Recent Advances in Policy Gradient Methods

Igor Sokolov
KAUST

December 11 2022

Abstract

Policy gradient (PG) is one of the most widely used model-free reinforcement learning (RL) strategy of finding optimal policies that maximize long-term rewards [43, 39, 2]. The success of PG methods is due to the simplicity and flexibility,and it can be easily implemented to solve many problems (including both non-Markov and partially observed environments), as well as efficiently combined with other tricks and techniques to achieve more sophisticated algorithms, such as the actor-critic [18, 25], natural PG [16], natural actor-critic [31, 4], policy mirror descent [40, 42], trust-region based variants [35, 36, 37], and variance-reduced methods [28, 38, 47, 49, 13, 32, 48, 12]. In practice, PG methods are known to be sample-inefficient due to the high variance affecting the gradient estimates, and many variants have been proposed to alleviate this issue, e.g., using baselines, variance-reduction techniques, and natural gradient variants. Unlike value-based methods (like TD-learning), a rigorous theoretical understanding of even the vanilla Policy Gradient method has been elusive for a long time. Recently, a complete theory of PG has been derived by exploring and exploiting the RL structure of the problem together with the theoretical techniques from non-convex optimization. In this project, we present an overview of the current theoretical state of the development of PG methods. In particular, in the first part of the project, we overview recent SOTA results unveiling the theoretical understanding of the vanilla policy gradient methods. In the second part, we gradually shift to the more advanced PG methods that improve sample complexity via different variance reduction techniques.

Contents

I A general sample complexity analysis of vanilla policy gradient **3**

1 Introduction **3**

2 Preliminaries **3**

3 General convergence results **5**
3.1 First-order stationary point guarantees 5
3.2 Global optimum guarantees under relaxed weak gradient domination 6

4 Applications **7**
4.1 Expected Lipschitz and smooth policies 7
4.2 Concrete policies and loss functions 9

5 Final remark **11**

II Variance-reduced PG methods **12**

6 Introduction **12**

7 Preliminaries **12**

8 Brief overview of recent SOTA methods **13**

9 Theoretical comparison **15**

Part I
A general sample complexity analysis of vanilla policy gradient

1 Introduction

This part is based on the recent results presented in [50]. In particular, we show the tightest known bounds on the sample complexity of PG for reaching a FOSP (first-order stationary point), global optimum under relaxed weak gradient domination for both exact and stochastic settings. Looking into details, [50] shows the variant of PG analysis under a very general assumption on the second moment bound of the truncated gradient estimation called the ABC assumption. As the gradient estimator, authors considered both REINFORCE [43] and GPOMDP [39, 2]. The first author's contribution is establishing the generic form of the convergence rate to the FOSP under ABC, smoothness, and truncation assumptions. As a direct result, they manage to recover optimal $\mathcal{O}(\epsilon^{-2})$ iteration complexity of exact PG and the $\widetilde{\mathcal{O}}(\epsilon^{-4})$ sample complexity of REINFORCE and GPOMDP (stochastic cases). The notable thing is that both of these results were obtained under weaker assumptions than had previously been done [53, 22, 45]. Furthermore, their analysis is less restrictive in terms of the hyper-parameter choices: it is allowed to launch a method within a broader range of step sizes without any restriction on the batch size m, even allowing for single trajectory sampling $(m = 1)$. Also, similar (in terms of generality) global convergence guarantees are obtained as well with the called relaxed weak gradient domination assumption. It turned out that the generality of the proposed in [50] assumption allows authors to unify much of the fragmented results in the literature under a single framework. Further, we illustrate that Lipschitz and smooth policies, Gaussian policies, and softmax tabular policies with or without a log barrier or an entropy regularize are all special cases satisfying the proposed generic assumption under a particular choice

2 Preliminaries

Before giving the rigorous formulations of previously announced results, we introduce the notation used throughout the report.

Markov decision process (MDP). We consider a MDP $M = \{\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, \rho\}$, where each symbol is defined as follows:

- $\mathcal{S}$ – state space;
- $\mathcal{A}$ – action space;
- $\mathcal{P}$ – Markovian transition model, where $\mathcal{P}(s' \mid s, a)$ is the transition density from state s to s' under action a;
- $\mathcal{R}$ – reward function, where $\mathcal{R}(s, a) \in [-\mathcal{R}_{\max}, \mathcal{R}_{\max}]$ is the uniformly-bounded reward;
- $\gamma \in [0,1)$ – discounted factor;
- ρ is the initial state distribution;

The agent's behavior in the environment is modeled via policy $\pi \in \Delta(\mathcal{A})^{\mathcal{S}}$, where $\pi(a \mid s)$ is the density of the distribution over actions at state $s \in \mathcal{S}$. We assume the infinite-horizon discounted setting. In particular, we consider episodic MDPs with effective horizon H. This differs from the finite-horizon setting, where the optimal policy is non-stationary. We can thus limit the attention to trajectories of length H.

Next, $p(\tau \mid \pi) \stackrel{\text{def}}{=} \rho(s_0) \prod_{t=0}^{\infty} \pi(a_t \mid s_t) \mathcal{P}(s_{t+1} \mid s_t, a_t)$ is the probability density of a single trajectory τ being sampled according to the policy π. Also, let $\mathcal{R}(\tau) \stackrel{\text{def}}{=} \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t)$ be

the total discounted reward summed along trajectory τ. We define the expected return of π as $J(\pi) \stackrel{\text{def}}{=} \mathbb{E}_{\tau \sim p(\cdot|\pi)}[\mathcal{R}(\tau)]$.

To compute the policy maximizing $J(\pi)$, one can use policy gradient methods.

Policy gradient.

Following [50], we introduce the necessary notion to define a policy gradient method. We firstly consider a set of parametrized policies$\{\pi_\theta : \theta \in \mathbb{R}^d\}$, with the assumption that π_θ is differentiable w.r.t. θ. We denote $J(\theta) = J(\pi_\theta)$ and $p(\tau \mid \theta) = p_\theta(\tau) = p(\tau \mid \pi_\theta)$. In general, $J(\theta)$ is a non-convex function. To find the policy that maximizes the expected return (i.e., $\theta^* \in \arg\sup_{\theta \in \mathbb{R}^d} J(\theta)$), the PG methods use gradient ascent in the space of θ. For simplicity, we consider $\Theta \subseteq \mathbb{R}^d$. We also define $J^* = \sup_{\theta \in \mathbb{R}^d} J(\theta)$ as the optimal expected total reward and the parameter of the optimal policy.

Finally, the gradient $\nabla J(\theta)$ of the expected return can be written the following way

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

However, in practice, we cannot compute this full gradient since computing the above expectation means averaging over all possible trajectories $\tau \sim p(\cdot \mid \theta)$. Because of that, the RL community, instead of the computation of full gradient, uses an empirical estimation based on sampling m truncated trajectories for some fixed horizon $H \in N$, i.e., $\tau_i = \left(s_0^i, a_0^i, r_0^i, s_1^i, \cdots, s_{H-1}^i, a_{H-1}^i, r_{H-1}^i\right)$ is obtained by executing π_θ where $r_t^i = \mathcal{R}(s_t^i, a_t^i)$.

