A general sample complexity analysis of vanilla policy gradient

Rui Yuan, Robert M Gower, Alessandro Lazaric

A general sample complexity analysis of vanilla policy gradient

Rui Yuan
Meta AI
LTCI, Télécom Paris
Institut Polytechnique de Paris

Robert M. Gower
CCM, Flatiron Institute, New York
LTCI, Télécom Paris
Institut Polytechnique de Paris

Alessandro Lazaric
Meta AI

Abstract

We adapt recent tools developed for the analysis of Stochastic Gradient Descent (SGD) in non-convex optimization to obtain convergence and sample complexity guarantees for the vanilla policy gradient (PG). Our only assumptions are that the expected return is smooth w.r.t. the policy parameters, that its H-step truncated gradient is close to the exact gradient, and a certain *ABC assumption*. This assumption requires the second moment of the estimated 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 that recovers the $\widetilde{\mathcal{O}}(\epsilon^{-4})$ sample complexity of PG. Our results also affords greater flexibility in the choice of hyper parameters such as the step size and places no restriction on the batch size m, including the single trajectory case (i.e., $m = 1$). We then instantiate our theorem in different settings, where we both recover existing results and obtained improved sample complexity, e.g., for convergence to the global optimum for Fisher-non-degenerated parameterized policies.

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; Baxter and Bartlett, 2001). 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 and Tsitsiklis, 2000; Mnih et al., 2016), natural PG (Kakade, 2002), policy mirror descent (Tomar et al., 2020; Vaswani et al., 2022), trust-region based variants (Schulman et al., 2015, 2017; Shani et al., 2020), and variance-reduced methods (Papini et al., 2018; Shen et al., 2019; Xu et al., 2020a; Yuan et al., 2020; Huang et al., 2020; Pham et al., 2020; Yang et al., 2021; Huang et al., 2022). 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 (see App. A for a thorough review).

In this paper, we first 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. Our first contribution is convergence guarantees and sample complexity for both REINFORCE (Williams, 1992) and GPOMDP (Sutton et al., 2000; Baxter and Bartlett, 2001) under the ABC and assumptions on the smoothness of the expected return and on its truncated gradient, 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 (Zhang et al., 2020b; Liu et al., 2020; Xiong et al., 2021). Furthermore, our analysis is less restrictive when it comes to the hyper-parameter choices. In fact, our results allow for a wide range of step sizes and

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

Table 1: Overview of different convergence results for vanilla PG methods. The darker cells contain our new results. The light 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. H.2**	$\widetilde{\mathcal{O}}(\epsilon^{-1})$	Recover linear convergence for the exact PG
	ABC + (14)	**Thm. C.1**	$\widetilde{\mathcal{O}}(\epsilon^{-3})$	Recover $\mathcal{O}(\epsilon^{-1})$ for the exact PG
	E-LS + FI + compatible	**Cor. 4.14**	$\widetilde{\mathcal{O}}(\epsilon^{-3})$	Weaker asm.; Improved by ϵ compared to Thm. 3.4
Sample complexity of stochastic PG for AR	LS + FI + compatible	Liu et al. (2020)	$\widetilde{\mathcal{O}}(\epsilon^{-4})$	
	Softmax + log barrier (28)	Zhang et al. (2021a) **Cor. 4.11**	$\widetilde{\mathcal{O}}(\epsilon^{-6})$	Constant step size; Wider range of parameters; Extra phased learning step unnecessary
Iteration complexity of the exact PG for GO	Softmax + log barrier (28)	Agarwal et al. (2021) **Cor. E.5**	$\mathcal{O}(\epsilon^{-2})$	Improved by $1-\gamma$
	Softmax (25)	Mei et al. (2020) **Thm. C.1**	$\mathcal{O}(\epsilon^{-1})$	
	Softmax + entropy (125)	Mei et al. (2020) **Thm. H.2**	linear	
	LS + bijection + PPG	Zhang et al. (2020a)	$\mathcal{O}(\epsilon^{-1})$	
	Tabular + PPG	Xiao (2022)	$\mathcal{O}(\epsilon^{-1})$	
	LQR	Fazel et al. (2018)	linear	

* **Type of convergence.** *PG*: policy gradient; *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; *Tabular*: direct parametrized policy; *LQR*: linear-quadratic regulator.

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 one guise. Indeed, we show that the analysis of Lipschitz and smooth policies, Gaussian polices, softmax tabular polices with or without a log barrier or an entropy 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). This usually requires more restrictive assumptions (Zhang et al., 2020a, 2021b), specific RL settings (e.g., linear-quadratic regulator (Fazel et al., 2018), tabular (Agarwal et al., 2021) and softmax tabular policy (Mei et al., 2020)), and it is often limited to exact PG. Inspired by the sample complexity analysis of the stochastic PG for the global optimum in Liu et al. (2020) and Ding et al. (2021a), our second contribution is to establish a novel global optimum convergence theory of PG when an additional *relaxed weak gradient domination* assumption is available. Our sample complexity analysis recovers the well known $\mathcal{O}(\epsilon^{-1})$ iteration complexity of the exact PG with the softmax tabular policy (Mei et al., 2020) as a special case and obtains a new improved $\widetilde{\mathcal{O}}(\epsilon^{-3})$ sample complexity compared to $\widetilde{\mathcal{O}}(\epsilon^{-4})$ in Liu et al. (2020), with the Fisher-non-degenerate parametrized policy (Liu et al., 2020; Ding et al., 2021a) as a special case. We also establish even faster global optimum convergence theory when replacing the relaxed weak gradient domination assumption by gradient domination in App. H. Table 1 provides a complete overview of our results.

2 Preliminaries

Markov decision process (MDP). We consider a MDP $M = \{\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, \rho\}$, where $\mathcal{S}$ is a state space; $\mathcal{A}$ is an action space; $\mathcal{P}$ is a Markovian transition model, where $\mathcal{P}(s' \mid s, a)$ 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 $\pi(a \mid s)$ is the density of the distribution over actions at 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

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

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

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

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 $p(\tau \mid \theta) = p_\theta(\tau) = p(\tau \mid \pi_\theta)$. In general, $J(\theta)$ is a non-convex function. The PG methods use gradient ascent in the space of θ to find the policy that maximizes the expected return, i.e., $\theta^* \in \arg\sup_{\theta \in \mathbb{R}^d} J(\theta)$. We denote the *optimal expected return* as $J^* \stackrel{\text{def}}{=} J(\theta^*)$.

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

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

In practice, we cannot compute this full gradient, since computing the above expectation requires 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 $\tau_i = \left(s_0^i, a_0^i, r_0^i, s_1^i, \cdots, s_{H-1}^i, a_{H-1}^i, r_{H-1}^i\right)$ with $r_t^i = \mathcal{R}(s_t^i, a_t^i)$ 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

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

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

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

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

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.

3 Non-convex optimization under ABC assumption

3.1 First-order stationary point convergence

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 exact gradient $\nabla J(\theta)$ when $H = m = \infty$, or the truncated gradient estimators 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

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

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

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

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

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 and Richtárik, 2020)[1] which bounds the second moment of the norm of the gradient estimators using the norm of the truncated full gradient, the suboptimality gap and an additive constant.

Assumption 3.3 (ABC)**.** There exists $A, B, C \geq 0$ such that the policy gradient estimator satisfies

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

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

The ABC assumption effectively summarizes a number of popular and more restrictive assumptions commonly used in non-convex optimization. Indeed, the bounded variance of the stochastic gradient assumption (Ghadimi and Lan, 2013), the gradient confusion assumption (Sankararaman et al., 2020), the sure-smoothness assumption (Lei et al., 2020), the convex expected smoothness assumption (Gower et al., 2019, 2021) and different variants of strong growth assumptions proposed by Schmidt and 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 and Richtárik (2020).

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

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

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

In particular if $A = 0$, we have

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

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

Thm. 3.4 provides a general characterization of the convergence of PG as a function of all the constants involved in the assumptions on the problem and the policy gradient estimator. Refer to App. A.1 for a discussion comparing the technical aspects of this result compared to Khaled and Richtárik (2020). From (11) we derive the sample complexity 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 satisfies

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

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 cases.

First, (13) recovers the $\mathcal{O}(\epsilon^{-2})$ iteration complexity of the exact 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), this shows that for any policy and MDP that satisfy the smoothness property (Asm. 3.1), the exact full PG converges to a ϵ-FOSP in $T = \mathcal{O}(\epsilon^{-2})$ 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).

Second, we recover sample complexity for stochastic vanilla PG. 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 con-

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

verge to an ϵ-stationary point. Our sample complexity matches the results of Papini (2020); Zhang et al. (2020b); Liu et al. (2020); Xiong et al. (2021), but improve upon them in generality, i.e., by recovering the exact PG analysis, providing wider range of parameter choices and using the weaker ABC assumption (see Sec. 4.1 for more details).

3.2 Global optimum convergence under relaxed weak gradient domination

In this section, we present a global optimum convergence of the vanilla PG when the relaxed weak gradient domination assumption is available, in addition to the (ABC) assumption.

Assumption 3.6 (Relaxed weak gradient domination)**.** We say that J satisfies the weak gradient domination condition if for all $\theta \in \mathbb{R}^d$, there exists $\mu > 0$ and $\epsilon' \geq 0$ such that

$$\epsilon' + \|\nabla J_H(\theta)\| \;\geq\; 2\sqrt{\mu}\,(J^* - J(\theta))\,. \tag{14}$$

The relaxed weak gradient domination is an extension of weak gradient domination[2] (Agarwal et al., 2021; Mei et al., 2020, 2021) where $\epsilon' = 0$. Equipped with this assumption, we obtain a new global optimum convergence guarantee (see Thm. C.1 in App. C.3 for the full details).

Corollary 3.7. Consider the setting of Thm. C.1. 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 $\min_{t\in\{0,1,\cdots,T\}} J^* - \mathbb{E}\left[J(\theta_t)\right] \leq \mathcal{O}(\epsilon) + \mathcal{O}(\epsilon')$.

Consequently, when $\epsilon' = \Theta(\epsilon)$ we have that the complexity of PG to reach a global optimum is $\mathcal{O}(\epsilon^{-3})$. Thus the relaxed weak gradient domination has afforded us a factor of ϵ^{-1} improvement as compared to the $\mathcal{O}(\epsilon^{-4})$ complexity in Corollary 3.5. The relaxed weak gradient domination is an assumption that is unique to PG methods. In Sec. 4.3, we show that the Fisher-non-degenerate parametrized policy satisfies this 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 match 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). Finally, we show that the relaxed weak gradient domination assumption holds for Fisher-non-degenerate parametrized policies, thus leading to new improved sample complexity result for this setting.

4.1 Expected Lipschitz and smooth policies

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

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

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

$$\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\left\|\nabla^2_\theta \log \pi_\theta(a \mid s)\right\|\right] \;\leq\; F. \tag{16}$$

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 more restrictive **Lipschitz and smooth policy** (LS) assumption

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

for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. The (LS) assumption is widely adopted in the analysis of vanilla PG (Zhang et al., 2020b) and variance-reduced PG methods, e.g. Shen et al. (2019); Xu et al. (2020b,a); Yuan et al. (2020); Huang et al. (2020); Pham et al. (2020); Liu et al. (2020); Zhang et al. (2021b). 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.

[2]The weak gradient domination is the special case of the Kurdyka-Łojasiewicz (KL) condition with KL exponent 1 (Kurdyka, 1998).

[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 (15) by Cauchy-Schwarz inequality.

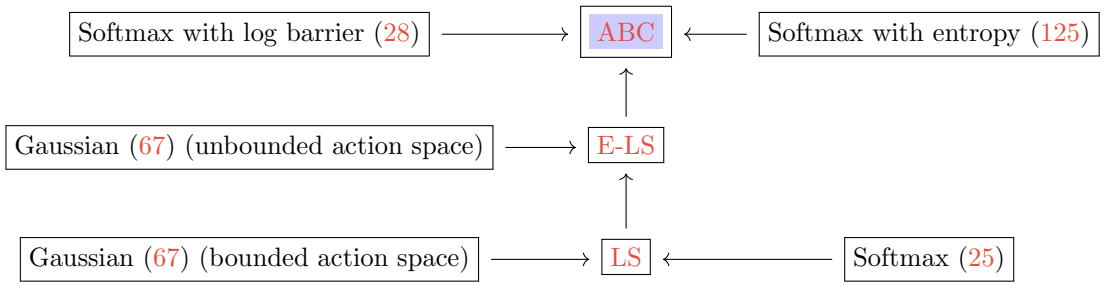

Figure 1: A hierarchy between the assumptions we present throughout the paper. An arrow indicates an implication.

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,

$$\mathbb{E}\left[\left\|\widehat{\nabla}_m J(\theta)\right\|^2\right] \leq \left(1 - \frac{1}{m}\right)\|\nabla J_H(\theta)\|^2 + \frac{\nu}{m}, \quad (17)$$

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

Bounded variance of the gradient estimator. Interestingly, from (17) we immediately obtain

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[\widehat{\nabla}_m J(\theta)\right] &= \mathbb{E}\left[\left\|\widehat{\nabla}_m J(\theta)\right\|^2\right] - \|\nabla J_H(\theta)\|^2 \\ &\overset{(17)}{\leq} \frac{\nu - \|\nabla J_H(\theta)\|^2}{m} \leq \frac{\nu}{m}, \quad (18)\end{aligned}$$

which was used as an assumption by Papini et al. (2018); Xu et al. (2020b,a); Yuan et al. (2020); Huang et al. (2020); Liu et al. (2020). Yet (18) needs not to 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.

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

$$L = \frac{\mathcal{R}_{\max}}{(1-\gamma)^2}\left(G^2 + F\right). \quad (19)$$

The smoothness constant (19) 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.3).

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

$$D = \frac{D'G\mathcal{R}_{\max}}{(1-\gamma)^{3/2}}, \quad (20)$$

$$D' = \frac{G\mathcal{R}_{\max}}{1-\gamma}\sqrt{\frac{1}{1-\gamma} + H}. \quad (21)$$

The coefficient D' in (21) got improved and is tighter by a factor of $(1-\gamma)^{1/2}$ as compared to the same term analysed in Lemma B.1 in Liu et al. (2020).

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. (2020a); 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

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

satisfies

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

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 (22) we have that $\eta \in (0, \infty)$, i.e., we place no restriction on the step size. In this case, we have that (23) 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

$$Tm \geq \frac{8\delta_0 L\nu}{\epsilon^4} = \begin{cases} \mathcal{O}\left(\frac{H}{(1-\gamma)^4\epsilon^4}\right) & \text{for REINFORCE} \\ \mathcal{O}\left(\frac{1}{(1-\gamma)^5\epsilon^4}\right) & \text{for GPOMDP} \end{cases} \quad (24)$$

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

Remark. Given the horizon $H = \mathcal{O}\left((1-\gamma)^{-1}\log(1/\epsilon)\right)$, we have that (24) shows that the sample complexity of GPOMDP is a factor of $\log(1/\epsilon)$ 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, Cor.4.4 in Zhang et al. (2020b), Prop.1 in Xiong et al. (2021) and Thm.E.1 in Liu et al. (2020) 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 $\widetilde{\mathcal{O}}\left(\epsilon^{-4}\right)$. Indeed, to achieve the optimal sample complexity of FOSP, Papini (2020); Zhang et al. (2020b); Xiong et al. (2021); Liu et al. (2020) do not allow a single trajectory sampled per iteration. They require the batch size m to be either ϵ^{-1} or ϵ^{-2}. The existing analysis for vanilla PG that allows $m = 1$ that we are aware of is Zhang et al. (2021a), which we compare with in Sec. 4.2.1 under the specific setting of softmax tabular policy with log barrier regularization for the average regret analysis.

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., 2021a) 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

$$\pi_\theta(a \mid s) \overset{\text{def}}{=} \frac{\exp(\theta_{s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta_{s,a'})}. \quad (25)$$

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

$$\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\|\nabla_\theta \log \pi_\theta(a \mid s)\|^2\right] \leq 1 - \frac{1}{|\mathcal{A}|}, \quad (26)$$

$$\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\left\|\nabla^2_\theta \log \pi_\theta(a \mid s)\right\|\right] \leq 1. \quad (27)$$

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 all assumptions including the (ABC) assumption of Thm. 3.4 are verified. Thus, as a consequence of Cor. 4.6 and 4.7, 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, an entropy-based regularization term is commonly used to keep the probabilities from getting too small (Williams and Peng, 1991; Mnih et al., 2016; Nachum et al., 2017; Haarnoja et al., 2018; Mei et al., 2019) . Here we study stochastic gradient ascent on a relative entropy regularized objective, softmax with log barrier regularization, which is defined as

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

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

Let $\widehat{\nabla}_m L_\lambda(\theta)$ be the stochastic gradient estimator of $L_\lambda(\theta)$ using REINFORCE or GPOMDP with batch size m. Thus $\widehat{\nabla}_m L_\lambda(\theta)$ is an unbiased estimate of the gradient of the truncated function

$$L_{\lambda,H}(\theta) \stackrel{\text{def}}{=} J_H(\theta) + \frac{\lambda}{|\mathcal{A}||\mathcal{S}|}\sum_{s,a}\log \pi_\theta(a \mid s) + \lambda \log|\mathcal{A}|. \tag{29}$$

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

We show in the following that $\widehat{\nabla}_m L_\lambda(\theta)$ satisfies the (ABC).

Lemma 4.10. Consider $\widehat{\nabla}_m L_\lambda(\theta)$ by using either RIENFORCE (4) or GPOMDP (6), Asm. 3.3 holds with $A = 0, B = 1 - \frac{1}{m}$ and $C = \frac{\nu}{m}$, that is,

$$\mathbb{E}\left[\left\|\widehat{\nabla}_m L_\lambda(\theta)\right\|^2\right] \leq \left(1 - \frac{1}{m}\right)\|\nabla L_{\lambda,H}(\theta)\|^2 + \frac{\nu}{m}, \tag{30}$$

where $\nu = 2\left(1 - \frac{1}{|\mathcal{A}|}\right)\left(\frac{H\mathcal{R}_{\max}^2}{(1-\gamma)^2} + \frac{\lambda^2}{|\mathcal{S}|}\right)$ when using REINFORCE or $\nu = 2\left(1 - \frac{1}{|\mathcal{A}|}\right)\left(\frac{\mathcal{R}_{\max}^2}{(1-\gamma)^3} + \frac{\lambda^2}{|\mathcal{S}|}\right)$ when using GPOMDP.

Similar to the softmax case, we show in App. E.3 that $L_\lambda(\theta)$ is also smooth and verifies Asm. 3.2. 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 approximately the global optimal solution of $J(\cdot)$ when the regularization parameter λ is is sufficiently small. 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 4.11. 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.4 . 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

$$J^* - \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[J(\theta_t)\right] = \mathcal{O}(\epsilon). \tag{31}$$

This result recovers the sample complexity $\widetilde{\mathcal{O}}(\epsilon^{-6})$ of Zhang et al. (2021a). However, Zhang et al. (2021a) do not study the vanilla policy gradient. Instead, they add an extra phased learning step to enforce the exploration of the MDP and use 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.

As Agarwal et al. (2021) mentioned, the regularizer (28) is more "aggressive" in penalizing small probabilities than the more commonly utilized entropy regularizer. We also show that entropy regularized softmax satisfies the (ABC) and provide its FOSP analysis in App. G, again thanks to the versatility of Thm. 3.4. Notice that for the FOSP convergence, only an asymptotic result was established in Lemma 4.4 in Ding et al. (2021b). Thus all proofs and implications in Fig. 1 are provided.

4.3 Fisher-non-degenerate parameterization

In this section, we study a general policy class that satisfies the following assumption.

Assumption 4.12 (Fisher-non-degenerate, Asm. 2.1 in Ding et al. (2021a)). For all $\theta \in \mathbb{R}^d$, there exists $\mu_F > 0$ s.t. the Fisher information matrix $F_\rho(\theta)$ induced by policy π_θ and initial state distribution ρ satisfies

$$F_\rho(\theta) \stackrel{\text{def}}{=} \mathbb{E}_{(s,a)\sim v_\rho^{\pi_\theta}} \left[\nabla_\theta \log \pi_\theta(a \mid s) \nabla_\theta \log \pi_\theta(a \mid s)^\top\right] \geq \mu_F \mathbf{I}_d, \tag{FI}$$

where $v_\rho^{\pi_\theta}$ is the state-action visitation measure defined as

$$v_\rho^{\pi_\theta}(s,a) \stackrel{\text{def}}{=} (1-\gamma)\mathbb{E}_{s_0\sim\rho} \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s, a_t = a | s_0, \pi_\theta).$$

This assumption is commonly used in the literatures (Liu et al., 2020; Ding et al., 2021a). Similar conditions of the Fisher-non-degeneracy is also required in other global optimum convergence framework (Asm. 6.5 in Agarwal et al. (2021) on the relative condition number). This assumption is satisfied by a wide families of policies, including the Gaussian policy (67) and certain neural policy. We refer to Sec. B.2 in Liu et al. (2020) and Sec. 8 in Ding et al. (2021a) for more discussions on the Fisher-non-degenerate setting.

We also need the following *compatible function approximation error* assumption[5].

Assumption 4.13 (Compatible, Asm. 4.6 in Ding et al. (2021a)). For all $\theta \in \mathbb{R}^d$, there exists $\epsilon_{bias} > 0$ s.t. the *transferred compatible function approximation error* with $(s,a) \sim v_\rho^{\pi_{\theta^*}}$ satisfies

$$\mathbb{E}\left[(A^{\pi_\theta}(s,a) - (1-\gamma)u^{*\top}\nabla_\theta \pi_\theta(a \mid s))^2\right] \leq \epsilon_{bias}, \tag{compatible}$$

where $v_\rho^{\pi_{\theta^*}}$ is the state-action distribution induced by an optimal policy π_{θ^*}, $u^* = (F_\rho(\theta))^\dagger \nabla J(\theta)$.

This is also a common assumption (Wang et al., 2020; Agarwal et al., 2021; Liu et al., 2020; Ding et al., 2021a). In particular, when π_θ is a softmax tabular policy (86), ϵ_{bias} is 0 (Ding et al., 2021a); when π_θ is a rich neural policy, ϵ_{bias} is small (Wang et al., 2020).

Combining Asm. (FI), (compatible) with Asm. E-LS, by Lemma 4.7 in Ding et al. (2021a), we know that $J(\cdot)$ satisfies the relaxed weak gradient domination property (14) with $\epsilon' = \frac{\mu_F\sqrt{\epsilon_{bias}}}{(1-\gamma)G}$ and $\mu = \frac{\mu_F^2}{4G^2}$. Consequently, we have the following new global optimum convergence result for the Fisher-non-degenerate parametrized policy.

