A general sample complexity analysis of vanilla policy gradient

Rui Yuan1 2 3  Alessandro Lazaric1  Robert M. Gower2 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 $\tilde{O}(\epsilon^{-4})$ 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., 2018) or with soft-max 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 $O(\epsilon^{-4})$ sample complexity guarantees, but for wider ranges of parameters (e.g., step size and batch size) with respect to previous literature. Notably,
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 identifying 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 \( \mathcal{R}(s, a) \defeq \mathbb{E}_{s' \sim \mathcal{P}(\cdot \mid s, a)}[\mathcal{R}(s, a, s')] \in [-\mathcal{R}_{\text{max}}, \mathcal{R}_{\text{max}}] \) is the expected reward for state-action pair \((s, a); \gamma \in [0, 1)\) is the discount factor; and \( \rho \) is the initial state distribution. The agent’s behavior is modeled as a policy \( \pi \in \mathcal{A}^\mathcal{S} \), where \( \pi(- \mid s) \) 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 \( \pi \) on the set \( \mathcal{T} \) of all possible trajectories, that is

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

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

\[
J(\pi) = \mathbb{E}_{\tau \sim p(\cdot \mid \pi, \mathcal{M})}[\mathcal{R}(\tau)] \defeq \mathbb{E}_{\tau \sim p(\cdot \mid \pi)}[\mathcal{R}(\tau)].
\]

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 \( \pi_\theta \) is differentiable w.r.t. \( \theta \). For simplicity, we consider \( \Theta \subseteq \mathbb{R}^d \). We denote \( J(\theta) = J(\pi_\theta) \) and \( p(\tau \mid \theta) = p(\tau \mid \pi_\theta) \).

We also define \( J^* = \sup_{\pi} J(\theta) \) the optimal expected total reward and \( \theta^* = \arg \sup_{\pi} J(\theta) \) 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:

\[
\nabla J(\theta) = \int \mathcal{R}(\tau) \nabla p(\tau \mid \theta) d\tau
\]

\[
= \int \mathcal{R}(\tau) \left( \nabla p(\tau \mid \theta)/p(\tau \mid \theta) \right) p(\tau \mid \theta) d\tau
\]

\[
= \mathbb{E}_{\tau \sim p(\cdot \mid \theta)} [\mathcal{R}(\tau) \nabla \log p(\tau \mid \theta) \mid \mathcal{M}]
\]

\[
(\text{1}) \mathbb{E}_{\tau} \left[ \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \sum_{t'=0}^{\infty} \nabla \theta \log \pi_\theta(a_{t'} \mid s_{t'}). \right].
\]

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 \( \tau_i = (s_0, a_0, s_0, a_0, \ldots, s_{H-1}, a_{H-1}, s_0) \) obtained by executing \( \pi_\theta \) for a fixed horizon \( H \). Then the gradient estimator is computed as

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

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

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

which leads to the following gradient estimator

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

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 \( J_H(\theta) \defeq \mathbb{E}_{\tau} \left[ \sum_{t=0}^{H-1} \gamma^t \mathcal{R}(s_t, a_t) \right] \).

Equipped with gradient estimators, vanilla policy gradient simply updates the policy parameters as

\[
\theta_{k+1} = \theta_k + \eta \hat{\nabla}_m J(\theta_k),
\]

with a step size \( \eta > 0 \).

3. Non-convex optimization under ABC assumption

We use \( \hat{\nabla}_m J(\theta) \) 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) 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

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

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

\[
\left\| (\nabla J_H(\theta), \nabla J_H(\theta) - \nabla J(\theta)) \right\| \leq D \gamma^H, \tag{10}
\]

\[
\left\| \nabla J_H(\theta) - \nabla J(\theta) \right\| \leq D' \gamma^H. \tag{11}
\]

