Convergence and Iteration Complexity of Policy Gradient Method for Infinite-horizon Reinforcement Learning

Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Başar

Abstract— **We focus on policy search in reinforcement learning problems over continuous spaces, where the value is defined by infinite-horizon discounted reward accumulation. This is the canonical setting proposed by Bellman [3]. Policy search, specifically, policy gradient (PG) method, scales gracefully to problems with continuous spaces and allows for deep network parameterizations; however, experimentally it is known to be volatile and its finite-time behavior is not well understood. A major source of this gap is that unbiased ascent directions are elusive, and hence only asymptotic convergence to stationarity can be shown via links to ordinary differential equations [4]. In this work, we propose a new variant of PG methods that uses a random rollout horizon for the Monte-Carlo estimation of the policy gradient, which we establish yields an *unbiased* policy search direction. Furthermore, we conduct global convergence analysis from a nonconvex optimization perspective: (i) we first recover the results of asymptotic convergence to the *stationary-point policies* in the literature through an alternative supermartingale argument; and (ii) we provide iteration complexity, i.e., convergence rate, of policy gradient in the infinite-horizon setting, showing that it exhibits comparable rates to stochastic gradient method in the nonconvex regime for diminishing and constant stepsize rules. Numerical experiments on the inverted pendulum demonstrate the validity of our results.**

I. Introduction

Reinforcement learning (RL) [5], [6] is a mathematical framework for data-driven control, where an autonomous agent interacts with an environment and endeavors to improve behavior according to sequentially observed incentives. This framework has gained attention in recent years with the success of AlphaGo [7], where an RL-based system outperformed the world champion in the game of Go. However, there is a significant gap between the highly engineered systems proposed in [7] and the theoretical foundations that guarantee the performance of the RL algorithms used. Within this gap, we propose to study the stability and complexity of RL algorithms.

In reinforcement learning, excluding lookahead approximations (model predictive control and tree search), methods roughly cluster into those based on search directions in policy space, i.e., "direct policy search," and approximate dynamic programming, which aims to solve Bellman fixed point equations [3]. Our emphasis in this work is to understand direct policy search in the infinite-horizon setting (Sec. II), the canonical example of which is policy gradient method [8]. Policy search has gained traction recently due to its ability to scale gracefully to continuous spaces [9], [10] and incorporate deep networks [11], [12].

Despite the popularity of policy gradient method, its global convergence in the infinite-horizon discounted setting [6] has not been established explicitly so far. This gap largely stems from the fact that obtaining unbiased estimates of the policy gradient through sampling is often missing (in contrast to optimization perspectives on approximate dynamic programming [13], [14]). As a result, one must quantify the stochastic descent directional error. Specifically, following the Policy Gradient Theorem [8], obtaining an unbiased estimate of the policy gradient requires two criteria: (1) the state-action pair is drawn from the ergodic distribution of the Markov chain under the policy; (2) the estimate of the action-value (or Q) function is unbiased, which is similar conceptually to the "double sampling" problem in approximate dynamic programming.

Monte-Carlo rollout, i.e., having the agent randomly explore and accumulate rewards up to some time horizon, may be used to obtain unbiased estimates of the action-value when one restricts focus to *episodic* reinforcement learning. However, this finite horizon rollout will be biased with respect to an infinite-horizon discounted value function. This bias has led most analyses to focus on asymptotic behavior via dynamical systems [4], although some efforts towards finite sample analysis of the infinite-horizon case have appeared recently for linear systems [15]. We propose to overcome this bias through the use of random geometric time rollout horizons, a technique first formulated in [16]. This allows us to obtain unbiased estimates of the Q function, using only rollouts with finite horizons. Moreover, the random rollout horizon also gives rise to an *unbiased sampling* of the state-action pair from the ergodic distribution. With these two challenges addressed, we obtain an unbiased estimate of the policy gradient, which allows us to link policy gradient method with the classical stochastic programming approach [17]. We refer to our algorithm (Sec. III) as *random-horizon* policy gradient (RPG), to emphasize that the horizon of the Monte-Carlo rollout is random.

With this connection established, we can bridge a noticeable gap in policy search: the convergence rate of policy search in a numerical optimization sense. In particular, it is well known in nonconvex optimization that with only first-

Kaiqing Zhang and Tamer Başar are with the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign. {kzhang66, basar1}illinois.edu. A. Koppel is with Computational and Information Sciences Directorate, U.S. Army Research Laboratory, Adelphi, MD, USA. {alec.e.koppel.civ@mail.mil}. H. Zhu is with the Dept. of Electrical and Computer Engineering, University of Texas at Austin. {haozhu@utexas.edu}

A technical report with extended proofs is available to reviewers as [1]. Unique to this submission is the asymptotic convergence and iteration complexity; however, enhancements of these results in the form of convergence to second-order stationary points (approximate local extrema) are spotlighted in a companion journal submission [2].

order information and no additional hypothesis, convergence to a stationary point is the best one may hope to achieve [18]. However, numerous analyses of policy gradient methods claim global convergence to local minimizers by invoking dynamical systems theory [4]. The validity of this analysis hinges on the existence of a strict Lyapunov function; however, for policy search, unless additional structure is present, only a *nonstrict* Lyapunov function can be defined, and hence *stationarity* would be the valid limit of the policy gradient method. Moreover, without a global descent property, only local convergence may be established. In summary, existing analyses for policy search are limited in that they (i) provide only one-step policy improvement; and ii) misuse the term *local-optimal policy* as the algorithm limit, when in fact it is a stationary point.

Last, we close these gaps by establishing a link between policy search and submartingales [19] (Sec. IV). This enables us to provide a unified asymptotic analysis, and consequently yields the first finite-iteration analysis and constant stepsize behavior. Experiments (Sec. V) corroborate our main findings: for the proposed RPG algorithm, the use of random rollout horizons avoids stochastic gradient bias and hence exhibits reliable convergence that matches the theoretically established rates, connecting policy search to stochastic gradient method for nonconvex optimization.

II. Problem Formulation

The mathematical formalism of reinforcement learning is encapsulated by a Markov decision process (MDP), which is a tuple $(\mathcal{S}, \mathcal{A}, \mathbb{P}, R, \gamma)$ with Markov kernel $\mathbb{P}(s' \,|\, s, a) : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$ that determines the transition probability to state s'. The autonomous agent's state s belongs to $\mathcal{S} \subset \mathbb{R}^q$ and it takes actions $a \in \mathcal{S} \subset \mathbb{R}^p$. Every time an agent selects action a_t, a random transition to state s'_t occurs according to $\mathbb{P}(s'_t \,|\, s_t, a_t)$ and a reward $r_t := R(s_t, a_t) \in \mathbb{R}$ is revealed which denotes the merit of a particular action. Here, $\gamma \in (0, 1)$ is a parameter of the problem to be defined shortly.

