A GENERAL SAMPLE COMPLEXITY ANALYSIS OF VANILLA POLICY GRADIENT

Anonymous authors
Paper under double-blind review

ABSTRACT

We adapt recent tools developed for the analysis of Stochastic Gradient Descent (SGD) in non-convex optimization to obtain convergence guarantees and sample complexities for the vanilla policy gradient (PG) – REINFORCE and GPOMDP. Our only assumptions are that the expected return is smooth w.r.t. the policy parameters and that the second moment of its gradient satisfies a certain *ABC assumption*. The ABC assumption allows for the second moment of the gradient to be bounded by $A \geq 0$ times the suboptimality gap, $B \geq 0$ times the norm of the full batch gradient and an additive constant $C \geq 0$, or any combination of aforementioned. We show that the ABC assumption is more general than the commonly used assumptions on the policy space to prove convergence to a stationary point. We provide a single convergence theorem under the ABC assumption, and show that, despite the generality of the ABC assumption, we recover the $\widetilde{\mathcal{O}}(\epsilon^{-4})$ sample complexity of PG. Our convergence theorem also affords greater flexibility in the choice of hyper parameters such as the step size and places no restriction on the batch size m. Even the single trajectory case (i.e., $m = 1$) fits within our analysis. We believe that the generality of the ABC assumption may provide theoretical guarantees for PG to a much broader range of problems that have not been previously considered.

1 INTRODUCTION

Policy gradient (PG) is one of the most popular reinforcement learning (RL) methods for computing 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 non-Markov and partially-observable environments) 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), and variance-reduced PG (Papini et al., 2018; Shen et al., 2019; Xu et al., 2020b; Yuan et al., 2020; Pham et al., 2020).

Unlike value-based methods, a solid theoretical understanding of even the "vanilla" PG has long been elusive. 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. Due to space constraints, we defer a thorough review of recent results to App. A.

In this paper, we focus on the sample complexity of PG for reaching a FOSP (first-order stationary point). We show how PG can be analysed under a very general assumption on the second moment of the estimated gradient called the *ABC* assumption, which includes most of the bounded gradient type assumptions as a special case. Under the ABC and a smoothness assumption on the expected return, we obtain convergence guarantees and the sample complexity for both REINFORCE (Williams, 1992) and GPOMDP (Sutton et al., 2000; Baxter & Bartlett, 2001). Our sample complexity analysis recovers both the well known $\mathcal{O}(\epsilon^{-2})$ iteration complexity of exact PG and the $\widetilde{\mathcal{O}}(\epsilon^{-4})$ sample complexity of REINFORCE and GPOMDP under weaker assumptions than had previously been explored. Furthermore, our analysis is less restrictive when it comes to the hyper-parameter choices. In fact, our results allow for wide range of step sizes and place almost no restriction on the batch size m, even allowing for single trajectory sampling ($m = 1$), which is uncommon in the literature. The generality of our assumption allows us to unify much of the fragmented results in the literature under

Table 1: Overview of different convergence results for vanilla PG methods. The darker shaded cells contain our new results. The medium shaded cells contain previously known results that we recover as special cases of our analysis, and extend the permitted parameter settings. White cells contain existing results that we could not recover under our general analysis.

Guarantee*	**Setting****	**Reference** (our results in bold)	**Bound**	**Remarks**
Sample complexity of stochastic PG for FOSP	ABC	**Thm. 3.4**	$\widetilde{\mathcal{O}}(\epsilon^{-4})$	Weakest asm.
	E-LS	Papini (2020) **Cor. 4.7**	$\widetilde{\mathcal{O}}(\epsilon^{-4})$	Weaker asm.; Wider range of parameters; Recover $\mathcal{O}(\epsilon^{-2})$ for exact PG; Improved smoothness constant
Sample complexity of stochastic PG for GO	ABC + PL	**Thm. G.2**	$\widetilde{\mathcal{O}}(\epsilon^{-1})$	Recover linear convergence for exact PG
	ABC + weak PL	**Thm. G.4**	$\widetilde{\mathcal{O}}(\epsilon^{-3})$	Recover $\mathcal{O}(\epsilon^{-1})$ for exact PG
Sample complexity of stochastic PG for AR	LS + FI + compatible	Liu et al. (2020)	$\widetilde{\mathcal{O}}(\epsilon^{-4})$	
	Softmax + log barrier (26)	Zhang et al. (2020b) **Cor. 4.10**	$\widetilde{\mathcal{O}}(\epsilon^{-6})$	Constant step size; Wider range of parameters; Extra phased learning step unnecessary
Iteration complexity of the exact PG for GO	Softmax + log barrier (26)	Agarwal et al. (2021) **Cor. E.6**	$\mathcal{O}(\epsilon^{-2})$	Improved by $1-\gamma$
	Softmax (23)	Mei et al. (2020) **Thm. G.4**	$\mathcal{O}(\epsilon^{-1})$	
	Softmax + entropy (84)	Mei et al. (2020) **Thm. G.2**	linear	
	LS + bijection + PPG	Zhang et al. (2020a)	$\mathcal{O}(\epsilon^{-1})$	
	Tabular + PPG	Xiao & Li (2021)	$\mathcal{O}(\epsilon^{-1})$	
	LQR	Fazel et al. (2018)	linear	