Corollary 4.14. If the policy π_θ satisfies Asm. 4.1, 4.12 and 4.13, consider the setting of Cor. 3.7 with $\epsilon' = \frac{\mu_F\sqrt{\epsilon_{bias}}}{(1-\gamma)G}$ and $\mu = \frac{\mu_F^2}{4G^2}$. Then $\min_{t\in\{0,1,\cdots,T\}} J^* - \mathbb{E}\left[J(\theta_t)\right] \leq \mathcal{O}(\epsilon) + \mathcal{O}(\sqrt{\epsilon_{bias}})$ and the sample complexity $T \times H = \widetilde{\mathcal{O}}(\epsilon^{-3})$ when $\epsilon_{bias} = 0$ or $T \times H = \widetilde{\mathcal{O}}((\epsilon_{bias}\cdot\epsilon)^{-1})$ when $\epsilon_{bias} > 0$.

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. Other interesting venues of investigation include whether it is possible to extend the analysis to projected PG, identify counterparts of the ABC assumption for variance-reduced versions of PG and for the improved analysis of Zhang et al. (2021b) leveraging composite optimization tools.

[5] We defer the definition of the advantage function A^{π_θ} in App.F.

References

Matteo Papini. Safe policy optimization. 2020.

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.

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 2021a.

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.

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.

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. In *Advances in Neural Information Processing Systems*, volume 33, pages 4572–4583. Curran Associates, Inc., 2020a.

Lin Xiao. On the convergence rates of policy gradient methods, 2022.

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

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

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.

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.

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

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.

Sham M Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2002.

Manan Tomar, Lior Shani, Yonathan Efroni, and Mohammad Ghavamzadeh. Mirror descent policy optimization, 2020.

Sharan Vaswani, Olivier Bachem, Simone Totaro, Robert Mueller, 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, 2022.

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.

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

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.

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.

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.

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

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

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.

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.

Long Yang, Yu Zhang, Gang Zheng, Qian Zheng, Pengfei Li, Jun Wen, and Gang Pan. Policy optimization with stochastic mirror descent, 2021.

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

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, 2020b. doi: 10.1137/19M1288012.

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.

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*, 2021b.

Yuhao Ding, Junzi Zhang, and Javad Lavaei. On the global convergence of momentum-based policy gradient, 2021a.

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

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.

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.

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.

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. SGD: General analysis and improved rates. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5200–5209. PMLR, 09–15 Jun 2019.

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.

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

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

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.

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

Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. *Annales de l'institut Fourier*, 48(3):769–783, 1998.

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.

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

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 2020b.

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.

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.

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

Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Bridging the gap between value and policy based reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1861–1870. PMLR, 10–15 Jul 2018.

Jincheng Mei, Chenjun Xiao, Ruitong Huang, Dale Schuurmans, and Martin Müller. On principled entropy exploration in policy optimization. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 3130–3136. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/434.

Yuhao Ding, Junzi Zhang, and Javad Lavaei. Beyond exact gradients: Convergence of stochastic soft-max policy gradient methods with entropy regularization, 2021b.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*, 2020.

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

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.

Lam M. Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2613–2621, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

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.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Quoc Tran-Dinh, Nhan H. Pham, Dzung T. Phan, and Lam M. Nguyen. A hybrid stochastic optimization framework for composite nonconvex optimization. *Mathematical Programming*, Jan 2021. ISSN 1436-4646. doi: 10.1007/s10107-020-01583-1.

Zhuoran Yang, Yongxin Chen, Mingyi Hong, and Zhaoran Wang. Provably global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Harshat Kumar, Alec Koppel, and Alejandro Ribeiro. On the sample complexity of actor-critic method for reinforcement learning with function approximation, 2021.

Tengyu Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for (natural) actor-critic algorithms. In *Advances in Neural Information Processing Systems*, volume 33, pages 4358–4369. Curran Associates, Inc., 2020c.

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

Stanisław Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. Equ. Derivees partielles, Paris 1962, Colloques internat. Centre nat. Rech. sci. 117, 87-89 (1963)., 1963.

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 $\widehat{\nabla}_m J(\theta_t)$
  Update $\theta_{t+1} = \theta_t + \eta_t \widehat{\nabla}_m J(\theta_t)$ and η_t
end for

Appendix

Table of Contents

Here we provide the related work discussion, the missing proofs from the main paper and some additional noteworthy observations made in the main paper.

A Related work

We provide an extended discussion for the context of our work, including a discussion comparing the technical novelty of the paper to the finite sum minimization result in Khaled and Richtárik (2020), a comparison of the convergence theories of vanilla PG and the problem dependent constants. We refer to Algorithm 1 as the vanilla PG with $\widehat{\nabla}_m J(\theta_t)$ defined as either the exact full gradient (3) and (5) or the stochastic PG estimator (4) or (6). Furthermore, we discuss future work to extend our general sample complexity analysis to other policy gradient methods and other RL settings.

A.1 Technical contribution and novelty compared to Khaled and Richtárik (2020)

Our technical novelty compared to Khaled and Richtárik (2020) is threefold. First, Theorem 3.4 is not a direct application of Theorem 2 in Khaled and Richtárik (2020), which requires unbiased estimators of the gradient. Yet in PG methods, we have to deal with biased estimators due to the truncation of the trajectories. The first technical challenge was to adapt the proof technique to allow for biased gradients and a truncation error. This also explains the need of Assumption 3.2. Similarly, we need to handle the same challenge for the proof of Theorem H.2 when adapting the proof of Theorem 3 in Khaled and Richtárik (2020). Second, when considering the results we derived in specific cases in Section 4, the difference between our work and Khaled and Richtárik (2020) is even more significant. All cases studied in Khaled and Richtárik (2020) (e.g., finite-sum structure)

are not applicable for PG methods and we had to derive specific analysis for our specialized settings (soft-max with different regularizers, expected Lipschitz and smooth policies, Fisher-non-degenerate parametrized policies). Furthermore, our focus is on deriving explicit sample complexity, whereas the results in Khaled and Richtárik (2020) are concerned with convergence rates in terms of number of iterations. These dimensions are where most of the technical work was done. Without this work of developing sample complexity and studying specific cases found in PG literatures, it was not clear at all that the (ABC) assumption proposed in Khaled and Richtárik (2020) would be relevant in RL. Finally, we also consider the setting where the relaxed weak gradient domination holds (Assumption 3.6 and Theorem C.1). This is an assumption that is unique to PG methods and had not been considered in Khaled and Richtárik (2020). Technically speaking, the proof of Theorem C.1 is unique and required a different approach (see the arguments following (53)).

A.2 Sample complexity analysis of the vanilla policy gradient

Despite the success of PG methods in practice, a comprehensive theoretical understanding was lacking until recently.

Global optimum convergence of vanilla PG with the exact full gradient. We refer to global optimum convergence as an analysis that guarantees that $J^* - J(\theta_T) \leq \epsilon$ after T iterations. The global optimum convergence results of PG with the exact full gradient have been developed under a number of different specific settings.

By using a gradient domination property of the expected return, which is also referred to as the Polyak-Lojasiewicz (PL) condition, Fazel et al. (2018) show that the linear-quadratic regulator (LQR) converges linearly to the global optimum for PG with the exact full gradient. However, in the LQR setting the function J is not smooth, and thus does not fit into the general setting we considered in this paper. More recently, Agarwal et al. (2021) leveraged a *weak* gradient domination property, also called the weak Polyak-Lojasiewicz condition which is exactly our condition (14) with $\epsilon' = 0$, to show that the projected PG converges to the global optimum with a $\mathcal{O}(\epsilon^{-2})$ convergence rate in tabular MDPs with tabular policies, also called direct policy parameterization. In later work, Xiao (2022) improve this result by a factor of ϵ, i.e., they establish a $\mathcal{O}(\epsilon^{-1})$ convergence rate for the projected PG in the tabular setting when the exact full gradient is available. At the moment, we could not adapt our general ABC structure to analyze and derive a sample complexity guarantee for the projected PG. The same convergence rate $\mathcal{O}(\epsilon^{-1})$ is developed by Zhang et al. (2020a) by leveraging the hidden convex structure of the cumulative reward and consequently showing that all local optima (i.e., stationary points) are in fact global optima under certain bijection assumptions based on the occupancy measure space (Assumption 1 in Zhang et al. (2020a)). Notice that the assumptions proposed by Zhang et al. (2020a) are satisfied in the specific case of the tabular setting. We do not cover this specific assumption in our current analysis.

The global optimum convergence analysis with exact PG is also investigated in the case of softmax tabular policy with or without regularization. Agarwal et al. (2021) first provide an asymptotic convergence for the softmax tabular without regularization and a $\mathcal{O}(\epsilon^{-2})$ convergence rate for the softmax tabular with log barrier regularization. Even though the gradient domination property ((PL) or (14)) is not globally satisfied for the softmax tabular, Mei et al. (2020) prove that it is available by following the path of the iterations with the exact full gradient updates. Such a property is called the non-uniform Lojasiewicz inequality. Consequently, Mei et al. (2020) show a $\mathcal{O}(\epsilon^{-1})$ convergence rate for the softmax tabular without regularization by the weak gradient domination condition and a linear convergence rate for the softmax tabular with entropy regularization by the gradient domination condition. Finally, Li et al. (2021) recently showed that the result of Mei et al. (2020) for softmax tabular policies may actually contain a term that is exponential in the discount factor, thus showing that exact PG may take an exponential time to converge.

Our Contributions. We provide a general sample complexity analysis which, when instantiated using specific settings given in the literature, recovers the same or even slightly improved convergence rates. Indeed, from Corollary E.5 we recover the $\mathcal{O}(\epsilon^{-2})$ convergence rate of Agarwal et al. (2021) for the softmax tabular with log barrier regularization and improve the rate by a factor of $1-\gamma$ through a better analysis of the smoothness constant. By leveraging the (relaxed weak) gradient domination properties which hold under the path of the iterations (Mei et al., 2020), we recover their results. That is, we recover the $\mathcal{O}(\epsilon^{-1})$ convergence rate for the softmax tabular without regularization in Theorem C.1 and the linear convergence rate for the softmax tabular with entropy regularization in Theorem H.2.

Sample complexity for FOSP convergence. The convergence rates derived for exact PG are representative of the behavior of the algorithm but do not take into account the additional errors due to the stochastic nature of the actual algorithm used in practice. In this paper we mostly focus on the sample complexity of the stochastic vanilla PG for FOSP convergence. The well known sample complexity for REINFORCE is $\widetilde{\mathcal{O}}(\epsilon^{-4})$ s.t. $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\widehat{\nabla}_m J(\theta_t)\right\|^2\right] \leq \epsilon^2$ after T iterations. However, as Papini (2020) mentioned, *"formal proofs of this result are surprisingly hard to find both in the policy optimization and in the nonconvex optimization literature."* Papini (2020) give a proof of the result under the expected Lipschitz and smooth policy assumption (E-LS) in Theorem 7.1. When an estimate of the Q-function is available, Zhang et al. (2020b) also establish the same dependency on ϵ for the sample complexity of FOSP convergence for the policy gradient theorem (Sutton et al., 2000) under more restrictive Lipschitz and smooth policy assumption (LS). By adding an additional uniform ergodicity assumption (Mitrophanov, 2005), Xiong et al. (2021) improve the sample complexity of (Zhang et al., 2020b) by some factors of $1-\gamma$ but still has the same dependency on ϵ.

Our Contributions. We establish the sample complexity analysis for the vanilla PG – REINFORCE (4) and GPOMDP (6). We improve the results of Papini (2020); Zhang et al. (2020b); Xiong et al. (2021) by using weaker assumptions and allowing much wider range of hyper parameters (the batch size m and the constant step size η) to achieve the optimal sample complexity. Overall, for both the exact and stochastic PG, our general sample complexity analysis recovers the state-of-the-art dependency on ϵ under the ABC assumption.

Sample complexity for global optimum convergence. We refer to sample complexity of global optimum convergence as an analysis that guarantees that $J^* - \mathbb{E}\left[J(\theta_T)\right] \leq \epsilon$ after T iterations. To the best of our knowledge, there is no existing analysis that considers this type of convergence result for the stochastic vanilla PG. As for variance-reduced PG, by using Assumption 1 in Zhang et al. (2020a) about occupancy distribution, Zhang et al. (2021b) establish a $\widetilde{\mathcal{O}}(\epsilon^{-2})$ sample complexity to achieve the global optimum.

Our Contributions. Under the ABC assumption, the smoothness and an additional gradient domination type assumptions (14) and (PL), we establish the faster sample complexity analysis for the global optimum convergence in Section 3.2 and Section H. More precisely, when the relaxed weak gradient domination assumption (14) is available, we establish $\widetilde{\mathcal{O}}(\epsilon^{-3})$ sample complexity in Theorem C.1. We also show that one wide family of policies, the Fisher-non-degenerate parametrized policies, satisfy this relaxed weak gradient domination assumption. When the gradient domination assumption (PL) is available, we establish $\widetilde{\mathcal{O}}(\epsilon^{-1})$ sample complexity for the global optimum in Theorem H.2. It remains an open question whether softmax or softmax with entropy still satisfy the (weak) gradient domination type of assumptions for the stochastic PG updates based on the exact PG analysis of Mei et al. (2020).

Sample complexity for the average regret convergence. We refer to the sample complexity for average regret as an analysis that guarantees that $J^* - \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[J(\theta_t)\right] \leq \epsilon$. Zhang et al. (2021a) show that with sample complexity $\widetilde{\mathcal{O}}(\epsilon^{-6})$, PG methods can converge to the average regret optimum by using as little as a single sampled trajectory per iteration (i.e., mini-batch size $m=1$) for softmax with log barrier regularization. However, their setting does not use "vanilla" PG but a modified version with re-projection meant to guarantee a sufficient level of policy randomization. Liu et al. (2020) obtain faster sample complexity $\widetilde{\mathcal{O}}(\epsilon^{-4})$ by assuming in addition a Fisher-non-degenerate parameterization, i.e. the Fisher information matrix is strictly lower bounded (Assumption 2.1 in Liu et al. (2020)), and the compatible function approximation assumption (see Assumption 4.4 in Liu et al. (2020) on function approximation error). Notice that the softmax with log barrier regularization does not satisfy all these assumptions and they require large batch sizes per iteration. We have not investigated this setting in our paper.

Our Contributions. We recover the sample complexity for the average regret convergence $\widetilde{\mathcal{O}}(\epsilon^{-6})$ of Zhang et al. (2021a) in the softmax with log barrier regularization with the vanilla PG setting . Compared to their results, we show that the extra phased learning step is unnecessary and the step size can be constant instead of using a decreasing step size. We also provide a wider range of parameter choices for the batch size and the step size with the same sample complexity.

Table 2: E-LS constants G, F (Assumption 4.1), smoothness constant L and Lipschitzness constant Γ for Gaussian and (regularized) Softmax tabular policies, where 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.

	Gaussian*	**Softmax**	**Softmax with log barrier**
G^2	$\frac{\varphi^2}{\sigma^2}$	$1 - \frac{1}{\lvert\mathcal{A}\rvert}$	✗**
F	$\frac{\varphi^2}{\sigma^2}$	1	✗
L	$\frac{2\mathcal{R}_{\max}\varphi^2}{(1-\gamma)^2\sigma^2}$	$\frac{\mathcal{R}_{\max}}{(1-\gamma)^2}\left(2 - \frac{1}{\lvert\mathcal{A}\rvert}\right)$	$\frac{\mathcal{R}_{\max}}{(1-\gamma)^2}\left(2 - \frac{1}{\lvert\mathcal{A}\rvert}\right) + \frac{\lambda}{\lvert\mathcal{S}\rvert}$
Γ	$\frac{\mathcal{R}_{\max}\varphi}{(1-\gamma)^{3/2}\sigma}$	$\frac{\mathcal{R}_{\max}}{(1-\gamma)^{3/2}}\sqrt{1 - \frac{1}{\lvert\mathcal{A}\rvert}}$	$\sqrt{2\left(1 - \frac{1}{\lvert\mathcal{A}\rvert}\right)\left(\frac{\mathcal{R}_{\max}^2}{(1-\gamma)^3} + \frac{\lambda^2}{\lvert\mathcal{S}\rvert}\right)}$

*The (E-LS) constants G^2 and F are provided in Lemma 15 in Papini et al. (2019).
**When there is a "✗", it means this is not applicable directly in such setting.

A.3 Better analysis of the problem dependent constants

Throughout the paper, we also provided tighter bounds on the smoothness constants, Lipschitzness constants, and the variance of the gradient estimators under Assumption (E-LS). Notice that the smoothness and Lipschitz constants we consider here are properties of the expected return $J(\cdot)$ in (2) or the regularized expected return $L_\lambda(\cdot)$ in (28). They depend only on the assumptions and are independent to the specific PG algorithm. For this reason, below we compare our bounds with work that studies variants of PG other than vanilla PG, where the bounds on the smoothness and Lipschitz constants are also needed. On the other hand, for the variance of the gradient estimators, we only consider the vanilla gradient estimators REINFORCE (4) and GPOMDP (6) with batch size m. A resume of the improved problem dependent constants – smoothness and Lipschitzness constants, is provided in Table 2.

Smoothness constant. The smoothness constant (19) provided in Lemma 4.4 is novel. It is tighter as compared to Lemma 6 in Papini et al. (2019) under Assumption (E-LS) and Proposition 4.2 (2) in Xu et al. (2020a) under more restrictive assumptions (LS). Compared to existing bounds, our result shows that when γ is close to 1, the smoothness constant (19) depends on $(1-\gamma)^{-2}$ instead of $(1-\gamma)^{-3}$ as derived in Papini et al. (2019) and Xu et al. (2020a). Consequently, the smoothness constant for softmax derived in Lemma E.1 and E.3 are also tighter than the one derived in Lemma 7 in Mei et al. (2020) and Lemma D.2 in Agarwal et al. (2021), which both have the dependency of $(1-\gamma)^{-3}$. Finally, compared to the smoothness constant in Shen et al. (2019) and Xu et al. (2020b), our result is independent to the horizon H .

Recent works, such as Proposition 1 in Huang et al. (2020), Lemma B.1 in Liu et al. (2020) and equation (17) in Yuan et al. (2020), have the dependency of $(1-\gamma)^{-2}$ for the smoothness constant under assumptions (LS). However, this is due to a recurring mistake in a crucial step in bounding the Hessian.[6]

Lipschitzness constant. The improved Lipschitzness constant under Assumption (E-LS) is provided in Lemma D.1 (iii) in Section D.5. Compared to the existing bounds, our result shows that when γ is close to 1, the Lipschitzness constant Γ depends on $(1-\gamma)^{-3/2}$ instead of $(1-\gamma)^{-2}$ derived in the proof of Lemma 6 in Papini et al. (2019) under the same Assumption (E-LS).

Upper bound of the variance of the gradient estimators. As for the result in Lemma 4.2, our bounds (18) on the variance of the gradient estimators REINFORCE and GPOMDP are slightly tighter than the one in Lemma 17 and 18 in Papini et al. (2019), see more details in Section D.1. Shen et al. (2019) and Pham et al. (2020) also showed that the variance of the vanilla gradient estimator with batch size $m = 1$ is bounded under more restrictive assumptions (LS). While their bounded variance depends on $(1-\gamma)^{-4}$ and they only consider the

[6] In a previous version of the proof in Sect. C, Xu et al. (2020a) rely on the identity $\nabla^2_\theta J(\theta) = \mathbb{E}_\tau\left[\nabla_\theta g(\tau \mid \theta)\right]$, which is incorrect since the operators ∇_θ and $\mathbb{E}[\cdot]$ are not commutative in this case as the density $p(\cdot \mid \theta)$ of $\mathbb{E}[\cdot]$ depends on θ as well. This error is recently fixed by Xu et al. (2020a) on https://arxiv.org/pdf/1909.08610.pdf in their original paper.

GPOMDP gradient estimator, ours (18) depends on $(1-\gamma)^{-3}$ for GPOMDP or $\frac{H}{(1-\gamma)^2}$ for REINFORCE which is tighter in both cases.

A.4 Future work

The main focus of this paper was the theoretical analysis of vanilla variants of the PG method. The results we have obtained open up several experimental questions related to parameter settings for PG. We leave such questions as an important future work to further support our theoretical findings.

One natural open question is whether the ABC assumption and the associated analysis can be extended to the projected PG. If the answer is positive, this might improve the sample complexity analysis of the direct policy parameterization setting in the stochastic case. Indeed, knowing that the direct policy parameterization satisfies a variant of (14) condition (Agarwal et al., 2021; Xiao, 2022) under the proximal framework, if the ABC assumption and the associated analysis can be extended, from Theorem C.1 which also uses the (14) condition, then it might be possible to establish the $\widetilde{\mathcal{O}}(\epsilon^{-3})$ sample complexity for the global optimum convergence for the direct policy parameterization and allow for a wider range of hyperparameter choices.

Similarly, we wonder if the ABC assumption and the associated analysis can be extended to the LQR setting. The challenge here will be the smoothness assumption and whether the ABC assumption is satisfied by the LQR when doing the stochastic PG updates. Indeed, the LQR only has an "almost" smoothness property (Fazel et al., 2018). One needs to investigate how this will affect the current ABC analysis by extending the smoothness property to the "almost" smoothness property.

Recently, variance reduced methods used to decrease the variance of SGD, such as SVRG (Johnson and Zhang, 2013), SARAH (Nguyen et al., 2017), SPIDER (Fang et al., 2018), STORM (Cutkosky and Orabona, 2019) and more (Tran-Dinh et al., 2021), have been applied to PG methods, such as SVRPG (Papini et al., 2018), SRVR-PG (Xu et al., 2020a), STORM-PG (Yuan et al., 2020), ProxHSPGA (Pham et al., 2020), VRMPO (Yang et al., 2021) and VR-BGPO (Huang et al., 2022). Leveraging these variance reduction techniques has led to an overall improved sample complexity of reaching a first-order stationary point (FOSP). However, all these works require either the exact full gradient updates or large batch sizes per iteration. It is interesting to understand whether the ABC assumption analysis can be applied to these algorithms and possibly allow for a wider range of hyperparameter choices, including the batch size. Furthermore, when the gradient domination type assumptions are available, it will be interesting to see if we can obtain faster sample complexity as we did for the vanilla PG.

