Sample Efficient Reinforcement Learning with REINFORCE

Junzi Zhang [*] Jongho Kim [†] Brendan O'Donoghue [‡] Stephen Boyd [§]

November 2, 2020

Abstract

Policy gradient methods are among the most effective methods for large-scale reinforcement learning, and their empirical success has prompted several works that develop the foundation of their global convergence theory. However, prior works have either required exact gradients or state-action visitation measure based mini-batch stochastic gradients with a *diverging* batch size, which limit their applicability in practical scenarios. In this paper, we consider classical policy gradient methods that compute an approximate gradient with a *single* trajectory or a *fixed size* mini-batch of trajectories, along with the widely-used REINFORCE gradient estimation procedure. By controlling the number of "bad" episodes and resorting to the classical doubling trick, we establish an anytime sub-linear high probability regret bound as well as almost sure global convergence of the average regret with an asymptotically sub-linear rate. These provide the first set of global convergence and sample efficiency results for the well-known REINFORCE algorithm and contribute to a better understanding of its performance in practice.

1 Introduction

In this paper, we study the global convergence rates of the REINFORCE algorithm [Wil92] for episodic reinforcement learning. REINFORCE is a vanilla policy gradient method that computes a stochastic approximate gradient with a single trajectory or a fixed size mini-batch of trajectories with particular choice of gradient estimator, where we use 'vanilla' here to disambiguate the method from more exotic variants such as natural policy gradient methods. REINFORCE and its variants are among the most widely used policy gradient methods in practice due to their good empirical performance and implementation simplicity [MG14, GLSM15, RMM+17, KvHW18, KvHW20]. Related methods include the actor-critic family [KT03, MBG+16] and deterministic and trust-region based variants [SLH+14, SWD+17, SLA+15].

[*]Institute for Computational & Mathematical Engineering, Stanford, USA. **Email:** junziz@stanford.edu
[†]Department of Electrical Engineering, Stanford, USA. **Email:** jkim22@stanford.edu
[‡]DeepMind, Google. **Email:** bodonoghue85@gmail.com
[§]Department of Electrical Engineering, Stanford, USA. **Email:** boyd@stanford.edu

The theoretical results for policy gradient methods have, up to recently, been restricted to convergence to local stationary points [AKLM19]. Lately, a series of works have established *global* convergence results. These recent developments cover a broad range of issues including global optimality characterization [FGKM18, BR19], convergence rates [ZKZB19, MXSS20, BR20, CCC+20], the use of function approximation [AKLM19, WCYW19, FYW20], and efficient exploration [AHKS20]. Nevertheless, prior work only guarantees convergence up to $\Theta(1/M^p)$ with a fixed mini-batch size $M > 0$ of trajectories collected when performing a single update (where $p > 0$ is $1/2$ in most cases), while global convergence is only achieved when the batch size M goes to infinity. By contrast, practical implementations of policy gradient methods typically use either a single or a fixed number of sample trajectories, which tends to perform well. In addition, prior theoretical results (for general MDPs) have used state-action visitation measure based gradient estimation (see *e.g.*, [WCYW19, (3.10)]), which are not sample efficient and are typically not used in practice.

The main purpose of this paper is to bridge this gap between theory and practice. We do this in two major ways. First, we derive performance bounds for the case of a fixed mini-batch size, rather than requiring diverging size. Second, we remove the need for the state-action visitation measure based gradient, instead using the REINFORCE gradient estimator. It is nontrivial to go from a diverging mini-batch size to a fixed one. In fact, by allowing for an arbitrarily large batch size, existing works in the literature were able to make use of IID samples to decouple the analysis into deterministic gradient descent/ascent and error control of stochastic gradient estimations. In contrast, with a single trajectory or a fixed batch size, such a decoupling is no longer feasible. In addition, the state-action visitation measure based gradient estimations are unbiased and unbounded, while REINFORCE gradient estimations are biased and bounded. Hence a key to the analysis is to deal with the bias while making better use of the boundedness. Our analysis not only addresses these challenges, but also leads to convergence results in almost sure and high probability senses, which are stronger than the expected convergence results that dominate the literature (for vanilla policy gradient methods). We also emphasize that the goal of this work is to provide a deeper understanding of a widely used algorithm, REINFORCE, with little or no modifications, rather than tweaking it to achieve near-optimal performance bounds. Lastly, our analysis is not the complete picture and several open questions about the performance of policy gradient methods remain. We discuss these issues in the conclusion.

1.1 Related work

Policy gradient methods are a large family of algorithms for reinforcement learning that directly operate on the agent policy, rather than on the action-value function [Gly86, SB18]. Examples of policy gradient methods include REINFORCE [Wil92], A3C [MBG+16], DPG [SLH+14], PPO [SWD+17], and TRPO [SLA+15], to name just a few. These methods seek to directly maximize the cumulative reward as a function of the policies, they are straightforward to implement and are amenable to function approximations. The asymptotic convergence of (some) policy gradient methods to a stationary point has long been established [SMSM00, KT03, MT01, BB01]. The rate of convergence is also known and has been improved more recently, with the help of variance reduction [PBC+18], Hessian information [SRH+19] and momentum techniques [XXLZ20, YLLZ20, PNP+20, HGPH20].