* **Type of convergence.** *FOSP*: first-order stationary point; *GO*: global optimum; *AR*: average regret to the global optimum.

** **Setting.** *bijection*: Asm.1 in Zhang et al. (2020a) about occupancy distribution; *PPG*: analysis also holds for the projected PG; *FI*: Asm. 2.1 in Liu et al. (2020) on Fisher information; *compatible*: Asm. 4.4 in Liu et al. (2020) on function approximation error; *Tabular*: direct parametrized policy; *LQR*: linear-quadratic regulator.

one guise. Indeed, we show that the analysis of Lipschitz and smooth policies, Gaussian polices, softmax tabular polices with or without a log barrier regularizer are all special cases of our general analysis (see hierarchy diagram further down in Figure 1).

Recently, there has also been much work on establishing the convergence of PG to a global optimum (i.e., the best-in-class policy) (Fazel et al., 2018; Agarwal et al., 2021; Zhang et al., 2020a; Mei et al., 2020; Liu et al., 2020; Zhang et al., 2020b; 2021). This usually requires more restrictive assumptions and specific RL settings (e.g., tabular). While our primary focus here is convergence to a stationary point, under the ABC and smoothness assumptions, we also establish the global optimum convergence theory when an additional (weak) gradient domination assumption is verified (App. G). Table 1 provides a complete overview of our results, how they recover existing results, as well as cases where we could not directly apply our general analysis.

We believe that the generality of the ABC assumption may provide theoretical guarantees for PG for a broader range of problems that have not been previously considered, and help unify our current understanding of PG and the many assumptions currently in use.

2 Preliminaries

Markov decision process (MDP). We consider a continuous MDP given by $\{\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, \rho\}$, where $\mathcal{S}$ is a state space; $\mathcal{A}$ is an action space; $\mathcal{P}$ is a Markovian transition model, where $\mathcal{P}(s' \mid s, a)$ is the transition density from state s to s' under action a; $\mathcal{R}$ is the reward function, where $\mathcal{R}(s, a) \in [-\mathcal{R}_{\max}, \mathcal{R}_{\max}]$ is the bounded reward for state-action pair (s, a) ; $\gamma \in [0, 1)$ is the discounted factor; and ρ is the initial state distribution. The agent's behaviour is modelled as a policy $\pi \in \Delta(\mathcal{A})^{\mathcal{S}}$, where $\pi(\cdot \mid s)$ is the density distribution over $\mathcal{A}$ in state $s \in \mathcal{S}$. We consider the infinite-horizon discounted setting.

Let $p(\tau \mid \pi)$ be the probability density of a single trajectory τ being sampled from π, that is

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

With a slight abuse of notation, let $\mathcal{R}(\tau) = \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t)$ be the total discounted reward accumulated along trajectory τ. We define the expected return of π as

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

Policy gradient. We introduce 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 $p(\tau \mid \theta) = p_\theta(\tau) = p(\tau \mid \pi_\theta)$. The PG methods use gradient ascent in the parametrized space of θ to find the policy that maximizes the expected return $J(\theta)$. That is, the policy with the *optimal parameters* $\theta^* \in \arg\sup_{\theta \in \mathbb{R}^d} J(\theta)$ would give the *optimal expected return* $J^* \overset{\text{def}}{=} J(\theta^*)$. In general, $J(\theta)$ is a non-convex function.