Another interesting venue of investigation might be whether the ABC assumption analysis can be extended to the sample complexity analysis of (natural) actor-critic (Yang et al., 2019; Kumar et al., 2021; Xu et al., 2020c) or natural policy gradient algorithms (Agarwal et al., 2021; Liu et al., 2020; Wang et al., 2020).

Finally we believe that the generality of Theorem 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, notably thinking about settings such that the constant A in Assumption (ABC) is *non-zero*, using additional assumptions such as the bijection assumptions based on the occupancy measure space (Zhang et al., 2020a) to not only get improved sample complexity for the global optimum convergence, but also allow a wider range of hyperparameter choices for the convergence.

B Auxiliary Lemmas

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

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

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

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

We have

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

Subtracting of the above two equations gives

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

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

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

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

Proof. Let

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

We have

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

Thus, the subtraction of the above two equations gives

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

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

Lemma B.3. Under Assumption 4.1, for all non negative integer t and any state-action pair $(s_t, a_t) \in \mathcal{S}\times\mathcal{A}$ at time t of a trajectory $\tau \sim p(\cdot \mid \theta)$ sampled under the parametrized policy π_θ, we have that

$$\mathbb{E}_{\tau\sim p(\cdot|\theta)}\left[\|\nabla_\theta \log \pi_\theta(a_t \mid s_t)\|^2\right] \;\leq\; G^2, \tag{32}$$

$$\mathbb{E}_{\tau\sim p(\cdot|\theta)}\left[\left\|\nabla^2_\theta \log \pi_\theta(a_t \mid s_t)\right\|\right] \;\leq\; F. \tag{33}$$

Proof. For $t > 0$ and $(s_t, a_t) \in \mathcal{S} \times \mathcal{A}$, we have

$$\mathbb{E}_\tau \left[\|\nabla_\theta \log \pi_\theta(a_t \mid s_t)\|^2 \right] = \mathbb{E}_{s_t} \left[\mathbb{E}_{a_t \sim \pi_\theta(\cdot|s_t)} \left[\|\nabla_\theta \log \pi_\theta(a_t \mid s_t)\|^2 \,|s_t \right] \right] \overset{(15)}{\leq} G^2,$$

where the first equality is obtained by the Markov property.

Similarly, we have

$$\mathbb{E}_\tau \left[\left\|\nabla^2_\theta \log \pi_\theta(a_t \mid s_t)\right\| \right] = \mathbb{E}_{s_t} \left[\mathbb{E}_{a_t \sim \pi_\theta(\cdot|s_t)} \left[\left\|\nabla^2_\theta \log \pi_\theta(a_t \mid s_t)\right\| \,|s_t \right] \right] \overset{(16)}{\leq} F.$$

□

Lemma B.4. For all non negative integers $0 \leq h < h'$, and any state-action pairs $(s_h, a_h), (s_{h'}, a_{h'}) \in \mathcal{S} \times \mathcal{A}$ at time h and h' respectively of the same trajectory $\tau \sim p(\cdot \mid \theta)$ sampled under the parametrized policy π_θ, we have

$$\mathbb{E}_\tau \left[\left(\nabla_\theta \log \pi_\theta(a_h \mid s_h)\right)^\top \nabla_\theta \log \pi_\theta(a_{h'} \mid s_{h'}) \right] = 0. \tag{34}$$

Proof. For $0 \leq h < h'$, we have

$$\begin{aligned}
&\mathbb{E}_\tau \left[\left(\nabla_\theta \log \pi_\theta(a_h \mid s_h)\right)^\top \nabla_\theta \log \pi_\theta(a_{h'} \mid s_{h'}) \right] \\
&= \mathbb{E}_{a_h, s_h, s_{h'}} \left[\mathbb{E}_{a_{h'}} \left[\left(\nabla_\theta \log \pi_\theta(a_h \mid s_h)\right)^\top \nabla_\theta \log \pi_\theta(a_{h'} \mid s_{h'}) \middle| s_h, a_h, s_{h'} \right] \right] \\
&= \mathbb{E}_{a_h, s_h, s_{h'}} \left[\left(\nabla_\theta \log \pi_\theta(a_h \mid s_h)\right)^\top \mathbb{E}_{a_{h'}} \left[\nabla_\theta \log \pi_\theta(a_{h'} \mid s_{h'}) \middle| s_h, a_h, s_{h'} \right] \right] \\
&= \mathbb{E}_{a_h, s_h, s_{h'}} \left[\left(\nabla_\theta \log \pi_\theta(a_h \mid s_h)\right)^\top \int_{a_{h'}} \pi_\theta(a_{h'} \mid s_{h'}) \nabla_\theta \log \pi_\theta(a_{h'} \mid s_{h'}) da_{h'} \right] \\
&= \mathbb{E}_{a_h, s_h, s_{h'}} \left[\left(\nabla_\theta \log \pi_\theta(a_h \mid s_h)\right)^\top \int_{a_{h'}} \nabla_\theta \pi_\theta(a_{h'} \mid s_{h'}) da_{h'} \right] \\
&= \mathbb{E}_{a_h, s_h, s_{h'}} \left[\left(\nabla_\theta \log \pi_\theta(a_h \mid s_h)\right)^\top \nabla_\theta \underbrace{\int_{a_{h'}} \pi_\theta(a_{h'} \mid s_{h'}) da_{h'}}_{=1} \right] = 0,
\end{aligned}$$

where the first and second equality is obtained by the Markov property. □

Lemma B.5. For all non negative integers $0 \leq t$, and any state-action pairs $(s_h, a_h) \in \mathcal{S} \times \mathcal{A}$ at time $0 \leq h \leq t$ of the same trajectory $\tau \sim p(\cdot \mid \theta)$ sampled under the parametrized policy π_θ, we have

$$\mathbb{E}_\tau \left[\left\| \sum_{h=0}^{t} \nabla_\theta \log \pi_\theta(a_h \mid s_h) \right\|^2 \right] = \sum_{h=0}^{t} \mathbb{E}_\tau \left[\|\log \pi_\theta(a_h \mid s_h)\|^2 \right]. \tag{35}$$

Proof. For $0 \leq t$, we have

$$\begin{aligned}
\mathbb{E}_\tau \left[\left\| \sum_{h=0}^{t} \nabla_\theta \log \pi_\theta(a_h \mid s_h) \right\|^2 \right] &= \sum_{h=0}^{t} \mathbb{E}_\tau \left[\|\log \pi_\theta(a_h \mid s_h)\|^2 \right] \\
&\quad + 2 \sum_{h=0}^{t-1} \sum_{h'=h+1}^{t} \mathbb{E}_\tau \left[\left(\nabla_\theta \log \pi_\theta(a_h \mid \theta_h)\right)^\top \nabla_\theta \log \pi_\theta(a_{h'} \mid \theta_{h'}) \right] \\
&\overset{(34)}{=} \sum_{h=0}^{t} \mathbb{E}_\tau \left[\|\log \pi_\theta(a_h \mid s_h)\|^2 \right].
\end{aligned}$$

□

C Proof of Section 3

C.1 Proof of Theorem 3.4

Proof. We start with L-smoothness of J from Assumption 3.1, which implies

$$\begin{aligned} J(\theta_{t+1}) &\geq J(\theta_t) + \langle \nabla J(\theta_t), \theta_{t+1} - \theta_t \rangle - \frac{L}{2} \|\theta_{t+1} - \theta_t\|^2 \\ &= J(\theta_t) + \eta \left\langle \nabla J(\theta_t), \widehat{\nabla}_m J(\theta_t) \right\rangle - \frac{L\eta^2}{2} \left\| \widehat{\nabla}_m J(\theta_t) \right\|^2. \end{aligned} \tag{36}$$

Taking expectations conditioned on θ_t, we get

$$\begin{aligned} \mathbb{E}_t \left[J(\theta_{t+1}) \right] &\geq J(\theta_t) + \eta \left\langle \nabla J(\theta_t), \nabla J_H(\theta_t) \right\rangle - \frac{L\eta^2}{2} \mathbb{E}_t \left[\left\| \widehat{\nabla}_m J(\theta_t) \right\|^2 \right] \\ &\overset{\text{(ABC)}}{\geq} J(\theta_t) + \eta \left\langle \nabla J_H(\theta_t) + (\nabla J(\theta_t) - \nabla J_H(\theta_t)), \nabla J_H(\theta_t) \right\rangle \\ &\quad - \frac{L\eta^2}{2} \left(2A(J^* - J(\theta_t)) + B \|\nabla J_H(\theta_t)\|^2 + C \right) \\ &= J(\theta_t) + \eta \left(1 - \frac{LB\eta}{2} \right) \|\nabla J_H(\theta_t)\|^2 - L\eta^2 A(J^* - J(\theta_t)) \\ &\quad - \frac{LC\eta^2}{2} + \eta \left\langle \nabla J_H(\theta_t), \nabla J(\theta_t) - \nabla J_H(\theta_t) \right\rangle \\ &\overset{(9)}{\geq} J(\theta_t) + \eta \left(1 - \frac{LB\eta}{2} \right) \|\nabla J_H(\theta_t)\|^2 - L\eta^2 A(J^* - J(\theta_t)) - \frac{LC\eta^2}{2} - \eta D \gamma^H. \end{aligned} \tag{37}$$

Subtracting J^* from both sides gives

$$-\left(J^* - \mathbb{E}_t \left[J(\theta_{t+1}) \right]\right) \geq -(1 + L\eta^2 A)(J^* - J(\theta_t)) + \eta \left(1 - \frac{LB\eta}{2} \right) \|\nabla J_H(\theta_t)\|^2 - \frac{LC\eta^2}{2} - \eta D \gamma^H. \tag{38}$$

Taking the total expectation and rearranging, we get

$$\mathbb{E} \left[J^* - J(\theta_{t+1}) \right] + \eta \left(1 - \frac{LB\eta}{2} \right) \mathbb{E} \left[\|\nabla J_H(\theta_t)\|^2 \right] \leq (1 + L\eta^2 A) \mathbb{E} \left[J^* - J(\theta_t) \right] + \frac{LC\eta^2}{2} + \eta D \gamma^H. \tag{39}$$

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

$$\eta \left(1 - \frac{LB\eta}{2} \right) r_t \leq (1 + L\eta^2 A)\delta_t - \delta_{t+1} + \frac{LC\eta^2}{2} + \eta D \gamma^H. \tag{40}$$

We now introduce a sequence of weights $w_{-1}, w_0, w_1, \cdots, w_{T-1}$ based on a technique developed by Stich (2019). Let $w_{-1} > 0$. Define $w_t \overset{\text{def}}{=} \frac{w_{t-1}}{1 + L\eta^2 A}$ for all $t \geq 0$. Notice that if $A = 0$, we have $w_t = w_{t-1} = \cdots = w_{-1}$. Multiplying (40) by w_t/η,

$$\begin{aligned} \left(1 - \frac{LB\eta}{2} \right) w_t r_t &\leq \frac{w_t(1 + L\eta^2 A)}{\eta} \delta_t - \frac{w_t}{\eta} \delta_{t+1} + \frac{LC\eta}{2} w_t + D\gamma^H w_t \\ &= \frac{w_{t-1}}{\eta} \delta_t - \frac{w_t}{\eta} \delta_{t+1} + \left(\frac{LC\eta}{2} + D\gamma^H \right) w_t. \end{aligned} \tag{41}$$

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

$$\begin{aligned} \left(1 - \frac{LB\eta}{2} \right) \sum_{t=0}^{T-1} w_t r_t &\leq \frac{w_{-1}}{\eta} \delta_0 - \frac{w_{T-1}}{\eta} \delta_T + \left(\frac{LC\eta}{2} + D\gamma^H \right) \sum_{t=0}^{T-1} w_t \\ &\leq \frac{w_{-1}}{\eta} \delta_0 + \left(\frac{LC\eta}{2} + D\gamma^H \right) \sum_{t=0}^{T-1} w_t. \end{aligned} \tag{42}$$

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

$$\left(1-\frac{LB\eta}{2}\right)\min_{0\le t\le T-1} r_t \;\le\; \frac{1}{W_T}\cdot\left(1-\frac{LB\eta}{2}\right)\sum_{t=0}^{T-1} w_t r_t \;\le\; \frac{w_{-1}}{W_T}\frac{\delta_0}{\eta}+\frac{LC\eta}{2}+D\gamma^H. \tag{43}$$

Note that,

$$W_T \;=\; \sum_{t=0}^{T-1} w_t \ge \sum_{t=0}^{T-1}\min_{0\le i\le T-1} w_i \;=\; T w_{T-1} \;=\; \frac{T w_{-1}}{(1+L\eta^2 A)^T}. \tag{44}$$

Using this in (43),

$$\left(1-\frac{LB\eta}{2}\right)\min_{0\le t\le T-1} r_t \;\le\; \frac{(1+L\eta^2A)^T}{\eta T}\delta_0+\frac{LC\eta}{2}+D\gamma^H. \tag{45}$$

However, we have

$$\begin{aligned}
\mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right] &= \mathbb{E}\left[\|\nabla J(\theta_t)-\nabla J_H(\theta_t)+\nabla J_H(\theta_t)\|^2\right]\\
&= \mathbb{E}\left[\|\nabla J_H(\theta_t)\|^2\right]+2\mathbb{E}\left[\langle \nabla J_H(\theta_t), \nabla J(\theta_t)-\nabla J_H(\theta_t)\rangle\right]+\mathbb{E}\left[\|\nabla J(\theta_t)-\nabla J_H(\theta_t)\|^2\right]\\
&\overset{(9)+(10)}{\le} \mathbb{E}\left[\|\nabla J_H(\theta_t)\|^2\right]+2D\gamma^H+D'^2\gamma^{2H}.
\end{aligned} \tag{46}$$

Substituting r_t in (45) by $\mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right]$ and using (46), we get

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

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

If $A=0$, we know that $\{w_t\}_{t\ge -1}$ is a constant sequence. In this case, $W_T = Tw_{-1}$. Dividing both sides of (42) by W_T, we have,

$$\left(1-\frac{LB\eta}{2}\right)\frac{1}{T}\sum_{t=0}^{T-1} r_t \;\le\; \frac{\delta_0}{\eta T}+\frac{LC\eta}{2}+D\gamma^H. \tag{47}$$

Similarly, substituting r_t in (47) by $\mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right]$ and using (46), we get

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

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

C.2 Proof of Corollary 3.5

Proof. Given $\epsilon>0$, from Corollary 1 in Khaled and Richtárik (2020), we know that if $\eta = \min\left\{\frac{1}{\sqrt{LAT}}, \frac{1}{LB}, \frac{\epsilon}{2LC}\right\}$ and the number of iterations T satisfies

$$T \;\ge\; \frac{12\delta_0 L}{\epsilon^2}\max\left\{B, \frac{12\delta_0 A}{\epsilon^2}, \frac{2C}{\epsilon^2}\right\},$$

we have

$$\frac{2\delta_0(1+L\eta^2A)^T}{\eta T(2-LB\eta)}+\frac{LC\eta}{2-LB\eta} \;\le\; \epsilon^2.$$

It remains to show

$$\left(\frac{2D(3-LB\eta)}{2-LB\eta}+D'^2\gamma^H\right)\gamma^H\leq\epsilon^2.$$

Besides, our choice of the step size $\eta\leq\frac{1}{LB}$ implies that $\frac{1}{2-LB\eta}\leq 1$, thus

$$\left(\frac{2D(3-LB\eta)}{2-LB\eta}+D'^2\gamma^H\right)\gamma^H\leq\left(6D+D'^2\gamma^H\right)\gamma^H.$$

Finally, it suffices to choose H such that

$$\gamma^H\;\leq\;\epsilon^2\quad\Longleftrightarrow\quad H\;\geq\;\frac{2\log\epsilon^{-1}}{\log\gamma^{-1}}\;=\;\mathcal{O}(\log\epsilon^{-1}),$$

to guarantee that $\min_{0\leq t\leq T-1}\mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right]=\mathcal{O}(\epsilon^{-2})$, which concludes the proof. □

Remark. When γ is close to 1, the horizon has the following property.

$$H\;=\;\frac{2\log\epsilon^{-1}}{\log\gamma^{-1}}\;=\;\mathcal{O}\left(\frac{\log\epsilon^{-1}}{1-\gamma}\right).$$

C.3 Global optimum convergence under the relaxed weak gradient domination assumption

In this section, we present the new global optimum convergence theory under the relaxed weak gradient domination assumption (14).

Theorem C.1. Suppose that Asm. 3.1, 3.2, 3.3 and 3.6 hold. Given $\epsilon>0$, define δ s.t. if $\epsilon'=0$, set $\delta=\epsilon$, if $\epsilon'>0$, set $\delta=\epsilon'$. Suppose that PG defined in (7) is run for $T>0$ iterations with step size $(\eta_t)_t$ chosen as

$$\eta_t=\begin{cases}\frac{1}{b} & \text{if } T\leq\frac{b}{\mu\delta} \text{ or } t\leq t_0\\ \frac{2}{2b+\mu\delta(t-t_0)} & \text{if } T\geq\frac{b}{\mu\delta} \text{ and } t>t_0\end{cases}\tag{48}$$

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

$$J^*-\mathbb{E}\left[J(\theta_T)\right]\leq 16\exp\left(-\frac{\mu\delta(T-1)}{2b}\right)(J^*-J(\theta_0))+\frac{12LC}{\mu^2\delta^2T}+\frac{26D\gamma^H}{\mu\delta}+\frac{12(\epsilon')^2(2b-LB)}{\mu^2\delta^2T}+\frac{2\epsilon'}{\mu},\tag{49}$$

otherwise, we have

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

Remark. Similar to the exact full gradient update in Thm. 3.4, notice that for the exact full gradient update, we have Asm. 3.2 and 3.3 hold with $A=C=D=0$ and $B=1$. Thus under the smoothness and the weak gradient domination assumption (i.e., $\epsilon'=0$), we have

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

With $T=\frac{1}{\epsilon}\log\left(\frac{1}{\epsilon}\right)$, we have $J^*-\mathbb{E}\left[J(\theta_T)\right]\leq\epsilon$. Thus we establish $\widetilde{\mathcal{O}}(\epsilon^{-1})$ convergence rate for the number of iterations to the global optimal. We recover the same rate for the softmax tabular policy in Theorem 4 in Mei et al. (2020) where the smoothness assumption holds and the weak gradient domination condition (14) holds on the path of the iterates in the exact case.

Proof. From (14), we obtain that

$$\begin{aligned}&(\epsilon')^2+\|\nabla J_H(\theta)\|^2\geq\frac{\left(\epsilon'+\|\nabla J_H(\theta)\|\right)^2}{2}\geq 2\mu(J^*-J(\theta))^2\\ \Longrightarrow\;&\|\nabla J_H(\theta)\|^2\geq 2\mu(J^*-J(\theta))^2-(\epsilon')^2.\end{aligned}\tag{50}$$

Let $t \in \{0, 1, \cdots, T-1\}$. Using the L-smoothness of J from Assumption 3.1,

$$\begin{aligned} J^* - J(\theta_{t+1}) &\leq J^* - J(\theta_t) - \langle \nabla J(\theta_t), \theta_{t+1} - \theta_t \rangle + \frac{L}{2} \|\theta_{t+1} - \theta_t\|^2 \\ &= J^* - J(\theta_t) - \eta_t \left\langle \nabla J(\theta_t), \widehat{\nabla}_m J(\theta_t) \right\rangle + \frac{L\eta_t^2}{2} \left\| \widehat{\nabla}_m J(\theta_t) \right\|^2. \end{aligned} \tag{51}$$

Taking expectation conditioned on θ_t and using Assumption 3.3 and 3.6,

$$\begin{aligned} \mathbb{E}_t \left[J^* - J(\theta_{t+1}) \right] &\leq J^* - J(\theta_t) - \eta_t \left\langle \nabla J(\theta_t), \nabla J_H(\theta_t) \right\rangle + \frac{L\eta_t^2}{2} \mathbb{E}_t \left[\left\| \widehat{\nabla}_m J(\theta_t) \right\|^2 \right] \\ &\overset{(ABC)}{\leq} J^* - J(\theta_t) - \eta_t \left\langle \nabla J_H(\theta_t) + (\nabla J(\theta_t) - \nabla J_H(\theta_t)), \nabla J_H(\theta_t) \right\rangle + \\ &\quad + \frac{L\eta_t^2}{2} \left(2A(J^* - J(\theta_t)) + B \|\nabla J_H(\theta_t)\|^2 + C \right) \\ &= (1 + L\eta_t^2 A)(J^* - J(\theta_t)) - \eta_t \left(1 - \frac{LB\eta_t}{2} \right) \|\nabla J_H(\theta_t)\|^2 + \frac{L\eta_t^2 C}{2} \\ &\quad - \eta_t \left\langle \nabla J(\theta_t) - \nabla J_H(\theta_t), \nabla J_H(\theta_t) \right\rangle \\ &\overset{(50)}{\leq} \left(1 + L\eta_t^2 A\right) (J^* - J(\theta_t)) - \mu\eta_t \left(2 - LB\eta_t\right) (J^* - J(\theta_t))^2 + \eta_t \left(1 - \frac{LB\eta_t}{2} \right) (\epsilon')^2 \\ &\quad + \frac{L\eta_t^2 C}{2} - \eta_t \left\langle \nabla J(\theta_t) - \nabla J_H(\theta_t), \nabla J_H(\theta_t) \right\rangle \\ &\overset{(9)}{\leq} \left(1 + L\eta_t^2 A\right) (J^* - J(\theta_t)) - \mu\eta_t \left(2 - LB\eta_t\right) (J^* - J(\theta_t))^2 + \eta_t \left(1 - \frac{LB\eta_t}{2} \right) (\epsilon')^2 \\ &\quad + \frac{L\eta_t^2 C}{2} + \eta_t D \gamma^H \\ &\leq \left(1 + L\eta_t^2 A\right) (J^* - J(\theta_t)) - \frac{3\mu}{2} \eta_t (J^* - J(\theta_t))^2 + \eta_t \left(1 - \frac{LB\eta_t}{2} \right) (\epsilon')^2 \\ &\quad + \frac{L\eta_t^2 C}{2} + \eta_t D \gamma^H, \end{aligned} \tag{52}$$

where the last line is obtained by the choice of the step size $\eta_t \leq \frac{1}{b}$ with $b \geq 2LB$.