While we specifically need those conditions to hold as an assumption, we notice that they are reasonable since we have \(|J(\theta) - J_H(\theta)| \leq \frac{\max}{1 - \gamma} \gamma^H\) 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 \(\theta\) such that \(\left\| \nabla J_H(\theta) \right\|\) 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 \(\eta\) such that \(\eta \in \left(0, \frac{2}{L^2 T}\right)\) where \(B\) can be zero.\(^a\) Let \(\delta_0 \overset{\text{def}}{=} J^* - J(\theta_0)\). If \(A > 0\), then PG satisfies

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

If \(A = 0\), we have

\[
\begin{align*}
\min_{0 \leq t \leq T-1} E \left[ \left\| \nabla J(\theta_t) \right\|^2 \right] & \leq \frac{2\delta_0}{\eta T(2 - LB\eta)} + \frac{LC\eta}{2 - LB\eta} + \left(\frac{2D(3 - LB\eta)}{2 - LB\eta} + D' \gamma^H\right) \gamma^H, \tag{13}
\end{align*}
\]

where \(\theta_U\) is uniformly sampled from \(\{\theta_0, \theta_1, \ldots, \theta_T\}\).

\(^a\)We set \(\frac{1}{\eta} = \infty\).

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

\[
\begin{align*}
\eta & = \min \left\{ \frac{1}{\sqrt{LAT}}, \frac{1}{LB}, \frac{1}{2LC} \right\}, \\
T & \geq \frac{12\delta_0 L}{\epsilon^2} \max \left\{ B, \frac{12\delta_0 A}{2C}, \frac{2C}{\epsilon^2} \right\}, \tag{14}
\end{align*}
\]

\[
H = O(\log \epsilon^{-1}),
\]

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

First, the iteration complexity (14) recovers the exact full gradient case. That is, considering \(H = \infty\) (i.e. \(J_H = J\)) and \(\nabla m J(\theta) = \nabla J(\theta)\) in (7), we have Asm. 3.2 and 3.3 hold automatically with \(A = C = D = D' = 0\) and \(B = 1\). Consequently, we require \(T = O(\epsilon^{-2})\) iterations to reach an \(\epsilon\)-stationary point. Thus, for any policy and MDP that satisfy the smoothness property (Asm. 3.1), the exact full PG converge in \(O(\epsilon^{-2})\) 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 \(O(\epsilon^{-2})\) rate of convergence for the exact full gradient in the constrained tabular parametrized policy. By leveraging the hidden convex structure using composite
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 $O(e^{-1})$ for the exact full gradient. Mei et al. (2020) also establishes the same convergence rate $O(e^{-1})$ 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 $TH = O(e^{-4})$ 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

$$\| \nabla_\theta \log \pi_\theta(a \mid s) \| \leq G, \quad \| \nabla^2_\theta \log \pi_\theta(a \mid s) \| \leq F. \quad (15), (16)$$

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 $\frac{\partial}{\partial \theta_0} \log \pi_\theta(a \mid s)$ and $\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log \pi_\theta(a \mid s)$ 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

$$\| \nabla^2 J(\theta) \| \leq L \quad \text{for all } \theta$$

of Asm. 3.1, with

$$L = \frac{2G^2R_{\max}}{(1 - \gamma)^3} + \frac{FR_{\max}}{(1 - \gamma)^2}. \quad (17)$$

**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. Compared to existing bounds, our result reveals an additional term depending on $(1 - \gamma)^{-3}$ which dominates the term $\frac{FR_{\max}}{(1 - \gamma)^2}$ derived in (Xu et al., 2020b) whenever $\gamma$ is close to 1.

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

$$E \left[ \| \nabla_m J(\theta) \|^2 \right] \leq \left( 1 - \frac{1}{m} \right) \| \nabla J_H(\theta) \|^2 + \frac{\Gamma^2_g}{m}, \quad (18)$$

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

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

$$\text{Var} \left[ \nabla_m J(\theta) \right] = E \left[ \| \nabla_m J(\theta) \|^2 \right] - \| \nabla J_H(\theta) \|^2 \leq \frac{\Gamma^2_g - \| \nabla J_H(\theta) \|^2}{m} \leq \frac{\Gamma^2_g}{m}, \quad (19)$$