Thus, the resulting gradient estimator (2) is known as the REINFORCE [43]

$$\hat{\nabla}_m J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \sum_{t=0}^{H-1} \gamma^t \mathcal{R}(s_t^i, a_t^i) \cdot \sum_{t'=0}^{H-1} \nabla_\theta \log \pi_\theta(a_{t'}^i \mid s_{t'}^i). \tag{2}$$

The REINFORCE estimator can be further simplified by exploiting the property that future actions do not depend on past rewards. This leads to the following gradient estimate (3) known as GPOMDP [2]

$$\hat{\nabla}_m J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \sum_{t=0}^{H-1} \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k^i \mid s_k^i)\right) \gamma^t \mathcal{R}(s_t^i, a_t^i). \tag{3}$$

The notable feature is that both REINFORCE and GPOMDP are unbiased gradient estimators [30] of the gradient of the truncated expected return $J_H(\theta) \stackrel{\text{def}}{=} \mathbb{E}_\tau \left[\sum_{t=0}^{H-1} \gamma^t \mathcal{R}(s_t, a_t)\right]$. [1]. Equipped with the necessary background (including concrete gradient estimators), one can finally state the vanilla policy gradient algorithm:

Algorithm 1 Vanilla policy gradient

Input: initialized θ_0, mini-batch size m, step size η_0
for $t = 0$ **to** $T - 1$ **do**
  Sample m trajectories following policy π_{θ_t} from the MDP
  Compute the policy gradient estimator $\hat{\nabla}_m J(\theta_t)$
  Update $\theta_{t+1} = \theta_t + \eta_t \hat{\nabla}_m J(\theta_t)$ and η_t,
end for

where $\widehat{\nabla}_m J(\theta)$ is the unbiased policy gradient estimator of $\nabla J_H(\theta)$. It can be either exact, or truncated (computed via REINFORCE and GPOMDP).

[1] If H is infinity, then $J_\infty(\cdot) = J(\cdot)$.

3 General convergence results

In this section, we state general convergence results for vanilla PG to the First-order stationary point (and some Global guarantees) as well as all necessary assumptions. All of the results are directly taken from [50].

3.1 First-order stationary point guarantees

Both reinforcement and supervised learning techniques in the smooth, non-convex setting often rely on the following standard assumption:

Assumption 1 (Smoothness)**.** *There exists $L > 0$ such that, for all $\theta, \theta' \in \mathbb{R}^d$, we have*

$$|J(\theta') - J(\theta) - \langle \nabla J(\theta), \theta' - \theta \rangle| \leq \frac{L}{2} \|\theta' - \theta\|^2 . \tag{4}$$

Further, one needs an additional assumption on the bound of the error due to the truncation of the horizon

Assumption 2 (Truncation)**.** *There exists $D, D' > 0$ such that, for all $\theta \in \mathbb{R}^d$, we have*

$$|\langle \nabla J_H(\theta), \nabla J_H(\theta) - \nabla J(\theta) \rangle| \;\leq\; D\gamma^H, \tag{5}$$

$$\|\nabla J_H(\theta) - \nabla J(\theta)\| \;\leq\; D'\gamma^H. \tag{6}$$

To the best of our knowledge, whereas assumption (6) has already appeared in the literature before [22], requirement (5) is one of the novelties of the paper.

The following assumptions (and the subsequent convergence results) are one of the critical contributions of this paper to the reinforcement learning field. [2]

Assumption 3 (ABC)**.** *There exists $A, B, C \geq 0$ such that the policy gradient estimator satisfies*

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

for all $\theta \in \mathbb{R}^d$.

This generic assumption summarizes several popular and more restrictive assumptions often used in non-convex optimization, such as bounded variance [9], the gradient confusion [33], the sure-smoothness [19], the expected smoothness [11, 10] and different variants of strong growth [34, 41, 5]. Indeed, all of the mentioned classical assumptions can be shown to be specific cases of Asm. 3. Moreover, the ABC assumption was shown to be currently the weakest among all existing ones to provide any of the considered in this report convergence guarantees.

We now state the main theorem that establishes the convergence to FOSP.

Theorem 1 ([50])**.** *Suppose that Asm. 1, 2 and 3 hold. Consider the iterates θ_t of the PG method (Algorithm 1) with stepsize $\eta_t = \eta \in \left(0, \frac{2}{LB}\right)$ where $B = 0$ means that $\eta \in (0, \infty)$. Let $\delta_0 \overset{def}{=} J^* - J(\theta_0)$. It follows that*

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

In particular if $A = 0$, we have

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

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

[2] The paper [50] did not invent it from scratch. Initially, it was proposed in [17] for smooth, non-convex supervised learning (optimization).

Theorem 1 provides a general form of the convergence result for PG as a function of all the constants involved in the assumptions characterizing the problem and the chosen policy. Next, from the Theorem 1, one can directly obtain sample complexity bounds.

Corollary 1 ([50]). *Consider the setting of Thh. 1 Given $\epsilon > 0$, let $\eta = \min\left\{\frac{1}{\sqrt{LAT}}, \frac{1}{LB}, \frac{\epsilon}{2LC}\right\}$ and $H = \mathcal{O}(\log \epsilon^{-1})$. If*

$$T \geq \frac{12\delta_0 L}{\epsilon^2} \max\left\{B, \frac{12\delta_0 A}{\epsilon^2}, \frac{2C}{\epsilon^2}\right\}, \tag{7}$$

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

We further show that Cor. 1 recovers the best-known iteration complexity for vanilla PG in several standard cases:

1. **Exact gradient method.** Let $H = m = \infty$ and $\hat{\nabla}_m J(\theta) = \nabla J(\theta)$, thus Asm. 2 and 3 hold with $A = C = D = D' = 0$ and $B = 1$. Plugging them to the (7), we recover $\mathcal{O}(\epsilon^{-2})$ iteration complexity. This is the optimal and non-improvable (without any other assumptions) convergence rate for the exact gradient descent on smooth, non-convex objectives [3].

2. **Stochastic vanilla PG** In this general case, A, C, D, D' are not all 0. Let $m = \mathcal{O}(1)$ (one can make such a choice since assumptions 1, 2, 3 do not impose any restrictions on the batch size m). Then, with the sample complexity $TH = \widetilde{\mathcal{O}}(\epsilon^{-4})$, once can reach stationary point. This matches sample complexity results from [27, 53, 22, 45], improving upon them in terms of generality. In particular, it uses weaker assumptions (1,2, 3) and provides a broader range of parameter choices.

3.2 Global optimum guarantees under relaxed weak gradient domination

The global optimum convergence of PG is an essential topic for RL literature. It's turned out that when the additional assumption is available (on top of the ABC assumption), one can get a global optimum convergence guarantee for vanilla PG.

As [1, 24, 44] did for the exact policy gradient update, relying on the following gradient domination assumption, [50] establishes a global optimum convergence guarantee and the sample complexity analysis for the stochastic vanilla PG.

Assumption 4 (Relaxed weak gradient domination). *We say that J satisfies the relaxed weak gradient domination condition if for all $\theta \in \mathbb{R}^d$, there exists $\mu > 0$ and $\epsilon' \geq 0$ such that*

$$\epsilon' + \|\nabla J_H(\theta)\| \;\geq\; 2\sqrt{\mu}\,(J^* - J(\theta))\,. \tag{8}$$

To the best of our knowledge, this assumption is unique to PG methods. Indeed, it can be seen as an extension of weak gradient domination [1, 24, 23] where $\epsilon' = 0$. With the same assumption, [50] gives a new global optimum convergence guarantee.