Taking total expectation and letting $r_t \overset{\text{def}}{=} \mathbb{E}\left[J^* - J(\theta_t)\right]$ on (52), we have

$$r_{t+1} \leq r_t + LA\eta_t^2 r_t - \frac{3\mu}{2} \eta_t r_t^2 + \eta_t \left(1 - \frac{LB\eta_t}{2} \right) (\epsilon')^2 + \frac{LC}{2} \eta_t^2 + \eta_t D \gamma^H. \tag{53}$$

If there exists $t \in \{0, 1, \cdots, T-1\}$ such that $r_t < \delta$, then we are done. Alternatively if $r_t \geq \delta$ for all $t \in \{0, 1, \cdots, T-1\}$, from (53), we have

$$\begin{aligned} r_{t+1} &\leq r_t + LA\eta_t^2 r_t - \frac{3\mu\delta}{2} \eta_t r_t + \eta_t \left(1 - \frac{LB\eta_t}{2} \right) (\epsilon')^2 + \frac{LC}{2} \eta_t^2 + \eta_t D \gamma^H \\ &\leq (1 - \mu\delta\eta_t) r_t + \eta_t \left(1 - \frac{LB\eta_t}{2} \right) (\epsilon')^2 + \frac{LC}{2} \eta_t^2 + \eta_t D \gamma^H, \end{aligned} \tag{54}$$

where the last line is obtained by the choice of the step size $\eta_t \leq \frac{1}{b}$ with $b \geq \frac{2LA}{\mu\delta}$. Here $1 - \mu\delta\eta_t \geq 0$ as $\eta_t \leq \frac{1}{b}$ with $b \geq \mu\delta$. We notice that (54) is similar to (142). The rest of the proof is similar to the one of Theorem H.2.

If $T \leq \frac{b}{\mu\delta}$, $\eta_t = \frac{1}{b}$. From (54), we have

$$\begin{aligned} r_T &\leq \left(1-\frac{\mu\delta}{b}\right) r_{T-1} + \frac{LC}{2b^2} + \frac{D\gamma^H}{b} + \frac{2b-LB}{2b^2}(\epsilon')^2 \\ &\overset{(54)}{\leq} \left(1-\frac{\mu\delta}{b}\right)^T r_0 + \left(\frac{LC}{2b^2} + \frac{D\gamma^H}{b} + \frac{2b-LB}{2b^2}(\epsilon')^2\right) \sum_{i=0}^{T-1} \left(1-\frac{\mu\delta}{b}\right)^i \\ &\leq \exp\left(-\frac{\mu\delta T}{b}\right) r_0 + \frac{LC}{2\mu\delta b} + \frac{D\gamma^H}{\mu\delta} + \frac{2b-LB}{2\mu\delta b}(\epsilon')^2 \qquad (55) \\ &\overset{T\leq\frac{b}{\mu\delta}}{\leq} \exp\left(-\frac{\mu\delta T}{b}\right) r_0 + \frac{LC}{2\mu^2\delta^2 T} + \frac{D\gamma^H}{\mu\delta} + \frac{2b-LB}{2\mu^2\delta^2 T}(\epsilon')^2. \qquad (56) \end{aligned}$$

If $T \geq \frac{b}{\mu\delta}$, as $\eta_t = \frac{1}{b}$ when $t \leq t_0$, from (55), we have

$$\begin{aligned} r_{t_0} &\leq \exp\left(-\frac{\mu\delta t_0}{b}\right) r_0 + \frac{LC}{2\mu\delta b} + \frac{D\gamma^H}{\mu\delta} + \frac{2b-LB}{2\mu\delta b}(\epsilon')^2 \\ &\leq \exp\left(-\frac{\mu\delta(T-1)}{2b}\right) r_0 + \frac{LC}{2\mu\delta b} + \frac{D\gamma^H}{\mu\delta} + \frac{2b-LB}{2\mu\delta b}(\epsilon')^2, \qquad (57) \end{aligned}$$

where the last line is obtained by $t_0 = \left[\frac{T}{2}\right] \geq \frac{T-1}{2}$.

For $t > t_0$,

$$\eta_t = \frac{2}{\mu\delta\left(\frac{2b}{\mu\delta} + t - t_0\right)}.$$

From (54), we have

$$\begin{aligned} r_t &\leq \frac{\frac{2b}{\mu\delta} + t - t_0 - 2}{\frac{2b}{\mu\delta} + t - t_0} r_{t-1} + \frac{2LC}{\mu^2\delta^2\left(\frac{2b}{\mu\delta} + t - t_0\right)^2} + \frac{2D\gamma^H}{\mu\delta\left(\frac{2b}{\mu\delta} + t - t_0\right)} \\ &\quad + \frac{2(\epsilon')^2}{\mu\delta\left(\frac{2b}{\mu\delta} + t - t_0\right)}\left(1 - \frac{LB}{\mu\delta\left(\frac{2b}{\mu\delta} + t - t_0\right)}\right). \qquad (58) \end{aligned}$$

Multiplying both sides by $\left(\frac{2b}{\mu\delta} + t - t_0\right)^2$, we have

$$\begin{aligned} \left(\frac{2b}{\mu\delta} + t - t_0\right)^2 r_t &\leq \left(\frac{2b}{\mu\delta} + t - t_0\right)\left(\frac{2b}{\mu\delta} + t - t_0 - 2\right) r_{t-1} + \frac{2LC}{\mu^2\delta^2} + \frac{2D\gamma^H}{\mu\delta}\left(\frac{2b}{\mu\delta} + t - t_0\right) \\ &\quad + \frac{2(\epsilon')^2}{\mu\delta}\left(\frac{2b-LB}{\mu\delta} + t - t_0\right) \\ &\leq \left(\frac{2b}{\mu\delta} + t - t_0 - 1\right)^2 r_{t-1} + \frac{2LC}{\mu^2\delta^2} + \frac{2D\gamma^H}{\mu\delta}\left(\frac{2b}{\mu\delta} + t - t_0\right) \\ &\quad + \frac{2(\epsilon')^2}{\mu\delta}\left(\frac{2b-LB}{\mu\delta} + t - t_0\right). \qquad (59) \end{aligned}$$

Let $w_t \overset{\text{def}}{=} \left(\frac{2b}{\mu\epsilon} + t - t_0\right)^2$. We have

$$w_t r_t \leq w_{t-1} r_{t-1} + \frac{2LC}{\mu^2\delta^2} + \frac{2D\gamma^H}{\mu\delta}\left(\frac{2b}{\mu\delta} + t - t_0\right) + \frac{2(\epsilon')^2}{\mu\delta}\left(\frac{2b-LB}{\mu\delta} + t - t_0\right). \qquad (60)$$

Summing up for $t = t_0 + 1, \cdots, T$ and telescoping, we get,

$$\begin{aligned} w_T r_T &\le w_{t_0} r_{t_0} + \frac{2LC(T-t_0)}{\mu^2\delta^2} + \frac{2D\gamma^H}{\mu\delta}\sum_{t=t_0+1}^{T}\left(\frac{2b}{\mu\delta} + t - t_0\right) + \frac{2(\epsilon')^2}{\mu\delta}\sum_{t=t_0+1}^{T}\left(\frac{2b-LB}{\mu\delta} + t - t_0\right) \\ &= \frac{4b^2}{\mu^2\delta^2} r_{t_0} + \frac{2LC(T-t_0)}{\mu^2\delta^2} + \frac{4bD(T-t_0)\gamma^H}{\mu^2\delta^2} + \frac{D\gamma^H}{\mu\delta}(T-t_0)(T-t_0+1) \\ &\quad + \frac{2(\epsilon')^2(2b-LB)(T-t_0)}{\mu^2\delta^2} + \frac{(\epsilon')^2}{\mu\delta}(T-t_0)(T-t_0+1). \end{aligned} \tag{61}$$

Dividing both sides by w_T and using that since

$$w_T \;=\; \left(\frac{2b}{\mu\delta} + T - t_0\right)^2 \;\ge\; (T-t_0)^2,$$

we have

$$\begin{aligned} r_T \;&\le\; \frac{4b^2}{\mu^2\delta^2 w_T} r_{t_0} + \frac{2LC(T-t_0)}{\mu^2\delta^2 w_T} + \frac{4bD(T-t_0)\gamma^H}{\mu^2\delta^2 w_T} + \frac{D\gamma^H}{\mu\delta w_T}(T-t_0)(T-t_0+1) \\ &\quad + \frac{2(\epsilon')^2(2b-LB)(T-t_0)}{\mu^2\delta^2 w_T} + \frac{(\epsilon')^2}{\mu\delta w_T}(T-t_0)(T-t_0+1) \\ &\le\; \frac{4b^2}{\mu^2\delta^2(T-t_0)^2} r_{t_0} + \frac{2LC}{\mu^2\delta^2(T-t_0)} + \frac{4bD\gamma^H}{\mu^2\delta^2(T-t_0)} + \frac{2D\gamma^H}{\mu\delta} + \frac{2(\epsilon')^2(2b-LB)}{\mu^2\delta^2(T-t_0)} + \frac{2(\epsilon')^2}{\mu\delta}. \end{aligned} \tag{62}$$

By the definition of t_0, we have $T - t_0 \ge \frac{T}{2}$. Plugging this estimate and notice that $\frac{(\epsilon')^2}{\delta} = \epsilon'$ by the definition of δ, we have

$$\begin{aligned} r_T \;&\le\; \frac{16b^2}{\mu^2\delta^2T^2} r_{t_0} + \frac{4LC + 8bD\gamma^H}{\mu^2\delta^2 T} + \frac{2D\gamma^H}{\mu\delta} + \frac{4(\epsilon')^2(2b-LB)}{\mu^2\delta^2T} + \frac{2\epsilon'}{\mu} \\ &\overset{T\ge\frac{b}{\mu\delta}}{\le}\; \frac{16b^2}{\mu^2\delta^2T^2} r_{t_0} + \frac{4LC}{\mu^2\delta^2T} + \frac{10D\gamma^H}{\mu\delta} + \frac{4(\epsilon')^2(2b-LB)}{\mu^2\delta^2T} + \frac{2\epsilon'}{\mu} \\ &\overset{(57)}{\le}\; \frac{16b^2}{\mu^2\delta^2T^2}\left(\exp\left(-\frac{\mu\delta(T-1)}{2b}\right) r_0 + \frac{LC}{2\mu\delta b} + \frac{D\gamma^H}{\mu\delta} + \frac{(\epsilon')^2(2b-LB)}{2\mu\delta b}\right) \\ &\quad + \frac{4LC}{\mu^2\delta^2T} + \frac{10D\gamma^H}{\mu\delta} + \frac{4(\epsilon')^2(2b-LB)}{\mu^2\delta^2T} + \frac{2\epsilon'}{\mu} \\ &\overset{T\ge\frac{b}{\mu\delta}}{\le}\; 16\exp\left(-\frac{\mu\delta(T-1)}{2b}\right) r_0 + \frac{8LC}{\mu^2\delta^2T} + \frac{16D\gamma^H}{\mu\delta} + \frac{8(\epsilon')^2(2b-LB)}{\mu^2\delta^2T} \\ &\quad + \frac{4LC}{\mu^2\delta^2T} + \frac{10D\gamma^H}{\mu\delta} + \frac{4(\epsilon')^2(2b-LB)}{\mu^2\delta^2T} + \frac{2\epsilon'}{\mu} \\ &=\; 16\exp\left(-\frac{\mu\delta(T-1)}{2b}\right) r_0 + \frac{12LC}{\mu^2\delta^2T} + \frac{26D\gamma^H}{\mu\delta} + \frac{12(\epsilon')^2(2b-LB)}{\mu^2\delta^2T} + \frac{2\epsilon'}{\mu}. \end{aligned} \tag{63}$$

It remains to take the maximum of the two bounds (56) and (63) with $b = \max\{\frac{2AL}{\mu\delta}, 2BL, \mu\delta\}$. □

C.4 Proof of Corollary 3.7

Proof. From Theorem C.1, when $H = \mathcal{O}(\log\epsilon^{-1})$, the dominant terms in (49) are $\frac{12LC}{\mu^2\delta^2T}$ and $\frac{2\epsilon'}{\mu}$. To guarantee that

$$\min_{t\in\{0,1,\cdots,T\}} J^* - \mathbb{E}\left[J(\theta_t)\right] \le \mathcal{O}(\epsilon) + \mathcal{O}(\epsilon'),$$

it suffices to choose $T = \mathcal{O}(\delta^{-2}\epsilon^{-1})$ such that $\frac{12LC}{\mu^2\delta^2T} = \mathcal{O}(\epsilon)$. Thus, by the definition of δ, when $\epsilon' = 0$, we have $T = \mathcal{O}(\epsilon^{-3})$; when $\epsilon' > 0$, we have $T = \mathcal{O}((\epsilon')^{-2}\epsilon^{-1})$. Otherwise, from Theorem C.1, notice that $\delta \le \epsilon + \epsilon'$, we have $\min_{t\in\{0,1,\cdots,T-1\}} J^* - \mathbb{E}\left[J(\theta_t)\right] \le \mathcal{O}(\epsilon) + \mathcal{O}(\epsilon')$, which concludes the proof. □

D Proof of Section 4.1

D.1 Proof of Lemma 4.2

Note that a similar result to Lemma 4.2 is given as Lemma 17 and 18 in (Papini et al., 2019). More precisely, Lemma 17 and 18 in (Papini et al., 2019) provide an upper bound of the variance of the PG estimator similar to the following result

$$\mathbb{V}\mathrm{ar}\left[\widehat{\nabla}_m J(\theta)\right] \leq \frac{\nu}{m}.$$

We derive a slightly tighter bound

$$\mathbb{V}\mathrm{ar}\left[\widehat{\nabla}_m J(\theta)\right] \leq \frac{\nu - \|\nabla J_H(\theta)\|}{m}.$$

This tighter bound is crucial for our work since it results in a tighter bound on $\mathbb{E}\left[\left\|\widehat{\nabla}_m J(\theta)\right\|^2\right]$ which still fits the format of (ABC). Here is the proof for Lemma 4.2.

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

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\nabla}_m J(\theta)\right\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=0}^{m-1} g(\tau_i \mid \theta)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=0}^{m-1} g(\tau_i \mid \theta) - \nabla J_H(\theta) + \nabla J_H(\theta)\right\|^2\right] \\
&= \|\nabla J_H(\theta)\|^2 + \mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=0}^{m-1} \left(g(\tau_i \mid \theta) - \nabla J_H(\theta)\right)\right\|^2\right] \\
&= \|\nabla J_H(\theta)\|^2 + \frac{1}{m^2}\sum_{i=0}^{m-1} \mathbb{E}\left[\|g(\tau_i \mid \theta) - \nabla J_H(\theta)\|^2\right] \\
&= \|\nabla J_H(\theta)\|^2 + \frac{1}{m}\mathbb{E}\left[\|g(\tau_1 \mid \theta) - \nabla J_H(\theta)\|^2\right] \\
&= \|\nabla J_H(\theta)\|^2 + \frac{\mathbb{E}\left[\|g(\tau_1 \mid \theta)\|^2 - \|\nabla J_H(\theta)\|^2\right]}{m},
\end{aligned} \tag{64}$$

where the third, the fourth and the fifth lines are all obtained by using $\nabla J_H(\theta) = \mathbb{E}\left[g(\tau_i \mid \theta)\right]$. It remains to show $\mathbb{E}_\tau\left[\|g(\tau \mid \theta)\|^2\right]$ is bounded under Assumption 4.1.

If $\widehat{\nabla}_m J(\theta)$ is a REINFORCE gradient estimator, then

$$\begin{aligned}
\mathbb{E}_\tau\left[\|g(\tau \mid \theta)\|^2\right] &\overset{(4)}{=} \mathbb{E}_\tau\left[\left\|\sum_{t'=0}^{H-1} \gamma^{t'} \mathcal{R}(s_{t'}, a_{t'}) \cdot \sum_{t=0}^{H-1} \nabla_\theta \log \pi_\theta(a_t \mid s_t)\right\|^2\right] \\
&\leq \frac{\mathcal{R}_{\max}^2}{(1-\gamma)^2}\mathbb{E}_\tau\left[\left\|\sum_{t=0}^{H-1} \nabla_\theta \log \pi_\theta(a_t \mid s_t)\right\|^2\right] \\
&\overset{(35)}{=} \frac{\mathcal{R}_{\max}^2}{(1-\gamma)^2}\sum_{t=0}^{H-1} \mathbb{E}_\tau\left[\|\nabla_\theta \log \pi_\theta(a_t \mid s_t)\|^2\right] \\
&\overset{(32)}{\leq} \frac{HG^2\mathcal{R}_{\max}^2}{(1-\gamma)^2},
\end{aligned} \tag{65}$$

where the second line is obtained by using $|\mathcal{R}(s_{t'}, a_{t'})| \leq \mathcal{R}_{\max}$.

Finally, the ABC assumption holds with

$$\mathbb{E}\left[\left\|\widehat{\nabla}_m J(\theta)\right\|^2\right] \overset{(64)+(65)}{\leq} \left(1 - \frac{1}{m}\right)\|\nabla J_H(\theta)\|^2 + \frac{HG^2\mathcal{R}_{\max}^2}{m(1-\gamma)^2}.$$

If $\widehat{\nabla}_m J(\theta)$ is a GPOMDP gradient estimator, then

$$\begin{aligned}
\mathbb{E}_\tau\left[\|g(\tau \mid \theta)\|^2\right] &\overset{(6)}{=} \mathbb{E}_\tau\left[\left\|\sum_{t=0}^{H-1}\gamma^{t/2}\mathcal{R}(s_t,a_t)\gamma^{t/2}\left(\sum_{k=0}^{t}\nabla_\theta \log \pi_\theta(a_k \mid s_k)\right)\right\|^2\right] \\
&\leq \mathbb{E}_\tau\left[\left(\sum_{t=0}^{H-1}\gamma^t\mathcal{R}(s_t,a_t)^2\right)\left(\sum_{k=0}^{H-1}\gamma^k\left\|\sum_{k'=0}^{k}\nabla_\theta \log \pi_\theta(a_{k'} \mid s_{k'})\right\|^2\right)\right] \\
&\leq \frac{\mathcal{R}_{\max}^2}{1-\gamma}\cdot\sum_{k=0}^{H-1}\gamma^k\mathbb{E}_\tau\left[\left\|\sum_{k'=0}^{k}\nabla_\theta \log \pi_\theta(a_{k'} \mid s_{k'})\right\|^2\right] \\
&\overset{(35)}{=} \frac{\mathcal{R}_{\max}^2}{1-\gamma}\cdot\sum_{k=0}^{H-1}\gamma^k\sum_{k'=0}^{k}\mathbb{E}_\tau\left[\|\nabla_\theta \log \pi_\theta(a_{k'} \mid s_{k'})\|^2\right] \\
&\overset{(32)}{\leq} \frac{G^2\mathcal{R}_{\max}^2}{1-\gamma}\cdot\sum_{k=0}^{H-1}\gamma^k(k+1) \\
&\leq \frac{G^2\mathcal{R}_{\max}^2}{(1-\gamma)^3},
\end{aligned} \tag{66}$$

where the second line is from the Cauchy-Schwarz inequality, the third line is obtained by using $|\mathcal{R}(s_t, a_t)| \leq \mathcal{R}_{\max}$ and the last line is obtained by Lemma B.1.

The above together with (64) imply that ABC assumption holds with

$$\mathbb{E}\left[\left\|\widehat{\nabla}_m J(\theta)\right\|^2\right] \overset{(64)+(66)}{\leq} \left(1 - \frac{1}{m}\right)\|\nabla J_H(\theta)\|^2 + \frac{G^2\mathcal{R}_{\max}^2}{m(1-\gamma)^3}.$$

□

D.2 Proof of Corollary 4.3

Proof. It is trivial that Assumption (LS) implies (E-LS). Now we show that (E-LS) is strictly weaker than (LS).

Consider a scalar-action, fixed-variance, Gaussian policy:

$$\pi_\theta(a \mid s) = \mathcal{N}\left(a \mid \theta^\top\phi(s), \sigma^2\right) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{a-\theta^\top\phi(s)}{\sigma}\right)^2\right\}, \tag{67}$$

where $\theta \in \mathbb{R}^d$, $\sigma > 0$ is the standard deviation, and $\phi : \mathcal{S} \to \mathcal{R}^d$ is a mapping from the state space to the feature space.

From Lemma 15 in Papini et al. (2019), the Gaussian policy (67) under the condition that the state feature vectors are bounded satisfies (E-LS). That is, under the condition that there exists $\varphi \geq 0$ such that $\sup_{s\in\mathcal{S}}\phi(s) \leq \varphi$. One does not require that the actions are bounded for the Gaussian policy. This is not the case in Xu et al. (2020a) in Section D under assumptions (LS).

Besides, from Lemma 4.2, we know that Assumption (E-LS) implies (ABC). This concludes the claim of the corollary. □

D.3 Proof of Lemma 4.4

Proof. We know that

$$
\begin{aligned}
\nabla^2 J(\theta) &\overset{(5)}{=} \nabla_\theta \mathbb{E}_\tau \left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \right) \right] \\
&= \nabla_\theta \int p(\tau \mid \theta) \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \right) d\tau \\
&= \int \nabla_\theta p(\tau \mid \theta) \left(\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \right) \right)^\top d\tau \\
&\quad + \int p(\tau \mid \theta) \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta^2 \log \pi_\theta(a_k \mid s_k) \right) d\tau \\
&= \int p(\tau \mid \theta) \nabla_\theta \log p(\tau \mid \theta) \left(\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \right) \right)^\top d\tau \\
&\quad + \int p(\tau \mid \theta) \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta^2 \log \pi_\theta(a_k \mid s_k) \right) d\tau \\
&= \mathbb{E}_\tau \left[\nabla_\theta \log p(\tau \mid \theta) \left(\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \right) \right)^\top \right] \\
&\quad + \mathbb{E}_\tau \left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta^2 \log \pi_\theta(a_k \mid s_k) \right) \right] \\
&\overset{(1)}{=} \underbrace{\mathbb{E}_\tau \left[\sum_{t'=0}^{\infty} \nabla_\theta \log \pi_\theta(a_{t'} \mid \theta_{t'}) \left(\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \right) \right)^\top \right]}_{①} \\
&\quad + \underbrace{\mathbb{E}_\tau \left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta^2 \log \pi_\theta(a_k \mid s_k) \right) \right]}_{②}.
\end{aligned} \tag{68}
$$

We now bound the above two terms separately. The second term can be bounded easily. That is,

$$
\begin{aligned}
\|②\| &\leq \mathbb{E}_\tau \left[\sum_{t=0}^{\infty} \gamma^t \left| \mathcal{R}(s_t, a_t) \right| \left(\sum_{k=0}^{t} \left\| \nabla_\theta^2 \log \pi_\theta(a_k \mid s_k) \right\| \right) \right] \\
&\leq \mathcal{R}_{\max} \sum_{t=0}^{\infty} \gamma^t \left(\sum_{k=0}^{t} \mathbb{E}_\tau \left[\left\| \nabla_\theta^2 \log \pi_\theta(a_k \mid s_k) \right\| \right] \right) \\
&\overset{(33)}{\leq} F \mathcal{R}_{\max} \sum_{t=0}^{\infty} \gamma^t (t+1) \\
&= \frac{F \mathcal{R}_{\max}}{(1-\gamma)^2},
\end{aligned} \tag{69}
$$

where the second line is obtained by using $|\mathcal{R}(s_t, a_t)| \leq \mathcal{R}_{\max}$ and the last line is obtained by Lemma B.1.