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

$$D = \frac{D' GR_{\max}}{(1 - \gamma)^2}, \quad (20)$$

$$D' = \left( \frac{1}{(1 - \gamma)^2} + \frac{H}{1 - \gamma} \right) GR_{\max}. \quad (21)$$

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.
**Corollary 4.5.** Suppose that Assumption 4.1 is satisfied. Let \( \delta_0 = J^* - J(\theta_0) \). Any PG method with a mini-batch sampling of size \( m \) and step size
\[
\eta \in \left( 0, \frac{2}{L(1 - 1/m)} \right),
\]
we have
\[
\mathbb{E} \left[ \|\nabla J(\theta_U)\|^2 \right] \leq \frac{2\delta_0}{\eta T \left( 2 - L\eta \left( 1 - \frac{1}{m} \right) \right)} + \frac{L\Gamma^2 \eta}{m \left( 2 - L\eta \left( 1 - \frac{1}{m} \right) \right)} + \frac{\left( 2D \left( 3 - L\eta \left( 1 - \frac{1}{m} \right) \right) \right)}{2 - L\eta \left( 1 - \frac{1}{m} \right)} + D'^2 \gamma^H, \tag{23}
\]
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 \( \epsilon \), if we set the parameters as (the detailed derivation is provided in App. F.1)
\[
m \in \left[ 1; \frac{2\Gamma^2 \eta}{\epsilon^2} \right],
\]
\[
T \text{ s.t. } Tm \geq \frac{8\delta_0 L\Gamma^2 \eta}{\epsilon^4},
\]
\[
\eta = \frac{\epsilon^2 m}{2L\Gamma^2 \eta},
\]
\[
H = \mathcal{O} \left( \log \left( 1/\epsilon \right) / \log \left( 1/\gamma \right) \right),
\]
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., \( \tilde{\mathcal{O}}(\epsilon^{-4}) \).

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


A general sample complexity analysis of vanilla policy gradient


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

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

$$= 1 + \sum_{t=1}^{\infty} (t + 1 - t) \gamma^t$$

$$= \sum_{t=0}^{\infty} \gamma^t$$

$$= \frac{1}{1 - \gamma}.$$

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

$$\gamma S = \sum_{t=0}^{\infty} (t + 1)^2 \gamma^{t+1} = \sum_{t=1}^{\infty} t^2 \gamma^t.$$
A general sample complexity analysis of vanilla policy gradient

Thus, the subtraction of the above two equations gives

\[(1 - \gamma)S = \sum_{t=0}^{\infty} (t+1)^2 \gamma^t - \sum_{t=1}^{\infty} t^2 \gamma^t
\]

\[= 1 + \sum_{t=1}^{\infty} ((t+1)^2 - t^2) \gamma^t
\]

\[= 1 + \sum_{t=1}^{\infty} (2t+1) \gamma^t
\]

\[= \sum_{t=0}^{\infty} (2t+1) \gamma^t
\]

\[= 2 \sum_{t=0}^{\infty} (t+1) \gamma^t - \sum_{t=0}^{\infty} \gamma^t
\]

Lemma A.1

\[= 2 \left( \frac{1}{(1 - \gamma)^2} - 1 \right)
\]

\[\leq 2 \left( \frac{1}{(1 - \gamma)^2} \right).
\]

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

\[J(\theta_{t+1}) \geq J(\theta_t) + \eta \left\langle \nabla J(\theta_t), \nabla H(\theta_t) \right\rangle - \frac{L \eta^2}{2} \left\| \nabla H(\theta_t) \right\|^2 .
\]