The gradient $\nabla J(\theta)$ of the expected return has the following structure

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

In practice, we cannot compute this full gradient, since computing the above expectation would require averaging over all possible trajectories $\tau \sim p(\cdot \mid \theta)$. We resort to an empirical estimate of the gradient by sampling m truncated trajectories $\tau_i = (s_0, a_0, r_0, s_1, \cdots, s_{H-1}, a_{H-1}, r_{H-1})$ obtained by executing π_θ for a given fixed horizon $H \in \mathbb{N}$. The resulting gradient estimator is

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

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

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 of the full gradient

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

which leads to the following estimate of the gradient known as GPOMDP (Baxter & Bartlett, 2001)

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

Notice that both REINFORCE and GPOMDP are the truncated versions of unbiased gradient estimators. More precisely, they are unbiased estimates of the gradient of the truncated expected return $J_H(\theta) \overset{\text{def}}{=} \mathbb{E}_\tau \left[\sum_{t=0}^{H-1} \gamma^t \mathcal{R}(s_t, a_t)\right]$ [1].

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

$$\theta_{t+1} = \theta_t + \eta_t \widehat{\nabla}_m J(\theta_t) \tag{7}$$

where $\eta_t > 0$ is the step size at the t-th iteration (see also Algorithm 1 in App. A).

[1] We allow H to be infinity so that $J_\infty(\cdot) = J(\cdot)$.

3 Non-convex optimization under ABC assumption

We use $\widehat{\nabla}_m J(\theta)$ to denote the unbiased policy gradient estimator of $\nabla J_H(\theta)$ used in (7). It can be the full gradient estimator $\nabla J(\theta)$ when $H = m = \infty$, or one of the truncated gradient estimators defined in (4) or (6). All our forthcoming analysis relies on the following common assumptions.

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

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

Assumption 3.2 (Truncation)**.** There exists $D, D' > 0$ such that, for all $\theta \in \mathbb{R}^d$, we have

$$|\langle \nabla J_H(\theta), \nabla J_H(\theta) - \nabla J(\theta) \rangle| \ \leq \ D\gamma^H, \tag{9}$$

$$\|\nabla J_H(\theta) - \nabla J(\theta)\| \ \leq \ D'\gamma^H. \tag{10}$$

We recall that given the boundedness of the reward function, we have $|J(\theta) - J_H(\theta)| \leq \frac{\mathcal{R}_{\max}}{1-\gamma}\gamma^H$ by the definition of $J(\cdot)$ and $J_H(\cdot)$. As such, when H is large, the difference between $J(\theta)$ and $J_H(\theta)$ is negligible. However, Asm. 3.2 is still necessary, since in our analysis we first prove that $\|\nabla J_H(\theta)\|^2$ is small, and then rely on (10) to show that $\|\nabla J(\theta)\|^2$ is also small.

We also make use of the recently introduced *ABC* assumption (Khaled & Richtárik, 2020)[2] 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 3.3 (ABC)**.** There exists $A, B, C \geq 0$ such that the policy gradient estimator satisfies

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

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 by Schmidt & Roux (2013); Vaswani et al. (2019) and Bottou et al. (2018) can all be seen as specific cases of Asm. 3.3. 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. A more detailed discussion of the assumption for non-convex optimization convergence theory can be found in Thm. 1 in (Khaled & Richtárik, 2020).

We state our main convergence theorem, that we will then develop into several corollaries.

Theorem 3.4. Suppose that Assumptions 3.1, 3.2 and 3.3 hold. Consider the iterates θ_t of the PG method (7) with stepsize $\eta_t = \eta \in (0, \frac{2}{LB})$ where $B = 0$ means that $\eta \in (0, \infty)$. Let $\delta_0 \overset{\text{def}}{=} J^* - J(\theta_0)$. It follows that

$$\min_{0 \leq t \leq T-1} \mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right] \leq \frac{2\delta_0(1 + LA\eta^2)^T}{\eta T(2 - LB\eta)} + \frac{LC\eta}{2 - LB\eta} + \left(\frac{2D(3 - LB\eta)}{2 - LB\eta} + D'^2\gamma^H\right)\gamma^H. \tag{11}$$

In particular if $A = 0$, we have

$$\mathbb{E}\left[\|\nabla J(\theta_U)\|^2\right] \leq \frac{2\delta_0}{\eta T(2 - LB\eta)} + \frac{LC\eta}{2 - LB\eta} + \left(\frac{2D(3 - LB\eta)}{2 - LB\eta} + D'^2\gamma^H\right)\gamma^H, \tag{12}$$

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

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

We give the proof of Thm. 3.4 in App. C.1. While Thm. 3.4 is based on Thm. 2 in (Khaled & Richtárik, 2020), our proof has to take care of the specific structure of PG estimators, notably the bias due to the truncation.