To bound the first term, we use the following notation $x_{0:t} \overset{\text{def}}{=} (x_0, x_1, \cdots, x_t)$ with $\{x_t\}_{t \geq 0}$ a sequence of random variables. Similar to the derivation of GPOMDP, we notice that future actions do not depend on past rewards

and past actions. That is, for $0 \le t < t'$ among terms of the two sums in ①, we have

$$
\begin{aligned}
&\mathbb{E}_\tau \left[\nabla_\theta \log \pi_\theta(a_{t'} \mid s_{t'}) \cdot \gamma^t \mathcal{R}(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \right)^\top \right] \\
&= \mathbb{E}_{s_{0:t'}, a_{0:t'}} \left[\nabla_\theta \log \pi_\theta(a_{t'} \mid s_{t'}) \cdot \gamma^t \mathcal{R}(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \right)^\top \right] \\
&= \mathbb{E}_{s_{0:t'}, a_{0:(t'-1)}} \left[\mathbb{E}_{a_{t'}} \left[\nabla_\theta \log \pi_\theta(a_{t'} \mid s_{t'}) \cdot \gamma^t \mathcal{R}(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \right)^\top \,\middle|\, s_{0:t'}, a_{0:(t'-1)} \right] \right] \\
&= \mathbb{E}_{s_{0:t'}, a_{0:(t'-1)}} \left[\mathbb{E}_{a_{t'}} \left[\nabla_\theta \log \pi_\theta(a_{t'} \mid s_{t'}) \,\middle|\, s_{t'} \right] \cdot \gamma^t \mathcal{R}(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \right)^\top \right] \\
&= \mathbb{E}_{s_{0:t'}, a_{0:(t'-1)}} \left[\int \pi_\theta(a_{t'} \mid s_{t'}) \nabla_\theta \log \pi_\theta(a_{t'} \mid s_{t'}) da_{t'} \cdot \gamma^t \mathcal{R}(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \right)^\top \right] \\
&= \mathbb{E}_{s_{0:t'}, a_{0:(t'-1)}} \left[\int \nabla_\theta \pi_\theta(a_{t'} \mid s_{t'}) da_{t'} \cdot \gamma^t \mathcal{R}(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \right)^\top \right] \\
&= \mathbb{E}_{s_{0:t'}, a_{0:(t'-1)}} \left[\nabla_\theta \underbrace{\int \pi_\theta(a_{t'} \mid s_{t'}) da_{t'}}_{=1} \cdot \gamma^t \mathcal{R}(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \right)^\top \right] \\
&= 0,
\end{aligned} \tag{70}
$$

where the third equality is obtained by the Markov property. Thus, ① can be simplified. We have

$$
\begin{aligned}
① \overset{(70)}{=} \; & \mathbb{E}_\tau \left[\sum_{t'=0}^{t} \nabla_\theta \log \pi_\theta(a_{t'} \mid \theta_{t'}) \left(\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \right) \right)^\top \right] \\
= \; & \mathbb{E}_\tau \left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \left(\sum_{t'=0}^{t} \nabla_\theta \log \pi_\theta(a_{t'} \mid \theta_{t'}) \right) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \right)^\top \right].
\end{aligned} \tag{71}
$$

Now we can bound ① easily. That is,

$$
\begin{aligned}
\|①\| \overset{(71)}{\le} \; & \mathbb{E}_\tau \left[\sum_{t=0}^{\infty} \gamma^t \left| \mathcal{R}(s_t, a_t) \right| \left\| \sum_{t'=0}^{t} \nabla_\theta \log \pi_\theta(a_{t'} \mid \theta_{t'}) \right\|^2 \right] \\
\le \; & \mathcal{R}_{\max} \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_\tau \left[\left\| \sum_{t'=0}^{t} \nabla_\theta \log \pi_\theta(a_{t'} \mid \theta_{t'}) \right\|^2 \right] \\
\overset{(35)}{=} \; & \mathcal{R}_{\max} \sum_{t=0}^{\infty} \gamma^t \sum_{t'=0}^{t} \mathbb{E}_\tau \left[\left\| \nabla_\theta \log \pi_\theta(a_{t'} \mid \theta_{t'}) \right\|^2 \right] \\
\overset{(32)}{\le} \; & G^2 \mathcal{R}_{\max} \sum_{t=0}^{\infty} \gamma^t (t+1) \\
= \; & \frac{G^2 \mathcal{R}_{\max}}{(1-\gamma)^2},
\end{aligned} \tag{72}
$$

where the second line is obtained by using $|\mathcal{R}(s_t, a_t)| \le \mathcal{R}_{\max}$ and the last line is obtained by Lemma B.1.

Finally,

$$\left\|\nabla^2 J(\theta)\right\| \overset{(68)+(72)+(69)}{\leq} \frac{\mathcal{R}_{\max}}{(1-\gamma)^2}(G^2+F).$$

□

D.4 Proof of Lemma 4.5

Proof. From (5), we have

$$\begin{aligned}
\|\nabla J(\theta) - \nabla J_H(\theta)\|^2 &= \left\|\mathbb{E}_\tau\left[\sum_{t=H}^{\infty}\gamma^t\mathcal{R}(s_t,a_t)\left(\sum_{k=0}^{t}\nabla_\theta\log\pi_\theta(a_k\mid s_k)\right)\right]\right\|^2 \\
&\leq \mathbb{E}_\tau\left[\left\|\sum_{t=H}^{\infty}\gamma^{t/2}\mathcal{R}(s_t,a_t)\gamma^{t/2}\left(\sum_{k=0}^{t}\nabla_\theta\log\pi_\theta(a_k\mid s_k)\right)\right\|^2\right] \\
&\leq \mathbb{E}_\tau\left[\left(\sum_{t=H}^{\infty}\gamma^t\mathcal{R}(s_t,a_t)^2\right)\left(\sum_{k=H}^{\infty}\gamma^k\left\|\sum_{k'=0}^{k}\nabla_\theta\log\pi_\theta(a_{k'}\mid s_{k'})\right\|^2\right)\right] \\
&\leq \frac{\mathcal{R}_{\max}^2\gamma^H}{1-\gamma}\mathbb{E}_\tau\left[\sum_{k=H}^{\infty}\gamma^k\left\|\sum_{k'=0}^{k}\nabla_\theta\log\pi_\theta(a_{k'}\mid s_{k'})\right\|^2\right] \\
&\overset{(35)}{=} \frac{\mathcal{R}_{\max}^2\gamma^H}{1-\gamma}\sum_{k=H}^{\infty}\gamma^k\sum_{k'=0}^{k}\mathbb{E}_\tau\left[\|\nabla_\theta\log\pi_\theta(a_{k'}\mid s_{k'})\|^2\right] \\
&\overset{(32)}{\leq} \frac{G^2\mathcal{R}_{\max}^2\gamma^H}{1-\gamma}\sum_{k=H}^{\infty}\gamma^k(k+1) \\
&= \frac{G^2\mathcal{R}_{\max}^2\gamma^{2H}}{1-\gamma}\sum_{k=0}^{\infty}\gamma^k(k+1+H) \\
&= \left(\frac{1}{1-\gamma}+H\right)\frac{G^2\mathcal{R}_{\max}^2\gamma^{2H}}{(1-\gamma)^2},
\end{aligned} \tag{73}$$

where the second and third lines are obtained by Jensen and Cauchy-Schwarz inequality respectively, the fourth line is obtained by using $|\mathcal{R}(s_t,a_t)|\leq\mathcal{R}_{\max}$ and the last line is obtained by Lemma B.1.

Thus

$$D' \overset{(73)}{=} \frac{G\mathcal{R}_{\max}}{1-\gamma}\sqrt{\frac{1}{1-\gamma}+H}.$$

Next, by inequality of Cauchy-Swartz we have

$$\begin{aligned}
|\langle\nabla J_H(\theta), \nabla J_H(\theta) - \nabla J(\theta)\rangle| &\leq \|\nabla J_H(\theta)\|\,\|\nabla J_H(\theta) - \nabla J(\theta)\| \\
&\overset{(10)}{\leq} \|\nabla J_H(\theta)\|\cdot D'\gamma^H \\
&\leq \frac{D'G\mathcal{R}_{\max}}{(1-\gamma)^{3/2}}\gamma^H,
\end{aligned} \tag{74}$$

where the last line is obtained by Lemma D.1 (iii). Thus

$$D \overset{(74)}{=} \frac{D'G\mathcal{R}_{\max}}{(1-\gamma)^{3/2}}.$$

□

D.5 Lipschitz continuity of $J(\cdot)$

In this section, we show that $J(\cdot)$ is Lipschitz-continuous under Assumption 4.1.

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

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

(ii) The norm of the gradient estimator squared in expectation is bounded, i.e. $\mathbb{E}\left[\left\|\widehat{\nabla}_m J(\theta)\right\|^2\right] \leq \Gamma_g^2$.

(iii) $J(\cdot)$ is Γ-Lipschitz, namely $\|\nabla J(\theta)\| \leq \Gamma$ with $\Gamma = \frac{G\mathcal{R}_{\max}}{(1-\gamma)^{3/2}}$. Similarly, we have $\|\nabla J_H(\theta)\| \leq \Gamma$ for the exact policy gradient of the truncated function $J_H(\cdot)$ for any horizon H.

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

Remark. The Lipschitzness constant proposed in Lemma D.1 (iii) is novel. See Section A.3 for more details.

The results in Lemma D.1 (ii) match the special case of Lemma 4.2 when the mini-batch size $m = 1$. It also implies Assumption (ABC) but with a looser upper bound, which is independent to the batch size m. We include a proof for completeness of the properties of a general vanilla policy gradient estimator. Notice that the bound of $\mathbb{E}\left[\left\|\widehat{\nabla}_m J(\theta)\right\|^2\right]$ with GPOMDP gradient estimator is a factor of $1-\gamma$ tighter as compared to Proposition 4.2 (3) in (Xu et al., 2020a) and equation (17) in (Yuan et al., 2020) under more restrictive assumptions (LS).

The result with GPOMDP gradient estimator in Lemma D.1 (i) was already proposed in Proposition 4.2 in (Xu et al., 2020a), but not with REINFORCE gradient estimator. We include a proof for both gradient estimators for the completeness.

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

$$\begin{aligned}
\left\|\nabla\left(\widehat{\nabla}_m J(\theta)\right)\right\| &= \left\|\frac{1}{m}\sum_{i=1}^{m}\sum_{t=0}^{H-1}\left(\sum_{t'=0}^{H-1}\gamma^{t'}\mathcal{R}(s_{t'}^i, a_{t'}^i)\right)\nabla_\theta^2 \log\pi_\theta(a_t^i \mid s_t^i)\right\| \\
&\leq \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{t'=0}^{H-1}\gamma^{t'}\left|\mathcal{R}(s_{t'}^i, a_{t'}^i)\right|\right)\sum_{t=0}^{H-1}\left\|\nabla_\theta^2 \log\pi_\theta(a_t^i \mid s_t^i)\right\| \\
&\leq \frac{\mathcal{R}_{\max}}{1-\gamma}\cdot\frac{1}{m}\sum_{i=1}^{m}\sum_{t=0}^{H-1}\left\|\nabla_\theta^2 \log\pi_\theta(a_t^i \mid s_t^i)\right\| \\
&\overset{\text{(LS)}}{\leq} \frac{HF\mathcal{R}_{\max}}{1-\gamma},
\end{aligned} \tag{75}$$

where the third line is obtained by using $\left|\mathcal{R}(s_{t'}^i, a_{t'}^i)\right| \leq \mathcal{R}_{\max}$. In this case, $L_g = \frac{HF\mathcal{R}_{\max}}{1-\gamma}$.

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

$$\begin{aligned}
\left\|\nabla\left(\widehat{\nabla}_m J(\theta)\right)\right\| &= \left\|\frac{1}{m}\sum_{i=1}^{m}\sum_{t=0}^{H-1}\gamma^t\mathcal{R}(s_t^i,a_t^i)\left(\sum_{k=0}^{t}\nabla_\theta^2\log\pi_\theta(a_k^i\mid s_k^i)\right)\right\| \\
&\leq \frac{1}{m}\sum_{i=1}^{m}\sum_{t=0}^{H-1}\gamma^t\left|\mathcal{R}(s_t^i,a_t^i)\right|\left(\sum_{k=0}^{t}\left\|\nabla_\theta^2\log\pi_\theta(a_k^i\mid s_k^i)\right\|\right) \\
&\leq \frac{\mathcal{R}_{\max}}{m}\sum_{i=1}^{m}\sum_{t=0}^{H-1}\gamma^t\left(\sum_{k=0}^{t}\left\|\nabla_\theta^2\log\pi_\theta(a_k^i\mid s_k^i)\right\|\right) \\
&\overset{\text{(LS)}}{\leq} F\mathcal{R}_{\max}\sum_{t=0}^{H-1}\gamma^t(t+1) \\
&\overset{\text{Lemma B.1}}{\leq} \frac{F\mathcal{R}_{\max}}{(1-\gamma)^2},
\end{aligned} \tag{76}$$

where similarly, the third line is obtained by using $\left|\mathcal{R}(s_t^i,a_t^i)\right|\leq\mathcal{R}_{\max}$. In this case, $L_g=\frac{F\mathcal{R}_{\max}}{(1-\gamma)^2}$.

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

$$\mathbb{E}\left[\left\|\widehat{\nabla}_m J(\theta)\right\|^2\right] \leq \mathbb{E}_\tau\left[\|g(\tau\mid\theta)\|^2\right].$$

If $\widehat{\nabla}_m J(\theta)$ is a REINFORCE gradient estimator, from (65), we have $\Gamma_g=\frac{\sqrt{H}G\mathcal{R}_{\max}}{1-\gamma}$. If $\widehat{\nabla}_m J(\theta)$ is a GPOMDP gradient estimator, from (66), we have $\Gamma_g=\frac{G\mathcal{R}_{\max}}{(1-\gamma)^{3/2}}$.

To prove (iii), we have

$$\begin{aligned}
\|\nabla J(\theta)\|^2 &\overset{(5)}{=} \left\|\mathbb{E}_\tau\left[\sum_{t=0}^{\infty}\gamma^t\mathcal{R}(s_t,a_t)\left(\sum_{k=0}^{t}\nabla_\theta\log\pi_\theta(a_k\mid s_k)\right)\right]\right\|^2 \\
&\leq \mathbb{E}_\tau\left[\left\|\sum_{t=0}^{\infty}\gamma^{t/2}\mathcal{R}(s_t,a_t)\gamma^{t/2}\left(\sum_{k=0}^{t}\nabla_\theta\log\pi_\theta(a_k\mid s_k)\right)\right\|^2\right] \\
&\leq \mathbb{E}_\tau\left[\left(\sum_{t=0}^{\infty}\gamma^t\mathcal{R}(s_t,a_t)^2\right)\left(\sum_{k=0}^{\infty}\gamma^k\left\|\sum_{k'=0}^{k}\nabla_\theta\log\pi_\theta(a_{k'}\mid s_{k'})\right\|^2\right)\right] \\
&\leq \frac{\mathcal{R}_{\max}^2}{1-\gamma}\mathbb{E}_\tau\left[\sum_{k=0}^{\infty}\gamma^k\left\|\sum_{k'=0}^{k}\nabla_\theta\log\pi_\theta(a_{k'}\mid s_{k'})\right\|^2\right] \\
&\overset{(35)}{=} \frac{\mathcal{R}_{\max}^2}{1-\gamma}\sum_{k=0}^{\infty}\gamma^k\sum_{k'=0}^{k}\mathbb{E}_\tau\left[\|\nabla_\theta\log\pi_\theta(a_{k'}\mid s_{k'})\|^2\right] \\
&\overset{(32)}{\leq} \frac{G^2\mathcal{R}_{\max}^2}{1-\gamma}\sum_{k=0}^{\infty}\gamma^k(k+1) \\
&= \frac{G^2\mathcal{R}_{\max}^2}{(1-\gamma)^3},
\end{aligned} \tag{77}$$

where the second and third lines are obtained by Jensen and Cauchy-Schwarz inequality respectively, the fourth line is obtained by using $|\mathcal{R}(s_t,a_t)|\leq\mathcal{R}_{\max}$ and the last line is obtained by Lemma B.1.

Thus,

$$\|\nabla J(\theta)\| \leq \Gamma \quad \text{with} \quad \Gamma=\frac{G\mathcal{R}_{\max}}{(1-\gamma)^{3/2}}.$$

Similarly, we also have

$$\|\nabla J_H(\theta)\| \;\leq\; \Gamma \quad \text{with} \quad \Gamma \;=\; \frac{G\mathcal{R}_{\max}}{(1-\gamma)^{3/2}}$$

for the exact policy gradient of the truncated function $J(\cdot)$ for any horizon H. □

D.6 Proof of Corollary 4.6

Proof. From Lemma 4.4, we know that J is L-smooth. Consider policy gradient with a mini-batch sampling of size m. From Lemma 4.2, we have Assumption 3.3 holds with $A=0$, $B=1-\frac{1}{m}$ and $C=\nu/m$. Assumption 3.2 is verified as well by Lemma 4.5 with appropriate D and D'. By Theorem 3.4, plugging $A=0$, $B=1-\frac{1}{m}$ and $C=\nu/m$ in (12) yields the corollary's claim with step size $\eta \in \left(0, \frac{2}{L\left(1-\frac{1}{m}\right)}\right)$. □

D.7 Proof of Corollary 4.7

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

$$\begin{aligned}
\mathbb{E}\left[\|\nabla J(\theta_U)\|^2\right] &\overset{(23)}{\leq} \frac{2\delta_0}{\eta T\left(2-L\eta\left(1-\frac{1}{m}\right)\right)} + \frac{L\nu\eta}{m\left(2-L\eta\left(1-\frac{1}{m}\right)\right)} \\
&\quad + \left(\frac{2D\left(3-L\eta\left(1-\frac{1}{m}\right)\right)}{2-L\eta\left(1-\frac{1}{m}\right)} + D'^2\gamma^H\right)\gamma^H \\
&\leq \frac{2\delta_0}{\eta T} + \frac{L\nu\eta}{m} + \left(6D + D'^2\gamma^H\right)\gamma^H,
\end{aligned}$$

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

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

$$\mathcal{O}(\epsilon^2) \;\geq\; \frac{2\delta_0}{\eta T} + \frac{L\nu\eta}{m} \tag{78}$$

and

$$\mathcal{O}(\epsilon^2) \;\geq\; \left(6D + D'^2\gamma^H\right)\gamma^H \tag{79}$$

respectively. To make the right hand side of (79) smaller than ϵ^2, we need $H\gamma^H = \mathcal{O}(\epsilon^2)$. Thus, we require

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

To make the right hand side of (78) smaller than ϵ^2, we require

$$\frac{L\nu\eta}{m} \;\leq\; \frac{\epsilon^2}{2} \quad\Longleftrightarrow\quad \eta \;\leq\; \frac{\epsilon^2 m}{2L\nu}. \tag{80}$$

Similarly, for the first term of the right hand side of (78), we require

$$\frac{2\delta_0}{\eta T} \;\leq\; \frac{\epsilon^2}{2} \quad\Longleftrightarrow\quad \frac{4\delta_0}{\epsilon^2 T} \;\leq\; \eta. \tag{81}$$

Combining the above two inequalities gives

$$\frac{4\delta_0}{\epsilon^2 T} \;\leq\; \eta \;\leq\; \frac{\epsilon^2 m}{2L\nu}. \tag{82}$$

This implies

$$Tm \;\geq\; \frac{8\delta_0 L\nu}{\epsilon^4}. \tag{83}$$

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

$$\frac{\epsilon^2 m}{2L\nu} \;\leq\; \frac{1}{L\left(1-\frac{1}{m}\right)} \;\Longrightarrow\; m \;\leq\; \frac{2\nu}{\epsilon^2}.$$

To conclude, it suffices to choose the step size $\eta = \frac{4\delta_0}{\epsilon^2 T} = \frac{\epsilon^2 m}{2L\nu}$, a mini-batch size m between 1 and $\frac{2\nu}{\epsilon^2}$, the number of iterations $T = \frac{8\delta_0 L\nu}{m\epsilon^4}$ and the fixed Horizon $H = \mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right)/\log\left(\frac{1}{\gamma}\right)\right)$ so that the inequalities (79), (80), (81), (82) and (83) hold, which guarantee $\mathbb{E}\left[\left\|\nabla J(\theta_U)\right\|^2\right] = \mathcal{O}(\epsilon^2)$.

Thus, the total sample complexity is

$$Tm \times H \;=\; \frac{8\delta_0 L\nu \log\left(\frac{1}{\epsilon}\right)}{\log\left(\frac{1}{\gamma}\right)\epsilon^4} \;=\; \widetilde{\mathcal{O}}(\epsilon^{-4}).$$

More precisely, from Lemma 4.4, $L = \frac{\mathcal{R}_{\max}}{(1-\gamma)^2}(G^2+F)$. When using REINFORCE gradient estimator (4), from Lemma 4.2, $\nu = \frac{HG^2\mathcal{R}_{\max}^2}{(1-\gamma)^2}$. Thus, when γ is close to 1, the sample complexity is

$$\frac{8\delta_0 H^2 G^2 \mathcal{R}_{\max}^3 (G^2+F)}{(1-\gamma)^4\epsilon^4} \;=\; \frac{8\delta_0 G^2 \mathcal{R}_{\max}^3 (G^2+F)\left(\log\left(\frac{1}{\epsilon}\right)\right)^2}{\left(\log\left(\frac{1}{\gamma}\right)\right)^2 (1-\gamma)^4\epsilon^4} \;=\; \mathcal{O}\left(\left(\log\left(\frac{1}{\epsilon}\right)\right)^2 (1-\gamma)^{-6}\epsilon^{-4}\right). \tag{84}$$

In this case, we can choose the mini-batch size $m \in \left[1; \frac{2\nu}{\epsilon^2}\right]$, i.e. from 1 to $\mathcal{O}\left(H(1-\gamma)^{-2}\epsilon^{-2}\right)$ and the constant step size $\eta = \frac{\epsilon^2 m}{2L\nu}$ varies from $\mathcal{O}\left((1-\gamma)^2\right)$ to $\mathcal{O}\left(H^{-1}(1-\gamma)^4\epsilon^2\right)$ accordingly.