Theorem 2 ([50]). *Suppose that Assumption 1, 2, 3 and 4 hold. Given $\epsilon > 0$, define δ s.t. if $\epsilon' = 0$, set $\delta = \epsilon$, if $\epsilon' > 0$, set $\delta = \epsilon'$. Suppose that PG (Algorithm 1) is run for $T > 0$ iterations with step size $(\eta_t)_t$ chosen as*

$$\eta_t = \begin{cases} \frac{1}{b} & \text{if } T \leq \frac{b}{\mu\delta} \;\text{ or }\; t \leq t_0 \\ \frac{2}{2b + \mu\delta(t - t_0)} & \text{if } T \geq \frac{b}{\mu\delta} \;\text{ and }\; t > t_0 \end{cases} \tag{9}$$

with $t_0 = \left[\frac{T}{2}\right]$ and $b = \max\{\frac{2AL}{\mu\delta}, 2BL, \mu\delta\}$. If $J^* - \mathbb{E}\left[J(\theta_t)\right] \geq \delta$ for all $t \in \{0, 1, \cdots, T-1\}$, then

$$J^* - \mathbb{E}\left[J(\theta_T)\right] \leq 16 \exp\left(-\frac{\mu\delta(T-1)}{2b}\right)(J^* - J(\theta_0)) + \frac{12LC}{\mu^2\delta^2 T} + \frac{26D\gamma^H}{\mu\delta} + \frac{12(\epsilon')^2(2b - LB)}{\mu^2\delta^2 T} + \frac{2\epsilon'}{\mu}, \tag{10}$$

otherwise, we have

$$\min_{t\in\{0,1,\cdots,T-1\}} J^* - \mathbb{E}\left[J(\theta_t)\right] \leq \delta.$$

An important special case of the Theorem 2 is the exact full gradient update (achieved with the constants $A = C = D = 0$ and $B = 1$ in the Assumption 3). Therefore, in this case (in addition when $\epsilon' = 0$), we have

$$J^* - \mathbb{E}\left[J(\theta_T)\right] \leq 16 \exp\left(-\frac{\mu\epsilon(T-1)}{2b}\right)(J^* - J(\theta_0)),$$

meaning that $T = \widetilde{\mathcal{O}}(\epsilon^{-1})$ implies $J^* - \mathbb{E}\left[J(\theta_T)\right] \leq \epsilon$. That establishes $\widetilde{\mathcal{O}}(\epsilon^{-1})$ iteration complexity, recovering the same rate as in [24] (however, authors establishes it for the softmax policy only).

As straightforward consequence of the Theorem 2, one can implicitly state the new global optimum convergence guarantee

Corollary 2 ([50])**.** *Consider the setting of Thm. 2. Given $\epsilon > 0$, let the horizon $H = \mathcal{O}(\log \epsilon^{-1})$. If $\epsilon' = 0$, we choose the number of iterations $T = \mathcal{O}(\epsilon^{-3})$; if $\epsilon' > 0$, we choose $T = \mathcal{O}((\epsilon')^{-2}\epsilon^{-1})$. Then $\min_{t\in\{0,1,\cdots,T\}} J^* - \mathbf{E}J(\theta_t) \leq \mathcal{O}(\epsilon) + \mathcal{O}(\epsilon')$.*

Corollary 2 says that when $\epsilon' = \Theta(\epsilon)$ the complexity of PG to reach a global optimum is $\mathcal{O}(\epsilon^{-3})$ showing a factor of ϵ^{-1} improvement as compared to the $\mathcal{O}(\epsilon^{-4})$ complexity in Corollary 1.

Next, with Assumption (8) it is straightforward to obtain the average regret to the global optimum convergence under the setting of Corollary 1.

Corollary 3 ([50])**.** *Suppose that Assumption 1, 2, 3 and 4 hold with $A = 0$. Given $\epsilon > 0$, let $\eta = \min\left\{\frac{1}{LB}, \frac{\epsilon}{2LC}\right\}$ and the horizon $H = \mathcal{O}(\log \epsilon^{-1})$. If the number of iterations T satisfies*

$$T \;\geq\; \frac{12\delta_0 L}{\epsilon^2}\max\left\{B, \frac{2C}{\epsilon^2}\right\}, \tag{11}$$

then $J^* - \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[J(\theta_t)\right] = \mathcal{O}(\epsilon) + \mathcal{O}(\epsilon')$.

4 Applications

Up to this point, we have been restating general convergence results for different settings without explicitly giving tighter rates for the concrete policy examples. This section is devoted to different applications for which we present convergence results (by establishing concrete constants A, B, C, D, D').

4.1 Expected Lipschitz and smooth policies

We start be considering the expected Lipschitz and smooth policy ((E-LS) assumption).

Assumption 5 (E-LS)**.** *There exists constants $G, F > 0$ such that for every state $s \in \mathcal{S}$, the expected gradient and Hessian of $\log \pi_\theta(\cdot \mid s)$ satisfy*

$$\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\left\|\nabla_\theta \log \pi_\theta(a \mid s)\right\|^2\right] \;\leq\; G^2, \tag{12}$$

$$\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\left\|\nabla^2_\theta \log \pi_\theta(a \mid s)\right\|\right] \;\leq\; F. \tag{13}$$

The notable feature of this assumption is that it can be considered as the relaxation of the LS assumption formulated as

$$\left\|\nabla_\theta \log \pi_\theta(a \mid s)\right\| \leq G \quad \text{and} \quad \left\|\nabla^2_\theta \log \pi_\theta(a \mid s)\right\| \leq F, \quad \text{for all } (s,a) \in \mathcal{S}\times\mathcal{A}. \tag{LS}$$

In fact, multiple analysis on vanilla PG [53] and variance-reduced PG methods [38, 46, 47, 49, 13, 32, 22, 51] relied upon this assumption. However, despite the popularity, the LS assumption is more restrictive than the expected counterpart.

Lemma 1 ([50]). *Under Asm. 5, consider a truncated gradient estimator defined either in* (2) *or* (3). *Asm. 3 holds with* $A = 0, B = 1 - \frac{1}{m}$ *and* $C = \frac{\nu}{m}$, *that is,*

$$\mathbb{E}\left[\left\|\hat{\nabla}_m J(\theta)\right\|^2\right] \leq \left(1 - \frac{1}{m}\right)\|\nabla J_H(\theta)\|^2 + \frac{\nu}{m}, \tag{14}$$

where m *is the mini-batch size, and* $\nu = \frac{HG^2\mathcal{R}_{\max}^2}{(1-\gamma)^2}$ *when using REINFORCE gradient estimator* (2) *or* $\nu = \frac{G^2\mathcal{R}_{\max}^2}{(1-\gamma)^3}$ *when using GPOMDP gradient estimator* (3).

Another notable feature of the ABC assumption 5 that implies tighter smoothness constant L and D, D' from the truncated gradient assumptions (see Lemmas 4.2, 4.4, 4.5 in[50]), in particular.

$$L = \frac{\mathcal{R}_{\max}}{(1-\gamma)^2}\left(G^2 + F\right), \tag{15}$$

$$D = \frac{D'G\mathcal{R}_{\max}}{(1-\gamma)^{3/2}}, \tag{16}$$

$$D' = \frac{G\mathcal{R}_{\max}}{1-\gamma}\sqrt{\frac{1}{1-\gamma} + H}. \tag{17}$$

When γ is close to 1, the Yuan et al improve the smoothness constant proposed in [29] by a factor of $1 - \gamma$ by a sharper analysis. Interestingly, [22] achieves a better smoothness constant $\frac{F\mathcal{R}_{\max}}{(1-\gamma)^2}$ of $J(\cdot)$ under more restrictive assumption (LS) without bounding the Hessian.

Further, plugging constants (15), (16), (17) to the Theorem 1 one can establish the sample complexity of vanilla PG

Corollary 4 ([50]). *Suppose that Asm. 5 is satisfied. Let* $\delta_0 \overset{def}{=} J^* - J(\theta_0)$. *The PG method (Algorithm 1) with a mini-batch sampling of size* m *and constant step size*

$$\eta \in \left(0, \frac{2}{L\left(1 - 1/m\right)}\right), \tag{18}$$

satisfies

$$\mathbb{E}\left[\|\nabla J(\theta_U)\|^2\right] \leq \frac{2\delta_0}{\eta T\left(2 - L\eta\left(1 - \frac{1}{m}\right)\right)} + \frac{L\nu\eta}{m\left(2 - L\eta\left(1 - \frac{1}{m}\right)\right)} + \left(\frac{2D\left(3 - L\eta\left(1 - \frac{1}{m}\right)\right)}{2 - L\eta\left(1 - \frac{1}{m}\right)} + D'^2\gamma^H\right)\gamma^H, \tag{19}$$

where ν, L *and* $D, D' > 0$ *are provided in Lemma 1 and Equations* (15), (16), (17), *respectively.*

Cor. 4 does not have any requirements on the batch size, allowing us to analyze both stochastic variants and exact full PG. In particular, for exact PG (i.e., when $H = m = \infty$), the iteration complexity $T = \mathcal{O}\left(\epsilon^{-2}\right)$.