We assume that the agent follows a stochastic stationary policy $\pi : \mathcal{S} \to \rho(\mathcal{S})$, i.e., actions at time t are chosen according to $\pi(s_t)$. For policy π, define the value $V_\pi : \mathcal{S} \to \mathbb{R}$ as

$$V_\pi(s) = \mathbb{E}_{a_t \sim \pi(\cdot \,|\, s_t), s_{t+1} \sim \mathbb{P}(\cdot \,|\, s_t, a_t)} \bigg(\sum_{t=0}^{\infty} \gamma^t r_t \,\bigg|\, s_0 = s \bigg),$$

which quantifies the long term expected accumulation of rewards discounted by γ. We can further define the value $V_\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ conditioned on a given initial action as the action-value, or Q-function as $Q_\pi(s, a) = \mathbb{E}\big(\sum_{t=0}^{\infty} \gamma^t r_t \,\big|\, s_0 = s, a_0 = a \big)$. We also define $A_\pi(s, a) = Q_\pi(s, a) - V_\pi(s)$ for any s, a to be the *advantage function*. Given any initial state s_0, the goal of the agent is to find the policy π that maximizes the value $V_\pi(s_0)$, i.e., to solve the following maximization

$$\max_{\pi \in \Pi} \quad V_\pi(s_0) \tag{1}$$

In this work, we investigate policy search to solve (1). In general, we must search over an arbitrarily complicated function class Π, including those which are unbounded and discontinuous. Instead, we *parameterize* policies π in Π by a vector $\theta \in \mathbb{R}^d$, i.e., $\pi = \pi_\theta$ to avoid this issue, which is typical of *policy gradient method* [8]. This parameterization reduces a search over arbitrarily complicated function class Π in (1) to one over the Euclidean space $\mathbb{R}^d$. For notational convenience, define $J(\theta) := V_{\pi_\theta}(s_0)$. Then (1) specializes to a vector-valued problem as

$$\max_{\theta \in \mathbb{R}^d} \quad J(\theta). \tag{2}$$

Generally, the value function is nonconvex with respect to the parameter θ, meaning that obtaining a globally optimal solution to (2) is out of reach unless the problem has additional structured properties, as in phase retrieval [20] and tensor decomposition [21]. Moreover, the conventional limit point of most approaches to nonconvex optimization is a stationary solution, which could either be a saddle point or a local optimum. Our goal in this work is to develop a stochastic gradient method to maximize $J(\theta)$, establish its limiting and finite-time behaviors, and clear up some misconceptions regarding its limit points.

III. Policy Gradient Method

In this section, we make a connection between the policy gradient method in RL and stochastic gradient in optimization. In particular, we develop an *unbiased* estimate of the policy gradient for *infinite-horizon* problems, with bounded absolute values. To this end, we first make the assumption below, on the MDP problem and the policy parameterized by π_θ.

Assumption 1. *Suppose the reward function R and the parameterized policy π_θ satisfy the following conditions:*

(i) The absolute value of the reward R is uniformly bounded, i.e., $|R(s, a)| \in [0, U_R]$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$.

(ii) The policy π_θ is differentiable with respect to θ, and $\nabla \log \pi_\theta(a \,|\, s)$, known as the score function *corresponding to the distribution $\pi_\theta(\cdot \,|\, s)$, is L_Θ-Lipschitz and has bounded norm, i.e., for any $(s, a) \in \mathcal{S} \times \mathcal{A}$,*

$$\|\nabla \log \pi_{\theta^1}(a|s) - \nabla \log \pi_{\theta^2}(a \,|\, s)\| \le L_\Theta \cdot \|\theta^1 - \theta^2\|,$$
$$\|\nabla \log \pi_\theta(a \,|\, s)\| \le B_\Theta, \quad \textit{for some constants } L_\Theta, B_\Theta.$$

for any θ^1, θ^2, and θ, respectively.

In the literature on policy gradient/actor-critic algorithms [22], [23], [24], [25], [26], the boundedness of the reward function in Assumption 1(i) is standard. Note that the uniform boundedness of R also results in the boundedness of the Q-function, whose absolute value is bounded by $U_R/(1-\gamma)$, since for any $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$|Q_{\pi_\theta}(s, a)| \le \sum_{t=0}^{\infty} \gamma^t \cdot U_R = U_R/(1 - \gamma). \tag{3}$$

Similarly, for any π_θ and $s \in \mathcal{S}$, we can bound $V_{\pi_\theta}(s)$, and thus $J(\theta)$ (since it is essentially $V_{\pi_\theta}(s_0)$), as follows

$$|V_{\pi_\theta}(s)| \le U_R/(1 - \gamma) \ , \quad |J(\theta)| \le U_R/(1-\gamma).$$

```
Algorithm 1: EstQ: Unbiasedly Estimating Q-function
Input: $s, a$, and $\theta$. Initialize $\hat{Q} \leftarrow 0$, $s_0 \leftarrow s$, $a_0 \leftarrow a$.
Draw $T \sim \mathrm{Geom}(1-\gamma^{1/2})$, i.e., $P(T=t)=(1-\gamma^{1/2})\gamma^{t/2}$.
for all $t = 0, \cdots, T-1$ do
    Collect & add reward $R(s_t, a_t)$: $\hat{Q} \leftarrow \hat{Q} + \gamma^{t/2} R(s_t, a_t)$.
    Simulate state $s_{t+1} \sim \mathbb{P}(\cdot \mid s_t, a_t)$, action $a_{t+1} \sim \pi(\cdot \mid s_{t+1})$.
end for
Collect $R(s_T, a_T)$ by $\hat{Q} \leftarrow \hat{Q} + \gamma^{T/2} \cdot R(s_T, a_T)$.
return $\hat{Q}$.
```

```
Algorithm 2: RPG: Random-horizon Policy Gradient
Input: $s_0$ and $\theta_0$, initialize $k \leftarrow 0$.
Repeat:
    Draw $T_{k+1}$ from $\mathrm{Geom}(1-\gamma)$.
    Draw $a_0 \sim \pi_{\theta_k}(\cdot \mid s_0)$
    for all $t = 0, \cdots, T_{k+1}-1$ do
        Simulate $s_{t+1} \sim \mathbb{P}(\cdot \mid s_t, a_t)$, action $a_{t+1} \sim \pi_{\theta_k}(\cdot \mid s_{t+1})$.
    end for
    Estimate $Q_{\pi_{\theta_k}}(s_{T_{k+1}}, a_{T_{k+1}})$ by Algorithm 1, i.e.,
        $\hat{Q}_{\pi_{\theta_k}}(s_{T_{k+1}}, a_{T_{k+1}}) \leftarrow \mathbf{EstQ}(s_{T_{k+1}}, a_{T_{k+1}}, \theta_k)$.
    Perform policy gradient update
        $\theta_{k+1} \leftarrow \theta_k + \frac{\alpha_k}{1-\gamma} \hat{Q}_{\pi_{\theta_k}}(s_{T_{k+1}}, a_{T_{k+1}})$
            $\times \nabla \log[\pi_{\theta_k}(a_{T_{k+1}} \mid s_{T_{k+1}})]$
    Update the iteration counter $k \leftarrow k+1$.
Until Convergence
```

Assumption 1(ii) has also been made in several recent works on the convergence analysis of policy gradient algorithms [24], [27], [28], which can be readily satisfied by common parametrized policies such as the Boltzmann policy [29] and the Gaussian policy [30]. For example, for Gaussian policy[1] in continuous spaces, $\pi_\theta(\cdot \mid s) = \mathcal{N}(\phi(s)^\top \theta, \sigma^2)$, where $\mathcal{N}(\mu, \sigma^2)$ denotes the Gaussian distribution with mean μ and variance σ^2. Thus, the score function can be written as $[a - \phi(s)^\top \theta]\phi(s)/\sigma^2$, which satisfies Assumption 1(ii) under the three conditions below: the norm of the feature vectors $\phi(s)$ is norm; the action $a \in \mathcal{A}$ is bounded; and the parameter θ lies in some bounded set.

