A general sample complexity analysis of vanilla policy gradient

Rui Yuan [1 2 3] **Alessandro Lazaric** [1] **Robert M. Gower** [2 3]

Abstract

The policy gradient (PG) is one of the most popular methods for solving reinforcement learning (RL) problems. However, a solid theoretical understanding of even the “vanilla” PG has remained elusive for long time. In this paper, we apply recent tools developed for the analysis of SGD in non-convex optimization to obtain convergence guarantees for both REINFORCE and GPOMDP under smoothness assumption on the objective function and weak conditions on the second moment of the norm of the estimated gradient. When instantiated under common assumptions on the policy space, our general result immediately recovers existing $\widetilde{\mathcal{O}}(\epsilon^{-4})$ sample complexity guarantees, but for wider ranges of parameters (e.g., step size and batch size m) with respect to previous literature. Notably, our result includes the single trajectory case (i.e., $m = 1$) and it provides a more accurate analysis of the dependency on problem-specific parameters by fixing previous results available in the literature. We believe that the integration of state-of-the-art tools from non-convex optimization may lead to identify a much broader range of problems where PG methods enjoy strong theoretical guarantees.

1. Introduction

The policy gradient (PG) is one of the most popular reinforcement learning (RL) methods to compute policies that maximize long-term rewards (Williams, 1992; Sutton et al., 2000). The success of PG methods is due to their simplicity and versatility, as they can be readily implemented to solve a wide range of problems (including in environments where the Markov assumption is not verified) and they can be effectively paired with other techniques to obtain more sophisticated algorithms such as the actor-critic (Konda & Tsitsiklis, 2000; Mnih et al., 2016), natural PG (Kakade, 2002), trust-region based variants (Schulman et al., 2015; 2017), variance-reduced PG (Papini et al., 2018; Shen et al., 2019; Xu et al., 2020b), etc.

Unlike value-based methods, a solid theoretical understanding of even the “vanilla” PG has remained elusive for long time. Recently, a more complete theory of PG has been derived by leveraging the RL structure of the problem together with tools from convex and non-convex optimization. By using a gradient domination property of the cumulative reward, the global convergence of PG with the exact full gradient is established for linear-quadratic regulators (Fazel et al., 2018) and Markov Decision Process (MDP) with constrained tabular parametrization (Agarwal et al., 2021) or with softmax tabular parametrization (Mei et al., 2020). Zhang et al. (2020b) also establishes the global convergence with the exact full gradient by leveraging the hidden convex structure of the cumulative reward and shows that all local optimums are in fact global optimums under certain assumptions. To improve sample efficiency, Papini et al. (2018); Xu et al. (2020a;b); Zhang et al. (2021) introduce stochastic variance reduced gradient techniques (Johnson & Zhang, 2013; Nguyen et al., 2017; Fang et al., 2018) to policy optimization, and they have studied the sample complexity of policy gradient methods to achieve a first-order stationary point (FOSP). However, these works require either the exact full gradient updates or large batch sizes per iteration. By doing a regret minimization analysis, Zhang et al. (2020a) shows that it is possible to allow a single sampled trajectory (i.e., mini-batch size $m = 1$) for the convergence. However, their setting is restricted to soft-max policy and does not use “vanilla” PG but a modified version with re-projection meant to guarantee a sufficient level of policy randomization.

In this paper, we apply recent tools developed for the analysis of stochastic gradient descent (SGD) in non-convex optimization (Khaled & Richtárik, 2020) to obtain FOSP convergence guarantees for both REINFORCE and GPOMDP under smoothness assumption on the objective function and weak conditions on the second moment of the norm of the estimated gradient. When instantiated under common assumptions on the policy space, our general result immediately recovers existing $\widetilde{\mathcal{O}}(\epsilon^{-4})$ sample complexity guarantees, but for wider ranges of parameters (e.g., step size and batch size) with respect to previous literature. Notably,

[1] Facebook AI Research [2] LTCI, Télécom Paris [3] Institut Polytechnique de Paris. Correspondence to: Rui Yuan <ruiyuan@fb.com>.

Workshop on "Reinforcement learning theory" at the 38^{th} International Conference on Machine Learning, 2021. Copyright 2021 by the author(s).

our result includes the single trajectory case (i.e., $m = 1$) and it provides a more accurate analysis of the dependency on problem-specific parameters by fixing previous results available in the literature. We believe that the integration of state-of-the-art tools from non-convex optimization may lead to identify a much broader range of problems where PG methods enjoy strong theoretical guarantees.

2. Preliminaries

Markov decision process (MDP). We consider a continuous MDP $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, \rho\}$, where $\mathcal{S}$ is a state space; $\mathcal{A}$ is an action space; $\mathcal{P}$ is a Markovian transition model, where $\mathcal{P}(s' \mid s, a)$ defines the transition density from state s to s' under action a; $\mathcal{R}$ is the reward function, where $\mathcal{R}(s,a) \overset{\text{def}}{=} \mathbb{E}_{s' \sim \mathcal{P}(\cdot \mid s,a)}\left[\mathcal{R}(s,a,s')\right] \in [-\mathcal{R}_{\max}, \mathcal{R}_{\max}]$ is the expected reward for state-action pair (s,a); $\gamma \in [0,1)$ is the discount factor; and ρ is the initial state distribution. The agent's behavior is modeled as a policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$, where $\pi(\cdot \mid s)$ is the density distribution over $\mathcal{A}$ in state $s \in \mathcal{S}$. We consider the infinite-horizon discounted setting.

Let $p(\tau \mid \pi)$ be the distribution induced by the policy π on the set $\mathcal{T}$ of all possible trajectories, that is

$$p(\tau \mid \pi) = \rho(s_0) \prod_{t=0}^{\infty} \pi(a_t \mid s_t) p(s_{t+1} \mid s_t, a_t). \quad (1)$$

Let $\mathcal{R}(\tau) = \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t)$ be the total discounted reward accumulated along trajectory τ, then we define the performance function

$$J(\pi) = \mathbb{E}_{\tau \sim p(\cdot \mid \pi, \mathcal{M})}\left[\mathcal{R}(\tau)\right] \overset{\text{def}}{=} \mathbb{E}_{\tau \sim p(\cdot \mid \pi)}\left[\mathcal{R}(\tau)\right]. \quad (2)$$

Policy gradient. PG is a class of methods designed to compute the policy maximizing the total reward $J(\pi)$ by gradient ascent. We introduce a class of parametrized policies $\Pi_\theta = \{\pi_\theta : \theta \in \Theta\}$, with the assumption that π_θ is differentiable w.r.t. θ. For simplicity, we consider $\Theta \subseteq \mathbb{R}^d$. We denote $J(\theta) = J(\pi_\theta)$ and $p(\tau \mid \theta) = p_\theta(\tau) = p(\tau \mid \pi_\theta)$. We also define $J^* = \sup_\pi J(\theta)$ the optimal expected total reward and $\theta^* \in \arg\sup_\pi J(\theta)$ the parameter of the optimal policy. In the most general case, $J(\theta)$ is a non-convex function of the parameter.

The gradient $\nabla J(\theta)$ is derived as follows:

$$\begin{aligned}\nabla J(\theta) &= \int \mathcal{R}(\tau) \nabla p(\tau \mid \theta) d\tau \quad (3)\\ &= \int \mathcal{R}(\tau) \left(\nabla p(\tau \mid \theta) / p(\tau \mid \theta)\right) p(\tau \mid \theta) d\tau \\ &= \mathbb{E}_{\tau \sim p(\cdot \mid \theta)}\left[\mathcal{R}(\tau) \nabla \log p(\tau \mid \theta) \mid \mathcal{M}\right] \\ &\overset{(1)}{=} \mathbb{E}_\tau \left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \sum_{t'=0}^{\infty} \nabla_\theta \log \pi_\theta(a_{t'} \mid s_{t'})\right].\end{aligned}$$

Since it is not possible to execute all possible trajectories up to infinity to compute the full gradient $\nabla J(\theta)$, one has to resort to an empirical estimate of the gradient by sampling m truncated trajectories $\tau_i = (s_0, a_0, r_0, s_1, \cdots, s_{H-1}, a_{H-1}, r_{H-1})$ obtained by executing π_θ for a fixed horizon H. Then the gradient estimator is computed as

$$\widehat{\nabla}_m J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \sum_{t=0}^{H-1} \gamma^t \mathcal{R}(s_t^i, a_t^i) \cdot \sum_{t'=0}^{H-1} \nabla_\theta \log \pi_\theta(a_{t'}^i \mid s_{t'}^i). \quad (4)$$

The estimator (4) is known as the REINFORCE gradient estimator (Williams, 1992).