The following corollary states an explicit sample complexity for stochastic PG methods.

Corollary 5 ([50]). *Consider the setting of Corollary 4. For a given* $\epsilon > 0$, *by choosing the mini-batch size* m *such that* $1 \leq m \leq \frac{2\nu}{\epsilon^2}$, *the step size* $\eta = \frac{\epsilon^2 m}{2L\nu}$, *the number of iterations* T *such that*

$$Tm \geq \frac{8\delta_0 L\nu}{\epsilon^4} = \begin{cases} \mathcal{O}\left(\frac{H}{(1-\gamma)^4\epsilon^4}\right) & \textit{for REINFORCE} \\ \mathcal{O}\left(\frac{1}{(1-\gamma)^5\epsilon^4}\right) & \textit{for GPOMDP} \end{cases} \tag{20}$$

and the horizon $H = \mathcal{O}\left((1-\gamma)^{-1}\log\left(1/\epsilon\right)\right)$, *then* $\mathbb{E}\left[\|\nabla J(\theta_U)\|^2\right] = \mathcal{O}(\epsilon^2)$.

The critical property of the sample complexity given by the corollary 5 is that it provides a more expansive (than in the existing literature) range of PG parameters, guaranteeing its convergence. Admittedly, Cor. (5) shows vanilla PG methods can converge with a mini-batch size per iteration from 1 to $\mathcal{O}(\epsilon^{-2})$ and a constant step size chosen accordingly between $\mathcal{O}(\epsilon^2)$ and $\mathcal{O}(1)$, while still achieving the $Tm \times H = \widetilde{\mathcal{O}}\left(\epsilon^{-4}\right)$ optimal complexity. In contrast,[53, 45, 22] establish $\widetilde{\mathcal{O}}\left(\epsilon^{-4}\right)$ for FOSP convergence by using the more restrictive assumptions (LS).

4.2 Concrete policies and loss functions

In this section, we instantiate the FOSP convergence results for softmax and gaussian policies with \ without log barrier and entropic regularization.

Definition 1 (Softmax tabular policy)**.** *Let $\mathcal{S}$ is a state space, $\mathcal{A}$ is a finite action space, then for all $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, softmax tabular policy is defined as*

$$\pi_\theta(a \mid s) \overset{def}{=} \frac{\exp(\theta_{s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta_{s,a'})}. \tag{21}$$

For this type of policy, [50] provides the following result:

Corollary 6 ([50])**.** *For any accuracy level ϵ, if we choose the mini-batch size m such that $1 \leq m \leq \frac{2\nu}{\epsilon^2}$, the step size $\eta = \frac{\epsilon^2 m}{2L\nu}$ with $L = \frac{\mathcal{R}_{\max}}{(1-\gamma)^2}\left(2 - \frac{1}{|\mathcal{A}|}\right)$ and*

$$\nu \;=\; \begin{cases} \frac{H\left(1-\frac{1}{|\mathcal{A}|}\right)\mathcal{R}_{\max}^2}{(1-\gamma)^2} & \textit{for REINFORCE} \\ \frac{\left(1-\frac{1}{|\mathcal{A}|}\right)\mathcal{R}_{\max}^2}{(1-\gamma)^3} & \textit{for GPOMDP} \end{cases},$$

the number of iterations T such that

$$Tm \;\geq\; \begin{cases} \frac{8\delta_0\mathcal{R}_{\max}^3\left(1-\frac{1}{|\mathcal{A}|}\right)\left(2-\frac{1}{|\mathcal{A}|}\right)}{(1-\gamma)^4\epsilon^4} \cdot H & \textit{for REINFORCE} \\ \frac{8\delta_0\mathcal{R}_{\max}^3\left(1-\frac{1}{|\mathcal{A}|}\right)\left(2-\frac{1}{|\mathcal{A}|}\right)}{(1-\gamma)^5\epsilon^4} & \textit{for GPOMDP} \end{cases}, \tag{22}$$

and the horizon $H = \mathcal{O}\left((1-\gamma)^{-1}\log(1/\epsilon)\right)$, then $\mathbb{E}\left[\left\|\nabla J(\theta_U)\right\|^2\right] = \mathcal{O}(\epsilon^2)$.

Definition 2 (Gaussian policy)**.** *Consider a scalar-action and fixed-variance σ^2, then Gaussian policy is defined as*

$$\pi_\theta(a \mid s) \;=\; \mathcal{N}\left(a \mid \theta^\top\phi(s), \sigma^2\right) \;=\; \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{a-\theta^\top\phi(s)}{\sigma}\right)^2\right\}, \tag{23}$$

where $\theta \in \mathbb{R}^d$, and $\phi : \mathcal{S} \to \mathcal{R}^d$ is a mapping from the state space to the feature space.

Definition 3 (Entropy regularized objective)**.**

$$\tilde{J}(\theta) \overset{def}{=} J(\theta) + \mathbb{H}(\theta) \tag{24}$$

where $\mathbb{H}(\theta)$ is the "discounted entropy" defined as

$$\mathbb{H}(\theta) \overset{def}{=} \mathbb{E}_{\tau\sim p(\cdot|\theta)}\left[\sum_{t=0}^{\infty} -\gamma^t\lambda\log\pi_{s_t,a_t}(\theta)\right].$$

[50] shows that the stochastic gradient estimator of $\nabla\tilde{J}(\theta)$ with mini-batch size m is

$$\hat{\nabla}_m\tilde{J}(\theta) \overset{\text{def}}{=} \hat{\nabla}_m J(\theta) - \frac{\lambda}{m}\sum_{i=1}^{m}\sum_{t=0}^{H-1}\gamma^t\left(\log\pi_{s_t^i,a_t^i}(\theta)\left(\sum_{k=0}^{t}\nabla_\theta\log\pi_{s_k^i,a_k^i}(\theta)\right) + \nabla_\theta\log\pi_{s_t^i,a_t^i}(\theta)\right). \tag{25}$$

The following result states that estimator (25) satisfies ABC assumption.

Lemma 2 ([50])**.** *The stochastic gradient estimator* (25) *satisfies Assumption* (ABC) *with*

$$\mathbf{E}\left\|\hat{\nabla}_m\tilde{J}(\theta)\right\|^2 \;\leq\; \left(1-\frac{1}{m}\right)\left\|\nabla\tilde{J}(\theta)\right\|^2 + \frac{2\left(1-\frac{1}{|\mathcal{A}|}\right)\mathcal{R}_{\max}^2}{m(1-\gamma)^3} + \frac{2\lambda^2}{m(1-\gamma^2)}\left(1-\frac{1}{|\mathcal{A}|}\right) + \frac{8H|\mathcal{A}|\lambda^2}{m(1-\gamma)^3}. \tag{26}$$

Using the result of the Lemma 2, [50] derives FOSP convergence guarantees.

Corollary 7 ([50])**.** *Consider the vanilla PG updates* (25) *for the softmax with entropy regularization* (24) *. For a given $\epsilon > 0$, by choosing the mini-batch size m such that $1 \leq m \leq \frac{2\nu}{\epsilon^2}$, the step size $\eta = \frac{\epsilon^2 m}{2L\nu}$, the horizon $H = \mathcal{O}\left((1-\gamma)^{-1}\log(1/\epsilon)\right)$ and the number of iterations T such that*

$$Tm \geq \frac{8\delta_0 L\nu}{\epsilon^4} = \mathcal{O}((1-\gamma)^{-6}\epsilon^{-4}) \tag{27}$$

with

$$L = \left(\frac{\mathcal{R}_{\max}}{(1-\gamma)^2}\left(2 - \frac{1}{|\mathcal{A}|}\right) + \frac{\lambda(4 + 8\log|\mathcal{A}|)}{(1-\gamma)^3}\right)$$

and

$$\nu = \frac{2\left(1-\frac{1}{|\mathcal{A}|}\right)\mathcal{R}_{\max}^2}{(1-\gamma)^3} + \frac{2\lambda^2}{(1-\gamma^2)}\left(1 - \frac{1}{|\mathcal{A}|}\right) + \frac{8H|\mathcal{A}|\lambda^2}{(1-\gamma)^3},$$

then $\mathbf{E}\left\|\nabla\tilde{J}(\theta_U)\right\|^2 = \mathcal{O}(\epsilon^2)$.