When using GPOMDP gradient estimator (6), from Lemma 4.2, $\nu = \frac{G^2\mathcal{R}_{\max}^2}{(1-\gamma)^3}$. Thus, when γ is close to 1, the sample complexity is

$$\frac{8\delta_0 H G^2 \mathcal{R}_{\max}^3 (G^2+F)}{(1-\gamma)^5\epsilon^4} \;=\; \frac{8\delta_0 G^2 \mathcal{R}_{\max}^3 (G^2+F)\log\left(\frac{1}{\epsilon}\right)}{\log\left(\frac{1}{\gamma}\right)(1-\gamma)^5\epsilon^4} \;=\; \mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right)(1-\gamma)^{-6}\epsilon^{-4}\right). \tag{85}$$

In this case, we can choose the mini-batch size $m \in \left[1; \frac{2\nu}{\epsilon^2}\right]$, i.e. from 1 to $\mathcal{O}\left((1-\gamma)^{-3}\epsilon^{-2}\right)$ and the constant step size $\eta = \frac{\epsilon^2 m}{2L\nu}$ varies from $\mathcal{O}\left((1-\gamma)^2\right)$ to $\mathcal{O}\left((1-\gamma)^5\epsilon^2\right)$ accordingly. □

Remark. Comparing (85) to (84), we have that the sample complexity of GPOMDP is a factor of $\log(1/\epsilon)$ smaller than that of REINFORCE.

E Proof of Section 4.2

In this section, $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and denote $\theta_s \equiv (\theta_{s,a})_{a\in\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}$. We also use the following notations

$$\pi_{s,a}(\theta) \;\stackrel{\text{def}}{=}\; \pi_\theta(a \mid s) \qquad \text{and} \qquad \pi_s(\theta) \;\stackrel{\text{def}}{=}\; \pi_\theta(\cdot \mid s) \;\in\; \Delta(\mathcal{A}) \;\in\; \mathbb{R}^{|\mathcal{A}|}.$$

E.1 Preliminaries for the softmax tabular policy

Recall the softmax tabular policy given by

$$\pi_{s,a}(\theta) \;\stackrel{\text{def}}{=}\; \frac{\exp(\theta_{s,a})}{\sum_{a'\in\mathcal{A}}\exp(\theta_{s,a'})}. \tag{86}$$

From (86), for any $(s, a, a') \in \mathcal{S} \times \mathcal{A} \times \mathcal{A}$ with $a' \neq a$, we have immediately the following partial derivatives for the softmax tabular policy

$$\frac{\partial \pi_{s,a}(\theta)}{\partial \theta_{s,a}} = \pi_{s,a}(\theta)(1 - \pi_{s,a}(\theta)), \tag{87}$$

$$\frac{\partial \pi_{s,a}(\theta)}{\partial \theta_{s,a'}} = -\pi_{s,a}(\theta)\pi_{s,a'}(\theta). \tag{88}$$

Notice that for $s' \in \mathcal{S}$ with $s' \neq s$, we have $\frac{\partial \pi_{s,a}(\theta)}{\partial \theta_{s',a}} = 0$. From (87) and (88), we obtain respectively the gradient of $\pi_{s,a}(\theta)$ and the Jacobian of $\pi_s(\theta)$ w.r.t. θ_s

$$\frac{\partial \pi_{s,a}(\theta)}{\partial \theta_s} = \left(\frac{\partial \pi_s(\theta)}{\partial \theta_{s,a}}\right)^\top = \pi_{s,a}(\theta)(\mathbf{1}_a - \pi_s(\theta)), \tag{89}$$

$$\frac{\partial \pi_s(\theta)}{\partial \theta_s} = \mathbf{Diag}\left(\pi_s(\theta)\right) - \pi_s(\theta)\pi_s(\theta)^\top \overset{\text{def}}{=} \mathbf{H}(\pi_s(\theta)), \tag{90}$$

where $\mathbf{1}_a \in \mathbb{R}^{|\mathcal{A}|}$ is a vector with zero entries except one non-zero entry 1 corresponding to the action a. Now from (89) and (90), we obtain respectively the gradient and the Hessian of $\log \pi_{s,a}(\theta)$ w.r.t. θ_s given by

$$\frac{\partial \log \pi_{s,a}(\theta)}{\partial \theta_s} = \mathbf{1}_a - \pi_s(\theta), \tag{91}$$

$$\frac{\partial^2 \log \pi_{s,a}(\theta)}{\partial \theta_s^2} = -\mathbf{H}(\pi_s(\theta)). \tag{92}$$

E.2 Stationary point convergence of the softmax tabular policy

First we provide the proof of Lemma 4.8.

Proof. For any state $s \in \mathcal{S}$ and any $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, from (91), we have

$$\begin{aligned}
\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\|\nabla_\theta \log \pi_\theta(a \mid s)\|^2\right] &= \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[1 + \|\pi_s(\theta)\|^2 - 2\pi_{s,a}(\theta)\right] \\
&= 1 + \|\pi_s(\theta)\|^2 - 2\sum_{a\in\mathcal{A}} \pi_{s,a}(\theta)^2 \\
&= 1 - \|\pi_s(\theta)\|^2 \\
&\leq 1 - \frac{1}{|\mathcal{A}|},
\end{aligned} \tag{93}$$

where the last line is obtained by using Cauchy-Schwarz inequality in the following

$$\|\pi_s(\theta)\|^2 = \sum_{a\in\mathcal{A}} \pi_{s,a}(\theta)^2 \geq \frac{1}{|\mathcal{A}|}\left(\sum_{a\in\mathcal{A}} \pi_{s,a}(\theta)\right)^2 = \frac{1}{|\mathcal{A}|}.$$

Thus we have $G^2 = 1 - \frac{1}{|\mathcal{A}|}$.

Besides, from Lemma 22 in Mei et al. (2020), we have $\|\mathbf{H}(\pi_s(\theta))\| \leq 1$. Thus from (92), we have $\left\|\nabla_\theta^2 \log \pi_\theta(a \mid s)\right\| \leq 1$. Taking expectation over action, we have

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

Thus we have $F = 1$. □

Remark. Without expectation, for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, (93) becomes

$$\|\nabla_\theta \log \pi_\theta(a \mid s)\|^2 = 1 + \|\pi_s(\theta)\|^2 - 2\pi_{s,a}(\theta) \leq 2, \tag{94}$$

where the inequality is obtained by

$$\|\pi_s(\theta)\|^2 \;=\; \sum_{a\in\mathcal{A}} \pi_{s,a}(\theta)^2 \;\leq\; \sum_{a\in\mathcal{A}} \pi_{s,a}(\theta) \;=\; 1 \tag{95}$$

with $\pi_{s,a}(\theta) \in [0,1]$. This means, the softmax tabular policy satisfies (LS) condition with a bigger constant $G^2 = 2$ instead of $1 - \frac{1}{|\mathcal{A}|}$ and $F = 1$.

Lemma 4.8 immediately implies that $J(\cdot)$ with the softmax tabular policy is smooth and Lipschitz as following.

Lemma E.1. $J(\cdot)$ with the softmax tabular policy is $\frac{\mathcal{R}_{\max}}{(1-\gamma)^2}\left(2 - \frac{1}{|\mathcal{A}|}\right)$-smooth and $\frac{\mathcal{R}_{\max}}{(1-\gamma)^{3/2}}\sqrt{1 - \frac{1}{|\mathcal{A}|}}$-Lipschitz.

Proof. From Lemma 4.8, we know that Assumption 4.1 is satisfied with $G^2 = 1 - \frac{1}{|\mathcal{A}|}$ and $F = 1$. Thus, $J(\cdot)$ with the softmax tabular policy is smooth and Lipschitz.

Indeed, from Lemma 4.4, we obtain the smoothness constant $\frac{\mathcal{R}_{\max}}{(1-\gamma)^2}\left(2 - \frac{1}{|\mathcal{A}|}\right)$ for $J(\cdot)$; and from Lemma D.1 (iii), we obtain the Lipschitzness constant $\frac{\mathcal{R}_{\max}}{(1-\gamma)^{3/2}}\sqrt{1 - \frac{1}{|\mathcal{A}|}}$ for $J(\cdot)$. □

Now we can provide the formal statement of Corollary 4.9.

Corollary E.2 (Formal)**.** 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

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

the number of iterations T such that

$$Tm \;\geq\; \begin{cases} \frac{8\delta_0\mathcal{R}_{\max}^3\left(1-\frac{1}{|\mathcal{A}|}\right)\left(2-\frac{1}{|\mathcal{A}|}\right)}{(1-\gamma)^4\epsilon^4}\cdot H & \text{for REINFORCE} \\ \frac{8\delta_0\mathcal{R}_{\max}^3\left(1-\frac{1}{|\mathcal{A}|}\right)\left(2-\frac{1}{|\mathcal{A}|}\right)}{(1-\gamma)^5\epsilon^4} & \text{for GPOMDP} \end{cases}, \tag{96}$$

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

Proof. From Lemma E.1, we know that $L = \frac{\mathcal{R}_{\max}}{(1-\gamma^2)}\left(2 - \frac{1}{|\mathcal{A}|}\right)$.

From Lemma 4.2 and 4.8, we know that

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

Plugging in L and ν in Corollary 4.7 yields the corollary's claim. □

E.3 Stationary point convergence of the softmax tabular policy with log barrier regularization

First we provide the proof of Lemma 4.10.

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

From (29), we have the following gradient estimator

$$\widehat{\nabla}_m L_\lambda(\theta) \;=\; \widehat{\nabla}_m J(\theta) + \frac{\lambda}{|\mathcal{A}||\mathcal{S}|}\sum_{s,a}\nabla_\theta \log \pi_{s,a}(\theta). \tag{97}$$

For a state $s \in \mathcal{S}$, from (91), we have

$$\begin{aligned}\frac{\lambda}{|\mathcal{A}||\mathcal{S}|}\sum_{a\in\mathcal{A}}\frac{\partial \log \pi_{s,a}(\theta)}{\partial \theta_s} &= \frac{\lambda}{|\mathcal{A}||\mathcal{S}|}\sum_{a\in\mathcal{A}}(\mathbf{1}_a - \pi_s(\theta)) \\ &= \frac{\lambda \mathbf{1}_{|\mathcal{A}|}}{|\mathcal{A}||\mathcal{S}|} - \frac{\lambda \pi_s(\theta)}{|\mathcal{S}|} \\ &= \frac{\lambda}{|\mathcal{S}|}\left(\frac{\mathbf{1}_{|\mathcal{A}|}}{|\mathcal{A}|} - \pi_s(\theta)\right), \end{aligned} \tag{98}$$

where $\mathbf{1}_{|\mathcal{A}|} \in \mathbb{R}^{|\mathcal{A}|}$ is a vector of all ones. Thus we have

$$\widehat{\nabla}_m L_\lambda(\theta) \overset{(97)+(98)}{=} \widehat{\nabla}_m J(\theta) + \frac{\lambda}{|\mathcal{S}|}\left(\frac{\mathbf{1}}{|\mathcal{A}|} - \left[\pi_s(\theta)\right]_{s\in\mathcal{S}}\right), \tag{99}$$

where $\mathbf{1} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and

$$\left[\pi_s(\theta)\right]_{s\in\mathcal{S}} = \left[\pi_{s_1}(\theta)\ ;\ \cdots\ ;\ \pi_{s_{|\mathcal{S}|}}(\theta)\right] \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$$

is the stacking[7] of the vectors $\pi_{s_i}(\theta)$.

Next, taking expectation on the trajectories, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{\nabla}_m L_\lambda(\theta)\right\|^2\right] &\overset{(99)}{=} \mathbb{E}\left[\left\|\widehat{\nabla}_m J(\theta) + \frac{\lambda}{|\mathcal{S}|}\left(\frac{\mathbf{1}}{|\mathcal{A}|} - \left[\pi_s(\theta)\right]_{s\in\mathcal{S}}\right)\right\|^2\right] \\ &= \mathbb{E}\left[\left\|\nabla J_H(\theta) + \frac{\lambda}{|\mathcal{S}|}\left(\frac{\mathbf{1}}{|\mathcal{A}|} - \left[\pi_s(\theta)\right]_{s\in\mathcal{S}}\right) + \widehat{\nabla}_m J(\theta) - \nabla J_H(\theta)\right\|^2\right] \\ &= \|\nabla L_{\lambda,H}(\theta)\|^2 + \mathbb{E}\left[\left\|\widehat{\nabla}_m J(\theta) - \nabla J_H(\theta)\right\|^2\right] \\ &\overset{(64)}{=} \|\nabla L_{\lambda,H}(\theta)\|^2 + \frac{\mathbb{E}\left[\|g(\tau_1 \mid \theta) - \nabla J_H(\theta)\|^2\right]}{m} \\ &= \|\nabla L_{\lambda,H}(\theta)\|^2 \\ &\quad + \frac{\mathbb{E}\left[\left\|g(\tau_1 \mid \theta) + \frac{\lambda}{|\mathcal{S}|}\left(\frac{\mathbf{1}}{|\mathcal{A}|} - \left[\pi_s(\theta)\right]_{s\in\mathcal{S}}\right) - \nabla J_H(\theta) - \frac{\lambda}{|\mathcal{S}|}\left(\frac{\mathbf{1}}{|\mathcal{A}|} - \left[\pi_s(\theta)\right]_{s\in\mathcal{S}}\right)\right\|^2\right]}{m} \\ &= \left(1 - \frac{1}{m}\right)\|\nabla L_{\lambda,H}(\theta)\|^2 + \frac{\mathbb{E}\left[\left\|g(\tau_1 \mid \theta) + \frac{\lambda}{|\mathcal{S}|}\left(\frac{\mathbf{1}}{|\mathcal{A}|} - \left[\pi_s(\theta)\right]_{s\in\mathcal{S}}\right)\right\|^2\right]}{m} \\ &\leq \left(1 - \frac{1}{m}\right)\|\nabla L_{\lambda,H}(\theta)\|^2 + \frac{2\mathbb{E}\left[\|g(\tau_1 \mid \theta)\|^2\right] + 2\left\|\frac{\lambda}{|\mathcal{S}|}\left(\frac{\mathbf{1}}{|\mathcal{A}|} - \left[\pi_s(\theta)\right]_{s\in\mathcal{S}}\right)\right\|^2}{m}. \end{aligned} \tag{100}$$

In particular, we have

$$\left\|\frac{\lambda}{|\mathcal{S}|}\left(\frac{\mathbf{1}}{|\mathcal{A}|} - \left[\pi_s(\theta)\right]_{s\in\mathcal{S}}\right)\right\|^2 \leq \frac{\lambda^2}{|\mathcal{S}|^2}\left(\frac{|\mathcal{S}||\mathcal{A}|}{|\mathcal{A}|^2} - 2\frac{|\mathcal{S}|}{|\mathcal{A}|} + |\mathcal{S}|\right) = \frac{\lambda^2}{|\mathcal{S}|}\left(1 - \frac{1}{|\mathcal{A}|}\right), \tag{101}$$

where the inequality is obtained by using $\|\pi_s(\theta)\|^2 \leq 1$ in (95).

As for $\mathbb{E}\left[\|g(\tau_1 \mid \theta)\|^2\right]$, if $\widehat{\nabla}_m J(\theta)$ is a REINFORCE gradient estimator, from (65), we have

$$\mathbb{E}\left[\|g(\tau_1 \mid \theta)\|^2\right] \leq \frac{HG^2\mathcal{R}_{\max}^2}{(1-\gamma)^2} = \frac{H\mathcal{R}_{\max}^2\left(1 - \frac{1}{|\mathcal{A}|}\right)}{(1-\gamma)^2}, \tag{102}$$

[7] Here vectors are columns by default, and given $x_1, \cdots, x_{|\mathcal{S}|} \in \mathbb{R}^{|\mathcal{A}|}$ we note $[x_1\ ;\ \ldots\ ;\ x_{|\mathcal{S}|}] \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ the (column) vector stacking the x_i's on top of each other.

where the equality is obtained by Lemma 4.8 with $G^2 = \left(1 - \frac{1}{|\mathcal{A}|}\right)$.

Combining (100), (101) and (102), we have that the REINFORCE gradient estimator $\widehat{\nabla}_m L_\lambda(\theta)$ satisfies (ABC) assumption with

$$\mathbb{E}\left[\left\|\widehat{\nabla}_m L_\lambda(\theta)\right\|^2\right] \;\leq\; \left(1 - \frac{1}{m}\right)\|\nabla L_{\lambda,H}(\theta)\|^2 + \frac{2}{m}\left(1 - \frac{1}{|\mathcal{A}|}\right)\left(\frac{H\mathcal{R}_{\max}^2}{(1-\gamma)^2} + \frac{\lambda^2}{|\mathcal{S}|}\right).$$

If $\widehat{\nabla}_m J(\theta)$ is a GPOMDP gradient estimator, from (66), we have

$$\mathbb{E}\left[\|g(\tau_1 \mid \theta)\|^2\right] \;\leq\; \frac{G^2\mathcal{R}_{\max}^2}{(1-\gamma)^3} \;=\; \frac{\mathcal{R}_{\max}^2\left(1 - \frac{1}{|\mathcal{A}|}\right)}{(1-\gamma)^3}. \tag{103}$$

Combining (100), (101) and (103), we have that the GPOMDP gradient estimator $\widehat{\nabla}_m L_\lambda(\theta)$ satisfies (ABC) assumption with

$$\mathbb{E}\left[\left\|\widehat{\nabla}_m L_\lambda(\theta)\right\|^2\right] \;\leq\; \left(1 - \frac{1}{m}\right)\|\nabla L_{\lambda,H}(\theta)\|^2 + \frac{2}{m}\left(1 - \frac{1}{|\mathcal{A}|}\right)\left(\frac{\mathcal{R}_{\max}^2}{(1-\gamma)^3} + \frac{\lambda^2}{|\mathcal{S}|}\right).$$

Thus $\widehat{\nabla}_m L_\lambda(\theta)$ satisfies the (ABC) assumption for both REINFORCE and GPOMDP gradient estimators, which concludes the proof. □

We also verify that $L_\lambda(\cdot)$ is smooth and Lipschitz in the following lemma.

Lemma E.3. $L_\lambda(\cdot)$ is $\left(\frac{\mathcal{R}_{\max}}{(1-\gamma)^2}\left(2 - \frac{1}{|\mathcal{A}|}\right) + \frac{\lambda}{|\mathcal{S}|}\right)$-smooth and $\sqrt{2\left(1 - \frac{1}{|\mathcal{A}|}\right)\left(\frac{\mathcal{R}_{\max}^2}{(1-\gamma)^3} + \frac{\lambda^2}{|\mathcal{S}|}\right)}$-Lipschitz.

Proof. For the smoothness constant, first, from Lemma E.1, we know that $J(\cdot)$ is $\frac{\mathcal{R}_{\max}}{(1-\gamma)^2}\left(2 - \frac{1}{|\mathcal{A}|}\right)$-smooth.

It remains to show the regularizer $R(\theta) \overset{\text{def}}{=} \frac{\lambda}{|\mathcal{A}||\mathcal{S}|}\sum_{s,a}\log\pi_\theta(a \mid s)$ is $\frac{\lambda}{|\mathcal{S}|}$-smooth. From (99), we have

$$\nabla R(\theta) \;=\; \frac{\lambda}{|\mathcal{S}|}\left(\frac{\mathbf{1}}{|\mathcal{A}|} - \left[\pi_s(\theta)\right]_{s\in\mathcal{S}}\right).$$

From (90), we have

$$\left\|\frac{\partial^2 R(\theta)}{\partial\theta_s^2}\right\| \;=\; \left\|-\frac{\lambda}{|\mathcal{S}|}\mathbf{H}(\pi_s(\theta))\right\| \;\leq\; \frac{\lambda}{|\mathcal{S}|},$$

where the inequality is obtained by using $\|\mathbf{H}(\pi_s(\theta))\| \leq 1$ from Lemma 22 in Mei et al. (2020).

Since $\frac{\partial^2 R(\theta)}{\partial\theta_s\partial\theta_{s'}} = 0$ for $s \neq s'$, we have that $\left\|\nabla^2 R(\theta)\right\| \leq \frac{\lambda}{|\mathcal{S}|}$, which yields the smoothness constant of $L_\lambda(\cdot)$.