However, the REINFORCE estimator can be simplified by leveraging the fact that future actions do not depend on past rewards. This leads to the alternative formulation

$$\nabla J(\theta) = \mathbb{E}_\tau \left[\sum_{t=0}^{\infty} \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k \mid s_k)\right) \gamma^t \mathcal{R}(s_t, a_t)\right], \quad (5)$$

which leads to the following gradient estimator

$$\widehat{\nabla}_m J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \sum_{t=0}^{H-1} \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k^i \mid s_k^i)\right) \gamma^t \mathcal{R}(s_t^i, a_t^i), \quad (6)$$

which is known as GPOMDP (Baxter & Bartlett, 2001).

Notice that both REINFORCE and GPOMDP are the truncated versions of unbiased gradient estimators. More precisely, they are unbiased for the gradient of the truncated performance function $J_H(\theta) \overset{\text{def}}{=} \mathbb{E}_\tau \left[\sum_{t=0}^{H-1} \gamma^t \mathcal{R}(s_t, a_t)\right]$. Equipped with gradient estimators, vanilla policy gradient simply updates the policy parameters as

$$\theta_{k+1} = \theta_k + \eta \widehat{\nabla}_m J(\theta_k), \quad (7)$$

with a step size $\eta > 0$.

3. Non-convex optimization under ABC assumption

We use $\widehat{\nabla}_m J(\theta)$ to denote either of the truncated gradient estimators defined in (4) or (6). All following analyses rely on the following smoothness assumption.

Assumption 3.1 (Smoothness)**.** There exists $L > 0$ such that, for all $\theta, \theta' \in \mathbb{R}^d$, we have

$$|J(\theta') - J(\theta) - \langle \nabla J(\theta), \theta' - \theta \rangle| \leq \frac{L}{2} \|\theta' - \theta\|^2. \quad (8)$$

We also make use of the recently introduced *ABC* assumption (Khaled & Richtárik, 2020)[1] which bounds the second moment of the norm of the gradient estimators using the norm of the truncated full gradient, the suboptimality gap and an additive constant.

Assumption 3.2 (ABC)**.** The stochastic gradient satisfies

$$\mathbb{E}\left[\left\|\widehat{\nabla}_m J(\theta)\right\|^2\right] \leq 2A(J^* - J(\theta)) + B\left\|\nabla J_H(\theta)\right\|^2 + C, \tag{9}$$

for some $A, B, C \geq 0$ and all $\theta \in \mathbb{R}^d$.

The ABC assumption effectively summarizes a number of popular and more restrictive assumptions commonly used in non-convex optimization. Indeed, the bounded variance of the stochastic gradient assumption (Ghadimi & Lan, 2013), the gradient confusion assumption (Sankararaman et al., 2020), the sure-smoothness assumption (Lei et al., 2020) and different variants of strong growth assumptions proposed in (Schmidt & Roux, 2013; Vaswani et al., 2019; Bottou et al., 2018) can all be seen as specific cases of Asm. 3.2. The ABC assumption has been shown to be the weakest among all existing assumptions to provide convergence guarantees for SGD for the minimization of non-convex smooth functions.

In order to apply this result to our case, we need an additional assumption to bound the error due to the truncation of the horizon as follows.

Assumption 3.3. There exists $D, D' > 0$ such that, for all $\theta \in \mathbb{R}^d$, we have

$$\begin{aligned} \left|\langle \nabla J_H(\theta), \nabla J_H(\theta) - \nabla J(\theta)\rangle\right| &\leq D\gamma^H, & (10)\\ \left\|\nabla J_H(\theta) - \nabla J(\theta)\right\| &\leq D'\gamma^H. & (11)\end{aligned}$$

While we specifically need those conditions to hold as an assumption, we notice that they are reasonable since we have $|J(\theta) - J_H(\theta)| \leq \frac{\mathcal{R}_{\max}}{1-\gamma}\gamma^H$ by the definition of $J(\cdot)$ and $J_H(\cdot)$. When H is large, the difference between $J(\theta)$ and $J_H(\theta)$ is negligible. However, Asm. 3.3 is still necessary. In fact, the forthcoming convergence results is built on the first-order stationary point. Once we find a stationary point θ such that $\|\nabla J_H(\theta)\|$ is closed to 0, we need (11) to claim the first-order stationary point convergence.

Equipped with these assumptions, we can adapt Thm. 2 in (Khaled & Richtárik, 2020) and obtain the following guarantee.

Proposition 3.4. Suppose that Assumption 3.1, 3.2 and 3.3 are satisfied. We choose a constant step size η such that $\eta \in \left(0, \frac{2}{LB}\right)$ where B can be zero.[a] Let $\delta_0 \overset{\text{def}}{=} J^* - J(\theta_0)$. If $A > 0$, then PG satisfies

$$\begin{aligned}\min_{0\leq t\leq T-1} \mathbb{E}\left[\left\|\nabla J(\theta_t)\right\|^2\right] &\leq \frac{2\delta_0(1+L\eta^2 A)^T}{\eta T(2-LB\eta)} \\ &+ \frac{LC\eta}{2-LB\eta} + \left(\frac{2D(3-LB\eta)}{2-LB\eta} + D'^2\gamma^H\right)\gamma^H.\end{aligned} \tag{12}$$

If $A = 0$, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla J(\theta_U)\right\|^2\right] &\leq \frac{2\delta_0}{\eta T(2-LB\eta)} \\ &+ \frac{LC\eta}{2-LB\eta} + \left(\frac{2D(3-LB\eta)}{2-LB\eta} + D'^2\gamma^H\right)\gamma^H,\end{aligned} \tag{13}$$

where θ_U is uniformly sampled from $\{\theta_0, \theta_1, \cdots, \theta_{T-1}\}$.

[a] We set $\frac{1}{0} = \infty$.

While the proof of Prop. 3.4 is integrating the bias coming from the truncated estimators in the proof of Thm. 2 in (Khaled & Richtárik, 2020), we provide the full proof in App. B for completeness. Prop. 3.4 provides a very general characterization of the performance of PG as a function of all the constants involved in the assumptions on the problem and policy gradient estimator.

From (12) we can derive the sample complexity of PG (see also Cor. 1 in (Khaled & Richtárik, 2020)). If we set the parameters as

$$\begin{aligned}\eta &= \min\left\{\frac{1}{\sqrt{LAT}}, \frac{1}{LB}, \frac{\epsilon}{2LC}\right\}, \\ T &\geq \frac{12\delta_0 L}{\epsilon^2}\max\left\{B, \frac{12\delta_0 A}{\epsilon^2}, \frac{2C}{\epsilon^2}\right\}, \\ H &= \mathcal{O}(\log \epsilon^{-1}),\end{aligned} \tag{14}$$

then $\min_{0\leq t\leq T-1} \mathbb{E}\left[\left\|\nabla J(\theta_t)\right\|^2\right] = \mathcal{O}(\epsilon^2)$.

First, the iteration complexity (14) recovers the exact full gradient case. That is, considering $H = \infty$ (i.e. $J_H = J$) and $\widehat{\nabla}_m J(\theta) = \nabla J(\theta)$ in (7), we have Asm. 3.2 and 3.3 hold automatically with $A = C = D = D' = 0$ and $B = 1$. Consequently, we require $T = \mathcal{O}(\epsilon^{-2})$ iterations to reach an ϵ-stationary point. Thus, for any policy and MDP that satisfy the smoothness property (Asm. 3.1), the exact full PG converge in $\mathcal{O}(\epsilon^{-2})$ iterations. Notice that this is the standard rate of convergence for gradient descent on nonconvex function minimizations without any other assumptions (Beck, 2017). Under special cases, Agarwal et al. (2021) also establishes a $\mathcal{O}(\epsilon^{-2})$ rate of convergence for the exact full gradient in the constrained tabular parametrized policy. By leveraging the hidden convex structure using composite

[1] While Khaled & Richtárik (2020) refers to this assumption as *expected smoothness*, we prefer the alternative name ABC to avoid confusion with the smoothness of J.

optimization tools with additional assumptions where the constrained tabular parametrized policy satisfies, Zhang et al. (2020b) and Zhang et al. (2021) obtain an improved convergence rate $\mathcal{O}(\epsilon^{-1})$ for the exact full gradient. Mei et al. (2020) also establishes the same convergence rate $\mathcal{O}(\epsilon^{-1})$ for the true gradient in the soft-max policy by using a gradient domination property (Lojasiewicz inequality). However, these convergence rates are only conceptual, as we can rarely access the exact full gradient for the update in practice.