Definition 4 (Log barrier regularized objective)**.**

$$\begin{aligned} L_\lambda(\theta) &\overset{def}{=} J(\theta) - \lambda\mathbb{E}_{s\sim Unif_{\mathcal{S}}}\left[KL(Unif_{\mathcal{A}}, \pi_\theta(\cdot \mid s))\right] \\ &= J(\theta) + \frac{\lambda}{|\mathcal{A}||\mathcal{S}|}\sum_{s,a}\log\pi_\theta(a \mid s) + \lambda\log|\mathcal{A}|, \end{aligned} \tag{28}$$

where the relative entropy for distributions p and q is defined as $KL(p,q) \overset{def}{=} \mathbb{E}_{x\sim p}\left[-\frac{\log q(x)}{\log p(x)}\right]$, $Unif_\chi$ denotes the uniform distribution over a set χ and $\lambda > 0$ determines the strength of the penalty.

The motivation of such a problem is that policies can rapidly become near deterministic when optimizing the softmax by PG. The optimal policy is usually obtained by sending some parameters to infinity due to the exponential scaling with parameters θ. Thus PG can result in extremely slow convergence. Indeed, [20] shows that PG can experience exponential convergence.

The following result illustrates the convergence behavior to the first-order stationary point for softmax policy with log-barrier regularization.

Corollary 8 ([50])**.** *Consider the vanilla PG (either REINFORCE or GPOMDP) applied in $L_\lambda(\cdot)$. Let $\delta_0 \overset{def}{=} L_\lambda^* - L_\lambda(\theta_0)$ with $L_\lambda^* \overset{def}{=} \max_{\theta\in\mathbb{R}^d} L_\lambda(\theta)$. For any accuracy level ϵ, if we choose the mini-batch size m such that $1 \leq m \leq \frac{2\nu}{\epsilon^2}$, the step size $\eta = \frac{\epsilon^2 m}{2L\nu}$ with $L = \frac{\mathcal{R}_{\max}}{(1-\gamma)^2}\left(2 - \frac{1}{|\mathcal{A}|}\right) + \frac{\lambda}{|\mathcal{S}|}$ and*

$$\nu \;=\; \begin{cases} 2\left(1-\frac{1}{|\mathcal{A}|}\right)\left(\frac{H\mathcal{R}_{\max}^2}{(1-\gamma)^2} + \frac{\lambda^2}{|\mathcal{S}|}\right) & \textit{when using REINFORCE} \\ 2\left(1-\frac{1}{|\mathcal{A}|}\right)\left(\frac{\mathcal{R}_{\max}^2}{(1-\gamma)^3} + \frac{\lambda^2}{|\mathcal{S}|}\right) & \textit{when using GPOMDP} \end{cases}, \tag{29}$$

the number of iterations T such that

$$Tm \;\geq\; \frac{8\delta_0 L\nu}{\epsilon^4} \;=\; \mathcal{O}((1-\gamma)^{-5}\epsilon^{-4}), \tag{30}$$

and the horizon $H = \mathcal{O}\left((1-\gamma)^{-1}\log(1/\epsilon)\right)$, then $\mathbb{E}\left[\|\nabla L_\lambda(\theta_U)\|^2\right] = \mathcal{O}(\epsilon^2)$.

In the following corollary, [50] shows that one can derive a guarantee on the average regret with respect to the global optimum.

Corollary 9 ([50]). *Given $\epsilon > 0$, consider the batch size m such that $1 \leq m \leq \frac{1}{(1-\gamma)^6\epsilon^3}$, the step size $\mathcal{O}(\epsilon^3) \leq \eta = \frac{(1-\gamma)^3\epsilon^3 m}{2L\nu} \leq \mathcal{O}(1)$ with L, ν in the setting of Cor. 8 . If the horizon $H = \mathcal{O}\left(\frac{\log(1/\epsilon)}{1-\gamma}\right)$ and the number of iterations T is such that*

$$Tm \times H \;\geq\; \widetilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^{12}\epsilon^6}\right),$$

we have $J^* - \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[J(\theta_t)\right] \;=\; \mathcal{O}(\epsilon)$.

This result shoes the same sample complexity $\widetilde{\mathcal{O}}(\epsilon^{-6})$ as in [52].

Finally, tables 1 and 1 summarizes characterizing constants and the relationship between assumptions.

.

	Gaussian*	Softmax	Softmax with log barrier	Softmax with entropy
G^2	$\frac{\varphi^2}{\sigma^2}$	$1 - \frac{1}{\lvert\mathcal{A}\rvert}$	✗**	✗
F	$\frac{\varphi^2}{\sigma^2}$	1	✗	✗
L	$\frac{2\mathcal{R}_{\max}\varphi^2}{(1-\gamma)^2\sigma^2}$	$\frac{\mathcal{R}_{\max}}{(1-\gamma)^2}\left(2 - \frac{1}{\lvert\mathcal{A}\rvert}\right)$	$\frac{\mathcal{R}_{\max}}{(1-\gamma)^2}\left(2 - \frac{1}{\lvert\mathcal{A}\rvert}\right) + \frac{\lambda}{\lvert\mathcal{S}\rvert}$	$\frac{\mathcal{R}_{\max}}{(1-\gamma)^2}\left(2 - \frac{1}{\lvert\mathcal{A}\rvert}\right) + \frac{\lambda(4+8\log\lvert\mathcal{A}\rvert)}{(1-\gamma)^3}$
Γ	$\frac{\mathcal{R}_{\max}\varphi}{(1-\gamma)^{3/2}\sigma}$	$\frac{\mathcal{R}_{\max}}{(1-\gamma)^{3/2}}\sqrt{1 - \frac{1}{\lvert\mathcal{A}\rvert}}$	$\sqrt{2\left(1 - \frac{1}{\lvert\mathcal{A}\rvert}\right)\left(\frac{\mathcal{R}_{\max}^2}{(1-\gamma)^3} + \frac{\lambda^2}{\lvert\mathcal{S}\rvert}\right)}$	not provided in [50]

*The (E-LS) constants G^2 and F are provided in Lemma 23 in [29].

**When there is a "✗", it means this is not applicable directly in such setting.

Table 1: E-LS constants G,F (Assumption 5), smoothness constant L and Lipschitzness constant Γ for Gaussian and (regularized) Softmax tabular policies. φ is an upper bound on the euclidean norm of the feature function for the Gaussian policy, $R_{\max}$ is the maximum absolute-valued reward, γ is the discount factor, σ is the standard deviation of the Gaussian policy. This table is based on a similar diagram from [50]

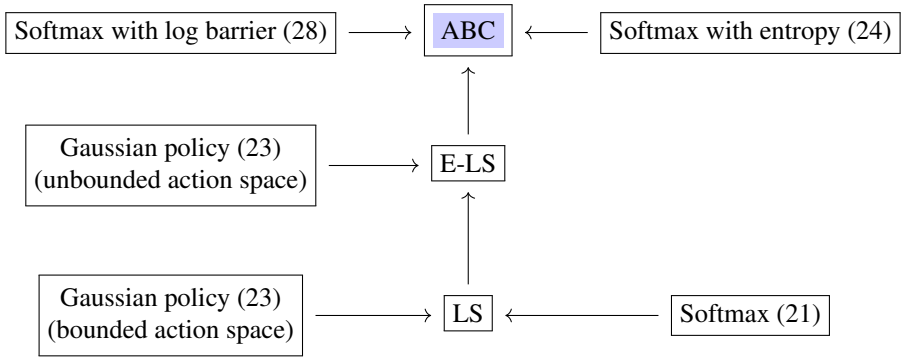

Figure 1: A relationship between the assumptions. An arrow indicates an implication. This diagram is taken from [50].