Taking expectations conditioned on \(\theta_t\), we get

\[E_t [J(\theta_{t+1})] \geq J(\theta_t) + \eta \left\langle \nabla J(\theta_t), \nabla H(\theta_t) \right\rangle - \frac{L \eta^2}{2} E_t \left[ \left\| \nabla H(\theta_t) \right\|^2 \right] .
\]

\[\geq J(\theta_t) + \eta \left\langle \nabla J(\theta_t), \nabla H(\theta_t) \right\rangle - \frac{L \eta^2}{2} \left( 2A J^* - J(\theta_t) \right) + B \left\| \nabla H(\theta_t) \right\|^2 + C
\]

\[= J(\theta_t) + \eta \left( 1 - \frac{LB \eta}{2} \right) \left\| \nabla H(\theta_t) \right\|^2 - L \eta^2 A J^* - J(\theta_t)
\]

\[- \frac{LC \eta^2}{2} + \eta \left\langle \nabla H(\theta_t), \nabla J(\theta_t) - \nabla H(\theta_t) \right\rangle
\]

\[\geq J(\theta_t) + \eta \left( 1 - \frac{LB \eta}{2} \right) \left\| \nabla H(\theta_t) \right\|^2 - L \eta^2 A J^* - J(\theta_t)
\]

\[- \frac{LC \eta^2}{2} - \eta D \gamma H .
\]

Subtracting \(J^*\) from both sides gives

\[- (J^* - E_t [J(\theta_{t+1})]) \geq -(1 + L \eta^2 A) J^* - J(\theta_t) + \eta \left( 1 - \frac{LB \eta}{2} \right) \left\| \nabla H(\theta_t) \right\|^2
\]

\[- \frac{LC \eta^2}{2} - \eta D \gamma H .
\]
Taking the total expectation and rearranging, we get

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

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

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

We now introduce a sequence of weights $w_{-1}, w_0, w_1, \ldots, w_T$ 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 $w_t/\eta$,

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

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

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

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

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

Note that,

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

Using this in (31),

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

However, we have

$$\mathbb{E} \left[ \Vert \nabla J(\theta_t) \Vert^2 \right] = \mathbb{E} \left[ \Vert \nabla J(\theta_t) - \nabla J_H(\theta_t) + \nabla J_H(\theta_t) \Vert^2 \right] = \mathbb{E} \left[ \Vert \nabla J_H(\theta_t) \Vert^2 \right] + 2 \mathbb{E} \left[ \Vert \nabla J_H(\theta_t) \Vert^2 \right] + \mathbb{E} \left[ \Vert \nabla J(\theta_t) - \nabla J_H(\theta_t) \Vert^2 \right] \leq \mathbb{E} \left[ \Vert \nabla J_H(\theta_t) \Vert^2 \right] + 2 D\gamma H + D^2\gamma^2 H. \quad (34)$$

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

$$\left( 1 - \frac{LB\eta}{2} \right) \min_{0 \leq t \leq T-1} \mathbb{E} \left[ \Vert \nabla J(\theta_t) \Vert^2 \right] \leq \frac{(1 + L\eta^2 A)^T}{\eta T} \delta_0 + \frac{LC\eta}{2} + D\gamma H + \left( 1 - \frac{LB\eta}{2} \right) (2 D\gamma H + D^2\gamma^2 H).$$
A general sample complexity analysis of vanilla policy gradient

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,

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

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

\[
\left( 1 - \frac{LB\eta}{2} \right) \mathbb{E} \left[ \|\nabla J(\theta_U)\|^2 \right] = \left( 1 - \frac{LB\eta}{2} \right) \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[ \|\nabla J(\theta_t)\|^2 \right] \leq \frac{\delta_0}{\eta T} + \frac{LC\eta}{2} + D\gamma^H + \left( 1 - \frac{LB\eta}{2} \right) \left( 2D\gamma^H + D^2\gamma^{2H} \right).
\]

Dividing both sides by \( 1 - \frac{LB\eta}{2} \) and rearranging yields the proposition’s claim. \( \square \)