In a more general case, i.e. A, C, D, D' are not all 0, the iteration complexity (14) shows that with $TH = \widetilde{\mathcal{O}}(\epsilon^{-4})$ samples (i.e., single-step interaction with the environment and single sampled trajectory per iteration), the vanilla policy gradient guarantees to converge to a first-order stationary point.

4. Convergence under the Lipschitz and smooth policy assumptions

In this section, we instantiate this general statement under (more restrictive) common assumptions on the policy space and we recover existing results for wider ranges of the parameters and more accurate dependencies.

4.1. Sufficient conditions for Asm. 3.1, 3.2 and 3.3

We consider a Lipschitz and smooth policy.

Assumption 4.1 (Lipschitz and smooth policy)**.** There exists constants $G, F > 0$ such that for every action $a \in \mathcal{A}$ and every state $s \in \mathcal{S}$, the gradient and Hessian of $\log \pi_\theta(a \mid s)$ satisfy

$$\begin{aligned} \|\nabla_\theta \log \pi_\theta(a \mid s)\| &\le G, & (15)\\ \left\|\nabla_\theta^2 \log \pi_\theta(a \mid s)\right\| &\le F. & (16) \end{aligned}$$

This assumption is widely adopted in the analysis of variance-reduced PG methods, e.g. (Xu et al., 2020a; Pham et al., 2020; Xu et al., 2020b; Zhang et al., 2021) and it is a relaxation of the one in (Papini et al., 2018), which assumes that $\left|\frac{\partial}{\partial\theta_i} \log \pi_\theta(a \mid s)\right|$ and $\left|\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log \pi_\theta(a \mid s)\right|$ are bounded element-wise. Such assumption is reasonable. For instance, Gaussian policy under the mild condition that the actions and the state feature vectors are bounded satisfies this assumption (Xu et al., 2020b).

Asm. 4.1 directly implies the smoothness of $J(\cdot)$ as well as the ABC and the truncated gradient assumptions for any PG estimator as illustrated in the following lemmas.

Lemma 4.2. Under Asm. 4.1, $J(\cdot)$ is L-smooth, namely $\left\|\nabla^2 J(\theta)\right\| \le L$ for all θ which is a sufficient condition of Asm. 3.1, with

$$L = \frac{2G^2\mathcal{R}_{\max}}{(1-\gamma)^3} + \frac{F\mathcal{R}_{\max}}{(1-\gamma)^2}. \qquad (17)$$

Remark. The smoothness constant (17) is different as compared to recent work such as Proposition 4.2 (2) in (Xu et al., 2020b) and Lemma 3.1 in (Pham et al., 2020). This difference is due to a recurring mistake in a crucial step in bounding the Hessian.[2] Compared to existing bounds, our result reveals an additional term depending on $(1-\gamma)^{-3}$ which dominates the term $\frac{F\mathcal{R}_{\max}}{(1-\gamma)^2}$ derived in (Xu et al., 2020b) whenever γ is close to 1.

Lemma 4.3. Under Asm. 4.1, Asm. 3.2 holds with $A = 0, B = 1 - \frac{1}{m}$ and $C = \frac{\Gamma_g^2}{m}$, that is,

$$\mathbb{E}\left[\left\|\widehat{\nabla}_m J(\theta)\right\|^2\right] \le \left(1 - \frac{1}{m}\right)\|\nabla J_H(\theta)\|^2 + \frac{\Gamma_g^2}{m}, \quad (18)$$

where m is the mini-batch size, and $\Gamma_g = \frac{HG\mathcal{R}_{\max}}{1-\gamma}$ when using REINFORCE gradient estimator or $\Gamma_g = \frac{G\mathcal{R}_{\max}}{(1-\gamma)^2}$ when using GPOMDP gradient estimator.

Bounded variance of the gradient estimator. Interestingly, from (18) we immediately obtain

$$\begin{aligned} \operatorname{Var}\left[\widehat{\nabla}_m J(\theta)\right] &= \mathbb{E}\left[\left\|\widehat{\nabla}_m J(\theta)\right\|^2\right] - \|\nabla J_H(\theta)\|^2 \\ &\overset{(18)}{\le} \frac{\Gamma_g^2 - \|\nabla J_H(\theta)\|^2}{m} \le \frac{\Gamma_g^2}{m}, \quad (19) \end{aligned}$$

which was used as an assumption in (Papini et al., 2018; Pham et al., 2020; Xu et al., 2020b), while it can be directly deduced from Asm. 4.1.

Lemma 4.4. Under Asm. 4.1, Asm. 3.3 holds with

$$\begin{aligned} D &= \frac{D' G\mathcal{R}_{\max}}{(1-\gamma)^2}, & (20)\\ D' &= \left(\frac{1}{(1-\gamma)^2} + \frac{H}{1-\gamma}\right) G\mathcal{R}_{\max}. & (21) \end{aligned}$$

As a by-product, in Lemma D.1 in the appendix we also show that $J(\cdot)$ is Lipschitz.

4.2. Sample complexity of the vanilla policy gradient

Now we can establish the sample complexity of policy gradient for Lipschitz and smooth policies as an immediate corollary of Proposition 3.4 and Lemma 4.2, 4.3 and 4.4.

[2] In the proof in Sect. C, Xu et al. (2020b) rely on the step $\nabla_\theta^2 J(\theta) = \mathbb{E}_\tau\left[\nabla_\theta g(\tau \mid \theta)\right]$, which is not correct since the operators ∇_θ and $\mathbb{E}\left[\cdot\right]$ are not commutative in this case as the density $p(\cdot \mid \theta)$ of $\mathbb{E}\left[\cdot\right]$ depends on θ as well.

Corollary 4.5. Suppose that Assumption 4.1 is satisfied. Let $\delta_0 \overset{\text{def}}{=} J^* - J(\theta_0)$. Any PG method with a mini-batch sampling of size m and step size

$$\eta \quad \in \quad \left(0,\ \frac{2}{L\left(1-1/m\right)}\right), \tag{22}$$

we have

$$\begin{aligned}\mathbb{E}\left[\|\nabla J(\theta_U)\|^2\right] &\leq \frac{2\delta_0}{\eta T\left(2-L\eta\left(1-\frac{1}{m}\right)\right)} \\ &+ \frac{L\Gamma_g^2\eta}{m\left(2-L\eta\left(1-\frac{1}{m}\right)\right)} \\ &+ \left(\frac{2D\left(3-L\eta\left(1-\frac{1}{m}\right)\right)}{2-L\eta\left(1-\frac{1}{m}\right)} + D'^2\gamma^H\right)\gamma^H, \end{aligned} \tag{23}$$

where $D, D' > 0$ are provided in Lemma 4.4.

We first notice that we impose no restriction on the batch size and when $m = 1$, by (22) we have that $\eta \in (0, \infty)$, i.e., the guarantee holds for any choice of the step size. This greatly extends the range of parameters for which PG is guaranteed to converge w.r.t. existing literature.

As in Prop. 3.4, we can then derive explicit sample complexity guarantees. For any accuracy level ϵ, if we set the parameters as (the detailed derivation is provided in App. F.1)

$$\begin{aligned} m &\in \left[1; \frac{2\Gamma_g^2}{\epsilon^2}\right], \\ T \text{ s.t. } Tm &\geq \frac{8\delta_0 L\Gamma_g^2}{\epsilon^4}, \\ \eta &= \frac{\epsilon^2 m}{2L\Gamma_g^2}, \\ H &= \mathcal{O}\left(\log\left(1/\epsilon\right)/\log\left(1/\gamma\right)\right), \end{aligned}$$

then $\mathbb{E}\left[\|\nabla J(\theta_U)\|^2\right] = \mathcal{O}(\epsilon^2)$. This shows that it is *possible* to have the vanilla policy gradient methods converge with a mini-batch size per iteration that can actually vary from 1 to $\mathcal{O}(\epsilon^{-2})$, while the sample complexity remains the same as known in the literature, i.e., $\widetilde{\mathcal{O}}(\epsilon^{-4})$.