5 Final remark

In conclusion, it is vital to highlight one of the critical limitations of the Theorem 1 and Corollary 1. In particular, it works only for vanilla PG methods, and it does not apply to more sophisticated mechanisms such as variance-reduced PG methods, which is the topic of the next part of our review.

Part II
Variance-reduced PG methods

6 Introduction

This part is based on the results presented in [28, 47, 49, 8]. Variance reduction is a well-known algorithmic tool in supervised learning to decrease the variance from stochastic approximation. There are plenty of famous algorithms for the non-convex setting, such as SVRG [14], SARAH [26], SPIDER [7], STORM [6], and PAGE [21]. Recently, various VR techniques have been applied to PG methods, such as SVRPG [28], SRVRPG [47], STORM-PG [49], ProxHSPGA [32], VRMPO [48], VR-BGPO [12] and PAGE-PG [8]. It has been found that variance reduction is helpful in PG methods since it helps to improve the overall sample complexity of reaching a first-order stationary point (FOSP).

In this part, we consider variance-reduced extensions of REINFORCE and GPOMDP stochastic estimators. In particular, we focus our attention on the recently proposed PAGE-PG estimator and compare it against SVRPG, SRVRPG, and STORM-PG.

PAGE-PG is based on the recently proposed PAGE estimator for supervised learning [21]. It was proven that PAGE-PG [8] has $\mathcal{O}\left(\epsilon^{-3}\right)$ sample complexity to achieve an ϵ-stationary policy that matches the bounds of the most competitive variance-reduced REINFORCE-type methods under the same setting.

Similar to STORM-PG, PAGE-PG shares a loopless structure, but it is done by replacing the double-loop structure typical of variance-reduced methods with a probabilistic switch between two types of updates. However, PAGE-PG achieves a better convergence rate compared to STORM-PG.

7 Preliminaries

In this part, we primarily use the same notation as it was introduced in the similar section 2 in part 1. In particular, we keep using the same MDP model, and our goal remained the same: find a policy π maximizing total expected return $J(\pi) \overset{\text{def}}{=} \mathbb{E}_{\tau\sim p(\cdot|\pi)}[\mathcal{R}(\tau)]$.

However, we need to introduce a new notion for REINFORCE and GPOMDP estimators.

$$\hat{\nabla}_\theta J^{\text{RF}}(\theta) \overset{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N}\sum_{h=0}^{H-1}\nabla_\theta \log \pi_\theta(a_h^i \mid s_h^i)\mathcal{R}(\tau_i)\,, \tag{31}$$

$$\hat{\nabla}_\theta J^{\text{GPOMDP}}(\theta) \overset{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N}\sum_{h=0}^{H-1}\gamma^h\mathcal{R}(s_h^i,a_h^i)\sum_{z=0}^{h}\nabla_\theta \log \pi_\theta(a_z^i \mid s_z^i) \tag{32}$$

One of the critical technical difficulties that variance-reduced PG methods often meet is the need for periodic computation of large batches of samples (e.g., N samples). Generally, its size and computation frequency is tuned to achieve the best trade-off between the simulation cost and variance of the estimate. This motivates the use of past gradients to reduce variance, leading to an improvement in terms of sample complexity.

The second difficulty comes from the fact that the data distribution changes over time. This is because data distribution depends on the parameter θ, which gets updated during training. This refers to the so-called *distribution shift* that requires the usage of importance weighting to reuse previous information without adding a bias to the gradient estimate. Following [8], for example, let us assume that we have two policies θ_1 and θ_2, where θ_2 is used before for the interaction with the system. At the same time, we aim to obtain an unbiased estimate of the gradient concerning θ_1. The unbiased off-policy extension of the REINFORCE estimator [8] is obtained by changing Equation (31) to the

following quantity

$$g^{\omega_{\theta_2}}(\tau_i\,|\theta_1) = \omega(\tau_i \,|\, \theta_2,\theta_1) \sum_{h=0}^{H-1} \nabla_\theta \log \pi_{\theta_1}(a_h^i \,|\, s_h^i)\mathcal{R}(\tau_i)\,, \tag{33}$$

where $\omega(\tau_i \,|\, \theta_2, \theta_1) = \Pi_{j=0}^{H-1} \frac{\pi_{\theta_1}(a_j^i \,|\, s_j^i)}{\pi_{\theta_2}(a_j^i \,|\, s_j^i)}$ is the importance weight for the full trajectory realization τ_i. Similarly, the off-policy extension of the GPOMDP estimator [28] is defined as

$$g^{\omega_{\theta_2}}(\tau_i\,|\theta_1) = \sum_{h=0}^{H-1} \omega_{0:h}(\tau_i \,|\, \theta_2,\theta_1)\gamma^h \mathcal{R}(s_h^i,a_h^i) Z_{\theta_1,h}\,, \tag{34}$$

where $\omega_{0:h}(\tau_i \,|\, \theta_2,\theta_1) = \Pi_{j=0}^{h} \frac{\pi_{\theta_1}(a_j^i \,|\, s_j^i)}{\pi_{\theta_2}(a_j^i \,|\, s_j^i)}$ is the importance weight for the trajectory realization τ_i truncated at time h.

It is easy to verify that for both the off-policy extensions of REINFORCE and GPOMDP are unbiased estimators of the gradient of the truncated expected return $J_H(\theta) \stackrel{\text{def}}{=} \mathbb{E}_\tau \left[\sum_{t=0}^{H-1} \gamma^t \mathcal{R}(s_t, a_t)\right]$. We refer to both the REINFORCE and GPOMDP estimators through $g(\tau \,|\, \theta)$ and $g^{\omega_{\theta_2}}(\tau \,|\, \theta_1)$.

8 Brief overview of recent SOTA methods

This section briefly reviews some of the state-of-the-art PG methods with variance reduction.

Stochastic Varaice-Reduced Policy Gradient (SVRPG)[28] is suitable for RL adaptation of the famous stochastic variance-reduced SVRG-method [15] designed for finite-sum problems. The method has by a double loop structure. At the s-th epoch (outer loop), a snapshot of the current θ_0^s is fixed. Then, $N >> 1$ trajectories $\{\tau_i\}_{i=1}^N$ are sampled based on the current policy, stored and used to compute the gradient estimator $v_0^s = \frac{1}{N}\sum_{i=1}^N g(\tau_i \,|\, \theta_0^s)$. On every epoch, m steps of the inner loop are performed. At the t-th iteration, the parameter θ_t^s is updated. Then $B << N$ trajectories $\{\tau_j\}_{j=1}^B$ are sampled and collected according $\pi(\theta_{t+1}^s)$ and the gradient estimate at point θ_{t+1}^s is computed After m iterations in the inner loop, the snapshot is refreshed by setting $\theta_0^{s+1} = \theta_m^s$, and the process is cycled for a fixed number of epochs.

The complete SVRPG algorithm is presented below.

Algorithm 2 SVRPG

Input: initial parameter θ_0, large batch-size N, small batch-size B, step-size $\eta > 0$, inner-loop length $m \in \mathbb{N}$, number of epochs $S \in \mathbb{N}$
initialize $\theta_m^0 = \theta_0$
for $s = 1$ **to** S **do**
  set $\theta_0^s = \theta_m^{s-1}$
  collect N trajectories with policy θ_0^s
  $v_0^s = \frac{1}{N}\sum_{i=1}^N g(\tau_i \,|\, \theta_0^s)$
  for $t = 0$ **to** $m-1$ **do**
    $\theta_{t+1}^s = \theta_t^s + \eta v_t^s$
    sample B trajectories with policy θ_{t+1}^s
    $v_{t+1}^s = \frac{1}{B}\sum_{i=1}^B g(\tau_i \,|\, \theta_{t+1}^s) + v_0^s - \frac{1}{B}\sum_{i=1}^B g^{\omega_{\theta_{t+1}^s}}(\tau_i \,|\, \theta_0^s)$
  end for
end for
Output: θ_{out} chosen uniformly at random from $\{\theta_t^s\}_{t=1,\,s=1}^{m,\,S}$

Stochastic Recursive Variance-Reduced Policy Gradient(SRVRPG)[47] is inspired by SARAH [26], and it is similar to SVRPG with only two differences in the estimator v_{t+1}^s update. Firstly,

unlike SVRPG, it reuses an estimate from the previous state (i.e., v_t^s). Secondly, it does not employ the same snapshot within the epoch as SVRPG but instead computes a batch based on the previous iterate.