By the Policy Gradient Theorem [8], the gradient of $J(\theta)$ with respect to the policy parameter θ can be written as

$$\nabla J(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{(s,a)\sim\rho_\theta(\cdot,\cdot)} \left[\nabla \log \pi_\theta(a \mid s) Q_{\pi_\theta}(s,a)\right], \quad (4)$$

where we denote the probability that state $s_t = s$ given initial state s_0 and policy parameter θ by $p(s_t = s \mid s_0, \pi_\theta)$, and define the distribution $\rho_{\pi_\theta}(s) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t p(s_t = s \mid s_0, \pi_\theta)$, a well-defined probability measure [8]. For notational convenience, we let $\rho_\theta(s,a) = \rho_{\pi_\theta}(s) \cdot \pi_\theta(a \mid s)$. Under Assumption 1, we can first obtain the Lipschitz continuity of the policy gradient $\nabla J(\theta)$, as shown in the lemma below, whose proof can be found in [1][Appendix A.1].

Lemma 1 (Lipschitz-Continuity of Policy Gradient)**.** *Under Assumption 1, the policy gradient $\nabla J(\theta)$ is Lipschitz continuous with constant $L > 0$, i.e., for any $\theta^1, \theta^2 \in \mathbb{R}^d$*

$$\|\nabla J(\theta^1) - \nabla J(\theta^2)\| \leq L \cdot \|\theta^1 - \theta^2\|,$$

where the value of the Lipschitz constant L is given as

$$L := \frac{U_R \cdot L_\Theta}{(1-\gamma)^2} + \frac{(1+\gamma) \cdot U_R \cdot B_\Theta^2}{(1-\gamma)^3}. \quad (5)$$

To align with the stochastic gradient method in optimization, unbiased samples of the gradient $\nabla J(\theta)$ are required. With only first-order information, the algorithm may converge to a stationary solution of the nonconvex optimization problem. In addition, with a more careful choice of stepsizes, attaining second-order stationary points [21], [31], which is equivalent to actual *local optima* under certain conditions, would also be possible; see [2] for more details. In the sequel, we focus on the convergence to first-order stationary points, which is not completely covered in [2].

Sampling the Policy Gradient: According to (4), the following two conditions are necessary to obtain an *unbiased* sample of $\nabla J(\theta)$: i) draw state-action pair $(s,a) \sim \rho_\theta(\cdot,\cdot)$; and ii) obtain an unbiased estimate of the Q-function $Q_{\pi_\theta}(s,a)$ for the drawn (s,a).

To this end, we use a random horizon T that follows certain geometric distribution in the sampling process. In particular, condition i) is ensured to be satisfied, if we evaluate both $Q_{\pi_\theta}(\cdot,\cdot)$ and $\nabla \log \pi_\theta(\cdot \mid \cdot)$ at the sample (s_T, a_T), which is the last sample of a finite sample trajectory $(s_0, a_0, s_1, \cdots, s_T, a_T)$. Here the horizon $T \sim \mathrm{Geom}(1-\gamma)$. This way, we can show that $(s_T, a_T) \sim \rho_\theta(\cdot,\cdot)$. In addition, given (s_T, a_T), we perform Monte-Carlo rollouts for another horizon $T' \sim \mathrm{Geom}(1-\gamma^{1/2})$ independent of T, and estimate the Q-function value $Q_{\pi_\theta}(s,a)$ as follows:

$$\hat{Q}_{\pi_\theta}(s,a) = \sum_{t=0}^{T'} \gamma^{t/2} \cdot R(s_t, a_t) \,\Big|\, s_0 = s, a_0 = a. \quad (6)$$

This subroutine of estimating the Q-function is summarized as **EstQ** in Algorithm 1. It can be shown that $\hat{Q}_{\pi_\theta}(s,a)$ unbiasedly estimates $Q_{\pi_\theta}(s,a)$ for any (s,a) (see the proof of Theorem 1, which can be found in [1]).

Remark 1. *In practice, finite-horizon rollouts have been widely used to approximate the infinite-horizon Q-function, e.g., in the REINFORCE algorithm. However, inevitable biases in Q-function estimates, and hence in the policy gradient estimates, are created. We also note that the proposed sampling technique improves the one in [16] that uses Geom$(1-\gamma)$ (instead of Geom$(1-\gamma^{1/2})$) to generate*

[1] Note that in practice, the action space $\mathcal{A}$ is bounded, and thus a truncated Gaussian policy over $\mathcal{A}$ is often used; see [28].

the rollout horizon T'. The proposed Q-function estimate is almost surely bounded thanks to the $\gamma^{1/2}$-discount factor in (6)*, which leads to almost sure boundedness of the stochastic policy gradient, and thus decreases its variance. This boundedness is also critical in the proof of convergence to second-order stationary points [2].*

We now propose the following stochastic estimate $\hat{\nabla} J(\theta)$ of the policy gradient $\nabla J(\theta)$ given by (4):

$$\hat{\nabla} J(\theta) = \frac{1}{1-\gamma} \cdot \hat{Q}_{\pi_\theta}(s_T, a_T) \cdot \nabla \log[\pi_\theta(a_T \mid s_T)]. \quad (7)$$

We then establish the following theorem, showing that the stochastic policy gradient $\hat{\nabla} J(\theta)$ indeed unbiasedly estimates $\nabla J(\theta)$. Additionally, we can also establish the boundedness of $\|\hat{\nabla} J(\theta)\|$ and $\|\nabla J(\theta)\|$ for any $\theta \in \Theta$ – see [1][Appendix B] for proof.

Theorem 1 (Properties of Stochastic Policy Gradient)**.** *For any θ, $\hat{\nabla} J(\theta)$ given in* (7) *is an unbiased estimate of* $\nabla J(\theta)$ *in* (4)*, i.e., for any θ,*

$$\mathbb{E}[\hat{\nabla} J(\theta) \mid \theta] = \nabla J(\theta).$$

where the expectation is with respect to the random horizon T', the trajectory along $(s_0, a_0, s_1, \cdots, s_{T'}, a_{T'})$, and the random sample (s_T, a_T). Moreover, the norm of the policy gradient $\nabla J(\theta)$ is bounded, and its stochastic estimate $\hat{\nabla} J(\theta)$ is almost surely (a.s.) bounded, i.e.,

$$\|\nabla J(\theta)\| \leq \frac{B_\Theta \cdot U_R}{(1-\gamma)^2}, \quad \|\hat{\nabla} J(\theta)\| \leq \hat{\ell} \ \ a.s.$$

for some constant $\hat{\ell} > 0$.

Now we are ready to present the policy gradient update based on (7). Let θ_k be the estimate for the policy parameter at iteration $k \geq 0$. Then the policy gradient update for step $k+1$ can be written as

$$\begin{aligned} \theta_{k+1} &= \theta_k + \alpha_k \hat{\nabla} J(\theta_k) \quad (8) \\ &= \theta_k + \frac{\alpha_k}{1-\gamma} \hat{Q}_{\pi_{\theta_k}}(s_{T_{k+1}}, a_{T_{k+1}}) \nabla \log[\pi_{\theta_k}(a_{T_{k+1}} | s_{T_{k+1}})], \end{aligned}$$

where $\{T_k\} \sim \text{Geom}(1-\gamma)$ are i.i.d., and $\{\alpha_k\}$ is the stepsize sequence that can be either diminishing or constant. We refer to the policy gradient method as the *random-horizon policy gradient*, with the details summarized in Algorithm 2.