This result is novel compared to (Papini et al., 2018; Xu et al., 2020b; Zhang et al., 2021) that do not allow a single trajectory sampled per iteration. The only existing analysis that allows $m = 1$ we are aware of is (Zhang et al., 2020a). However, Zhang et al. (2020a) does not study the vanilla policy gradient. Instead, they add an extra phased learning step to enforce the exploration of the MDP and used a decreasing step size. Moreover, their result is restricted to the soft-max policy parametrization with a log-barrier regularization, which makes their analysis less general. Our results show that such extra phased learning step is unnecessary, the step size can be constant and our convergence theory is satisfied for a much larger class of parametrized policies.

5. Discussion

We believe the generality of Prop. 3.4 opens the possibility to identify a broader set of configurations (i.e., MDP and policy space) for which PG is guaranteed to converge. In particular, we notice that Asm. 4.1 despite being very common, is somehow restrictive, as general policy spaces defined by e.g., a multi-layer neural network, may not satisfy it, unless some restriction on the parameters is imposed. Another interesting venue of investigation is whether it is possible to identify counterparts of the ABC assumption for variance-reduced versions of PG and for the improved analysis of (Zhang et al., 2020b; 2021) leveraging composite optimization tools. For those better sample complexity results, it remains an open question whether we still have convergence guarantee for one single sampled trajectory per iteration with a constant step size.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

Baxter, J. and Bartlett, P. L. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, Nov 2001. ISSN 1076-9757. doi: 10.1613/jair.806.

Beck, A. *First-Order Methods in Optimization*. SIAM-Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2017. ISBN 1611974984.

Bottou, L., Curtis, F., and Nocedal, J. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. ISSN 0036-1445. doi: 10.1137/16M1080173.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, volume 31, pp. 689–699, 2018.

Fazel, M., Ge, R., Kakade, S., and Mesbahi, M. Global convergence of policy gradient methods for the linear quadratic regulator. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1467–1476. PMLR, 10–15 Jul 2018.

Ghadimi, S. and Lan, G. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013. ISSN 1052-6234.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Kakade, S. M. A natural policy gradient. In Dietterich, T., Becker, S., and Ghahramani, Z. (eds.), *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2002.

Khaled, A. and Richtárik, P. Better theory for sgd in the nonconvex world, 2020.

Konda, V. and Tsitsiklis, J. Actor-critic algorithms. In Solla, S., Leen, T., and Müller, K. (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 2000.

Lei, Y., Hu, T., Li, G., and Tang, K. Stochastic gradient descent for nonconvex learning without bounded gradient assumptions. *IEEE Transactions on Neural Networks and Learning Systems*, 31(10):4394–4400, 2020. doi: 10.1109/TNNLS.2019.2952219.

Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. On the global convergence rates of softmax policy gradient methods. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6820–6829. PMLR, 13–18 Jul 2020.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1928–1937, New York, New York, USA, 20–22 Jun 2016. PMLR.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2613–2621, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

Papini, M., Binaghi, D., Canonaco, G., Pirotta, M., and Restelli, M. Stochastic variance-reduced policy gradient. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 4026–4035. PMLR, 2018.

Pham, N., Nguyen, L., Phan, D., Nguyen, P. H., van Dijk, M., and Tran-Dinh, Q. A hybrid stochastic policy gradient algorithm for reinforcement learning. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 374–385. PMLR, 26–28 Aug 2020.

Sankararaman, K. A., De, S., Xu, Z., Huang, W. R., and Goldstein, T. The impact of neural network overparameterization on gradient confusion and stochastic gradient descent. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8469–8479. PMLR, 13–18 Jul 2020.

Schmidt, M. and Roux, N. L. Fast convergence of stochastic gradient descent under a strong growth condition, 2013.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897, Lille, France, 07–09 Jul 2015. PMLR.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms, 2017.

Shen, Z., Ribeiro, A., Hassani, H., Qian, H., and Mi, C. Hessian aided policy gradient. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5729–5738. PMLR, 09–15 Jun 2019.

Stich, S. U. Unified optimal analysis of the (stochastic) gradient method, 2019.

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In Solla, S. A., Leen, T. K., and Müller, K. (eds.), *Advances in Neural Information Processing Systems 12*, pp. 1057–1063. MIT Press, 2000.

Vaswani, S., Bach, F., and Schmidt, M. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1195–1204. PMLR, 16–18 Apr 2019.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

Xu, P., Gao, F., and Gu, Q. An improved convergence analysis of stochastic variance-reduced policy gradient. In Adams, R. P. and Gogate, V. (eds.), *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pp. 541–551. PMLR, 22–25 Jul 2020a.

Xu, P., Gao, F., and Gu, Q. Sample efficient policy gradient methods with recursive variance reduction. In *International Conference on Learning Representations*, 2020b.

Zhang, J., Kim, J., O'Donoghue, B., and Boyd, S. Sample efficient reinforcement learning with reinforce, 2020a.

Zhang, J., Koppel, A., Bedi, A. S., Szepesvari, C., and Wang, M. Variational policy gradient method for reinforcement learning with general utilities. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 4572–4583. Curran Associates, Inc., 2020b.

Zhang, J., Ni, C., Yu, Z., Szepesvari, C., and Wang, M. On the convergence and sample efficiency of variance-reduced policy gradient method, 2021.

Supplementary material

Here we provide the missing proofs from the main paper and some additional noteworthy observations made in the main paper. Each proposition and lemma have a respect section with its proof.

A. Auxiliary Lemmas

Lemma A.1. For all $\gamma \in [0,1)$ and any strictly positive integer H, we have that

$$\sum_{t=0}^{H-1}(t+1)\gamma^t \;\leq\; \sum_{t=0}^{\infty}(t+1)\gamma^t \;=\; \frac{1}{(1-\gamma)^2}.$$

Proof. The first part of the inequality is trivial. We now prove the second part of the inequality. Let

$$S \;\overset{\text{def}}{=}\; \sum_{t=0}^{\infty}(t+1)\gamma^t.$$

We have

$$\gamma S \;=\; \sum_{t=0}^{\infty}(t+1)\gamma^{t+1} \;=\; \sum_{t=1}^{\infty}t\gamma^t.$$

Thus, the subtraction of the above two equations gives

$$\begin{aligned}(1-\gamma)S &= \sum_{t=0}^{\infty}(t+1)\gamma^t - \sum_{t=1}^{\infty}t\gamma^t \\ &= 1+\sum_{t=1}^{\infty}(t+1-t)\gamma^t \\ &= \sum_{t=0}^{\infty}\gamma^t \\ &= \frac{1}{1-\gamma}.\end{aligned}$$

Finally, the proof follows by dividing $1-\gamma$ on both hand side. □

Lemma A.2. For all $\gamma \in [0,1)$ and any strictly positive integer H, we have that

$$\sum_{t=0}^{\infty}(t+1)^2\gamma^t \;\leq\; \frac{2}{(1-\gamma)^3}.$$

Proof. Let

$$S \;\overset{\text{def}}{=}\; \sum_{t=0}^{\infty}(t+1)^2\gamma^t.$$

We have

$$\gamma S \;=\; \sum_{t=0}^{\infty}(t+1)^2\gamma^{t+1} \;=\; \sum_{t=1}^{\infty}t^2\gamma^t.$$

Thus, the subtraction of the above two equations gives

$$\begin{aligned}
(1-\gamma)S &= \sum_{t=0}^{\infty}(t+1)^2\gamma^t - \sum_{t=1}^{\infty} t^2\gamma^t \\
&= 1 + \sum_{t=1}^{\infty}((t+1)^2 - t^2)\gamma^t \\
&= 1 + \sum_{t=1}^{\infty}(2t+1)\gamma^t \\
&= \sum_{t=0}^{\infty}(2t+1)\gamma^t \\
&= 2\sum_{t=0}^{\infty}(t+1)\gamma^t - \sum_{t=0}^{\infty}\gamma^t \\
&\overset{\text{Lemma A.1}}{=} \frac{2}{(1-\gamma)^2} - \frac{1}{1-\gamma} \\
&\leq \frac{2}{(1-\gamma)^2}.
\end{aligned}$$

Finally, the proof follows by dividing $1-\gamma$ on both hand side. □

B. Proof of Proposition 3.4

Proof. We start with L-smoothness of J, which implies

$$\begin{aligned}
J(\theta_{t+1}) &\geq J(\theta_t) + \langle \nabla J(\theta_t), \theta_{t+1} - \theta_t \rangle - \frac{L}{2}\left\|\theta_{t+1} - \theta_t\right\|^2 \\
&= J(\theta_t) + \eta\left\langle \nabla J(\theta_t), \widehat{\nabla}_m J(\theta_t)\right\rangle - \frac{L\eta^2}{2}\left\|\widehat{\nabla}_m J(\theta_t)\right\|^2. \qquad (24)
\end{aligned}$$