Thm. 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 the policy gradient estimator. From (11) we can derive the sample complexity of PG as follows.

Corollary 3.5. Consider the setting of Thm. 3.4. Given $\epsilon > 0$, let $\eta = \min\left\{\frac{1}{\sqrt{LAT}}, \frac{1}{LB}, \frac{\epsilon}{2LC}\right\}$ and the horizon $H = \mathcal{O}(\log \epsilon^{-1})$. If the number of iterations T satisfy

$$T \geq \frac{12\delta_0 L}{\epsilon^2} \max\left\{B, \frac{12\delta_0 A}{\epsilon^2}, \frac{2C}{\epsilon^2}\right\}, \tag{13}$$

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

Despite the generality of the ABC assumption, Cor. 3.5 recovers the best known iteration complexity for vanilla PG in several well known special cases. For instance (13) recovers the $\mathcal{O}(\epsilon^{-2})$ iteration complexity of the full gradient method as a special case. To see this, let $H = m = \infty$ and $\widehat{\nabla}_m J(\theta) = \nabla J(\theta)$ in (7), thus Asm. 3.2 and 3.3 hold automatically with $A = C = D = D' = 0$ and $B = 1$. By (13) we require $T = \mathcal{O}(\epsilon^{-2})$ iterations to reach an ϵ-stationary point. Thus, for any policy and MDP that satisfy the smoothness property (Asm. 3.1), the exact full PG converges to a FOSP in $\mathcal{O}(\epsilon^{-2})$ iterations. This is the state-of-the-art convergence rate for the exact gradient descent on non-convex objectives without any other assumptions (Beck, 2017). As we can rarely access the exact full gradient in practice, in general A, C, D, D' are not all 0.

From Cor. 3.5, notice that there is no restriction on the batch size m. By choosing $m = \mathcal{O}(1)$, Eq. (13) shows that with $TH = \widetilde{\mathcal{O}}(\epsilon^{-4})$ samples (i.e., single-step interaction with the environment and single sampled trajectory per iteration), the vanilla PG either with updates (4) or (6) is guaranteed to converge to a stationary point. Our sample complexity to achieve an ϵ-FOSP for the stochastic vanilla PG is the same as (Papini, 2020; Zhang et al., 2020c; Xiong et al., 2021) but improve upon them by recovering the exact full PG analysis, providing wider range of parameter choices and using the weaker ABC assumption (see Sec. 4.1 for more details). In short, for both the exact and stochastic PG, we recover the state-of-the-art dependency on ϵ under the ABC assumption.

4 Applications

In this section we show how the ABC assumption can be used to unify many of the current assumptions used in the literature. In Figure 1 we collect all these special cases in a hierarchy tree. Then for each special case we give the sample complexity of PG as a corollary of Thm 3.4. Each of our corollaries matches the best known results in these special cases, while also providing a wider range of parameter choices and, in some cases, improving the dependency on some terms in the bound (e.g., the discount factor γ).

4.1 Expected Lipschitz and smooth policies

We consider the recently introduced **expected Lipschitz and smooth policy** (E-LS) assumptions proposed by Papini et al. (2019)[3].

[3] While Papini et al. (2019) refers to this assumption as *smoothing policy*, we prefer the alternative name expected Lipschitz and smooth policy, as they not only induce the smoothness of J (see Lemma 4.4), but also the Lipschitzness (see Lemma D.1). In Papini et al. (2019), they also assume that $\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\|\nabla_\theta \log \pi_\theta(a \mid s)\|\right]$ is bounded, while it is a direct consequence of (14) by Cauchy-Schwarz inequality.

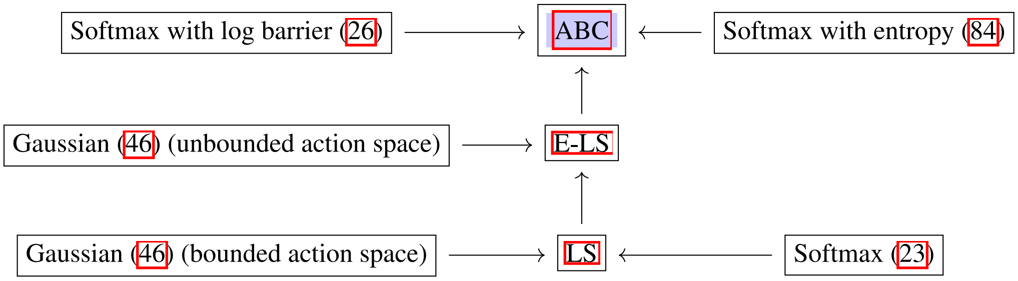
Softmax with log barrier (26)
ABC
Softmax with entropy (84)
Gaussian (46) (unbounded action space)
E-LS
Gaussian (46) (bounded action space)
LS
Softmax (23)

Figure 1: A hierarchy between the assumptions we present throughout the paper. An arrow indicates an implication.