C. Proof of Lemma 4.2

**Proof.** We know that

\[
\nabla^2 J(\theta)_t \overset{(5)}{=} \nabla_\theta \mathbb{E}_\tau \left[ \sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \left( \sum_{k=0}^{t} \nabla_\theta \log p(\pi_\theta | s_k) \right) \right]
\]

\[
= \nabla_\theta \int p(\tau | \theta) \sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \left( \sum_{k=0}^{t} \nabla_\theta \log p(\pi_\theta | s_k) \right) d\tau
\]

\[
= \int \nabla_\theta p(\tau | \theta) \left( \sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \left( \sum_{k=0}^{t} \nabla_\theta \log p(\pi_\theta | s_k) \right) \right)^{\top} d\tau
\]

\[
+ \int p(\tau | \theta) \sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \left( \sum_{k=0}^{t} \nabla_\theta^2 \log p(\pi_\theta | s_k) \right) d\tau
\]

\[
= \mathbb{E}_\tau \left[ \nabla_\theta p(\tau | \theta) \left( \sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \left( \sum_{k=0}^{t} \nabla_\theta \log p(\pi_\theta | s_k) \right) \right)^{\top} \right]
\]

\[
+ \mathbb{E}_\tau \left[ \sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \left( \sum_{k=0}^{t} \nabla_\theta^2 \log p(\pi_\theta | s_k) \right) \right]
\]

\[
\overset{(1)}{=} \mathbb{E}_\tau \left[ \sum_{t=0}^{\infty} \nabla_\theta \log p(\pi_\theta | a_t) \left( \sum_{t=0}^{t} \gamma^t R(s_t, a_t) \left( \sum_{k=0}^{t} \nabla_\theta \log p(\pi_\theta | s_k) \right) \right)^{\top} \right]
\]

\[
+ \mathbb{E}_\tau \left[ \sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \left( \sum_{k=0}^{t} \nabla_\theta^2 \log p(\pi_\theta | s_k) \right) \right].
\]
We aim to bound two terms separately. The second term can be bounded easily. That is,

\[
\left\| \mathcal{O} \right\| \leq E_T \left[ \sum_{t=0}^{\infty} \gamma^t \left| \mathcal{R}(s_t, a_t) \right| \left( \sum_{k=0}^{t} \left\| \nabla_\theta \log \pi_\theta(a_k | s_k) \right\| \right) \right]
\]

\[
\leq E_T \left[ F\mathcal{R}_{\text{max}} \sum_{t=0}^{\infty} (t+1) \gamma^t \right]
\]

\[
= \frac{F\mathcal{R}_{\text{max}}}{(1-\gamma)^2},
\]

(37)

where the second line is obtained by using $|\mathcal{R}(s_t, a_t)| \leq \mathcal{R}_{\text{max}}$ and $\|\nabla_\theta^2 \log \pi_\theta(a_k | s_k)\| \leq F$ 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 $\{x_t\}_{t \geq 0}$ 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 $0 \leq t < t'$ among terms of the two sums in (1), we have

\[
E_T \left[ \nabla_\theta \log \pi_\theta(a_{t'} | s_{t'}) \cdot \gamma^t \mathcal{R}(s_t, a_t) \left( \sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k | s_k) \right)^T \right]
\]

\[
= E_{s_0, t', a_0, t'} \left[ \nabla_\theta \log \pi_\theta(a_{t'} | s_{t'}) \cdot \gamma^t \mathcal{R}(s_t, a_t) \left( \sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k | s_k) \right)^T \right]
\]

\[
= E_{s_0, t', a_0, (t'-1)} \left[ E_{a_{t'}} \left[ \nabla_\theta \log \pi_\theta(a_{t'} | s_{t'}) \cdot \gamma^t \mathcal{R}(s_t, a_t) \left( \sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k | s_k) \right)^T \right]_{s_0, t', a_{t'(t'-1)}} \right]
\]

\[
= E_{s_0, t', a_0, (t'-1)} \left[ \int \nabla_\theta \pi_\theta(a_{t'} | s_{t'}) \cdot \gamma^t \mathcal{R}(s_t, a_t) \left( \sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k | s_k) \right)^T \right]
\]

\[
= E_{s_0, t', a_0, (t'-1)} \left[ \int \nabla_\theta \pi_\theta(a_{t'} | s_{t'}) \cdot \gamma^t \mathcal{R}(s_t, a_t) \left( \sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k | s_k) \right)^T \right]
\]

\[
= 0.
\]