Next, we analyze the convergence of the aforementioned policy gradient method, establishing both their asymptotic and finite-time performances, using tools from optimization.

IV. Convergence Analysis

In this section, we provide convergence results for the proposed policy gradient algorithms. To start, we first introduce the following assumption for the diminishing stepsize α_k, which is standard in stochastic approximation.

Assumption 2. *The sequence of stepsizes $\{\alpha_k\}_{k \geq 0}$ satisfies the Robbins-Monro condition*

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \quad \sum_{k=0}^{\infty} \alpha_k^2 < \infty.$$

Now we can obtain asymptotic convergence of Algorithm 2 under Assumptions 1-2 as follows.

Theorem 2 (Asymptotic Convergence of Algorithm 2)**.** *Let $\{\theta_k\}_{k \geq 0}$ be the sequence of policy parameters π_{θ_k} given by Algorithm 2. Under Assumptions 1-2, we have $\lim_{k \to \infty} \theta_k \in \Theta^*$, with Θ^* being the stationary points set of* (2)*.*

Proof. At each iteration k, we define the random horizon used in estimating $\hat{Q}_{\pi_{\theta_k}}(s_{T_{k+1}}, a_{T_{k+1}})$ in the inner-loop of Algorithm 1 as T'_{k+1}. We then introduce a probability measure space $(\Omega, \mathcal{F}, P)$ and let $\{\mathcal{F}_k\}_{k \geq 0}$ denote a sequence of increasing sigma-algebras $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \mathcal{F}_\infty \subset \mathcal{F}$, where $\mathcal{F}_k$ denotes the one generated by all the random variables before iteration k. We also define the following auxiliary random variable W_k, which is essential to the analysis of Algorithm 2:

$$W_k = J(\theta_k) - L\hat{\ell}^2 \sum_{j=k}^{\infty} \alpha_k^2, \quad (9)$$

where we recall that L is the Lipchitz constant of $\nabla J(\theta)$ as defined in (5), and $\hat{\ell}$ is the upper bound of $\|\hat{\nabla} J(\theta_k)\|$ in Theorem 1. Noting that $J(\theta)$ is bounded and $\{\alpha_k\}$ is square-summable, we conclude that W_k is bounded for any $k \geq 0$. In fact, we can show that $\{W_k\}$ is a bounded submartingale, as stated in the following lemma.

Lemma 2. *The objective function sequence defined by Algorithm 2 satisfies the following stochastic ascent property:*

$$\begin{aligned} &\mathbb{E}[J(\theta_{k+1}) \mid \mathcal{F}_k] \\ &\quad \geq J(\theta_k) + \mathbb{E}[(\theta_{k+1} - \theta_k) \mid \mathcal{F}_k]^\top \nabla J(\theta_k) - L\alpha_k^2 \hat{\ell}^2 \quad (10) \end{aligned}$$

Moreover, the sequence $\{W_k\}$ defined in (9) *is a bounded submartingale:*

$$\mathbb{E}(W_{k+1} \mid \mathcal{F}_k) \geq W_k + \alpha_k \|\nabla J(\theta_k)\|^2. \quad (11)$$

Proof. Note that W_k is adapted to the sigma-algebra $\mathcal{F}_k$. Consider the first-order Taylor expansion of $J(\theta_{k+1})$ at θ_k. Then there exists some $\widetilde{\theta}_k = \lambda \theta_k + (1-\lambda)\theta_{k+1}$ for some $\lambda \in [0,1]$ such that W_{k+1} can be written as

$$\begin{aligned} W_{k+1} &= J(\theta_k) + (\theta_{k+1} - \theta_k)^\top \nabla J(\widetilde{\theta}_k) - L\hat{\ell}^2 \sum_{j=k+1}^{\infty} \alpha_k^2 \\ &\geq J(\theta_k) + (\theta_{k+1} - \theta_k)^\top \nabla J(\theta_k) \\ &\qquad - L\|\theta_{k+1} - \theta_k\|^2 - L\hat{\ell}^2 \sum_{j=k+1}^{\infty} \alpha_k^2 \end{aligned}$$

where the inequality follows from applying Lipschitz continuity of the gradient (Lemma 1), i.e.

$$\begin{aligned} &(\theta_{k+1} - \theta_k)^\top [\nabla J(\widetilde{\theta}_k) - \nabla J(\theta_k)] \\ &\quad \geq -\|\theta_{k+1} - \theta_k\| \cdot \|\nabla J(\widetilde{\theta}_k) - \nabla J(\theta_k)\| \\ &\quad \geq -\|\theta_{k+1} - \theta_k\| \cdot L\|\widetilde{\theta}_k - \theta_k\| \\ &\quad = -\|\theta_{k+1} - \theta_k\| \cdot L(1-\lambda) \cdot \|\theta_{k+1} - \theta_k\| \\ &\quad \geq -L \cdot \|\theta_{k+1} - \theta_k\|^2, \end{aligned}$$

with the constant L being defined in (5). By taking conditional expectation over $\mathcal{F}_k$ on both sides, we further obtain

$$\begin{aligned}\mathbb{E}[W_{k+1}\,|\,\mathcal{F}_k] &\geq J(\theta_k)+\mathbb{E}[(\theta_{k+1}-\theta_k)\,|\,\mathcal{F}_k]^\top\nabla J(\theta_k)\\ &\quad - L\mathbb{E}(\|\theta_{k+1}-\theta_k\|^2\,|\,\mathcal{F}_k)-L\hat{\ell}^2\sum_{j=k+1}^{\infty}\alpha_k^2\\ &\geq J(\theta_k)+\mathbb{E}[(\theta_{k+1}-\theta_k)\,|\,\mathcal{F}_k]^\top\nabla J(\theta_k)\\ &\quad - L\alpha_k^2\hat{\ell}^2-L\hat{\ell}^2\sum_{j=k+1}^{\infty}\alpha_k^2, \qquad (12)\end{aligned}$$

where the first inequality comes from substituting $\theta_{k+1}-\theta_k=\alpha_k\hat{\nabla}J(\theta_k)$ and the second one uses the fact that $\mathbb{E}[\|\hat{\nabla}J(\theta_k)\|^2]\leq\hat{\ell}^2$. By definition of $J(\theta)$, we have

$$\begin{aligned}&\mathbb{E}[J(\theta_{k+1})\,|\,\mathcal{F}_k]\\ &\quad\geq J(\theta_k)+\mathbb{E}[(\theta_{k+1}-\theta_k)\,|\,\mathcal{F}_k]^\top\nabla J(\theta_k)-L\alpha_k^2\hat{\ell}^2,\end{aligned}$$

which establishes the first argument of the lemma.