Assumption 4.1 (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

$$\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\|\nabla_\theta \log \pi_\theta(a \mid s)\|^2\right] \;\leq\; G^2, \tag{14}$$

$$\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\left\|\nabla_\theta^2 \log \pi_\theta(a \mid s)\right\|\right] \;\leq\; F. \tag{15}$$

We call the above *Expected* Lipschitz and Smooth (E-LS), due to the expectation of $a \sim \pi_\theta(\cdot \mid s)$, in contrast to the following more restrictive **Lipschitz and smooth policy** (LS) assumption without expectation

$$\|\nabla_\theta \log \pi_\theta(a \mid s)\| \;\leq\; G \quad \text{and} \quad \left\|\nabla_\theta^2 \log \pi_\theta(a \mid s)\right\| \;\leq\; F, \tag{LS}$$

for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. This more restrictive (LS) assumption is widely adopted in the analysis of vanilla PG (Zhang et al., 2020c) and variance-reduced PG methods, e.g. (Shen et al., 2019; Xu et al., 2020a;b; Yuan et al., 2020; Huang et al., 2020; Pham et al., 2020; Liu et al., 2020; Zhang et al., 2021). It is also a relaxation of the element-wise boundness of $\left|\frac{\partial}{\partial\theta_i} \log \pi_\theta(a \mid s)\right|$ and $\left|\frac{\partial^2}{\partial\theta_i\partial\theta_j} \log \pi_\theta(a \mid s)\right|$ assumed by Pirotta et al. (2015) and Papini et al. (2018).

4.1.1 Expected Lipschitz and smooth policy is a special case of ABC

In the following lemma we show that (E-LS) implies the ABC assumption.

Lemma 4.2. Under Asm. 4.1, consider a truncated gradient estimator defined either in (4) or (6). Asm. 3.3 holds with $A = 0, B = 1 - \frac{1}{m}$ and $C = \frac{\nu}{m}$, that is,

$$\mathbb{E}\left[\left\|\widehat{\nabla}_m J(\theta)\right\|^2\right] \leq \left(1 - \frac{1}{m}\right)\|\nabla J_H(\theta)\|^2 + \frac{\nu}{m}, \tag{16}$$

where m is the mini-batch size, and $\nu = \frac{HG^2\mathcal{R}^2_{\max}}{(1-\gamma)^2}$ when using REINFORCE gradient estimator (4) or $\nu = \frac{G^2\mathcal{R}^2_{\max}}{(1-\gamma)^3}$ when using GPOMDP gradient estimator (6).

Bounded variance of the gradient estimator. Interestingly, from (16) we immediately obtain

$$\mathbb{V}\mathrm{ar}\left[\widehat{\nabla}_m J(\theta)\right] \;=\; \mathbb{E}\left[\left\|\widehat{\nabla}_m J(\theta)\right\|^2\right] - \|\nabla J_H(\theta)\|^2 \overset{(16)}{\leq} \frac{\nu - \|\nabla J_H(\theta)\|^2}{m} \;\leq\; \frac{\nu}{m}, \tag{17}$$

which was used as an assumption by Papini et al. (2018); Xu et al. (2020a;b); Yuan et al. (2020); Huang et al. (2020); Liu et al. (2020). Yet (17) need not be an additional assumption since it is a direct consequence of Asm. 4.1.

The (LS) and (E-LS) form the backbone of our hierarchy of assumptions in Figure 1. In particular, (LS) implies (E-LS), and thus ABC is the weaker (and most general) assumption of the three. We formalize this statement in Cor. 4.3.

Corollary 4.3. The (ABC) assumption is the weakest condition compared to (LS) and (E-LS).

4.1.2 Sample complexity analysis for stationary point convergence

Of independent interest to the ABC assumption, Asm. 4.1 also implies the smoothness of $J(\cdot)$ and the truncated gradient assumptions as reported in the following lemmas.