(38)

Thus, (1) can be simplified. We have

\[
\mathcal{O} = E_T \left[ \sum_{t' = 0}^{t} \nabla_\theta \log \pi_\theta(a_{t'} | \theta_{t'}) \left( \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \left( \sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k | s_k) \right)^T \right) \right]
\]

\[
= E_T \left[ \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \left( \sum_{t' = 0}^{t} \nabla_\theta \log \pi_\theta(a_{t'} | \theta_{t'}) \left( \sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k | s_k) \right)^T \right) \right].
\]

(39)
Now we can bound (1) easily. That is,

\[
\|\mathbb{E}_T \left[ \sum_{t=0}^{\infty} \gamma^t |R(s_t, a_t)| \left\| \sum_{t'=0}^t \nabla_{\theta} \log \pi_{\theta}(a_{t'} \mid \theta_{t'}) \right\|^2 \right] \|
\]

\[
\leq \mathbb{E}_T \left[ \sum_{t=0}^{\infty} \gamma^t |R(s_t, a_t)| \left( \sum_{t'=0}^t \| \nabla_{\theta} \log \pi_{\theta}(a_{t'} \mid \theta_{t'}) \| \right)^2 \right]
\]

\[
\leq \mathbb{E}_T \left[ G^2 \mathcal{R}_{\max} \sum_{t=0}^{\infty} (t+1)^2 \gamma^t \right]
\]

\[
\leq \frac{2G^2 \mathcal{R}_{\max}}{(1-\gamma)^3}
\]

(40)

where the third line is obtained by using \(|R(s_t, a_t)| \leq \mathcal{R}_{\max}\) and \(\| \nabla_{\theta} \log \pi_{\theta}(a_{t'} \mid s_{t'}) \| \leq G\) 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. \(\square\)

**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 \(\tau_i\) and \(\theta \in \mathbb{R}^d\), we have

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

(ii) The gradient estimator is bounded, i.e. \(\|\hat{\nabla}_m J(\theta)\| \leq \Gamma_g\).

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

Furthermore, if \(\hat{\nabla}_m J(\theta)\) is a REINFORCE gradient estimator, then \(L_g = \frac{H \mathcal{R}_{\max}}{1-\gamma}\) and \(\Gamma_g = \frac{H \mathcal{R}_{\max}}{1-\gamma}\); if \(\hat{\nabla}_m J(\theta)\) is a GPOMDP gradient estimator, then \(L_g = \frac{\mathcal{R}_{\max}}{(1-\gamma)^2}\) 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 \(\hat{\nabla}_m J(\theta)\) be a REINFORCE gradient estimator. From (4), we have

\[
\| \nabla \left( \hat{\nabla}_m J(\theta) \right) \| = \left\| \frac{1}{m} \sum_{i=1}^{m} \sum_{t=0}^{H-1} \sum_{t'=0}^{H-1} \gamma^{t'} R(s_{t'}, a_{t'}) \nabla_{\theta} \log \pi_{\theta}(a_{t'} \mid s_{t'}) \right\|
\]

\[
\leq \frac{1}{m} \sum_{i=1}^{m} \sum_{t=0}^{H-1} \sum_{t'=0}^{H-1} \gamma^{t'} \left| R(s_{t'}, a_{t'}) \right| \sum_{t=0}^{H-1} \| \nabla_{\theta} \log \pi_{\theta}(a_t \mid s_t) \|
\]

\[
\leq H F \mathcal{R}_{\max} \sum_{t'=0}^{H-1} \gamma^{t'}
\]

\[
\leq \frac{H F \mathcal{R}_{\max}}{1-\gamma},
\]

