}$, 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
\[
\min_{\theta \in \mathcal{O}} \mathbb{E} \| \nabla J(\theta) \|^2 \leq \epsilon + O(\kappa^2),
\]
by using the total number of samples $\mathcal{O}$ ($\epsilon^2$).

In Corollary [2], the system error $\kappa$ 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 with constant step-sizes. 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 $\| w_{t+1} - w_{\theta_{t+1}}^* \|^2$ and $\| w_t - w_{\theta_t}^* \|^2$ by their coupling with actor’s updates.

Recall that $w_{\theta_t}^*$ is the global optimum of TD given a fixed policy $\mu_{\theta_t}$, or equivalently the unique root of $\bar{g}_{\theta_t}(w_t) := \mathbb{E}_{d_{\theta_t}} \left[ \sum_{j=0}^{M-1} \delta_{t,j} \phi(x_{t,j}) \right] = 0$. 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] and replace the stationary distribution $d_\theta$ by $d_{\mu_\theta}$, we readily obtain the convergence result for NoiDPG-TD as follows.

\[
\text{Bhandari et al. [2018] Xiong et al. [2020].}
\]

**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 $\theta_t$ in Algorithm [1] we have $\theta_{t+1} - \theta_t = \frac{\alpha}{M} \sum_{j=0}^{M-1} \nabla \theta \mu_{\theta_t}(s_{t,j}) \nabla \theta \mu_{\theta_t}(s_{t,j})^T w_t := \alpha \theta h_{\theta_t}(w_t, B_t)$. Since $h_{\theta_t}(w_t, B_t)$ 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 ?? for the proof).

\[
\mathbb{E} \left[ \left\| h_{\theta_t}(w_t, B_t) - \nabla J(\theta_t) \right\| ^2 \right] \leq 3L^2 \mathbb{E} \| w_t - w_{\theta_t}^* \|^2 + 3L^2 \kappa^2 + \frac{6L^4 C^2 w_{\ell}}{M}.
\]

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

\[
\mathbb{E} \left[ \left\| w_{t+1} - w_{\theta_{t+1}}^* \right\| ^2 \right] \leq \left( 1 - \frac{\lambda \alpha w}{8} \right) \mathbb{E} \left[ \left\| w_t - w_{\theta_t}^* \right\| ^2 + \frac{48 \alpha w^2 (C_A C_w^2 + C_b^2)}{M} + \frac{8L^2 w_{\ell}}{\lambda \alpha w} \mathbb{E} \left[ \left\| \nabla J(\theta_t) \right\| ^2 \right] \right] + \frac{48 \alpha w^2 (C_A C_w^2 + C_b^2)}{M}.
\]
We first bound the cumulative policy gradient by the cumulative progress towards a stationary point. This also explains the fact that the above critic’s tracking error can decay sufficiently fast by decoupling it from actor’s update, which need to use the importance sampling to adjust the sampling distribution [Silver et al., 2014]. To overcome such a divergence issue, Silver et al. 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 $\beta$.
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 $s_{t+1,j} \sim P(s_{t+1,j} | s_{t,j}, a_{t,j})$ and $r_{t,j}$. Generate $a_{t+1,j} = \mu_\theta(s_{t+1,j})$.
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: $w_{t+1} = w_t + \frac{\alpha_w}{M} \sum_{j=0}^{M-1} \delta_{t,j} \phi_{t,j} - \gamma \phi_{t+1,j}^T \phi_{t,j} w_t$.
10: $u_{t+1} = u_t + \frac{\alpha_u}{M} \sum_{j=0}^{M-1} \delta_{t,j} \phi_{t,j} - \gamma \phi_{t+1,j}^T \phi_{t,j} w_t$.
11: for $j = 0, 1, \ldots, M - 1$ do
12: Sample $s'_{t,j} \sim \nu_\theta$.
13: end for
14: $\theta_{t+1} = \theta_t + \frac{\alpha_\theta}{M} \sum_{j=0}^{M-1} \nabla_\theta \mu_\theta(s'_{t,j}) \nabla_\phi \phi_{s'_{t,j}} T w_t$.
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 $\beta$, 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 $\theta_t$ 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 $L_{J_\beta}$, i.e., $\forall \theta_1, \theta_2 \in \mathbb{R}^d,$

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

where $L_{J_\beta} = \frac{L_r}{1-\gamma} \left( L_C L_\mu + \frac{1}{2} \gamma L_r^2 R_{\max} L_\mu C_\nu + \frac{3L_r L_\mu C_\nu}{1-\gamma} \right) + \frac{L_\phi}{1-\gamma} \left( L_r + \frac{2R_{\max} L_\mu}{1-\gamma} \right)$.

Compared with Lemma 1 for the on-policy case, the Lipschitz parameter $L_{J_\beta}$ 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 ?? of the appendix.

**Theorem 2.** Suppose that Assumptions 1–3 hold. Let $\alpha_w \leq \frac{\lambda}{2C_\phi^2}; M \geq \frac{48\alpha_w C_\phi^2}{\lambda} \alpha_\theta \leq \min \left\{ \frac{1}{3L_J}, \frac{1}{24L_\mu C_\nu} \right\}$; $\eta > \max \left\{ 0, \sigma_{\min}(D - \frac{\lambda + \lambda T}{2}) \right\}$ 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 $c_4 = \frac{8R_{\max}}{\alpha_\theta (1-\gamma)} + \frac{144L_\mu^2}{\lambda \alpha_w} \|z_0 - z_0^*\|^2$, $c_5 = \left[ 48\alpha_w^2 C_\phi^2 C_\omega^2 + C_\bar{\phi}^2 \right]$, $c_6 = 36L_\mu^2 C_\phi^2 C_\omega^2$, $c_7 = 18L_r^2 + \frac{24L_r^2 L_\mu^2 C_\phi^2}{\lambda \alpha_w}$ with $C_\bar{\phi}^2 = \frac{1}{2} (1 + \eta^2) C_\phi^4, C_\bar{\phi}^2 = (1 + \eta^2) R_{\max}^2 C_\phi^2 L_\nu^2 = \frac{L_r}{\lambda \nu} + \frac{2L_\mu L_C \nu}{\lambda \nu (1-\gamma)}$, $\kappa$ given by (7). $L_{J_\beta}$ 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 $2MT = O\left(1/\epsilon^2\right)$.

In Corollary 3, the system error $\kappa$ 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 $\epsilon$-accurate stationary policy with a sample complexity of $O(\epsilon^{-2})$. 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.

Acknowledgements

The work of H. Xiong, T. Xu and Y. Liang was supported in part by U.S. National Science Foundation under the grants CCF-1761506 and CCF-1900145. The work of L. Zhao was supported in part by the Singapore Ministry of Education Academic Research Fund Tier 1 under the grant R-263-000-E60-133. The work of W. Zhang is supported in part by the National Natural Science Foundation of China under Grants 62073159 and 62003155, the Shenzhen Science and Technology Program under Grant JCYJ20200109141601708, the Science, Technology, and Innovation Commission of Shenzhen Municipality under grant ZDSYS2020081114360104.

References


Gabriel Barth-Maron, Matthew W. Hoffman, David Budden, Will Dabney, Dan Horgan, Dhruva TB, Alistair Muldal,