Lemma 4.4. Under Asm. 4.1, $J(\cdot)$ is L-smooth, namely $\left\|\nabla^2 J(\theta)\right\| \leq L$ for all θ which is a sufficient condition of Asm. 3.1, with

$$L \quad = \quad \frac{\mathcal{R}_{\max}}{(1-\gamma)^2}\left(G^2+F\right). \tag{18}$$

The smoothness constant (18) is tighter by a factor of $1-\gamma$ as compared to the smoothness constant proposed in (Papini et al., 2019). This is the tightest upper bound of $\nabla^2 J(\cdot)$ we are aware of in the existing literature (see App. A).

Lemma 4.5. Under Asm. 4.1, Asm. 3.2 holds with

$$D \;=\; \frac{D' G \mathcal{R}_{\max}}{(1-\gamma)^{3/2}} \qquad \text{and} \qquad D' \;=\; \frac{G\mathcal{R}_{\max}}{1-\gamma}\sqrt{\frac{1}{1-\gamma}+H}. \tag{19}$$

As a by-product, in Lemma D.1 in the appendix, we also show that $J(\cdot)$ is Lipschitz under Asm. 4.1 with a tighter Lipschitzness constant, as compared to (Papini et al., 2019; Xu et al., 2020b; Yuan et al., 2020). See more details in App. D.5.

Now we can establish the sample complexity of vanilla PG for the expected Lipschitz and smooth policy assumptions as a corollary of Thm. 3.4 and Lemmas 4.2, 4.4, and 4.5.

Corollary 4.6. Suppose that Asm. 4.1 is satisfied. Let $\delta_0 \overset{\text{def}}{=} J^* - J(\theta_0)$. The PG method applied in (7) with a mini-batch sampling of size m and constant step size

$$\eta \quad \in \quad \left(0,\; \frac{2}{L(1-1/m)}\right), \tag{20}$$

satisfies

$$\begin{aligned} \mathbb{E}\left[\|\nabla J(\theta_U)\|^2\right] \;\leq\; & \frac{2\delta_0}{\eta T\left(2-L\eta\left(1-\frac{1}{m}\right)\right)} + \frac{L\nu\eta}{m\left(2-L\eta\left(1-\frac{1}{m}\right)\right)} \\ & + \left(\frac{2D\left(3-L\eta\left(1-\frac{1}{m}\right)\right)}{2-L\eta\left(1-\frac{1}{m}\right)} + D'^2\gamma^H\right)\gamma^H, \end{aligned} \tag{21}$$

where ν, L and $D, D' > 0$ are provided in Lemmas 4.2, 4.4 and 4.5, respectively.

We first note that Cor. 4.6 imposes no restriction on the batch size, allowing us to analyse both exact full PG and its stochastic variants REINFORCE and GPOMDP. For exact PG, i.e., $H = m = \infty$, we recover the $\mathcal{O}(1/T)$ convergence. This translates to an iteration complexity $T = \mathcal{O}\left(\frac{1}{\epsilon^2}\right)$ with a constant step size $\eta = \frac{1}{L}$ to guarantee $\mathbb{E}\left[\|\nabla J(\theta_U)\|^2\right] = \mathcal{O}(\epsilon^2)$. On the other extreme, when $m = 1$, by (20) we have that $\eta \in (0, \infty)$, i.e., we place no restriction on the step size. In this case, we have that (21) reduces to

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

Thus the stepsize η controls the trade-off between the rate of convergence $\frac{1}{\eta T}$ and leading constant term $\frac{L\nu\eta}{2}$. Using Cor. 4.6, next we develop an explicit sample complexity for PG methods.

Corollary 4.7. Consider the setting of Corollary 4.6. 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

$$Tm \geq \frac{8\delta_0 L\nu}{\epsilon^4} = \begin{cases} \mathcal{O}\left(\frac{H}{(1-\gamma)^4\epsilon^4}\right) & \text{for REINFORCE} \\ \mathcal{O}\left(\frac{1}{(1-\gamma)^5\epsilon^4}\right) & \text{for GPOMDP} \end{cases} \tag{22}$$

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

Remark. Given the horizon $H = \mathcal{O}\left((1-\gamma)^{-1}\log\left(1/\epsilon\right)\right)$, we have that (22) shows that the sample complexity of GPOMDP is a factor of $\log\left(1/\epsilon\right)$ smaller than that of REINFORCE.