(41)

where the third line is obtained by using \(|R(s_{t'}, a_{t'})| \leq \mathcal{R}_{\max}\) and \(\| \nabla_{\theta} \log \pi_{\theta}(a_{t'} \mid s_{t'}) \| \leq F\) from Assumption 4.1. In this case, \(L_g = \frac{H F \mathcal{R}_{\max}}{1-\gamma}\).
Let \( \hat{\nabla}_m J(\theta) \) be a GPOMDP gradient estimator. From (6), we have

\[
\| \nabla \left( \hat{\nabla}_m J(\theta) \right) \| = \frac{1}{m} \sum_{i=1}^{m} \left( \sum_{t=0}^{H-1} \gamma^t \mathcal{R}(s_t, a_t^i) \left( \sum_{k=0}^{t} \nabla^2 \log \pi_\theta(a_k^i | s_k) \right) \right) \leq \frac{F \mathcal{R}_{\max}}{(1-\gamma)^2},
\]

where similarly, the third line is obtained by using \( |\mathcal{R}(s_t^i, a_t^i)| \leq \mathcal{R}_{\max} \) and \( \| \nabla^2 \log \pi_\theta(a_k^i | s_k) \| \leq F \) 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 \( \nabla \left( \hat{\nabla}_m J(\theta) \right) \) by \( \hat{\nabla}_m J(\theta) \), \( \nabla^2 \log \pi_\theta(a_k^i | s_k^i) \) by \( \nabla \log \pi_\theta(a_t^i | s_t^i) \) and \( F \) by \( G \). If \( \hat{\nabla}_m J(\theta) \) is a REINFORCE gradient estimator, we have \( \Gamma_g = \frac{H \mathcal{R}_{\max}}{(1-\gamma)} \); if \( g(\tau | \cdot) \) is a GPOMDP gradient estimator, then \( \Gamma_g = \frac{G \mathcal{R}_{\max}}{(1-\gamma)} \).

To prove (iii), notice that

\[
\| \nabla J(\theta) \| \overset{(5)}{=} \quad \mathbb{E}_\tau \left[ \sum_{t=0}^{\infty} \left( \sum_{k=0}^{t} \nabla \log \pi_\theta(a_k | s_k) \right) \gamma^t \mathcal{R}(s_t, a_t) \right] \leq \quad \mathbb{E}_\tau \left[ \sum_{t=0}^{\infty} \left( \sum_{k=0}^{t} \nabla \log \pi_\theta(a_k | s_k) \| \right) \gamma^t |\mathcal{R}(s_t, a_t)| \right] \leq \quad \mathbb{E}_\tau \left[ G \mathcal{R}_{\max} \sum_{t=0}^{\infty} (t+1) \gamma^t \right] \]

\[
\overset{\text{Lemma A.1}}{=} \frac{G \mathcal{R}_{\max}}{(1-\gamma)(1-\gamma^2)},
\]

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

\[\square\]

D.2. The proof

Now we show the proof of Lemma 4.3.

\[\text{Proof.} \] We denote \( g(\tau | \theta) \) a stochastic gradient estimator of one single sampled trajectory \( \tau \). Thus \( \hat{\nabla}_m J(\theta) = \frac{1}{m} \sum_{i=1}^{m} g(\tau_i | \theta) \). Both \( \hat{\nabla}_m J(\theta) \) and \( g(\tau | \theta) \) are unbiased estimators of \( J_H(\theta) \). We have
\[ E \left[ \left\| \nabla_m J(\theta) \right\|^2 \right] = E \left[ \left\| \frac{1}{m} \sum_{i=0}^{m-1} g(\tau_i | \theta) \right\|^2 \right] \\
= E \left[ \left\| \frac{1}{m} \sum_{i=0}^{m-1} g(\tau_i | \theta) - \nabla_J H(\theta) + \nabla_J H(\theta) \right\|^2 \right] \\
= \| \nabla_J H(\theta) \|^2 + E \left[ \left\| \frac{1}{m} \sum_{i=0}^{m-1} g(\tau_i | \theta) - \nabla_J H(\theta) \right\|^2 \right] \\
= \| \nabla_J H(\theta) \|^2 + \frac{1}{m^2} \sum_{i=0}^{m-1} E \left[ \| g(\tau_i | \theta) - \nabla_J H(\theta) \|^2 \right] \\
\leq \| \nabla_J H(\theta) \|^2 + \frac{\Gamma g^2 - \| \nabla_J H(\theta) \|^2}{m}, \quad (44) \]