Taking expectations conditioned on θ_t, we get

$$\begin{aligned}
\mathbb{E}_t\left[J(\theta_{t+1})\right] &\geq J(\theta_t) + \eta\left\langle \nabla J(\theta_t), \nabla J_H(\theta_t)\right\rangle - \frac{L\eta^2}{2}\mathbb{E}_t\left[\left\|\widehat{\nabla}_m J(\theta_t)\right\|^2\right] \\
&\overset{(9)}{\geq} J(\theta_t) + \eta\left\langle \nabla J_H(\theta_t) + \left(\nabla J(\theta_t) - \nabla J_H(\theta_t)\right), \nabla J_H(\theta_t)\right\rangle \\
&\quad - \frac{L\eta^2}{2}\left(2A(J^* - J(\theta_t)) + B\left\|\nabla J_H(\theta_t)\right\|^2 + C\right) \\
&= J(\theta_t) + \eta\left(1 - \frac{LB\eta}{2}\right)\left\|\nabla J_H(\theta_t)\right\|^2 - L\eta^2 A(J^* - J(\theta_t)) \\
&\quad - \frac{LC\eta^2}{2} + \eta\left\langle \nabla J_H(\theta_t), \nabla J(\theta_t) - \nabla J_H(\theta_t)\right\rangle \\
&\overset{(10)}{\geq} J(\theta_t) + \eta\left(1 - \frac{LB\eta}{2}\right)\left\|\nabla J_H(\theta_t)\right\|^2 - L\eta^2 A(J^* - J(\theta_t)) \\
&\quad - \frac{LC\eta^2}{2} - \eta D\gamma^H.
\end{aligned}$$

(25)

Subtracting J^* from both sides gives

$$\begin{aligned}
-\left(J^* - \mathbb{E}_t\left[J(\theta_{t+1})\right]\right) &\geq -(1 + L\eta^2 A)(J^* - J(\theta_t)) + \eta\left(1 - \frac{LB\eta}{2}\right)\left\|\nabla J_H(\theta_t)\right\|^2 \\
&\quad - \frac{LC\eta^2}{2} - \eta D\gamma^H. \qquad (26)
\end{aligned}$$

Taking the total expectation and rearranging, we get

$$\mathbb{E}\left[J^* - J(\theta_{t+1})\right] + \eta\left(1 - \frac{LB\eta}{2}\right)\mathbb{E}\left[\left\|\nabla J_H(\theta_t)\right\|^2\right] \quad \leq \quad (1 + L\eta^2 A)\mathbb{E}\left[J^* - J(\theta_t)\right] + \frac{LC\eta^2}{2} + \eta D\gamma^H. \tag{27}$$

Letting $\delta_t \overset{\text{def}}{=} \mathbb{E}\left[J^* - J(\theta_t)\right]$ and $r_t \overset{\text{def}}{=} \mathbb{E}\left[\left\|\nabla J_H(\theta_t)\right\|^2\right]$, we can rewrite the last inequality as

$$\eta\left(1 - \frac{LB\eta}{2}\right) r_t \quad \leq \quad (1 + L\eta^2 A)\delta_t - \delta_{t+1} + \frac{LC\eta^2}{2} + \eta D\gamma^H. \tag{28}$$

We now introduce a sequence of weights $w_{-1}, w_0, w_1, \cdots, w_{T-1}$ based on a technique developed by (Stich, 2019). Let $w_{-1} > 0$. Define $w_t = \frac{w_{t-1}}{1+L\eta^2 A}$ for all $t \geq 0$. Notice that if $A = 0$, we have $w_t = w_{t-1} = \cdots = w_{-1}$. Multiplying (28) by w_t/η,

$$\begin{aligned}\left(1 - \frac{LB\eta}{2}\right) w_t r_t \quad & \leq \quad \frac{w_t(1 + L\eta^2 A)}{\eta}\delta_t - \frac{w_t}{\eta}\delta_{t+1} + \frac{LC\eta}{2}w_t + D\gamma^H w_t \\ & = \quad \frac{w_{t-1}}{\eta}\delta_t - \frac{w_t}{\eta}\delta_{t+1} + \left(\frac{LC\eta}{2} + D\gamma^H\right) w_t. \end{aligned} \tag{29}$$

Summing up both sides as $t = 0, 1, \cdots, T-1$ and using telescopic sum, we have,

$$\begin{aligned}\left(1 - \frac{LB\eta}{2}\right)\sum_{t=0}^{T-1} w_t r_t \quad & \leq \quad \frac{w_{-1}}{\eta}\delta_0 - \frac{w_{T-1}}{\eta}\delta_T + \left(\frac{LC\eta}{2} + D\gamma^H\right)\sum_{t=0}^{T-1} w_t \\ & \leq \quad \frac{w_{-1}}{\eta}\delta_0 + \left(\frac{LC\eta}{2} + D\gamma^H\right)\sum_{t=0}^{T-1} w_t. \end{aligned} \tag{30}$$

Let $W_T = \sum_{t=0}^{T-1} w_t$. Dividing both sides by W_T, we have,

$$\left(1 - \frac{LB\eta}{2}\right)\min_{0\leq t\leq T-1} r_t \;\leq\; \frac{1}{W_T}\cdot\left(1 - \frac{LB\eta}{2}\right)\sum_{t=0}^{T-1} w_t r_t \;\leq\; \frac{w_{-1}}{W_T}\frac{\delta_0}{\eta} + \frac{LC\eta}{2} + D\gamma^H. \tag{31}$$

Note that,

$$W_T \;=\; \sum_{t=0}^{T-1} w_t \geq \sum_{t=0}^{T-1} \min_{0\leq i\leq T-1} w_i \;=\; T w_{T-1} \;=\; \frac{T w_{-1}}{(1 + L\eta^2 A)^T}. \tag{32}$$

Using this in (31),

$$\left(1 - \frac{LB\eta}{2}\right)\min_{0\leq t\leq T-1} r_t \quad \leq \quad \frac{(1 + L\eta^2 A)^T}{\eta T}\delta_0 + \frac{LC\eta}{2} + D\gamma^H. \tag{33}$$

However, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla J(\theta_t)\right\|^2\right] \quad & = \quad \mathbb{E}\left[\left\|\nabla J(\theta_t) - \nabla J_H(\theta_t) + \nabla J_H(\theta_t)\right\|^2\right] \\ & = \quad \mathbb{E}\left[\left\|\nabla J_H(\theta_t)\right\|^2\right] + 2\mathbb{E}\left[\langle\nabla J_H(\theta_t), \nabla J(\theta_t) - \nabla J_H(\theta_t)\rangle\right] + \mathbb{E}\left[\left\|\nabla J(\theta_t) - \nabla J_H(\theta_t)\right\|^2\right] \\ & \overset{(10)+(11)}{\leq} \quad \mathbb{E}\left[\left\|\nabla J_H(\theta_t)\right\|^2\right] + 2D\gamma^H + D'^2\gamma^{2H}. \end{aligned} \tag{34}$$

Substituting r_t in (33) by $\mathbb{E}\left[\left\|\nabla J(\theta_t)\right\|^2\right]$ and using (34), we get

$$\left(1 - \frac{LB\eta}{2}\right)\min_{0\leq t\leq T-1}\mathbb{E}\left[\left\|\nabla J(\theta_t)\right\|^2\right] \quad \leq \quad \frac{(1 + L\eta^2 A)^T}{\eta T}\delta_0 + \frac{LC\eta}{2} + D\gamma^H + \left(1 - \frac{LB\eta}{2}\right)\left(2D\gamma^H + D'^2\gamma^{2H}\right).$$

Our choice of step size guarantees that no matter $B > 0$ or $B = 0$, we have $1 - \frac{LB\eta}{2} > 0$. Dividing both sides by $1 - \frac{LB\eta}{2}$ and rearranging yields the proposition's claim.