In addition, note that

$$\mathbb{E}[(\theta_{k+1}-\theta_k)\,|\,\mathcal{F}_k]=\alpha_k\mathbb{E}(\hat{\nabla}J(\theta_k)\,|\,\mathcal{F}_k)=\alpha_k\nabla J(\theta_k),$$

which we may substitute into the right-hand side of (12), and upper-bound the negative constant terms by null to obtain

$$\mathbb{E}(W_{k+1}\,|\,\mathcal{F}_k)\geq W_k+\alpha_k\|\nabla J(\theta_k)\|^2.$$

This concludes the proof. □

Now we are ready to show that as $k\to\infty$, $\|\nabla J(\theta_k)\|$ converges to zero almost surely. Note that W_k is bounded by J^*, the global maximum of $J(\theta)$. Hence, (11) can be re-written as

$$\mathbb{E}(J^*-W_{k+1}\,|\,\mathcal{F}_k)\leq(J^*-W_k)-\alpha_k\|\nabla J(\theta_k)\|^2,$$

where $\{J^*-W_k\}$ is a nonnegative sequence. By the supermartingale convergence theorem [32], we have

$$\sum_{k=1}^{\infty}\alpha_k\|\nabla J(\theta_k)\|^2<\infty,\quad \text{a.s.} \qquad (13)$$

Due to Assumption 2 that the stepsize $\{\alpha_k\}$ is non-summable, (13) holds if the following is satisfied:

$$\liminf_{k\to\infty}\ \|\nabla J(\theta_k)\|=0. \qquad (14)$$

Now it suffices to show that $\limsup_{k\to\infty}\|\nabla J(\theta_k)\|=0$. We show this by contradiction. We construct a sequence of $\{\theta_k\}$ that has two sub-sequences lying in two disjoint sets. We aim to establish a contradiction on the sum of the distances between the points in the two sets. To this end, suppose that for some random realization $\omega\in\Omega$

$$\limsup_{k\to\infty}\|\nabla J(\theta_k)\|=\epsilon>0; \qquad (15)$$

then this implies that for infinitely many k, $\|\nabla J(\theta_k)\|\geq 2\epsilon/3$. On the other hand, (14) implies that $\|\nabla J(\theta_k)\|\leq\epsilon/3$ for infinitely many k. Therefore, we can define the sets $\mathcal{N}_1$ and $\mathcal{N}_2$ as follows:

$$\mathcal{N}_1=\{\theta_k:\|\nabla J(\theta_k)\|\geq 2\epsilon/3\},\quad \mathcal{N}_2=\{\theta_k:\|\nabla J(\theta_k)\|\leq\epsilon/3\}.$$

Note that by Lemma 1 $\|\nabla J(\theta)\|$ is continuous. Thus, both sets are closed in the Euclidean space. We define the distance between the two sets as

$$D(\mathcal{N}_1,\mathcal{N}_2)=\inf_{\theta^1\in\mathcal{N}_1}\inf_{\theta^2\in\mathcal{N}_2}\|\theta^1-\theta^2\|.$$

Then $D(\mathcal{N}_1,\mathcal{N}_2)$ must be a positive number, because the sets $\mathcal{N}_1$ and $\mathcal{N}_2$ are disjoint and closed. Moreover, there exists an index set $\mathcal{I}$ such that the subsequence $\{\theta_k\}_{k\in\mathcal{I}}$ of $\{\theta_k\}_{k\geq 0}$ crosses the two sets infinitely often. In particular, there exist two sequences of indices $\{s_i\}_{i\geq 0}$ and $\{t_i\}_{i\geq 0}$ such that

$$\{\theta_k\}_{k\in\mathcal{I}}=\{\theta_{s_i},\cdots,\theta_{t_i-1}\}_{i\geq 0},$$

with $\{\theta_{s_i}\}_{i\geq 0}\subseteq\mathcal{N}_1,\{\theta_{t_i}\}_{i\geq 0}\subseteq\mathcal{N}_2$, and for any indices $k=s_i+1,\cdots,t_i-1\in\mathcal{I}$ (not including s_i) in between the indices $\{s_i\}$ and $\{t_i\}$, we have

$$\frac{\epsilon}{3}\leq\|\nabla J(\theta_k)\|\leq\frac{2\epsilon}{3}\leq\|\nabla J(\theta_{s_i})\|.$$

Setting aside this expression for now, let us analyze the norm-difference of iterates θ_k associated with indices in $\mathcal{I}$. By the triangle inequality, we may write

$$\begin{aligned}\sum_{k\in\mathcal{I}}\|\theta_{k+1}-\theta_k\|&=\sum_{i=0}^{\infty}\sum_{k=s_i}^{t_i-1}\|\theta_{k+1}-\theta_k\|\\ &\geq\sum_{i=0}^{\infty}\|\theta_{s_i}-\theta_{t_i}\|\geq\sum_{i=0}^{\infty}D(\mathcal{N}_1,\mathcal{N}_2)=\infty. \qquad (16)\end{aligned}$$

Moreover, (13) implies that

$$\infty>\sum_{k\in\mathcal{I}}\alpha_k\|\nabla J(\theta_k)\|^2\geq\sum_{k\in\mathcal{I}}\alpha_k\cdot\frac{\epsilon^2}{9},$$

using the definition of ϵ in (15). We may therefore conclude that $\sum_{k\in\mathcal{I}}\alpha_k<\infty$. Also from Theorem 1, we have that the stochastic policy gradient has a finite first moment: $\mathbb{E}(\|\hat{\nabla}J(\theta_k)\|)<\infty$. Taken together, we therefore have

$$\sum_{k\in\mathcal{I}}\mathbb{E}(\|\theta_{k+1}-\theta_k\|)=\sum_{k\in\mathcal{I}}\alpha_k\mathbb{E}(\|\hat{\nabla}J(\theta_k)\|)<\infty.$$

The monotone convergence theorem then implies that $\sum_{k\in\mathcal{I}}\|\theta_{k+1}-\theta_k\|<\infty$ almost surely, which contradicts (16). Therefore, (16) must be false, which implies that the hypothesis that the limsup is bounded away from zero, as in (15), is invalid. As a consequence, its negation must be true: the set of sample paths for which this condition holds has measure zero. This allows us to conclude

$$\limsup_{k\to\infty}\|\nabla J(\theta_k)\|=0,\quad a.s.$$

This statement together with (14) allows us to conclude that $\lim_{k\to\infty}\|\nabla J(\theta_k)\|=0$ a.s., which completes the proof. □

Theorem 2 verifies that the proposed RPG algorithm converges almost surely to the (first-order) stationary points of $J(\theta)$. This result is established from an optimization perspective using supermartingale convergence theorem [32], which differs from existing techniques based on the theory of dynamical systems (or ODE method) [4]. This optimization perspective can be established due to the unbiasedness of the stochastic policy gradients.

An additional virtue of this perspective is that we can also establish convergence rate of the policy gradient method. In fact, the finite-iteration analysis for actor-critic algorithms is known to be quite challenging [33], [34], [35]. Here we choose the stepsize to be either $\alpha_k = k^{-a}$ for some parameter $a \in (0,1)$ or constant $\alpha_k = \alpha > 0$. This way, the choice of the stepsize is more general than that in Assumption 2. Following the literature of nonconvex optimization, we consider the norm of the gradient $\|\nabla J(\theta_k)\|$ as the metric to establish the convergence rate. The results using both diminishing and constant stepsizes are formally stated in the theorem below and the corollary that follows it.