For the Lipschitzness constant, from (99), we know that

$$\begin{aligned}
\|\nabla L_\lambda(\theta)\|^2 \quad &= \quad \left\|\nabla J(\theta) + \frac{\lambda}{|\mathcal{S}|}\left(\frac{\mathbf{1}}{|\mathcal{A}|} - \left[\pi_s(\theta)\right]_{s\in\mathcal{S}}\right)\right\|^2 \\
&\leq \quad 2\|\nabla J(\theta)\|^2 + 2\left\|\frac{\lambda}{|\mathcal{S}|}\left(\frac{\mathbf{1}}{|\mathcal{A}|} - \left[\pi_s(\theta)\right]_{s\in\mathcal{S}}\right)\right\|^2 \\
&\overset{\text{Lemma E.1}}{\leq} \quad 2\left(1 - \frac{1}{|\mathcal{A}|}\right)\frac{\mathcal{R}_{\max}^2}{(1-\gamma)^3} + 2\left\|\frac{\lambda}{|\mathcal{S}|}\left(\frac{\mathbf{1}}{|\mathcal{A}|} - \left[\pi_s(\theta)\right]_{s\in\mathcal{S}}\right)\right\|^2 \\
&\overset{(101)}{\leq} \quad 2\left(1 - \frac{1}{|\mathcal{A}|}\right)\frac{\mathcal{R}_{\max}^2}{(1-\gamma)^3} + \frac{2\lambda^2}{|\mathcal{S}|}\left(1 - \frac{1}{|\mathcal{A}|}\right) \\
&= \quad 2\left(1 - \frac{1}{|\mathcal{A}|}\right)\left(\frac{\mathcal{R}_{\max}^2}{(1-\gamma)^3} + \frac{\lambda^2}{|\mathcal{S}|}\right).
\end{aligned} \tag{104}$$

Thus,

$$\|\nabla L_\lambda(\theta)\| \leq \sqrt{2\left(1-\frac{1}{|\mathcal{A}|}\right)\left(\frac{\mathcal{R}_{\max}^2}{(1-\gamma)^3}+\frac{\lambda^2}{|\mathcal{S}|}\right)}.$$

□

The truncated gradient assumption in the case of $L_{\lambda,H}(\cdot)$. As $L_\lambda(\theta)$ and $L_{\lambda,H}(\theta)$ use the same regularizer, the bias due to the truncation does not affect the regularization. Besides, from Lemma 4.8, we have that Assumption (E-LS) holds. Thus, from Lemma 4.5, Assumption 3.2 holds for $L_\lambda(\theta)$ and $L_{\lambda,H}(\theta)$ with the same constant D and D' in Lemma 4.5 and the constant G in Lemma 4.8. That is,

$$|\langle \nabla L_{\lambda,H}(\theta), L_{\lambda,H}(\theta) - L_\lambda(\theta)\rangle| \leq D\gamma^H, \tag{105}$$
$$\|\nabla L_{\lambda,H}(\theta) - L_\lambda(\theta)\| \leq D'\gamma^H, \tag{106}$$

with

$$D = \frac{D'\mathcal{R}_{\max}}{(1-\gamma)^{3/2}}\sqrt{1-\frac{1}{|\mathcal{A}|}}, \tag{107}$$
$$D' = \frac{\mathcal{R}_{\max}}{1-\gamma}\sqrt{\left(\frac{1}{1-\gamma}+H\right)\left(1-\frac{1}{|\mathcal{A}|}\right)}. \tag{108}$$

Similar to Corollary E.2, now we can provide the FOSP convergence of $L_\lambda(\theta)$.

Corollary E.4. Consider the vanilla PG (either REINFORCE or GPOMDP) applied in $L_\lambda(\cdot)$. Let $\delta_0 \overset{\text{def}}{=} L_\lambda^* - L_\lambda(\theta_0)$ with $L_\lambda^* \overset{\text{def}}{=} \max_{\theta\in\mathbb{R}^d} L_\lambda(\theta)$. 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

$$\nu = \begin{cases} 2\left(1-\frac{1}{|\mathcal{A}|}\right)\left(\frac{H\mathcal{R}_{\max}^2}{(1-\gamma)^2}+\frac{\lambda^2}{|\mathcal{S}|}\right) & \text{when using REINFORCE} \\ 2\left(1-\frac{1}{|\mathcal{A}|}\right)\left(\frac{\mathcal{R}_{\max}^2}{(1-\gamma)^3}+\frac{\lambda^2}{|\mathcal{S}|}\right) & \text{when using GPOMDP} \end{cases}, \tag{109}$$

the number of iterations T such that

$$Tm \geq \frac{8\delta_0 L\nu}{\epsilon^4} = \mathcal{O}((1-\gamma)^{-5}\epsilon^{-4}), \tag{110}$$

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

Proof. From Lemma E.3, we know that $L = \frac{\mathcal{R}_{\max}}{(1-\gamma)^2}\left(2-\frac{1}{|\mathcal{A}|}\right)+\frac{\lambda}{|\mathcal{S}|}$.

From Lemma 4.10, we know that

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

Plugging in L and ν in Corollary 4.7 yields the corollary's claim. □

E.4 Sample complexity for high probability global optimum convergence

In this section, we provide the sample complexity to reach a global optimum convergence of the expected return $J(\cdot)$ in high probability for the softmax tabular policy with log barrier regularization.

Before the results, we introduce the stationary distribution

$$d_{\rho,s}(\pi^*) \overset{\text{def}}{=} \mathbb{E}_{s_0\sim\rho(\cdot),\tau\sim p(\cdot|\pi^*)}\left[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t\mathbb{P}(s_t=s)\right],$$

where π^* is the optimal policy. We refer to $\left\| \frac{d_\rho(\pi^*)}{\rho} \right\|_\infty \overset{\text{def}}{=} \max_{s \in \mathcal{S}} \frac{d_{\rho,s}(\pi^*)}{\rho(s)}$ as the distribution mismatch coefficient of π under ρ (Agarwal et al., 2021)[8]. We assume that the initial state distribution ρ satisfies $\min_s \rho(s) > 0$. This assumption was adapted by Agarwal et al. (2021) to ensure that the distribution mismatch coefficient is finite.

Corollary E.5. For any accuracy level $\epsilon > 0$, any probability accuracy level $\delta \in (0,1)$ and any starting state distribution ρ, consider the vanilla PG (either REINFORCE or GPOMDP) applied to $L_\lambda(\cdot)$. If we chose the horizon $H = \mathcal{O}\left((1-\gamma)^{-1} \log\left(1/\epsilon_{opt}\right) \log\left(1/\delta\right)\right)$, the batch size $1 \leq m \leq \frac{2\nu}{\delta \epsilon_{opt}^2}$ and the number of iterations T such that $Tm \geq \frac{8(L_\lambda^* - L_\lambda(\theta_0))L\nu}{\delta^2 \epsilon_{opt}^4}$, the regularization parameter $\lambda = \frac{(1-\gamma)\epsilon}{2\left\| \frac{d_\rho(\pi^*)}{\rho} \right\|_\infty}$ and

$$\epsilon_{opt} \;=\; \frac{\lambda}{2|\mathcal{S}||\mathcal{A}|} \;=\; \frac{(1-\gamma)\epsilon}{4|\mathcal{S}||\mathcal{A}| \left\| \frac{d_\rho(\theta^*)}{\rho} \right\|_\infty} \tag{111}$$

with L, ν in the setting of Corollary E.4, then we have an upper bound of the sample complexity

$$Tm \times H = \mathcal{O}\left(\frac{|\mathcal{S}|^4 |\mathcal{A}|^4 \left\| \frac{d_\rho(\theta^*)}{\rho} \right\|_\infty^4}{\delta^2 \epsilon^4 (1-\gamma)^{10}} \cdot \log\left(1/\epsilon\right) \log\left(1/\delta\right) \right) \tag{112}$$

guarantees that $J^* - J(\theta_T) \leq \epsilon$ with probability at least $1 - \delta$.

The above high probability global optimum sample complexity holds with a wide range of parameters (e.g. batch size m and step size η) thanks to Corollary E.4.

We need the following result to link the stationary point convergence of $L_\lambda(\cdot)$ to the suboptimality gap convergence $J^* - J(\cdot)$ when the norm of the gradient of a stationary point and the regularization parameter λ are sufficiently small.

Proposition E.6 (Theorem 5.2 in Agarwal et al. (2021))**.** Suppose θ is such that $\|\nabla L_\lambda(\theta)\| \leq \frac{\lambda}{2|\mathcal{S}||\mathcal{A}|}$, then for every initial distribution ρ, we have

$$J^* - J(\theta) \;\leq\; \frac{2\lambda}{1-\gamma} \left\| \frac{d_\rho(\theta^*)}{\rho} \right\|_\infty . \tag{113}$$

By leveraging Proposition E.6, we now derive the proof for Corollary E.5.

Proof. From Corollary E.4 we have that $\mathbb{E}\left[\|\nabla L_\lambda(\theta_U)\|^2 \right] \leq \delta \epsilon_{opt}^2$,

Thus, there exists $t_0 \in \{0, \cdots, T-1\}$ s.t. $\mathbb{E}\left[\|\nabla L_\lambda(\theta_{t_o})\|^2 \right] \leq \mathbb{E}\left[\|\nabla L_\lambda(\theta_U)\|^2 \right] \leq \delta \epsilon_{opt}^2$.

From Proposition E.6, we know that if $\|\nabla L_\lambda(\theta_{t_o})\| \leq \epsilon_{opt}$, we have

$$J^* - J(\theta_{t_0}) \;\leq\; \frac{2\lambda}{1-\gamma} \left\| \frac{d_\rho(\theta^*)}{\rho} \right\|_\infty \;=\; \epsilon.$$

Thus, we have

$$\mathbb{P}(J^* - J(\theta_{t_0}) \leq \epsilon) \;\geq\; \mathbb{P}\left(\|\nabla L_\lambda(\theta_{t_o})\| \leq \epsilon_{opt} \right). \tag{114}$$

[8] For simplicity, we assume that the sampling for the initial state distribution is the same as the initial state distribution appeared in the expected return $J(\cdot)$. There is no difference, compared to our results, to impose a different initial state distribution $\mu \neq \rho$ for the stochastic vanilla PG. In this case, the distribution mismatch coefficient will be $\left\| \frac{d_\rho(\pi^*)}{\mu} \right\|_\infty$.

Consequently, we have

$$\begin{aligned}
\mathbb{P}(J^* - J(\theta_{t_0}) \geq \epsilon) &= 1 - \mathbb{P}(J^* - J(\theta_{t_0}) \leq \epsilon) \\
&\overset{(114)}{\leq} 1 - \mathbb{P}\left(\|\nabla L_\lambda(\theta_{t_o})\| \leq \epsilon_{opt}\right) \\
&= \mathbb{P}\left(\|\nabla L_\lambda(\theta_{t_o})\| \geq \epsilon_{opt}\right) \\
&= \mathbb{P}\left(\|\nabla L_\lambda(\theta_{t_o})\|^2 \geq \epsilon_{opt}^2\right) \\
&\leq \frac{\mathbb{E}\left[\|\nabla L_\lambda(\theta_{t_o})\|^2\right]}{\epsilon_{opt}^2} \qquad \text{(by Markov's inequality)} \\
&\leq \delta. \qquad (115)
\end{aligned}$$

Since $t_0 m \leq Tm$, we conclude that the upper bound of the sample complexity is

$$Tm \times H \;\geq\; \frac{8(J^* - J(\theta_0))L\nu}{\delta^2 \epsilon_{opt}^4} \times H \;=\; \mathcal{O}\left(\frac{|\mathcal{S}|^4 |\mathcal{A}|^4 \left\|\frac{d_\rho(\theta^*)}{\rho}\right\|_\infty^4}{\delta^2 \epsilon^4 (1-\gamma)^{10}} \cdot \log(1/\epsilon) \log(1/\delta)\right).$$

□

Remark. Following the proof of Corollary E.5, we can also deduce the iteration complexity of the exact full gradient updates for the global optimum convergence.

Indeed, from Lemma E.3, $L_\lambda(\cdot)$ is smooth. From Theorem 3.4, we know that with the number of iterations

$$T \;\geq\; \frac{12\delta_0 L}{\epsilon_{opt}^2} \;=\; \mathcal{O}\left(\frac{\delta_0}{(1-\gamma)^4 \epsilon^2}\right), \qquad (116)$$

we have $\min_{0 \leq t \leq T-1} \|\nabla L_\lambda(\theta_t)\|^2 \leq \epsilon_{opt}^2$ for the exact full gradient updates.

From Proposition E.6, we have $\min_{0 \leq t \leq T-1} J^* - J(\theta_t) \leq \epsilon$.

Compared to the iteration complexity in Corollary 5.1 in Agarwal et al. (2021), ours (116) is improved by a factor of $1 - \gamma$ thanks to an improved analysis of the smoothness constant in Lemma E.3.

E.5 Sample complexity for the average regret convergence

By leveraging Proposition E.6, we now derive the proof for Corollary 4.11.

Proof. We define the following set of "bad" iterates based on a technique developed by Zhang et al. (2021a)

$$I^+ \;\overset{\text{def}}{=}\; \left\{ t \in \{0, \cdots, T-1\} \;\middle|\; \|\nabla L_\lambda(\theta_t)\| \;\geq\; \frac{\lambda}{2|\mathcal{S}||\mathcal{A}|} \right\} \qquad (117)$$

with

$$\lambda \;=\; \frac{(1-\gamma)\epsilon}{2\left\|\frac{d_\rho(\theta^*)}{\mu}\right\|_\infty}. \qquad (118)$$

We have

$$\begin{aligned} J^* - \frac{1}{T}\sum_{t=0}^{T-1} J(\theta_t) &= \frac{1}{T}\sum_{t\in I^+} J^* - J(\theta_t) + \frac{1}{T}\sum_{t\notin I^+} J^* - J(\theta_t) \\ &\leq \frac{|I^+|}{T}\cdot\frac{2\mathcal{R}_{\max}}{1-\gamma} + \frac{1}{T}\sum_{t\notin I^+} J^* - J(\theta_t) \\ &\leq \frac{|I^+|}{T}\cdot\frac{2\mathcal{R}_{\max}}{1-\gamma} + \frac{T-|I^+|}{T}\cdot\frac{2\lambda}{1-\gamma}\left\|\frac{d_\rho(\theta^*)}{\rho}\right\|_\infty \\ &\leq \frac{|I^+|}{T}\cdot\frac{2\mathcal{R}_{\max}}{1-\gamma} + \frac{2\lambda}{1-\gamma}\left\|\frac{d_\rho(\theta^*)}{\rho}\right\|_\infty \\ &\overset{(118)}{=} \frac{|I^+|}{T}\cdot\frac{2\mathcal{R}_{\max}}{1-\gamma} + \epsilon. \end{aligned} \tag{119}$$

where the second line is obtained as $|J(\cdot)| \leq \frac{\mathcal{R}_{\max}}{1-\gamma}$ and the third line is obtained by Proposition E.6.

It remains to bound $|I^+|$. In fact,

$$\begin{aligned} \sum_{t=0}^{T-1}\|\nabla L_\lambda(\theta_t)\|^2 &\geq \sum_{t\in I^+}\|\nabla L_\lambda(\theta_t)\|^2 \\ &\geq \frac{|I^+|\lambda^2}{4|\mathcal{S}|^2|\mathcal{A}|^2}. \end{aligned} \tag{120}$$

Thus, we have

$$\begin{aligned} \frac{|I^+|}{T} &\leq \frac{4|\mathcal{S}|^2|\mathcal{A}|^2}{\lambda^2}\cdot\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla L_\lambda(\theta_t)\|^2 \\ &\overset{(118)}{=} \frac{16\left\|\frac{d_\rho(\theta^*)}{\mu}\right\|_\infty^2|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^2\epsilon^2}\cdot\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla L_\lambda(\theta_t)\|^2. \end{aligned} \tag{121}$$

Thus, we have

$$J^* - \frac{1}{T}\sum_{t=0}^{T-1} J(\theta_t) \overset{(119)+(121)}{\leq} \frac{32\mathcal{R}_{\max}\left\|\frac{d_\rho(\theta^*)}{\rho}\right\|_\infty^2|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^3\epsilon^2}\cdot\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla L_\lambda(\theta_t)\|^2 + \epsilon. \tag{122}$$

Taking expectation over the iterations on both side, we have

$$J^* - \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[J(\theta_t)\right] \overset{(119)+(121)}{\leq} \frac{32\mathcal{R}_{\max}\left\|\frac{d_\rho(\theta^*)}{\rho}\right\|_\infty^2|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^3\epsilon^2}\cdot\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla L_\lambda(\theta_t)\|^2\right] + \epsilon. \tag{123}$$

It suffices to have $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla L_\lambda(\theta_t)\|^2\right] \leq (1-\gamma)^3\epsilon^3$ to guarantee that $J^* - \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[J(\theta_t)\right] \leq \mathcal{O}(\epsilon)$.

From Corollary 4.7, consider the batch size m such that $1 \leq m \leq \frac{2\nu}{(1-\gamma)^3\epsilon^3} = \mathcal{O}\left(\frac{1}{(1-\gamma)^6\epsilon^3}\right)$, 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 Corollary E.4 . If the horizon $H = \mathcal{O}\left(\frac{\log(1/\epsilon)}{1-\gamma}\right)$ and the number of iterations T is such that

$$Tm \times H \geq \frac{8(J^* - J(\theta_0))L\nu}{(1-\gamma)^6\epsilon^6} \times H = \widetilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^{12}\epsilon^6}\right),$$

we have $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla L_\lambda(\theta_t)\|^2\right] \leq (1-\gamma)^3\epsilon^3$, which conclude the proof. □

F Proof of Section 4.3

First, we give the definition of the advantage function A^{π_θ} induced by the policy π_θ appeared in the transferred compatible function approximation error in Assumption 4.13. To do this, given a policy π, we define the state-action value function $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as

$$Q^\pi(s,a) \stackrel{\text{def}}{=} \mathbb{E}_{a_t \sim \pi(\cdot|s_t), s_{t+1} \sim \mathcal{P}(\cdot|s_t,a_t)} \left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t,a_t) \;\middle|\; s_0 = s, a_0 = a \right].$$

From this, the state-value function $V^\pi : \mathcal{S} \to \mathbb{R}$ and the advantage function $A^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, under the policy π, can be defined as

$$V^\pi(s) \stackrel{\text{def}}{=} \mathbb{E}_{a \sim \pi(\cdot|s)} \left[Q^\pi(s,a) \right],$$
$$A^\pi(s,a) \stackrel{\text{def}}{=} Q^\pi(s,a) - V^\pi(s).$$

Before presenting the proof of Corollary 4.14, we need the following result to show that Fisher-non-degenerate parametrized policy satisfies the relaxed weak gradient domination assumption.

Proposition F.1 (Lemma 4.7 in Ding et al. (2021a)). If the policy π_θ satisfies Assumption 4.1, 4.12 and 4.13, then

$$\frac{\mu_F \sqrt{\epsilon_{bias}}}{(1-\gamma)G} + \|\nabla J_H(\theta)\| \geq \frac{\mu_F}{G}(J^* - J(\theta)). \tag{124}$$

Remark. Here we use the weaker assumption (E-LS) instead of (LS) compared to the original Lemma 4.7 in Ding et al. (2021a). The relaxed weak gradient domination property still holds. The proof essentially follows the same arguments and thus is omitted here.

Now we provide the proof of Corollary 4.14.

Proof. From Proposition F.1, we have that Assumption 3.6 holds. Also because of Assumption (E-LS), we have Lemmas 4.2, 4.4 and 4.5 hold. Finally, by Corollary 3.7, this directly concludes the proof. □

G FOSP convergence analysis for the softmax with entropy regularization.

In this section, we study stochastic gradient ascent on the softmax tabular policy with entropy regularization, which is

$$\tilde{J}(\theta) \stackrel{\text{def}}{=} J(\theta) + \mathbb{H}(\theta) \tag{125}$$

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

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

Using the same technique to derive the full gradient of the expected return (3), we have

$$\begin{aligned}
\nabla \tilde{J}(\theta) &= \nabla J(\theta) - \lambda \mathbb{E}_\tau \left[\nabla \log p(\tau \mid \theta) \sum_{t=0}^{\infty} \gamma^t \log \pi_{s_t,a_t}(\theta) \right] - \lambda \mathbb{E}_\tau \left[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta \log \pi_{s_t,a_t}(\theta) \right] \\
&\overset{(1)}{=} \nabla J(\theta) - \lambda \mathbb{E}_\tau \left[\sum_{k=0}^{\infty} \nabla_\theta \log \pi_{s_k,a_k}(\theta) \sum_{t=0}^{\infty} \gamma^t \log \pi_{s_t,a_t}(\theta) \right] - \lambda \mathbb{E}_\tau \left[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta \log \pi_{s_t,a_t}(\theta) \right] \\
&= \nabla J(\theta) - \lambda \mathbb{E}_\tau \left[\sum_{t=0}^{\infty} \gamma^t \log \pi_{s_t,a_t}(\theta) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_{s_k,a_k}(\theta) \right) \right] - \lambda \mathbb{E}_\tau \left[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta \log \pi_{s_t,a_t}(\theta) \right] \\
&\overset{(5)}{=} \mathbb{E}_\tau \left[\sum_{t=0}^{\infty} \gamma^t \left((\mathcal{R}(s_t,a_t) - \lambda \log \pi_{s_t,a_t}(\theta)) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_{s_k,a_k}(\theta) \right) - \lambda \nabla_\theta \log \pi_{s_t,a_t}(\theta) \right) \right],
\end{aligned} \tag{126}$$

where the third line is obtained by using the fact that for any $0 \leq t < k$, we have

$$\mathbb{E}_\tau \left[\log \pi_{s_t,a_t}(\theta) \nabla_\theta \log \pi(s_k, a_k)(\theta)\right] \;=\; 0. \tag{127}$$

Equation (127) is derived by following the same proof technique of Lemma B.4.

Thus, the stochastic gradient estimator of $\nabla \tilde{J}(\theta)$ with mini-batch size m is

$$\widehat{\nabla}_m \tilde{J}(\theta) \;\stackrel{\text{def}}{=}\; \widehat{\nabla}_m J(\theta) - \frac{\lambda}{m} \sum_{i=1}^{m} \sum_{t=0}^{H-1} \gamma^t \left(\log \pi_{s_t^i,a_t^i}(\theta) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_{s_k^i,a_k^i}(\theta) \right) + \nabla_\theta \log \pi_{s_t^i,a_t^i}(\theta) \right). \tag{128}$$

Notice that $\widehat{\nabla}_m \tilde{J}(\cdot)$ is the unbiased gradient estimator of the truncated function

$$\tilde{J}_H(\theta) \;\stackrel{\text{def}}{=}\; \mathbb{E}_\tau \left[\sum_{t=0}^{H-1} \gamma^t \left(\mathcal{R}(s_t, a_t) - \lambda \log \pi_{s_t,a_t}(\theta) \right) \right]. \tag{129}$$

We show that $\widehat{\nabla}_m \tilde{J}(\cdot)$ satisfies the (ABC) assumption as following.

Lemma G.1. The stochastic gradient estimator (128) satisfies Assumption (ABC) with

$$\mathbb{E}\left[\left\| \widehat{\nabla}_m \tilde{J}(\theta) \right\|^2 \right] \;\leq\; \left(1 - \frac{1}{m}\right) \left\| \nabla \tilde{J}(\theta) \right\|^2 + \frac{2\left(1 - \frac{1}{|\mathcal{A}|}\right) \mathcal{R}_{\max}^2}{m(1-\gamma)^3} + \frac{2\lambda^2}{m(1-\gamma^2)} \left(1 - \frac{1}{|\mathcal{A}|}\right) + \frac{8H|\mathcal{A}|\lambda^2}{m(1-\gamma)^3}. \tag{130}$$

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

Similarly, let $\tilde{g}(\tau \mid \theta)$ be a stochastic gradient estimator of one single sampled trajectory τ of $\nabla \tilde{J}_H(\theta)$. Thus $\widehat{\nabla}_m \tilde{J}(\theta) = \frac{1}{m} \sum_{i=1}^m \tilde{g}(\tau_i \mid \theta)$, and $\widehat{\nabla}_m \tilde{J}(\theta)$ and $\tilde{g}(\tau \mid \theta)$ are unbiased estimators of $\tilde{J}_H(\theta)$.