If $A = 0$, we know that $\{w_t\}_{t \geq -1}$ is a constant sequence. In this case, $W_T = Tw_{-1}$. Dividing both sides of (30) by W_T, we have,

$$\left(1 - \frac{LB\eta}{2}\right) \frac{1}{T} \sum_{t=0}^{T-1} r_t \;\leq\; \frac{\delta_0}{\eta T} + \frac{LC\eta}{2} + D\gamma^H. \tag{35}$$

Similarly, substituting r_t in (35) by $\mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right]$ and using (34), we get

$$\begin{aligned}\left(1 - \frac{LB\eta}{2}\right) \mathbb{E}\left[\|\nabla J(\theta_U)\|^2\right] &= \left(1 - \frac{LB\eta}{2}\right) \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right] \\ &\leq \frac{\delta_0}{\eta T} + \frac{LC\eta}{2} + D\gamma^H + \left(1 - \frac{LB\eta}{2}\right)\left(2D\gamma^H + D'^2\gamma^{2H}\right).\end{aligned}$$

Dividing both sides by $1 - \frac{LB\eta}{2}$ and rearranging yields the proposition's claim. □

C. Proof of Lemma 4.2

Proof. We know that

$$\begin{aligned}\nabla^2 J(\theta) &\overset{(5)}{=} \nabla_\theta \mathbb{E}_\tau \left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k \mid s_k)\right)\right] \\ &= \nabla_\theta \int p(\tau \mid \theta) \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k \mid s_k)\right) d\tau \\ &= \int \nabla_\theta p(\tau \mid \theta) \left(\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k \mid s_k)\right)\right)^\top d\tau \\ &\quad + \int p(\tau \mid \theta) \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta^2 \log \pi_\theta(a_k \mid s_k)\right) d\tau \\ &= \int p(\tau \mid \theta) \nabla_\theta \log p(\tau \mid \theta) \left(\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k \mid s_k)\right)\right)^\top d\tau \\ &\quad + \int p(\tau \mid \theta) \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta^2 \log \pi_\theta(a_k \mid s_k)\right) d\tau \\ &= \mathbb{E}_\tau \left[\nabla_\theta \log p(\tau \mid \theta) \left(\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k \mid s_k)\right)\right)^\top\right] \\ &\quad + \mathbb{E}_\tau \left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta^2 \log \pi_\theta(a_k \mid s_k)\right)\right] \\ &\overset{(1)}{=} \underbrace{\mathbb{E}_\tau \left[\sum_{t'=0}^{\infty} \nabla_\theta \log \pi_\theta(a_{t'} \mid \theta_{t'}) \left(\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k \mid s_k)\right)\right)^\top\right]}_{①} \\ &\quad + \underbrace{\mathbb{E}_\tau \left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta^2 \log \pi_\theta(a_k \mid s_k)\right)\right]}_{②}.\end{aligned} \tag{36}$$

We aim to bound two terms separately. The second term can be bounded easily. That is,

$$\begin{aligned} \|\textcircled{2}\| &\le \mathbb{E}_\tau\left[\sum_{t=0}^{\infty}\gamma^t\left|\mathcal{R}(s_t,a_t)\right|\left(\sum_{k=0}^{t}\left\|\nabla_\theta^2\log\pi_\theta(a_k\mid s_k)\right\|\right)\right] \\ &\le \mathbb{E}_\tau\left[F\mathcal{R}_{\max}\sum_{t=0}^{\infty}(t+1)\gamma^t\right] \\ &= \frac{F\mathcal{R}_{\max}}{(1-\gamma)^2}, \end{aligned} \tag{37}$$

where the second line is obtained by using $|\mathcal{R}(s_t,a_t)|\le\mathcal{R}_{\max}$ and $\left\|\nabla_\theta^2\log\pi_\theta(a_k\mid s_k)\right\|\le F$ from Assumption 4.1; the last line is obtained by Lemma A.1.

To bound the first term, we use the following notation $x_{0:t}\overset{\text{def}}{=}(x_0,x_1,\cdots,x_t)$ with $\{x_t\}_{t\ge0}$ a sequence of random variables. Similar to the derivation of GPOMDP, we notice that future actions do not depend on past rewards and past actions. That is, for $0\le t<t'$ among terms of the two sums in ①, we have

$$\begin{aligned} & \mathbb{E}_\tau\left[\nabla_\theta\log\pi_\theta(a_{t'}\mid s_{t'})\cdot\gamma^t\mathcal{R}(s_t,a_t)\left(\sum_{k=0}^{t}\nabla_\theta\log\pi_\theta(a_k\mid s_k)\right)^\top\right] \\ = \;& \mathbb{E}_{s_{0:t'},a_{0:t'}}\left[\nabla_\theta\log\pi_\theta(a_{t'}\mid s_{t'})\cdot\gamma^t\mathcal{R}(s_t,a_t)\left(\sum_{k=0}^{t}\nabla_\theta\log\pi_\theta(a_k\mid s_k)\right)^\top\right] \\ = \;& \mathbb{E}_{s_{0:t'},a_{0:(t'-1)}}\left[\mathbb{E}_{a_{t'}}\left[\nabla_\theta\log\pi_\theta(a_{t'}\mid s_{t'})\cdot\gamma^t\mathcal{R}(s_t,a_t)\left(\sum_{k=0}^{t}\nabla_\theta\log\pi_\theta(a_k\mid s_k)\right)^\top\;\middle|\; s_{0:t'},a_{0:(t'-1)}\right]\right] \\ = \;& \mathbb{E}_{s_{0:t'},a_{0:(t'-1)}}\left[\mathbb{E}_{a_{t'}}\left[\nabla_\theta\log\pi_\theta(a_{t'}\mid s_{t'})\;\middle|\; s_{t'}\right]\cdot\gamma^t\mathcal{R}(s_t,a_t)\left(\sum_{k=0}^{t}\nabla_\theta\log\pi_\theta(a_k\mid s_k)\right)^\top\right] \\ = \;& \mathbb{E}_{s_{0:t'},a_{0:(t'-1)}}\left[\int\pi_\theta(a_{t'}\mid s_{t'})\nabla_\theta\log\pi_\theta(a_{t'}\mid s_{t'})da_{t'}\cdot\gamma^t\mathcal{R}(s_t,a_t)\left(\sum_{k=0}^{t}\nabla_\theta\log\pi_\theta(a_k\mid s_k)\right)^\top\right] \\ = \;& \mathbb{E}_{s_{0:t'},a_{0:(t'-1)}}\left[\int\nabla_\theta\pi_\theta(a_{t'}\mid s_{t'})da_{t'}\cdot\gamma^t\mathcal{R}(s_t,a_t)\left(\sum_{k=0}^{t}\nabla_\theta\log\pi_\theta(a_k\mid s_k)\right)^\top\right] \\ = \;& \mathbb{E}_{s_{0:t'},a_{0:(t'-1)}}\left[\nabla_\theta\underbrace{\int\pi_\theta(a_{t'}\mid s_{t'})da_{t'}}_{=1}\cdot\gamma^t\mathcal{R}(s_t,a_t)\left(\sum_{k=0}^{t}\nabla_\theta\log\pi_\theta(a_k\mid s_k)\right)^\top\right] \\ = \;& 0. \end{aligned} \tag{38}$$

Thus, ① can be simplified. We have

$$\begin{aligned} \textcircled{1} &= \mathbb{E}_\tau\left[\sum_{t'=0}^{t}\nabla_\theta\log\pi_\theta(a_{t'}\mid\theta_{t'})\left(\sum_{t=0}^{\infty}\gamma^t\mathcal{R}(s_t,a_t)\left(\sum_{k=0}^{t}\nabla_\theta\log\pi_\theta(a_k\mid s_k)\right)\right)^\top\right] \\ &= \mathbb{E}_\tau\left[\sum_{t=0}^{\infty}\gamma^t\mathcal{R}(s_t,a_t)\left(\sum_{t'=0}^{t}\nabla_\theta\log\pi_\theta(a_{t'}\mid\theta_{t'})\right)\left(\sum_{k=0}^{t}\nabla_\theta\log\pi_\theta(a_k\mid s_k)\right)^\top\right]. \end{aligned} \tag{39}$$

Now we can bound ① easily. That is,

$$
\begin{aligned}
\left\|\textcircled{1}\right\| &\le \mathbb{E}_\tau\left[\sum_{t=0}^{\infty}\gamma^t\left|\mathcal{R}(s_t,a_t)\right|\left\|\sum_{t'=0}^{t}\nabla_\theta\log\pi_\theta(a_{t'}\mid\theta_{t'})\right\|^2\right] \\
&\le \mathbb{E}_\tau\left[\sum_{t=0}^{\infty}\gamma^t\left|\mathcal{R}(s_t,a_t)\right|\left(\sum_{t'=0}^{t}\left\|\nabla_\theta\log\pi_\theta(a_{t'}\mid\theta_{t'})\right\|\right)^2\right] \\
&\le \mathbb{E}_\tau\left[G^2\mathcal{R}_{\max}\sum_{t=0}^{\infty}(t+1)^2\gamma^t\right] \\
&\le \frac{2G^2\mathcal{R}_{\max}}{(1-\gamma)^3} \qquad (40)
\end{aligned}
$$

where the third line is obtained by using $|\mathcal{R}(s_t,a_t)|\le\mathcal{R}_{\max}$ and $\|\nabla_\theta\log\pi_\theta(a_{t'}\mid s_{t'})\|\le G$ from Assumption 4.1; the last line is obtained by Lemma A.2.