Theorem 3 (Convergence Rate of Algorithm 2 with Diminishing Stepsize)**.** *Let $\{\theta_k\}_{k\geq 0}$ be the sequence of parameters of the policy π_{θ_k} given by Algorithm 2. Let the stepsize be $\alpha_k = k^{-a}$ where $a \in (0,1)$. Let*

$$K_\epsilon = \min\big\{k : \inf_{0\leq m\leq k} \mathbb{E}\|\nabla J(\theta_m)\|^2 \leq \epsilon\big\}.$$

Then, under Assumption 1, we have $K_\epsilon \leq O(\epsilon^{-1/p})$, where p is defined as $p = \min\{1-a, a\}$. By optimizing over a, we obtain $K_\epsilon \leq O(\epsilon^{-2})$ with $a = 1/2$.

Proof. By Lemma 2, we can write

$$\begin{aligned}
&\mathbb{E}[J(\theta_{k+1}) \,|\, \mathcal{F}_k] \\
&\quad \geq J(\theta_k) + \mathbb{E}[(\theta_{k+1} - \theta_k) \,|\, \mathcal{F}_k]^\top \nabla J(\theta_k) - L\alpha_k^2 \hat{\ell}^2 \\
&\quad = J(\theta_k) + \alpha_k \|\nabla J(\theta_k)\|^2 - L\alpha_k^2 \hat{\ell}^2. \qquad (17)
\end{aligned}$$

Let $U(\theta) = J^* - J(\theta)$, where J^* is the global optimum[2] of $J(\theta)$. Then, we immediately have $0 \leq U(\theta) \leq 2U_R/(1-\gamma)$, since $|J(\theta)| \leq U_R/(1-\gamma)$ for any θ. Moreover, we may write (17) as

$$\mathbb{E}[U(\theta_{k+1}) \,|\, \mathcal{F}_k] \leq U(\theta_k) - \alpha_k \|\nabla J(\theta_k)\|^2 + L\alpha_k^2 \hat{\ell}^2. \quad (18)$$

Let $N > 0$ be an arbitrary positive integer. By re-ordering the terms in (18) and summing over $k-N, \cdots, k$, we have

$$\begin{aligned}
&\sum_{m=k-N}^{k} \mathbb{E}\|\nabla J(\theta_m)\|^2 \\
&\quad \leq \sum_{m=k-N}^{k} \frac{1}{\alpha_m} \cdot \big\{\mathbb{E}[U(\theta_m)] - \mathbb{E}[U(\theta_{m+1})]\big\} + \sum_{m=k-N}^{k} L\alpha_m \hat{\ell}^2 \\
&\quad = \sum_{m=k-N}^{k} \left(\frac{1}{\alpha_m} - \frac{1}{\alpha_{m-1}}\right) \cdot \mathbb{E}[U(\theta_m)] - \frac{1}{\alpha_k} \cdot \mathbb{E}[U(\theta_{k+1})] \\
&\qquad + \frac{1}{\alpha_{k-N-1}} \cdot \mathbb{E}[U(\theta_{k-N})] + \sum_{m=k-N}^{k} L\alpha_m \hat{\ell}^2 \qquad (19)
\end{aligned}$$

where the equality follows from adding and subtracting an additional term ${\alpha_{k-N-1}}^{-1} \cdot \mathbb{E}[U(\theta_{k-N})]$. Now, using the fact that the value sub-optimality is bounded by $0 \leq U(\theta) \leq 2U_R/(1-\gamma)$, we can further bound the right-hand side of (19) as

$$\begin{aligned}
&\sum_{m=k-N}^{k} \left(\frac{1}{\alpha_m} - \frac{1}{\alpha_{m-1}}\right) \cdot \mathbb{E}[U(\theta_m)] - \frac{1}{\alpha_k} \cdot \mathbb{E}[U(\theta_{k+1})] \\
&\qquad + \frac{1}{\alpha_{k-N-1}} \cdot \mathbb{E}[U(\theta_{k-N})] + \sum_{m=k-N}^{k} L\alpha_m \hat{\ell}^2 \\
&\quad \leq \sum_{m=k-N}^{k} \left(\frac{1}{\alpha_m} - \frac{1}{\alpha_{m-1}}\right) \cdot \frac{2U_R}{1-\gamma} \\
&\qquad + \frac{1}{\alpha_{k-N-1}} \cdot \frac{2U_R}{1-\gamma} + \sum_{m=k-N}^{k} L\alpha_m \hat{\ell}^2 \\
&\quad \leq \frac{1}{\alpha_k} \cdot \frac{2U_R}{1-\gamma} + \sum_{m=k-N}^{k} L\alpha_m \hat{\ell}^2, \qquad (20)
\end{aligned}$$

where we drop the nonpositive term $-\mathbb{E}[U(\theta_{k+1})]/\alpha_k$ and upper-bound $\mathbb{E}[U(\theta_m)]$ by $2U_R/(1-\gamma)$ for all $m = k - N, \cdots, k$. We use the fact that the stepsize is non-increasing $\alpha_m \leq \alpha_{m+1}$, such that $1/\alpha_m \geq 1/\alpha_{m+1}$. By substituting $\alpha_k = k^{-a}$ into (20) and then (19), we further have

$$\begin{aligned}
&\sum_{m=k-N}^{k} \mathbb{E}\|\nabla J(\theta_m)\|^2 \\
&\quad \leq O\bigg(k^a \cdot \frac{2U_R}{1-\gamma} + L\hat{\ell}^2 \cdot [k^{1-a} - (k-N)^{1-a}]\bigg). \quad (21)
\end{aligned}$$

Setting $N = k-1$ and dividing both sides by k on both sides of (21), we obtain

$$\begin{aligned}
&\frac{1}{k}\sum_{m=1}^{k} \mathbb{E}\|\nabla J(\theta_m)\|^2 \qquad (22) \\
&\quad \leq O\bigg(k^{a-1} \cdot \frac{2U_R}{1-\gamma} + L\hat{\ell}^2 \cdot [k^{-a} - k^{-1}]\bigg) \leq O(k^{-p}),
\end{aligned}$$

where $p = \min\{1-a, a\}$. By definition of K_ϵ, we have

$$\mathbb{E}\|\nabla J(\theta_k)\|^2 > \epsilon, \qquad \text{for any} \quad k < K_\epsilon,$$

which together with (22) gives us

$$\epsilon \leq \frac{1}{K_\epsilon}\sum_{m=1}^{K_\epsilon} \mathbb{E}\|\nabla J(\theta_m)\|^2 \leq O(K_\epsilon^{-p}).$$

This shows that $K_\epsilon \leq O(\epsilon^{-1/p})$. Note that $\max_{a\in(0,1)} p(a) = 1/2$ with $a = 1/2$, which concludes the proof. □

Corollary 1 (Convergence Rate of Algorithm 2 with Constant Stepsize)**.** *Let $\{\theta_k\}_{k\geq 0}$ be the sequence of parameters of the policy π_{θ_k} given by Algorithm 2. Let the stepsize be $\alpha_k = \alpha > 0$. Then, under Assumption 1, we have*

$$\frac{1}{k}\sum_{m=1}^{k} \mathbb{E}\|\nabla J(\theta_m)\|^2 \leq O(\alpha L \hat{\ell}^2),$$

where L is the Lipschitz constant of the policy gradient as defined in (5) *in Lemma 1.*

[2] Such an optimum is assumed to exist for the parameterization π_θ.