The complete SRVRPG algorithm is presented below.

Algorithm 3 SRVRPG

Input: initial parameter θ_0, large batch-size N, small batch-size B, step-size $\eta > 0$, inner-loop length $m \in \mathbb{N}$, number of epochs $S \in \mathbb{N}$
initialize $\theta_m^0 = \theta_0$
for $s = 1$ **to** S **do**
  set $\theta_0^s = \theta_m^{s-1}$
  collect N trajectories with policy θ_0^s
  $v_0^s = \frac{1}{N}\sum_{i=1}^{N} g(\tau_i \,|\, \theta_0^s)$
  for $t = 0$ **to** $m - 1$ **do**
    $\theta_{t+1}^s = \theta_t^s + \eta v_t^s$
    sample B trajectories with policy θ_{t+1}^s
    $v_{t+1}^s = \frac{1}{B}\sum_{i=1}^{B} g(\tau_i \,|\, \theta_{t+1}^s) + v_t^s - \frac{1}{B}\sum_{i=1}^{B} g^{\omega_{\theta_{t+1}^s}}(\tau_i \,|\, \theta_t^s)$
  end for
end for
Output: θ_{out} chosen uniformly at random from $\{\theta_t^s\}_{t=1,\, s=1}^{m,\, S}$

Stochastic Recursive Momentum Policy Gradient(STORM-PG)[49] is suitable for RL adaptation of the famous stochastic variance-reduced STORM-method [6] for finite-sum problems. Unlike SRVRPG and SVRPG, it does not have a double loop structure deploying an exponential moving averaging mechanism instead. The complete STORM-PG algorithm is presented below.

Algorithm 4 STORM-PG

Input: initial parameter θ_0, large batch-size N, small batch-size B, step-size $\eta > 0$, momentum parameter $\alpha \in (0,1)$
collect N trajectories with policy θ_0
$v_0 = \frac{1}{N}\sum_{i=1}^{N} g(\tau_i \,|\, \theta_0)$
for $t = 0$ **to** $T - 1$ **do**
  $\theta_{t+1} = \theta_t + \eta v_t$
  collect B trajectories with policy θ_{t+1}
  $v_{t+1} = \frac{1}{B} g(\tau_i \,|\, \theta_{t+1}) + (1-\alpha)\left[v_t - \frac{1}{B}\sum_{i=1}^{B} g^{\omega_{\theta_{t+1}}}(\tau_i \,|\, \theta_t)\right]$
end for
Output: θ_{out} chosen uniformly at random from $\{\theta_t\}_{t=1}^{T}$

ProbAbilistic Gradient Estimation for Policy Gradient(PAGE-PG)[8] is inspired by PAGE-algorithm [21] for supervised learning, which is based on a small adjustment to the mini-batch gradient estimator. Initially, it samples N trajectories with the initial policy θ_0 and to build a gradient estimate $v_0 = \frac{1}{N}\sum_{i=1}^{N} g(\tau_i \,|\, \theta_0)$. For any $t > 0$, PAGE-PG either computes SARAH-type update or large-batch estimate.

The complete PAGE-PG algorithm is presented below.

Algorithm 5 PAGE-PG

Input: initial parameter θ_0, large batch-size N, small batch-size B, step-size $\eta > 0$, probability $p \in (0,1]$

collect N trajectories with policy θ_0

$v_0 = \frac{1}{N}\sum_{i=1}^{N} g(\tau_i \,|\, \theta_0)$

for $t = 0$ **to** $T-1$ **do**

$\theta_{t+1} = \theta_t + \eta v_t$

$$v_{t+1} = \begin{cases} \frac{1}{N}\sum\limits_{i=1}^{N} g(\tau_i \,|\, \theta_t) & \text{prob. } p \\ \frac{1}{B}\sum\limits_{i=1}^{B} g(\tau_i \,|\, \theta_t) + v_{t-1} - \frac{1}{B}\sum\limits_{i=1}^{B} g^{\omega_{\theta_t}}(\tau_i \,|\, \theta_{t-1}) & \text{prob. } 1-p \end{cases}$$

end for

Output: θ_{out} chosen uniformly at random from $\{\theta_t\}_{t=1}^{T}$

9 Theoretical comparison

This section states the theoretical convergence results for the PAGE-PG method. We will show that it reaches the best-known convergence rate among VR PG methods while preserving optimal $\mathcal{O}(\epsilon^{-3})$ sample complexity bound. Following [8], we start by listing the necessary assumption.

Assumption 6 (Bounded log-policy gradient norm)**.** *For any $a \in \mathcal{A}$ and $s \in \mathcal{S}$ there exists a constant $G > 0$ such that $\|\nabla_\theta \log \pi_\theta(a \,|\, s)\| \leq G$ for all $\theta \in \mathbb{R}^d$.*

Assumption 7 (Smoothness)**.** *π_θ is twice differentiable and for any $a \in \mathcal{A}$ and $s \in \mathcal{S}$ there exists a constant $M > 0$ such that $\|\nabla^2_\theta \log \pi_\theta(a \,|\, s)\| \leq M$ for all $\theta \in \mathbb{R}^d$.*

Assumption 8 (Finite variance)**.** *There exists a constant $\sigma > 0$ such that Var$(g(\tau \,|\, \theta)) \leq \sigma^2$ for all $\theta \in \mathbb{R}^d$.*

Assumption 9 (Finite importance weight variance)**.** *For any policy pair $\theta_a, \theta_b \in \mathbb{R}^d$ and with $\tau \sim p(\cdot \,|\, \theta_b)$, the importance weight $\omega(\tau \,|\, \theta_b, \theta_a) = \frac{p(\tau \,|\, \theta_a)}{p(\tau \,|\, \theta_b)}$ is well-defined. In addition, there exists a constant $W > 0$ such that Var$(\omega(\tau \,|\theta_b, \theta_a)) \leq W$.*

Next, we can formulate main convergence result.

Theorem 3 ([8])**.** *Suppose that Assumptions 6-9 hold and select $\eta > 0$, $p \in (0,1]$ and $B \in \mathbb{N}$ such that $\eta^2 \leq \min\left\{p/(1-p) \cdot B/2C, 1/4L^2\right\}$. The average expected squared gradient norm after T iterations of PAGE-PG satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla_\theta V(\theta_t)\|^2\right] \leq \frac{2\left(V^* - V(\theta_0)\right)}{\eta T} + \frac{\sigma^2}{N} + \frac{\sigma^2}{pNT}$$

where $C \overset{def}{=} 2(L^2 + C_\omega)$, $L \overset{def}{=} MR/(1-\gamma)^2 + 2G^2R/(1-\gamma)^3$ and $C_\omega \overset{def}{=} 24RG^2(2G^2+M)(W+1)\gamma/(1-\gamma)^5$.

Theorem 3 states the $\mathcal{O}\left(\frac{1}{T} + \frac{1}{NT} + \frac{1}{N}\right)$ convergence rate. It improves the rate $\mathcal{O}\left(\frac{1}{T} + \frac{1}{B} + \frac{1}{N}\right)$ of SVRPG [28] and $\mathcal{O}\left(\frac{1}{T} + \frac{1}{B} + \frac{1}{TN}\right)$ of STORM-PG [49], by avoiding the dependency on the batch-size B. However, compared to the rate of SRVRPG $\mathcal{O}\left(\frac{1}{T} + \frac{1}{N}\right)$, PAGE-PG's one has an extra $\mathcal{O}\left(\frac{1}{TN}\right)$ term. Nonetheless, by choosing $p = \frac{B}{N}$ and $B = \mathcal{O}(1)$, one can recover the rate $\mathcal{O}\left(\frac{1}{T} + \frac{1}{N}\right)$ for PAGE-PG.