where the third and fourth lines are obtained by \( \nabla_J H(\theta) = \mathbb{E} [g(\tau_i | \theta)] \), and the last line is obtained by Lemma D.1 (ii). If \( \nabla_m J(\theta) \) is a REINFORCE gradient estimator, then \( \Gamma g = \frac{HG_{\text{max}}}{1 - \gamma} \); if \( \nabla_m J(\theta) \) is a GPOMDP gradient estimator, then \( \Gamma g = \frac{G_{\text{max}}}{(1 - \gamma)^2} \). By rearranging, we obtain the lemma’s claim. \( \Box \)

**E. Proof of Lemma 4.4**

**Proof.** From (5), we have

\[
\| \nabla J(\theta) - \nabla_J H(\theta) \| = \left\| \mathbb{E}_t \left[ \sum_{t=0}^{\infty} \left( \sum_{k=0}^{t} \nabla_{\theta} \log \pi_{\theta}(a_k | s_k) \right) \gamma^t \mathcal{R}(s_t, a_t) \right] \right\|
\]

\[
\leq \mathbb{E}_t \left[ \sum_{t=0}^{\infty} \gamma^t |\mathcal{R}(s_t, a_t)| \left( \sum_{k=0}^{t} \| \nabla_{\theta} \log \pi_{\theta}(a_k | s_k) \| \right) \right]
\]

\[
\leq \mathbb{E}_t \left[ GR_{\text{max}} \sum_{t=0}^{\infty} (t+1)\gamma^t \right]
\]

\[
= GR_{\text{max}} \gamma \sum_{t=0}^{\infty} (t+1+H)\gamma^t
\]

\[
= GR_{\text{max}} \gamma \left( \frac{1}{(1-\gamma)^2} + \frac{H}{1-\gamma} \right) GRE_{\text{max}} \gamma^H, \quad (45)
\]

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

Next, by inequality of Cauchy-Swarz we have

\[
|\langle \nabla_J H(\theta), \nabla_J H(\theta) - \nabla_J(\theta) \rangle | \leq \| \nabla_J H(\theta) \| \| \nabla_J H(\theta) - \nabla_J(\theta) \| \leq (1) \| \nabla_J H(\theta) \| \cdot D' \gamma^H
\]

\[
\leq \frac{D' G R_{\text{max}} \gamma^H}{(1 - \gamma)^2}, \quad (46)
\]

where the last line is obtained by Lemma D.1 (iii). Thus \( D = \frac{D' G R_{\text{max}} \gamma^H}{(1 - \gamma)^2} \). \( \Box \)
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 \( \eta \in \left( 0, \frac{2}{L(1 - \frac{1}{m})} \right) \).

F.1. Sample complexity

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

\[
\mathbb{E} \left[ \| \nabla J(\theta_U) \|^2 \right] \leq \frac{2\delta_0}{\eta T} \left( 2 - L \eta \left( 1 - \frac{1}{m} \right) \right) + \frac{L \Gamma_g^2 \eta}{m} \left( 2 - L \eta \left( 1 - \frac{1}{m} \right) \right) + \left( \frac{2D}{2 - L \eta \left( 1 - \frac{1}{m} \right)} \right) \gamma H
\]

\[
\leq \frac{2\delta_0}{\eta T} + \frac{L \Gamma_g^2 \eta}{m} + (6D + D^2 \gamma H) \gamma H, \tag{47}
\]

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

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

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

and

\[
O(\epsilon^2) \geq (6D + D^2 \gamma H) \gamma H \tag{49}
\]

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

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

To make the right hand side of (48) smaller than \( \epsilon^2 \), 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

\[
T m \geq \frac{8\delta_0 \Gamma_g^2}{\epsilon^4}. \tag{53}
\]

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

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