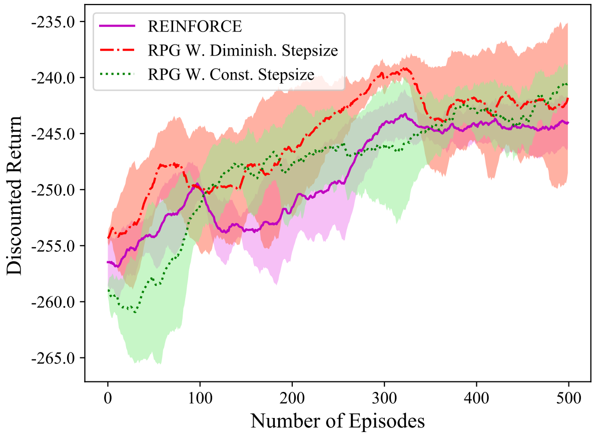

Fig. 1: The convergence of discounted return $J(\theta)$ when REINFORCE and the proposed RPG (Algorithm 2) are used. Both diminishing and constant stepsizes are used for RPG.

Proof. From (19) in the proof of Theorem 3, we obtain that for any $k > 0$ and $0 \le N < k$,

$$\begin{aligned}&\sum_{m=k-N}^{k}\mathbb{E}\|\nabla J(\theta_m)\|^2\\ &\le \sum_{m=k-N}^{k}\frac{1}{\alpha}\{\mathbb{E}[U(\theta_m)]-\mathbb{E}[U(\theta_{m+1})]\}+\sum_{m=k-N}^{k}L\alpha\hat{\ell}^2\\ &\le \frac{1}{\alpha}\cdot\frac{2U_R}{1-\gamma}+(N+1)\cdot L\alpha\hat{\ell}^2, \qquad (23)\end{aligned}$$

where the second inequality follows from the fact that $0 \le U(\theta) \le 2U_R/(1-\gamma)$. By choosing $N = k-1$ and dividing both sides of (23) by k, we obtain

$$\frac{1}{k}\sum_{m=1}^{k}\mathbb{E}\|\nabla J(\theta_m)\|^2 \le \frac{1}{k\alpha}\cdot\frac{2U_R}{1-\gamma}+L\alpha\hat{\ell}^2 \le O(\alpha L\hat{\ell}^2), \quad (24)$$

which completes the proof. □

By Theorem 3, we know that with diminishing stepsize, a $1/\sqrt{k}$ rate for the convergence of the expected gradient norm square $\|\nabla J(\theta_k)\|^2$ is established. By Corollary 1, we know that the average of the gradient norm square will converge to a neighborhood around zero with the rate of $1/k$. Note that the neighborhood size is characterized by the stepsize α. Specifically, (24) illustrates that a smaller stepsize may reduce the neighborhood size, but also sacrifice the convergence speed. In fact, both results recover the convergence properties of stochastic gradient descent for nonconvex optimization problems, which are standard in the literature [17], [36].

V. Simulations

In this section, we provide some numerical results to validate our theory in Sec. IV. We use the Pendulum environment in the OpenAI gym [37] to be the test environment, since it can be modeled within a discounted infinite-horizon setting. Specifically, the pendulum is initialized to be in a random position, and the objective of the task is to swing it up so that it stays upright. The state s_t here has dimension 3, which is defined as $s_t = (\cos(\theta_t), \sin(\theta_t), \dot{\theta}_t)^\top$. Here θ_t denotes the angle between the pendulum and the upright direction, and $\dot{\theta}_t$ denotes the derivative of θ_t. The action a_t is a one-dimensional scalar representing the joint effort. The reward $R(s_t, a_t)$ is defined as

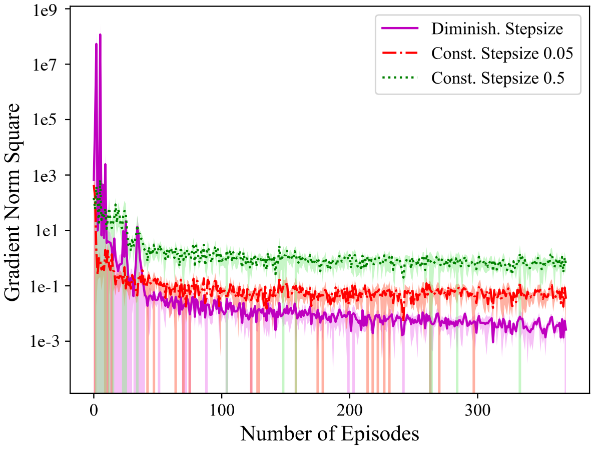

Fig. 2: The convergence of gradient norm square $\mathbb{E}\|\nabla J(\theta_m)\|^2$ when the proposed RPG (Algorithm 2) with different stepsizes are used.

$$R(s_t, a_t) := -(\theta^2 + 0.1 * \dot{\theta}^2 + 0.001 * a_t^2) - 0.5. \quad (25)$$

Since θ is normalized between $[-\pi, \pi]$ and a_t lies in $[-20, 20]$, the reward $R(s_t, a_t)$ lies in $[-17.1736044, -0.5]$. The transition probability is determined by the physical rules of Newton's Second Law. The discounted factor γ is 0.97. We use Gaussian policy π_θ, which is parameterized as $\pi_\theta(\cdot \mid s) = \mathcal{N}(\mu_\theta(s), \sigma^2)$, and truncated over $[-20, 20]$. Here we choose $\sigma = 1.0$ and $\mu_\theta(s) : \mathcal{S} \to \mathcal{A}$ to be a neural network that has two hidden layers. 10 neurons are included in each hidden layer, and *softmax* is used as activation functions. At the output layer of $\mu_\theta(s)$, $\tanh$ is used as the activation function.

First, we compare the performance of our RPG with that of the popular REINFORCE algorithm [38]. Recall that REINFORCE creates bias in the policy gradient estimate. To make a fair comparison, we set the rollout horizon of REINFORCE to be the expected value of the geometric distribution with success probability $1-\gamma^{1/2}$, the same distribution that the rollout horizon for Q-function estimate in Algorithm 1 is drawn from, i.e., $T = \gamma^{1/2}/(1-\gamma^{1/2}) = 66$. For RPG, we test both diminishing and constant stepsizes, where the former is set as $\alpha_k = 1/\sqrt{k}$ and the latter is set as $\alpha_k = 0.05$ for all $k \ge 0$. Fig. 1 plots the discounted return obtained along the iterations of REINFORCE and our proposed RPG algorithms. The return is estimated by running the algorithms 30 times. The bar areas represent the standard deviation region calculated using the 30 simulations. It is shown that our proposed algorithms perform slightly better than REINFORCE in terms of discounted return, but with higher variance. This is expected since our policy gradient estimates are unbiased, compared to REINFORCE. Moreover, the higher variance possibly comes from the additional randomness of the rollout horizon in RPG.