Finally, combining the bounds of ① and ② yields the lemma's claim. □

D. Proof of Lemma 4.3

In this section, we aim to prove Lemma 4.3. It is beneficial to first show that J is Lipschitz.

D.1. Lipschitz continuity of $J(\cdot)$

Lemma D.1. If Assumption 4.1 holds, for any m trajectories τ_i and $\theta\in\mathbb{R}^d$, we have

(i) $\widehat{\nabla}_m J(\theta)$ is L_g-Lipschitz continuous ;

(ii) The gradient estimator is bounded, i.e. $\left\|\widehat{\nabla}_m J(\theta)\right\|\le\Gamma_g$.

(iii) $J(\cdot)$ is Γ-Lipschitz, namely $\|\nabla J(\theta)\|\le\Gamma$ with $\Gamma=\frac{G\mathcal{R}_{\max}}{(1-\gamma)^2}$.

Furthermore, if $\widehat{\nabla}_m J(\theta)$ is a REINFORCE gradient estimator, then $L_g=\frac{HF\mathcal{R}_{\max}}{1-\gamma}$ and $\Gamma_g=\frac{HG\mathcal{R}_{\max}}{1-\gamma}$; if $\widehat{\nabla}_m J(\theta)$ is a GPOMDP gradient estimator, then $L_g=\frac{F\mathcal{R}_{\max}}{(1-\gamma)^2}$ and $\Gamma_g=\Gamma$.

The results with GPOMDP gradient estimator were already proposed in Proposition 4.2 in (Xu et al., 2020b). We include them for the completeness of the properties of a general vanilla policy gradient estimator.

Proof. To prove (i), let $\widehat{\nabla}_m J(\theta)$ be a REINFORCE gradient estimator. From (4), we have

$$
\begin{aligned}
\left\|\nabla\left(\widehat{\nabla}_m J(\theta)\right)\right\| &= \left\|\frac{1}{m}\sum_{i=1}^{m}\sum_{t=0}^{H-1}\left(\sum_{t'=0}^{H-1}\gamma^{t'}\mathcal{R}(s^i_{t'},a^i_{t'})\right)\nabla^2_\theta\log\pi_\theta(a^i_t\mid s^i_t)\right\| \\
&\le \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{t'=0}^{H-1}\gamma^{t'}\left|\mathcal{R}(s^i_{t'},a^i_{t'})\right|\right)\sum_{t=0}^{H-1}\left\|\nabla^2_\theta\log\pi_\theta(a^i_t\mid s^i_t)\right\| \\
&\le HF\mathcal{R}_{\max}\sum_{t'=0}^{H-1}\gamma^{t'} \\
&\le \frac{HF\mathcal{R}_{\max}}{1-\gamma}, \qquad (41)
\end{aligned}
$$

where the third line is obtained by using $\left|\mathcal{R}(s^i_{t'},a^i_{t'})\right|\le\mathcal{R}_{\max}$ and $\left\|\nabla^2_\theta\log\pi_\theta(a^i_t\mid s^i_t)\right\|\le F$ from Assumption 4.1. In this case, $L_g=\frac{HF\mathcal{R}_{\max}}{1-\gamma}$.

Let $\widehat{\nabla}_m J(\theta)$ be a GPOMDP gradient estimator. From (6), we have

$$\begin{aligned}
\left\|\nabla\left(\widehat{\nabla}_m J(\theta)\right)\right\| &= \left\|\frac{1}{m}\sum_{i=1}^{m}\sum_{t=0}^{H-1}\gamma^t\mathcal{R}(s_t^i,a_t^i)\left(\sum_{k=0}^{t}\nabla_\theta^2\log\pi_\theta(a_k^i\mid s_k^i)\right)\right\| \\
&\le \frac{1}{m}\sum_{i=1}^{m}\sum_{t=0}^{H-1}\gamma^t\left|\mathcal{R}(s_t^i,a_t^i)\right|\left(\sum_{k=0}^{t}\left\|\nabla_\theta^2\log\pi_\theta(a_k^i\mid s_k^i)\right\|\right) \\
&\le F\mathcal{R}_{\max}\sum_{t=0}^{H-1}(t+1)\gamma^t \\
&\overset{\text{Lemma A.1}}{\le} \frac{F\mathcal{R}_{\max}}{(1-\gamma)^2},
\end{aligned} \tag{42}$$

where similarly, the third line is obtained by using $\left|\mathcal{R}(s_t^i,a_t^i)\right|\le\mathcal{R}_{\max}$ and $\left\|\nabla_\theta^2\log\pi_\theta(a_k^i\mid s_k^i)\right\|\le F$ from Assumption 4.1. In this case, $L_g=\frac{F\mathcal{R}_{\max}}{(1-\gamma)^2}$.

The proof for (ii) is verbatim. We simply replace $\nabla\left(\widehat{\nabla}_m J(\theta)\right)$ by $\widehat{\nabla}_m J(\theta)$, $\nabla_\theta^2\log\pi_\theta(a_t^i\mid s_t^i)$ by $\nabla_\theta\log\pi_\theta(a_t^i\mid s_t^i)$ and F by G. If $\widehat{\nabla}_m J(\theta)$ is a REINFORCE gradient estimator, we have $\Gamma_g=\frac{HG\mathcal{R}_{\max}}{1-\gamma}$; if $g(\tau\mid\cdot)$ is a GPOMDP gradient estimator, then $\Gamma_g=\frac{G\mathcal{R}_{\max}}{(1-\gamma)^2}$.

To prove (iii), notice that

$$\begin{aligned}
\|\nabla J(\theta)\| &\overset{(5)}{=} \left\|\mathbb{E}_\tau\left[\sum_{t=0}^{\infty}\left(\sum_{k=0}^{t}\nabla_\theta\log\pi_\theta(a_k\mid s_k)\right)\gamma^t\mathcal{R}(s_t,a_t)\right]\right\| \\
&\le \mathbb{E}_\tau\left[\sum_{t=0}^{\infty}\left(\sum_{k=0}^{t}\|\nabla_\theta\log\pi_\theta(a_k\mid s_k)\|\right)\gamma^t\left|\mathcal{R}(s_t,a_t)\right|\right] \\
&\le \mathbb{E}_\tau\left[G\mathcal{R}_{\max}\left(\sum_{t=0}^{\infty}(t+1)\gamma^t\right)\right] \\
&\overset{\text{Lemma A.1}}{=} \frac{G\mathcal{R}_{\max}}{(1-\gamma)^2},
\end{aligned} \tag{43}$$

where similarly, the third line is obtained by using $|\mathcal{R}(s_t,a_t)|\le\mathcal{R}_{\max}$ and $\|\nabla_\theta\log\pi_\theta(a_k\mid s_k)\|\le G$ from Assumption 4.1. Thus, $\|\nabla J(\theta)\|\le\Gamma$ with $\Gamma=\frac{G\mathcal{R}_{\max}}{(1-\gamma)^2}$. □

D.2. The proof

Now we show the proof of Lemma 4.3.