Similar to (64), from (128) we have

$$\begin{aligned}
\mathbb{E}\left[\left\| \widehat{\nabla}_m \tilde{J}(\theta) \right\|^2 \right] &= \mathbb{E}\left[\left\| \widehat{\nabla}_m \tilde{J}(\theta) + \nabla \tilde{J}_H(\theta) - \nabla \tilde{J}_H(\theta) \right\|^2 \right] \\
&= \left\| \nabla \tilde{J}_H(\theta) \right\|^2 + \mathbb{E}\left[\left\| \widehat{\nabla}_m \tilde{J}(\theta) - \nabla \tilde{J}_H(\theta) \right\|^2 \right] \\
&= \left\| \nabla \tilde{J}_H(\theta) \right\|^2 + \mathbb{E}\left[\left\| \frac{1}{m} \sum_{i=1}^m (\tilde{g}(\tau_i \mid \theta) - \nabla \tilde{J}_H(\theta)) \right\|^2 \right] \\
&= \left\| \nabla \tilde{J}_H(\theta) \right\|^2 + \frac{1}{m} \mathbb{E}\left[\left\| \tilde{g}(\tau_1 \mid \theta) - \nabla \tilde{J}_H(\theta) \right\|^2 \right] \\
&= \left(1 - \frac{1}{m}\right) \left\| \nabla \tilde{J}(\theta) \right\|^2 + \frac{1}{m} \mathbb{E}\left[\left\| \tilde{g}(\tau_1 \mid \theta) \right\|^2 \right].
\end{aligned} \tag{131}$$

It remains to show $\mathbb{E}_\tau \left[\left\| \tilde{g}(\tau \mid \theta) \right\|^2 \right]$ is bounded. From (128) we have

$$\begin{aligned}
\mathbb{E}\left[\|\tilde{g}(\tau \mid \theta)\|^2\right] &= \mathbb{E}_\tau\left[\left\|g(\tau \mid \theta) - \lambda \sum_{t=0}^{H-1} \gamma^t \log \pi_{s_t,a_t}(\theta)\left(\sum_{k=0}^{t} \nabla_\theta \log \pi_{s_k,a_k}(\theta)\right) - \lambda \sum_{t=0}^{H-1} \gamma^t \nabla_\theta \log \pi_{s_t,a_t}(\theta)\right\|^2\right] \\
&\leq 2\mathbb{E}\left[\|g(\tau \mid \theta)\|^2\right] + 2\lambda^2 \mathbb{E}\left[\left\|\sum_{t=0}^{H-1} \gamma^t \log \pi_{s_t,a_t}(\theta)\left(\sum_{k=0}^{t} \nabla_\theta \log \pi_{s_k,a_k}(\theta)\right)\right\|^2\right] \\
&\quad + 2\lambda^2 \mathbb{E}\left[\left\|\sum_{t=0}^{H-1} \gamma^t \nabla_\theta \log \pi_{s_t,a_t}(\theta)\right\|^2\right] \\
&\leq \frac{2\left(1 - \frac{1}{|\mathcal{A}|}\right)\mathcal{R}_{\max}^2}{(1-\gamma)^3} + 2\lambda^2 \underbrace{\mathbb{E}\left[\left\|\sum_{t=0}^{H-1} \gamma^t \log \pi_{s_t,a_t}(\theta)\left(\sum_{k=0}^{t} \nabla_\theta \log \pi_{s_k,a_k}(\theta)\right)\right\|^2\right]}_{①} \\
&\quad + 2\lambda^2 \underbrace{\mathbb{E}\left[\left\|\sum_{t=0}^{H-1} \gamma^t \nabla_\theta \log \pi_{s_t,a_t}(\theta)\right\|^2\right]}_{②},
\end{aligned} \tag{132}$$

where the last inequality is obtained by Lemma 4.2 with GPOMDP estimator and the constant $G^2 = 1 - \frac{1}{|\mathcal{A}|}$ provided from Lemma 4.8.

Now we will bound ① and ② separately.

From Lemma B.5, we know that

$$\begin{aligned}
② &= \sum_{t=0}^{H-1} \gamma^{2t} \mathbb{E}\left[\|\nabla_\theta \log \pi_{s_t,a_t}(\theta)\|^2\right] \\
&\overset{\text{Lemma 4.8}}{\leq} \left(1 - \frac{1}{|\mathcal{A}|}\right) \sum_{t=0}^{H-1} \gamma^{2t} \\
&\leq \frac{1}{1-\gamma^2}\left(1 - \frac{1}{|\mathcal{A}|}\right).
\end{aligned} \tag{133}$$

As for ①, we have

$$\begin{aligned}
① &\leq H \sum_{t=0}^{H-1} \gamma^{2t} \mathbb{E}\left[\left(\log \pi_{s_t,a_t}(\theta)\right)^2 \left\|\sum_{k=0}^{t} \nabla_\theta \log \pi_{s_k,a_k}(\theta)\right\|^2\right] \\
&\leq H \sum_{t=0}^{H-1} \gamma^{2t} \mathbb{E}\left[\left(\log \pi_{s_t,a_t}(\theta)\right)^2 \left\|\sum_{k=0}^{t} \nabla_\theta \log \pi_{s_k,a_k}(\theta)\right\|^2\right] \\
&\leq H \sum_{t=0}^{H-1} \gamma^{2t} \mathbb{E}\left[\left(\log \pi_{s_t,a_t}(\theta)\right)^2 (t+1) \sum_{k=0}^{t} \|\nabla_\theta \log \pi_{s_k,a_k}(\theta)\|^2\right] \\
&\overset{(94)}{\leq} 2H \sum_{t=0}^{H-1} \gamma^{2t} (t+1)^2 \mathbb{E}\left[\left(\log \pi_{s_t,a_t}(\theta)\right)^2\right] \\
&\leq 2H|\mathcal{A}| \sum_{t=0}^{H-1} \gamma^{2t}(t+1)^2 \qquad (134) \\
&\leq \frac{4H|\mathcal{A}|}{(1-\gamma)^3}, \qquad (135)
\end{aligned}$$

where (134) is obtained by using

$$\mathbb{E}\left[\left(\log \pi_{s_t,a_t}(\theta)\right)^2\right] \;=\; \mathbb{E}_{s_t}\left[\sum_{a\in\mathcal{A}} \pi_{s_t,a}(\theta)\left(\log \pi_{s_t,a}(\theta)\right)^2\right] \;\leq\; |\mathcal{A}|,$$

and the last line is obtained by $\gamma^{2t} \leq \gamma^t$ and Lemma B.2.

Combining (131), (132), (133) and (135) yields the claim of the lemma. □

By adopting Lemma 14 in Mei et al. (2020), we show that $\tilde{J}(\cdot)$ is smooth as following.

Lemma G.2. $\tilde{J}(\cdot)$ is $\left(\frac{\mathcal{R}_{\max}}{(1-\gamma)^2}\left(2-\frac{1}{|\mathcal{A}|}\right)+\frac{\lambda(4+8\log|\mathcal{A}|)}{(1-\gamma)^3}\right)$-smooth.

Proof. From (125), we have

$$\tilde{J}(\theta) \;=\; J(\theta) - \lambda\mathbb{E}_\tau\left[\sum_{t=0}^{\infty}\gamma^t \log \pi_{s_t,a_t}(\theta)\right].$$

From Lemma E.1, we know that $J(\cdot)$ is $\left(\frac{\mathcal{R}_{\max}}{(1-\gamma)^2}\left(2-\frac{1}{|\mathcal{A}|}\right)\right)$-smooth.

From Lemma 14 in Mei et al. (2020), we know that $\mathbb{E}_\tau\left[\sum_{t=0}^{\infty}\gamma^t \log \pi_{s_t,a_t}(\theta)\right]$ is $\left(\frac{\lambda(4+8\log|\mathcal{A}|)}{(1-\gamma)^3}\right)$-smooth.

Combining the two smoothness constants yields the claim of the lemma. □

From Lemma G.1 and Lemma G.2 we can also establish a similar FOSP convergence as for Corollary 4.7.

Corollary G.3. Consider the vanilla PG updates (128) for the softmax with entropy regularization (125) . 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

$$Tm \geq \frac{8\delta_0 L\nu}{\epsilon^4} = \mathcal{O}((1-\gamma)^{-6}\epsilon^{-4}) \tag{136}$$

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

Remark. The sample complexity $Tm \times H$ is $\mathcal{O}((1-\gamma)^{-8}\epsilon^{-4})$ instead of $\mathcal{O}((1-\gamma)^{-6}\epsilon^{-4})$ as in Corollary 4.7 due to the $(1-\gamma)^{-3}$ dependency on the smoothness constant L and the $(1-\gamma)^{-4}$ dependency on the bounded variance constant ν.

Proof. From Lemma G.2, we know that

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

From Lemma G.1, we know that

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

Plugging in L and ν in Corollary 4.7 yields the corollary's claim. □

H Global optimum convergence under the gradient domination assumption

As Fazel et al. (2018); Mei et al. (2020) did for the exact policy gradient update, relying on the following gradient domination assumption, we establish a global optimum convergence guarantee and the sample complexity analysis for the stochastic vanilla PG.

Assumption H.1 (Gradient domination)**.** We say that a differentiable function J satisfies the gradient domination condition if for all $\theta \in \mathbb{R}^d$, there exists $\mu > 0$ such that

$$\frac{1}{2}\left\|\nabla J_H(\theta)\right\|^2 \;\geq\; \mu\left(J^* - J(\theta)\right). \tag{PL}$$

The gradient domination condition is also known as the Polyak-Lojasiewicz (PL) condition (Łojasiewicz, 1963). Equipped with this additional assumption, we can adapt Theorem 3 in Khaled and Richtárik (2020) and obtain the following global optimum convergence guarantee.

Theorem H.2. Suppose that Assumptions 3.1, 3.2, 3.3 and H.1 hold. Suppose that PG defined in (7) (Alg. 1) is run for $T > 0$ iterations with step size $(\eta_t)_t$ chosen as

$$\eta_t \;=\; \begin{cases} \frac{1}{b} & \text{if } T \leq \frac{b}{\mu} \;\text{ or }\; t \leq t_0 \\ \frac{2}{2b+\mu(t-t_0)} & \text{if } T \geq \frac{b}{\mu} \;\text{ and }\; t > t_0 \end{cases} \tag{137}$$

with $t_0 = \left[\frac{T}{2}\right]$ and $b = \max\{2AL/\mu, 2BL, \mu\}$. Then

$$J^* - \mathbb{E}\left[J(\theta_T)\right] \leq 16\exp\left(-\frac{\mu(T-1)}{2\max\{\frac{2AL}{\mu}, 2BL, \mu\}}\right)\left(J^* - J(\theta_0)\right) + \frac{12LC}{\mu^2 T} + \frac{26D\gamma^H}{\mu}. \tag{138}$$

Remark. Notice that for the exact full gradient update, we have Assumption 3.2 and 3.3 hold with $A = C = D = 0$ and $B = 1$. Thus under the smoothness assumption and the (PL) condition , we establish a linear convergence rate for the number of iterations to the global optimal. We recover the linear convergence rate for the softmax with entropy regularization in Theorem 6 in Mei et al. (2020) where the smoothness assumption holds and the (PL) condition holds under the path of the iterations in the exact case.

As for the stochastic vanilla PG, the dominant term in (138) is $\frac{12LC}{\mu^2 T}$. This implies that the sample complexity is $T \times H = \widetilde{\mathcal{O}}(\epsilon^{-1})$ with $T = \mathcal{O}(\epsilon^{-1})$ and $H = \log \epsilon^{-1}$.

Proof. Using the L-smoothness of J from Assumption 3.1,

$$\begin{aligned} J^* - J(\theta_{t+1}) \;&\leq\; J^* - J(\theta_t) - \langle \nabla J(\theta_t), \theta_{t+1} - \theta_t \rangle + \frac{L}{2}\left\|\theta_{t+1} - \theta_t\right\|^2 \\ &=\; J^* - J(\theta_t) - \eta_t \left\langle \nabla J(\theta_t), \widehat{\nabla}_m J(\theta_t) \right\rangle + \frac{L\eta_t^2}{2}\left\|\widehat{\nabla}_m J(\theta_t)\right\|^2. \end{aligned} \tag{139}$$

Taking expectation conditioned on θ_t and using Assumption 3.3 and H.1,

$$\begin{aligned}
\mathbb{E}_t\left[J^* - J(\theta_{t+1})\right] &\leq J^* - J(\theta_t) - \eta_t \left\langle \nabla J(\theta_t), \nabla J_H(\theta_t)\right\rangle + \frac{L\eta_t^2}{2}\mathbb{E}_t\left[\left\|\widehat{\nabla}_m J(\theta_t)\right\|^2\right] \\
&\overset{\text{(ABC)}}{\leq} J^* - J(\theta_t) - \eta_t \left\langle \nabla J_H(\theta_t) + (\nabla J(\theta_t) - \nabla J_H(\theta_t)), \nabla J_H(\theta_t)\right\rangle + \\
&\quad + \frac{L\eta_t^2}{2}\left(2A(J^* - J(\theta_t)) + B\left\|\nabla J_H(\theta_t)\right\|^2 + C\right) \\
&= (1 + L\eta_t^2 A)(J^* - J(\theta_t)) - \eta_t\left(1 - \frac{LB\eta_t}{2}\right)\left\|\nabla J_H(\theta_t)\right\|^2 + \frac{L\eta_t^2 C}{2} \\
&\quad -\eta_t \left\langle \nabla J(\theta_t) - \nabla J_H(\theta_t), \nabla J_H(\theta_t)\right\rangle \\
&\overset{\text{(PL)}}{\leq} \left(1 - 2\eta_t\mu\left(1 - \frac{LB\eta_t}{2}\right) + L\eta_t^2 A\right)(J^* - J(\theta_t)) + \frac{L\eta_t^2 C}{2} \\
&\quad -\eta_t \left\langle \nabla J(\theta_t) - \nabla J_H(\theta_t), \nabla J_H(\theta_t)\right\rangle \\
&\leq \left(1 - \frac{3\eta_t\mu}{2} + L\eta_t^2 A\right)(J^* - J(\theta_t)) + \frac{L\eta_t^2 C}{2} \\
&\quad -\eta_t \left\langle \nabla J(\theta_t) - \nabla J_H(\theta_t), \nabla J_H(\theta_t)\right\rangle \qquad (140) \\
&\overset{(9)}{\leq} \left(1 - \frac{3\eta_t\mu}{2} + L\eta_t^2 A\right)(J^* - J(\theta_t)) + \frac{L\eta_t^2 C}{2} + \eta_t D\gamma^H \\
&\leq (1 - \eta_t\mu)(J^* - J(\theta_t)) + \frac{L\eta_t^2 C}{2} + \eta_t D\gamma^H, \qquad (141)
\end{aligned}$$

where (140) is obtained by the inequality $1 - \frac{LB\eta_t}{2} \geq \frac{3}{4}$, and (141) is obtained by the inequality $L\eta_t A \leq \frac{\mu}{2}$, due to the choice of step size $\eta_t \leq \frac{1}{b}$ for all $t \geq 0$ with $b \geq 2BL, 2AL/\mu$, respectively. Here, $1 - \eta_t\mu \geq 0$ as $\eta_t \leq \frac{1}{b}$ and $b \geq \mu$.

Taking total expectation and letting $r_t \overset{\text{def}}{=} \mathbb{E}\left[J^* - J(\theta_t)\right]$ on (141), we have

$$r_{t+1} \leq (1 - \eta_t\mu)r_t + \frac{L\eta_t^2 C}{2} + \eta_t D\gamma^H. \qquad (142)$$

If $T \leq \frac{b}{\mu}$, we have $\eta_t = \frac{1}{b}$. Recursing the above inequality, we get

$$\begin{aligned}
r_T &\leq \left(1 - \frac{\mu}{b}\right) r_{T-1} + \frac{LC}{2b^2} + \frac{D\gamma^H}{b} \\
&\overset{(142)}{\leq} \left(1 - \frac{\mu}{b}\right)^T r_0 + \left(\frac{LC}{2b^2} + \frac{D\gamma^H}{b}\right)\sum_{i=0}^{T-1}\left(1 - \frac{\mu}{b}\right)^i \\
&\leq \exp\left(-\frac{\mu T}{b}\right) r_0 + \frac{LC}{2\mu b} + \frac{D\gamma^H}{\mu} \qquad (143) \\
&\overset{T \leq \frac{b}{\mu}}{\leq} \exp\left(-\frac{\mu T}{b}\right) r_0 + \frac{LC}{2\mu^2 T} + \frac{D\gamma^H}{\mu}. \qquad (144)
\end{aligned}$$

If $T \geq \frac{b}{\mu}$, as $\eta_t = \frac{1}{b}$ when $t \leq t_0$, from (143), we have

$$\begin{aligned}
r_{t_0} &\leq \exp\left(-\frac{\mu t_0}{b}\right) r_0 + \frac{LC}{2\mu b} + \frac{D\gamma^H}{\mu} \\
&\leq \exp\left(-\frac{\mu(T-1)}{2b}\right) r_0 + \frac{LC}{2\mu b} + \frac{D\gamma^H}{\mu}, \qquad (145)
\end{aligned}$$

where the last line is obtained by $t_0 = \left[\frac{T}{2}\right] \geq \frac{T-1}{2}$.

For $t > t_0$,

$$\eta_t = \frac{2}{\mu\left(\frac{2b}{\mu} + t - t_0\right)}.$$

From (142), we have

$$\begin{aligned} r_t &\leq (1-\eta_t\mu)r_{t-1} + \frac{L\eta_t^2 C}{2} + \eta_t D\gamma^H \\ &= \frac{\frac{2b}{\mu}+t-t_0-2}{\frac{2b}{\mu}+t-t_0} r_{t-1} + \frac{2LC}{\mu^2\left(\frac{2b}{\mu}+t-t_0\right)^2} + \frac{2D\gamma^H}{\mu\left(\frac{2b}{\mu}+t-t_0\right)}. \end{aligned} \tag{146}$$

Multiplying both sides by $\left(\frac{2b}{\mu}+t-t_0\right)^2$, we have

$$\begin{aligned} \left(\frac{2b}{\mu}+t-t_0\right)^2 r_t &\leq \left(\frac{2b}{\mu}+t-t_0\right)\left(\frac{2b}{\mu}+t-t_0-2\right) r_{t-1} + \frac{2LC}{\mu^2} + \frac{2D\gamma^H}{\mu}\left(\frac{2b}{\mu}+t-t_0\right) \\ &\leq \left(\frac{2b}{\mu}+t-t_0-1\right)^2 r_{t-1} + \frac{2LC}{\mu^2} + \frac{2D\gamma^H}{\mu}\left(\frac{2b}{\mu}+t-t_0\right). \end{aligned} \tag{147}$$

Let $w_t \stackrel{\text{def}}{=} \left(\frac{2b}{\mu}+t-t_0\right)^2$. Then,

$$w_t r_t \leq w_{t-1} r_{t-1} + \frac{2LC}{\mu^2} + \frac{2D\gamma^H}{\mu}\left(\frac{2b}{\mu}+t-t_0\right). \tag{148}$$

Summing up for $t = t_0+1, \cdots, T$ and telescoping, we get,

$$\begin{aligned} w_T r_T &\leq w_{t_0} r_{t_0} + \frac{2LC(T-t_0)}{\mu^2} + \frac{2D\gamma^H}{\mu}\sum_{t=t_0+1}^{T}\left(\frac{2b}{\mu}+t-t_0\right) \\ &= \frac{4b^2}{\mu^2} r_{t_0} + \frac{2LC(T-t_0)}{\mu^2} + \frac{4bD(T-t_0)\gamma^H}{\mu^2} + \frac{D\gamma^H}{\mu}(T-t_0)(T-t_0+1). \end{aligned} \tag{149}$$

Dividing both sides by w_T and using that since

$$w_T = \left(\frac{2b}{\mu}+T-t_0\right)^2 \geq (T-t_0)^2,$$

we have

$$\begin{aligned} r_T &\leq \frac{4b^2}{\mu^2 w_T} r_{t_0} + \frac{2LC(T-t_0)}{\mu^2 w_T} + \frac{4bD(T-t_0)\gamma^H}{\mu^2 w_T} + \frac{D\gamma^H}{\mu w_T}(T-t_0)(T-t_0+1) \\ &\leq \frac{4b^2}{\mu^2(T-t_0)^2} r_{t_0} + \frac{2LC}{\mu^2(T-t_0)} + \frac{4bD\gamma^H}{\mu^2(T-t_0)} + \frac{2D\gamma^H}{\mu}. \end{aligned} \tag{150}$$

By the definition of t_0, we have $T - t_0 \geq \frac{T}{2}$. Plugging this estimate, we have

$$\begin{aligned} r_T &\leq \frac{16b^2}{\mu^2 T^2} r_{t_0} + \frac{4LC + 8bD\gamma^H}{\mu^2 T} + \frac{2D\gamma^H}{\mu} \\ &\stackrel{T\geq\frac{b}{\mu}}{\leq} \frac{16b^2}{\mu^2 T^2} r_{t_0} + \frac{4LC}{\mu^2 T} + \frac{10D\gamma^H}{\mu} \\ &\stackrel{(145)}{\leq} \frac{16b^2}{\mu^2 T^2}\left(\exp\left(-\frac{\mu(T-1)}{2b}\right) r_0 + \frac{LC}{2\mu b} + \frac{D\gamma^H}{\mu}\right) + \frac{4LC}{\mu^2 T} + \frac{10D\gamma^H}{\mu} \\ &\stackrel{T\geq\frac{b}{\mu}}{\leq} 16\exp\left(-\frac{\mu(T-1)}{2b}\right) r_0 + \frac{8LC}{\mu^2 T} + \frac{16D\gamma^H}{\mu} + \frac{4LC}{\mu^2 T} + \frac{10D\gamma^H}{\mu} \\ &= 16\exp\left(-\frac{\mu(T-1)}{2b}\right) r_0 + \frac{12LC}{\mu^2 T} + \frac{26D\gamma^H}{\mu}. \end{aligned} \tag{151}$$

It remains to take the maximum of the two bounds (144) and (151) with $b = \max\{2AL/\mu, 2BL, \mu\}$. □