The next Corollary states previously announced sample complexity of PAGE-PG method.

Corollary 10 ([8])**.** *Under the conditions of Theorem 3, set $\eta = \sqrt{B}/\sqrt{2CN}$, $p = 1/N$, $N = \mathcal{O}\left(\epsilon^{-2}\right)$ and $B = \mathcal{O}(1)$. Then $\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\nabla_\theta V(\theta_t)\right\|^2\right] \leq \epsilon^2$ within $\mathcal{O}\left(\epsilon^{-3}\right)$ trajectories on average with $\epsilon \to 0$.*

Under the considered assumptions, vanilla PG methods, (based on REINFORCE, GPOMDP estimators) require $\mathcal{O}(\epsilon^{-4})$ samples to achieve an ϵ-stationary solution. However, Corollary 10 shows that by incorporating stochastic variance reduction techniques the complexity can be reduced to $\mathcal{O}(\epsilon^{-3})$. This result matches best-known sample complexity previously achieved by SRVRPG[47] and STORM-PG [49], whereas SVRPG[28] could reach only $\mathcal{O}(\epsilon^{-10/3})$.

References

[1] Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

[2] J. Baxter and P. L. Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, Nov 2001.

[3] Amir Beck. *First-Order Methods in Optimization*. SIAM-Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2017.

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

[5] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[6] Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems*, pages 15236–15245, 2019.

[7] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, volume 31, pages 689–699, 2018.

[8] Matilde Gargiani, Andrea Zanelli, Andrea Martinelli, Tyler Summers, and John Lygeros. Page-pg: A simple and loopless variance-reduced policy gradient method with probabilistic gradient estimation. *arXiv preprint arXiv:2202.00308*, 2022.

[9] Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

[10] Robert M. Gower, Peter Richtárik, and Francis Bach. Stochastic quasi-gradient methods: variance reduction via Jacobian sketching. *Mathematical Programming*, 188(1):135–192, Jul 2021.

[11] Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. SGD: General analysis and improved rates. In *International Conference on Machine Learning*, pages 5200–5209. PMLR, 2019.

[12] Feihu Huang, Shangqian Gao, and Heng Huang. Bregman gradient policy optimization. In *International Conference on Learning Representations*, 2022.

[13] Feihu Huang, Shangqian Gao, Jian Pei, and Heng Huang. Momentum-based policy gradient methods. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4422–4433. PMLR, 13–18 Jul 2020.

[14] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[15] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26:315–323, 2013.

[16] Sham M Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.

[17] Ahmed Khaled and Peter Richtárik. Better theory for SGD in the nonconvex world. *arXiv preprint arXiv:2002.03329*, 2020.

[18] Vijay Konda and John Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 2000.

[19] Yunwen Lei, Ting Hu, Guiying Li, and Ke Tang. Stochastic gradient descent for nonconvex learning without bounded gradient assumptions. *IEEE Transactions on Neural Networks and Learning Systems*, 31(10):4394–4400, 2020.

[20] Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Softmax policy gradient methods can take exponential time to converge. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3107–3110. PMLR, 15–19 Aug 2021.

[21] Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtárik. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International Conference on Machine Learning*, pages 6286–6295. PMLR, 2021.

[22] Yanli Liu, Kaiqing Zhang, Tamer Basar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. In *Advances in Neural Information Processing Systems*, volume 33, pages 7624–7636. Curran Associates, Inc., 2020.

[23] Jincheng Mei, Yue Gao, Bo Dai, Csaba Szepesvari, and Dale Schuurmans. Leveraging non-uniformity in first-order non-convex optimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7555–7564. PMLR, 18–24 Jul 2021.

[24] Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6820–6829. PMLR, 13–18 Jul 2020.

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

[26] Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621. PMLR, 2017.

[27] Matteo Papini. Safe policy optimization, 2020. Manuscript.

[28] Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 4026–4035. PMLR, 2018.

[29] Matteo Papini, Matteo Pirotta, and Marcello Restelli. Smoothing policies and safe policy gradients. *Machine Learning*, Oct 2022.

[30] J. Peters and S. Schaal. Reinforcement learning of motor skills with policy gradients. *Neural Networks*, 21(4):682–697, May 2008.

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

[32] Nhan Pham, Lam Nguyen, Dzung Phan, Phuong Ha Nguyen, Marten van Dijk, and Quoc Tran-Dinh. A hybrid stochastic policy gradient algorithm for reinforcement learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 374–385. PMLR, 26–28 Aug 2020.

[33] Karthik Abinav Sankararaman, Soham De, Zheng Xu, W. Ronny Huang, and Tom Goldstein. The impact of neural network overparameterization on gradient confusion and stochastic gradient descent. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8469–8479. PMLR, 13–18 Jul 2020.

[34] Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition, 2013.

[35] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1889–1897, Lille, France, 07–09 Jul 2015. PMLR.

[36] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

[37] Lior Shani, Yonathan Efroni, and Shie Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 5668–5675, 2020.

[38] Zebang Shen, Alejandro Ribeiro, Hamed Hassani, Hui Qian, and Chao Mi. Hessian aided policy gradient. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5729–5738. PMLR, 09–15 Jun 2019.

[39] Richard S Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems 12*, pages 1057–1063. MIT Press, 2000.

[40] Manan Tomar, Lior Shani, Yonathan Efroni, and Mohammad Ghavamzadeh. Mirror descent policy optimization. In *International Conference on Learning Representations*, 2022.

[41] Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1195–1204. PMLR, 2019.

[42] Sharan Vaswani, Olivier Bachem, Simone Totaro, Robert Müller, Shivam Garg, Matthieu Geist, Marlos C. Machado, Pablo Samuel Castro, and Nicolas Le Roux. A general class of surrogate functions for stable and efficient reinforcement learning. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 8619–8649. PMLR, 28–30 Mar 2022.

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

[44] Lin Xiao. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282):1–36, 2022.

[45] Huaqing Xiong, Tengyu Xu, Yingbin Liang, and Wei Zhang. Non-asymptotic convergence of adam-type reinforcement learning algorithms under markovian sampling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):10460–10468, May 2021.

[46] Pan Xu, Felicia Gao, and Quanquan Gu. An improved convergence analysis of stochastic variance-reduced policy gradient. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 541–551. PMLR, 22–25 Jul 2020.

[47] Pan Xu, Felicia Gao, and Quanquan Gu. Sample efficient policy gradient methods with recursive variance reduction. In *International Conference on Learning Representations*, 2020.

[48] Long Yang, Yu Zhang, Gang Zheng, Qian Zheng, Pengfei Li, Jianhang Huang, and Gang Pan. Policy optimization with stochastic mirror descent. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8):8823–8831, Jun. 2022.

[49] Huizhuo Yuan, Xiangru Lian, Ji Liu, and Yuren Zhou. Stochastic recursive momentum for policy gradient methods, 2020.

[50] Rui Yuan, Robert M Gower, and Alessandro Lazaric. A general sample complexity analysis of vanilla policy gradient. In *International Conference on Artificial Intelligence and Statistics*, pages 3332–3380. PMLR, 2022.

[51] Junyu Zhang, Chengzhuo Ni, Zheng Yu, Csaba Szepesvari, and Mengdi Wang. On the convergence and sample efficiency of variance-reduced policy gradient method. In *Advances in Neural Information Processing Systems*, 2021.

[52] Junzi Zhang, Jongho Kim, Brendan O'Donoghue, and Stephen Boyd. Sample efficient reinforcement learning with reinforce. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):10887–10895, May 2021.

[53] Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Başar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6):3586–3612, 2020.