Proof. We denote $g(\tau\mid\theta)$ a stochastic gradient estimator of one single sampled trajectory τ. Thus $\widehat{\nabla}_m J(\theta)=\frac{1}{m}\sum_{i=1}^{m}g(\tau_i\mid\theta)$. Both $\widehat{\nabla}_m J(\theta)$ and $g(\tau\mid\theta)$ are unbiased estimators of $J_H(\theta)$. We have

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\nabla}_m J(\theta)\right\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=0}^{m-1} g(\tau_i \mid \theta)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=0}^{m-1} g(\tau_i \mid \theta) - \nabla J_H(\theta) + \nabla J_H(\theta)\right\|^2\right] \\
&= \left\|\nabla J_H(\theta)\right\|^2 + \mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=0}^{m-1} \left(g(\tau_i \mid \theta) - \nabla J_H(\theta)\right)\right\|^2\right] \\
&= \left\|\nabla J_H(\theta)\right\|^2 + \frac{1}{m^2}\sum_{i=0}^{m-1} \mathbb{E}\left[\left\|g(\tau_i \mid \theta) - \nabla J_H(\theta)\right\|^2\right] \\
&= \left\|\nabla J_H(\theta)\right\|^2 + \frac{\mathbb{E}\left[\left\|g(\tau_1 \mid \theta)\right\|^2 - \left\|\nabla J_H(\theta)\right\|^2\right]}{m} \\
&\leq \left\|\nabla J_H(\theta)\right\|^2 + \frac{\Gamma_g^2 - \left\|\nabla J_H(\theta)\right\|^2}{m},
\end{aligned} \tag{44}$$

where the third and the fourth lines are obtained by $\nabla J_H(\theta) = \mathbb{E}\left[g(\tau_i \mid \theta)\right]$, and the last line is obtained by Lemma D.1 (ii). If $\widehat{\nabla}_m J(\theta)$ is a REINFORCE gradient estimator, then $\Gamma_g = \frac{HG\mathcal{R}_{\max}}{1-\gamma}$; if $\widehat{\nabla}_m J(\theta)$ is a GPOMDP gradient estimator, then $\Gamma_g = \frac{G\mathcal{R}_{\max}}{(1-\gamma)^2}$. By rearranging, we obtain the lemma's claim. □

E. Proof of Lemma 4.4

Proof. From (5), we have

$$\begin{aligned}
\left\|\nabla J(\theta) - \nabla J_H(\theta)\right\| &= \left\|\mathbb{E}_\tau\left[\sum_{t=H}^{\infty}\left(\sum_{k=0}^{t}\nabla_\theta \log \pi_\theta(a_k \mid s_k)\right)\gamma^t \mathcal{R}(s_t, a_t)\right]\right\| \\
&\leq \mathbb{E}_\tau\left[\sum_{t=H}^{\infty}\gamma^t \left|\mathcal{R}(s_t, a_t)\right|\left(\sum_{k=0}^{t}\left\|\nabla_\theta \log \pi_\theta(a_k \mid s_k)\right\|\right)\right] \\
&\leq \mathbb{E}_\tau\left[G\mathcal{R}_{\max}\sum_{t=H}^{\infty}(t+1)\gamma^t\right] \\
&= G\mathcal{R}_{\max}\gamma^H\sum_{t=0}^{\infty}(t+1+H)\gamma^t \\
&\overset{\text{Lemma A.1}}{=} \left(\frac{1}{(1-\gamma)^2} + \frac{H}{1-\gamma}\right)G\mathcal{R}_{\max}\gamma^H,
\end{aligned} \tag{45}$$

where the third line is obtained by using $|\mathcal{R}(s_t, a_t)| \leq \mathcal{R}_{\max}$ and $\|\nabla_\theta \log \pi_\theta(a_k \mid s_k)\| \leq G$ from Assumption 4.1. Thus $D' = \left(\frac{1}{(1-\gamma)^2} + \frac{H}{1-\gamma}\right)G\mathcal{R}_{\max}$.

Next, by inequality of Cauchy-Swartz we have

$$\begin{aligned}
\left|\langle \nabla J_H(\theta), \nabla J_H(\theta) - \nabla J(\theta)\rangle\right| &\leq \left\|\nabla J_H(\theta)\right\|\left\|\nabla J_H(\theta) - \nabla J(\theta)\right\| \\
&\overset{(11)}{\leq} \left\|\nabla J_H(\theta)\right\| \cdot D'\gamma^H \\
&\leq \frac{D'G\mathcal{R}_{\max}}{(1-\gamma)^2}\gamma^H,
\end{aligned} \tag{46}$$

where the last line is obtained by Lemma D.1 (iii). Thus $D = \frac{D'G\mathcal{R}_{\max}}{(1-\gamma)^2}$. □

F. Proof of Corollary 4.5

Proof. From Lemma 4.2, we know that J is L-smooth. Consider policy gradient with a mini-batch sampling of size m. From Lemma 4.3, we have Assumption 3.2 holds with $A = 0$, $B = 1 - \frac{1}{m}$ and $C = \Gamma_g^2/m$. Assumption 3.3 is verified as well by Lemma 4.4 with appropriate D and D'. By Proposition 3.4, plugging $A = 0$, $B = 1 - \frac{1}{m}$ and $C = \Gamma_g^2/m$ in (13) yields the corollary's claim with step size $\eta \in \left(0, \frac{2}{L\left(1-\frac{1}{m}\right)}\right)$. □

F.1. Sample complexity

Consider vanilla policy gradient with step size $\eta \in \left(0, \frac{1}{L\left(1-\frac{1}{m}\right)}\right)$ and a mini-batch sampling of size m. From (23), we have

$$\begin{aligned}\mathbb{E}\left[\|\nabla J(\theta_U)\|^2\right] &\leq \frac{2\delta_0}{\eta T\left(2 - L\eta\left(1-\frac{1}{m}\right)\right)} + \frac{L\Gamma_g^2\eta}{m\left(2 - L\eta\left(1-\frac{1}{m}\right)\right)} + \left(\frac{2D\left(3 - L\eta\left(1-\frac{1}{m}\right)\right)}{2 - L\eta\left(1-\frac{1}{m}\right)} + D'^2\gamma^H\right)\gamma^H \\ &\leq \frac{2\delta_0}{\eta T} + \frac{L\Gamma_g^2\eta}{m} + \left(6D + D'^2\gamma^H\right)\gamma^H, \end{aligned} \tag{47}$$

where the second inequality is obtained by $\frac{1}{2-L\eta\left(1-\frac{1}{m}\right)} \leq 1$ with $\eta \in \left(0, \frac{1}{L\left(1-\frac{1}{m}\right)}\right)$.

To get $\mathbb{E}\left[\|\nabla J(\theta_U)\|^2\right] = \mathcal{O}(\epsilon^2)$, it suffices to have

$$\mathcal{O}(\epsilon^2) \quad \geq \quad \frac{2\delta_0}{\eta T} + \frac{L\Gamma_g^2\eta}{m} \tag{48}$$

and

$$\mathcal{O}(\epsilon^2) \quad \geq \quad \left(6D + D'^2\gamma^H\right)\gamma^H \tag{49}$$

respectively. To make the right hand side of (49) smaller than ϵ^2, we need $H\gamma^H = \mathcal{O}(\epsilon^2)$. Thus, we require

$$H = \mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right) / \log\left(\frac{1}{\gamma}\right)\right).$$

To make the right hand side of (48) smaller than ϵ^2, we require

$$\frac{L\Gamma_g^2\eta}{m} \leq \frac{\epsilon^2}{2} \iff \eta \leq \frac{\epsilon^2 m}{2L\Gamma_g^2}. \tag{50}$$

Similarly, for the first term of the right hand side of (48), we require

$$\frac{2\delta_0}{\eta T} \leq \frac{\epsilon^2}{2} \iff \frac{4\delta_0}{\epsilon^2 T} \leq \eta. \tag{51}$$

Combine the two inequality, we get

$$\frac{4\delta_0}{\epsilon^2 T} \leq \eta \leq \frac{\epsilon^2 m}{2L\Gamma_g^2}. \tag{52}$$

This implies

$$Tm \quad \geq \quad \frac{8\delta_0 L\Gamma_g^2}{\epsilon^4}. \tag{53}$$

The condition on the step size $\eta \in \left(0, \frac{1}{L\left(1-\frac{1}{m}\right)}\right)$ requires the mini-batch size satisfy

$$\frac{\epsilon^2 m}{2L\Gamma_g^2} < \frac{1}{L\left(1-\frac{1}{m}\right)} \implies m \leq \frac{2\Gamma_g^2}{\epsilon^2}.$$

To conclude, it suffices to choose step size $\eta = \frac{4\delta_0}{\epsilon^2 T} = \frac{\epsilon^2 m}{2L\Gamma_g^2}$, a mini-batch size m between 1 and $\frac{2\Gamma_g^2}{\epsilon^2}$, the number of iterations $T = \frac{8\delta_0 L\Gamma_g^2}{m\epsilon^4}$ and the fixed Horizon $H = \mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right) / \log\left(\frac{1}{\gamma}\right)\right)$ such that the equalities of (49), (50), (51), (52) and (53) hold, which guarantee $\mathbb{E}\left[\left\|\nabla J(\theta_U)\right\|^2\right] = \mathcal{O}(\epsilon^2)$. Here the total sample complexity is $Tm \times H = \widetilde{\mathcal{O}}(\epsilon^{-4})$.