Cor. 4.7 greatly extends the range of parameters for which PG is guaranteed to converge within the existing literature. It shows that it is *possible* for vanilla policy gradient methods to 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 particular, both Cor.4.4 in Zhang et al. (2020c) and Prop.1 in Xiong et al. (2021) establish $\widetilde{\mathcal{O}}\left(\epsilon^{-4}\right)$ for FOSP convergence by using the more restrictive assumption (LS). Papini (2020) obtain the same results with the weaker assumption (E-LS), which is also our case. However, we improve upon all of them by recovering the exact full PG analysis, allowing much wider range of choices for the batch size m and the constant step size η to achieve the same optimal sample complexity $\widetilde{\mathcal{O}}\left(\epsilon^{-4}\right)$.

In terms of the freedom of the hyperparameter choices, our result is novel. Indeed, to achieve the optimal sample complexity, Papini et al. (2018); Shen et al. (2019); Xu et al. (2020a;b); Yuan et al. (2020); Liu et al. (2020); Zhang et al. (2021) do not allow a single trajectory sampled per iteration. They require the batch size m to be either ϵ^{-1} or ϵ^{-2}. Otherwise, when $m = 1$, their analysis would not return the optimal rate of convergence. The existing analysis that allow $m = 1$ that we are aware of are (Zhang et al., 2020b) and (Huang et al., 2020). However, when $m > \mathcal{O}(1)$, the analysis of Huang et al. (2020) does not benefit from larger batch sizes and thus fail to match the optimal sample complexity for large batch sizes. The comparison with (Zhang et al., 2020b) will be detailed in Sec. 4.2.1 under the specific setting of softmax tabular policy with log barrier regularization.

4.2 Softmax tabular policy

In this section, we instantiate the FOSP convergence results of Cor. 4.6 and 4.7 in the case of the softmax tabular policy. Combined with the specific properties of the softmax, our general theory also recovers the average regret of the global optimum convergence analysis for the softmax with log barrier regularization (Zhang et al., 2020b) and brings new insights of the theory by leveraing the ABC assumption analysis.

Here, the state space $\mathcal{S}$ and the action space $\mathcal{A}$ are finite. For all $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and any state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, consider the following softmax tabular policy

$$\pi_\theta(s \mid a) \overset{\text{def}}{=} \frac{\exp(\theta_{s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta_{s,a'})}. \tag{23}$$

We show that the softmax tabular policy satisfies (E-LS) as illustrated in the following lemma.

Lemma 4.8. The softmax tabular policy satisfies Asm. 4.1 with $G^2 = 1 - \frac{1}{|\mathcal{A}|}$ and $F = 1$, that is, for all $s \in \mathcal{S}$, we have

$$\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\|\nabla_\theta \log \pi_{s,a}(\theta)\|^2\right] \leq 1 - \frac{1}{|\mathcal{A}|}, \tag{24}$$

$$\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\left\|\nabla_\theta^2 \log \pi_{s,a}(\theta)\right\|\right] \leq 1. \tag{25}$$

Remark. The softmax tabular policy also satisfies (LS) but with a bigger constant (see App. E.2).

Lemma 4.8 and the results in Section 4.1 immediately imply that Asm. 3.1, 3.2 and 3.3 are verified. Thus, as a consequence of Cor. 4.6, we have the following sample complexity for the softmax tabular policy.[4]

Corollary 4.9 (Informal)**.** Given $\epsilon > 0$, there exists a range of parameter choices for the batch size m s.t. $1 \leq m \leq \mathcal{O}(\epsilon^{-2})$, the step size η s.t. $\mathcal{O}(\epsilon^2) \leq \eta \leq \mathcal{O}(1)$, the number of iterations T and the horizon H such that the sample complexity of the vanilla PG (either REINFORCE or GPOMDP) is $Tm \times H = \widetilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^6\epsilon^4}\right)$ to achieve $\mathbb{E}\left[\|\nabla J(\theta_U)\|^2\right] = \mathcal{O}(\epsilon^2)$.

4.2.1 Global optimum convergence of softmax with log barrier regularization

Leveraging the work of Agarwal et al. (2021) and our Thm. 3.4, we can establish a global optimum convergence analysis for softmax policies with log barrier regularization.