We also evaluate the convergence of the expected gradient norm square studied in Theorem 3 and Corollary 1. Fig. 2 plots the empirical estimates of $\mathbb{E}\|\nabla J(\theta_m)\|^2$ after 30 runs of the algorithms. It is verified that using diminishing stepsize results in convergence of the gradient norm to zero a.s. (the curve keeps decreasing), while using constant stepsizes leads to an error that is lower-bounded above zero (the curves stay mostly unchanged after certain episodes). Moreover, it is shown that a smaller constant stepsize indeed creates a smaller size of the error neighborhood. Convergence rates under both diminishing and constant stepsize choices are sublinear, as identified in our theoretical results.

VI. Concluding Remarks

In this paper, we have studied the convergence properties of a variant of policy gradient methods for infinite-horizon RL problems. Thanks to the random rollout horizon used, our algorithm provides an *unbiased* estimate of policy gradients, which enables the use of analysis tools from nonconvex optimization to establish both almost sure convergence and convergence rates. More interestingly, such a link between PG methods in RL and nonconvex optimization facilitates the improvements of PG methods from the nonconvex optimization perspective in the future, such as rate improvements through acceleration and Quasi-Newton methods, and convergence to second-order stationary points [2].

References

[1] K. Zhang, A. Koppel, H. Zhu, and T. Başar, “Global convergence of policy gradient methods: A nonconvex optimization perspective,” *U.S. Army Research Laboratory/University of Illinois Urbana-Champaign Technical Report*, 2019. Pdf: https://koppel.netlify.com/assets/papers/2019_report_zhang_etal.pdf.

[2] ——, “Global convergence of policy gradient methods: A nonconvex optimization perspective,” *SIAM Journal on Control and Optimization (SICON) (submitted)*, Jan 2019.

[3] R. E. Bellman, *Dynamic Programming*. Courier Dover, 1957.

[4] V. S. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

[5] R. S. Sutton, A. G. Barto *et al.*, *Reinforcement Learning: An Introduction*, 2nd ed., 2017.

[6] D. P. Bertsekas, *Dynamic Programming and Optimal Control*. Athena scientific Belmont, MA, 2005, vol. 1, no. 2.

[7] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton *et al.*, “Mastering the game of go without human knowledge,” *Nature*, vol. 550, no. 7676, p. 354, 2017.

[8] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, “Policy gradient methods for reinforcement learning with function approximation,” in *Advances in Neural Information Processing Systems*, 2000, pp. 1057–1063.

[9] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, “Deterministic policy gradient algorithms,” in *International Conference on Machine Learning*, 2014.

[10] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, “Trust region policy optimization,” in *International Conference on Machine Learning*, 2015, pp. 1889–1897.

[11] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, “Continuous control with deep reinforcement learning,” *arXiv preprint arXiv:1509.02971*, 2015.

[12] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, “Asynchronous methods for deep reinforcement learning,” in *International Conference on Machine Learning*, 2016, pp. 1928–1937.

[13] A. Koppel, G. Warnell, E. Stump, and A. Ribeiro, “Policy evaluation in continuous MDPs with efficient kernelized gradient temporal difference,” 2017.

[14] E. Tolstaya, A. Koppel, E. Stump, and A. Ribeiro, “Nonparametric stochastic compositional gradient descent for Q-learning in continuous markov decision problems,” in *2018 Annual American Control Conference (ACC)*. IEEE, 2018, pp. 6608–6615.

[15] M. Fazel, R. Ge, S. M. Kakade, and M. Mesbahi, “Global convergence of policy gradient methods for linearized control problems,” in *International Conference on Machine Learning*, 2018, pp. 1466–1475.

[16] S. Paternain, “Stochastic control foundations of autonomous behavior,” Ph.D. dissertation, University of Pennsylvania, 2018.

[17] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2009.

[18] S. Wright and J. Nocedal, “Numerical Optimization,” *Springer Science*, vol. 35, no. 67-68, p. 7, 1999.

[19] R. Durrett, *Probability: Theory and Examples*. Cambridge University Press, 2010.

[20] J. Sun, Q. Qu, and J. Wright, “A geometric analysis of phase retrieval,” in *Information Theory (ISIT), 2016 IEEE International Symposium on*. IEEE, 2016, pp. 2379–2383.

[21] R. Ge, F. Huang, C. Jin, and Y. Yuan, “Escaping from saddle points–online stochastic gradient for tensor decomposition,” in *Conference on Learning Theory*, 2015, pp. 797–842.

[22] S. Bhatnagar, M. Ghavamzadeh, M. Lee, and R. S. Sutton, “Incremental natural actor-critic algorithms,” in *Advances in Neural Information Processing Systems*, 2008, pp. 105–112.

[23] S. Bhatnagar, R. Sutton, M. Ghavamzadeh, and M. Lee, “Natural actor-critic algorithms,” *Automatica*, vol. 45, no. 11, pp. 2471–2482, 2009.

[24] D. D. Castro and R. Meir, “A convergent online single-time-scale actor-critic algorithm,” *Journal of Machine Learning Research*, vol. 11, no. Jan, pp. 367–410, 2010.

[25] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Başar, “Fully decentralized multi-agent reinforcement learning with networked agents,” in *International Conference on Machine Learning*, 2018, pp. 5872–5881.

[26] K. Zhang, Z. Yang, and T. Başar, “Networked multi-agent reinforcement learning in continuous spaces,” in *Proceedings of IEEE Conference on Decision and Control*, 2018, pp. 5872–5881.

[27] M. Pirotta, M. Restelli, and L. Bascetta, “Policy gradient in Lipschitz Markov Decision processes,” *Machine Learning*, vol. 100, no. 2-3, pp. 255–283, 2015.

[28] M. Papini, D. Binaghi, G. Canonaco, M. Pirotta, and M. Restelli, “Stochastic variance-reduced policy gradient,” in *International Conference on Machine Learning*, 2018, pp. 4026–4035.

[29] V. R. Konda and V. S. Borkar, “Actor-critic–type learning algorithms for markov decision processes,” *SIAM Journal on Control and Optimization*, vol. 38, no. 1, pp. 94–123, 1999.

[30] K. Doya, “Reinforcement learning in continuous time and space,” *Neural Computation*, vol. 12, no. 1, pp. 219–245, 2000.

[31] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, “How to escape saddle points efficiently,” *arXiv preprint arXiv:1703.00887*, 2017.

[32] H. Robbins and D. Siegmund, “A convergence theorem for non-negative almost supermartingales and some applications,” in *Herbert Robbins Selected Papers*. Springer, 1985, pp. 111–135.

[33] G. Dalal, B. Szorenyi, G. Thoppe, and S. Mannor, “Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning,” *arXiv preprint arXiv:1703.05376*, 2017.

[34] Z. Yang, K. Zhang, M. Hong, and T. Başar, “A finite sample analysis of the actor-critic algorithm,” in *Proceedings of IEEE Conference on Decision and Control*, 2018, pp. 5872–5881.

[35] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Başar, “Finite-sample analyses for fully decentralized multi-agent reinforcement learning,” *arXiv preprint arXiv:1812.02783*, 2018.

[36] S. Ghadimi and G. Lan, “Stochastic first-and zeroth-order methods for nonconvex stochastic programming,” *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2341–2368, 2013.

[37] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, S. John, T. Jie, and Z. Wojciech, “Openai gym,” *arXiv preprint arXiv:1606.01540*, 2016.

[38] R. J. Williams, “Simple statistical gradient-following algorithms for connectionist reinforcement learning,” *Machine Learning*, vol. 8, no. 3-4, pp. 229–256, 1992.