Log barrier regularization is often used to prevent the policy from becoming deterministic. Indeed, when optimizing the softmax by PG, policies can rapidly become near deterministic and the optimal policy is usually obtained by sending some parameters to infinity. This can result in an extremely slow convergence of PG. Li et al. (2021) show that PG can even take exponential time to converge. To prevent the parameters from becoming too large and to ensure enough exploration, a log barrier regularization term is commonly used to keep the probabilities from getting too small (Williams & Peng, 1991; Mnih et al., 2016). The regularized objective is defined as

$$L_\lambda(\theta) \overset{\text{def}}{=} J(\theta) + \frac{\lambda}{|\mathcal{A}||\mathcal{S}|}\sum_{s,a} \log \pi_\theta(a \mid s) + \lambda \log |\mathcal{A}|. \tag{26}$$

Similar to the softmax, we show in App. E.3 that $L_\lambda(\theta)$ is smooth and satisfies the (ABC) assumption. Thus, from Thm. 3.4, we have $\{\theta_t\}_{t\geq 0}$ converges to a FOSP of $L_\lambda(\cdot)$. We postpone the formal statement of this result to App. E.3 for the sake of space. Besides, thanks to Thm. 5.2 in (Agarwal et al., 2021), the FOSP of $L_\lambda(\cdot)$ is directly linked to the global optimum of $J(\cdot)$. As a by-product, we can also establish a high probability global optimum convergence analysis (App. E.4).

In the following corollary, we show that we can leverage the versatility of Thm. 3.4 to derive yet another type of result: a guarantee on the average regret w.r.t. the global optimum.

Corollary 4.10. 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. E.5, the horizon $H = \mathcal{O}\left(\frac{\log(1/\epsilon)}{1-\gamma}\right)$ and the number of iterations T 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 recovers the sample complexity $\widetilde{\mathcal{O}}(\epsilon^{-6})$ of (Zhang et al., 2020b). However, Zhang et al. (2020b) do not study vanilla policy gradient. Instead, they add an extra phased learning step to enforce the exploration of the MDP and used a decreasing step size. Our result shows that such extra phased learning step is unnecessary and the step size can be constant. We also provide a wider range of parameter choices for the batch size and the step size with the same sample complexity.

5 Discussion

We believe the generality of Thm. 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., 2021) leveraging composite optimization tools.

[4] The exact statement is similar to Cor. 4.7. For the sake of space here we report a more compact statement.

REFERENCES

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.

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

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

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

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

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

Feihu Huang, Shangqian Gao, Jian Pei, and Heng Huang. Momentum-based policy gradient methods. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4422–4433. PMLR, 13–18 Jul 2020.

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

Ahmed Khaled and Peter Richtárik. Better theory for sgd in the nonconvex world, 2020.

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

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. doi: 10.1109/TNNLS.2019.2952219.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Softmax policy gradient methods can take exponential time to converge. In Mikhail Belkin and Samory Kpotufe (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 3107–3110. PMLR, 15–19 Aug 2021.

Yanli Liu, Kaiqing Zhang, Tamer Basar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 7624–7636. Curran Associates, Inc., 2020.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6820–6829. PMLR, 13–18 Jul 2020.

A. Yu. Mitrophanov. Sensitivity and convergence of uniformly ergodic markov chains. *Journal of Applied Probability*, 42(4):1003–1014, 2005. ISSN 00219002.

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 Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1928–1937, New York, New York, USA, 20–22 Jun 2016. PMLR.

Matteo Papini. Safe policy optimization. 2020.

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, pp. 4026–4035. PMLR, 2018.

Matteo Papini, Matteo Pirotta, and Marcello Restelli. Smoothing policies and safe policy gradients, 2019.

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 Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 374–385. PMLR, 26–28 Aug 2020.

Matteo Pirotta, Marcello Restelli, and Luca Bascetta. Policy gradient in lipschitz markov decision processes. *Machine Learning*, 100(2):255–283, Sep 2015. ISSN 1573-0565. doi: 10.1007/s10994-015-5484-1.

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 Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8469–8479. PMLR, 13–18 Jul 2020.

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

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897, Lille, France, 07–09 Jul 2015. PMLR.

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

Zebang Shen, Alejandro Ribeiro, Hamed Hassani, Hui Qian, and Chao Mi. Hessian aided policy gradient. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5729–5738. PMLR, 09–15 Jun 2019.

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

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

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

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

Ronald J. Williams and Jing Peng. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3(3):241–268, 1991. doi: 10.1080/09540099108946587.

Lin Xiao and Lihong Li. A tutorial on policy gradient methods. In *SIAM Conference on Optimization*, 2021.

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.

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

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

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

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

Junyu Zhang, Chengzhuo Ni, Zheng Yu, Csaba Szepesvari, and Mengdi Wang. On the convergence and sample efficiency of variance-reduced policy gradient method, 2021.

Junzi Zhang, Jongho Kim, Brendan O'Donoghue, and Stephen Boyd. Sample efficient reinforcement learning with reinforce, 2020b.

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, 2020c. doi: 10.1137/19M1288012.