In the past two years, a line of research on the global convergence theory for policy gradient methods has emerged. By using a gradient domination property of the cumulative reward, global convergence of policy gradient methods was first established for linear-quadratic regulators [FGKM18]. For general Markov Decision Processes (MDPs), [ZKZB19] establishes convergence to approximately locally optimal (*i.e.*, second-order stationary) solutions for (vanilla) policy gradient methods. The global optimality of stationary points for general MDPs is shown in [BR19], and rates of convergence towards globally optimal solutions for (both vanilla and natural) policy gradient methods with (neural network) function approximation are derived in [AKLM19, WCYW19]. These convergence results are then improved by several very recent works focusing on exact gradient settings. In particular, [MXSS20] focuses on the more practically relevant soft-max parametrization and vanilla policy gradient methods and improves the results of [AKLM19] by removing the requirement of the relative entropy regularization and obtaining faster convergence rates; [BR20] obtains linear convergence for a general class of policy gradient methods; [CCC$^+$20] shows local quadratic convergence of natural policy gradient methods; and [ZKB$^+$20] extends the results to reinforcement learning with general utilities. For more modern policy gradient methods, [NJG17] and [ZLW19] establish the asymptotic global convergence of of TRPO, while [LCYW19] further derives the global convergence rates for PPO and TRPO. These rates are then improved in [SEM19] for TRPO with adaptive regularization terms. Very recently, [FYW20] extends these results to obtain the global convergence rates of single-timescale actor-critic methods with PPO actor updates, and [AHKS20] derives global convergence rates of a new policy gradient algorithm combining natural policy gradient methods with a policy cover technique and show that the algorithm entails better exploration behavior and hence removes the necessity for the access to a fully supported initial distribution ρ, which is assumed in most other works on global convergence of policy gradient methods (including our work). All the above works either require exact and deterministic updates or mini-batch updates with a diverging mini-batch size.

Lately, [JSW20] studies vanilla policy gradient methods using the REINFORCE gradient estimators computed with a single trajectory in each episode and obtains high probability sample complexity results, but the setting is restricted to linear-quadratic regulators and their bounds have polynomial dependency on $1/\delta$ (in contrast to our logarithmic dependency on $1/\delta$), where δ is the probability that the bounds are violated. The authors of [AYBB$^+$19] study natural policy gradient methods with a general high probability estimation oracle for state-action value functions (*i.e.*, Q-functions) in the average reward settings, and establishes high probability regret bounds for these algorithms. However, those regret bounds are not anytime (*i.e.*, requiring the total number of steps as the input) and contains a linear term proportional to the Q-function estimation errors, which is in contrast to our anytime and sub-linear regret bounds. Finally, we remark that there are also some recent results on the global convergence rates of natural policy gradient methods in adversarial settings (with full-information feedback) [CYJW19], model based natural policy gradient methods [ESRM20] as well as extensions to non-stationary [FYWX20] and multi-agent game settings [ZYB19, MRJS19, CLT19, FYCW19, GHXZ20], which are beyond the scope of this paper.

1.2 Contribution

Our major contribution can be summarized as follows. We establish the first set of global convergence results for the REINFORCE algorithm. In particular, we establish a high probability and anytime regret bound as well as almost sure global convergence of the average regret with an asymptotically sub-linear rate for REINFORCE. To our knowledge, these (almost sure and high probability) results are stronger than existing global convergence results for (vanilla) policy gradient methods in the literature. Moreover, our convergence results remove the non-vanishing $\Theta(1/M)$ term (with $M > 0$ being the mini-batch size of the trajectories) and hence show for the first time that policy gradient estimations with a single or finite number of trajectories also enjoy global convergence properties. This also leads to the first sub-linear and anytime regret bound for model-free policy gradient methods in the stochastic MDP setting. Finally, the widely-used REINFORCE gradient estimation procedure is studied, as opposed to the state-action visitation measure based estimators typically studied in the literature but rarely used in practice.

1.3 Outline

In §2 we introduce the problem setting and some preliminaries on policy gradient methods and performance criteria. In §3, we present the assumptions we need for our results and verify that they hold for REINFORCE gradient estimations with appropriate choices of hyper-parameters. Then in §4, convergence results in terms of regret bounds are established and finally, we discuss extensions of our results in §5 followed by a brief discussion about open problems and future work in §6.

2 Problem setting and preliminaries

Below we begin with our problem setting and some preliminaries on MDPs and policy optimization. For brevity we restrict ourselves to the stationary infinite-horizon discounted setting. We briefly discuss potential extensions beyond this setting in §6.

2.1 Problem setting

We consider a finite Markov decision process (MDP) $\mathcal{M}$, which is characterized by a finite state space $\mathcal{S} = \{1, \dots, S\}$, a finite action space $\mathcal{A} = \{1, \dots, A\}$, a transition probability p (with $p(s'|s,a)$ being the probability of transitioning to state s' given the current state s and action a), a reward function r (with $r(s,a)$ being the instantaneous reward when taking action a at state s), a discount factor $\gamma \in [0,1)$ and a initial state distribution $\rho \in \Delta(\mathcal{S})$. Here $\Delta(\mathcal{X})$ denotes the probability simplex over a finite set $\mathcal{X}$. A (stationary, stochastic) policy π is a mapping from $\mathcal{S}$ to $\Delta(\mathcal{A})$. We will use $\pi(a|s)$, $\pi(s,a)$ or $\pi_{s,a}$ alternatively to denote the probability of taking action a at state s following policy π. The policy π can also be viewed as an SA dimensional vector in

$$\Pi = \left\{ \pi \in \mathbf{R}^{SA} \,\middle|\, \sum\nolimits_{a=1}^{A} \pi_{s,a} = 1 \, (\forall s \in \mathcal{S}), \; \pi_{s,a} \geq 0 \, (\forall s \in \mathcal{S}, \; a \in \mathcal{A}) \right\}. \tag{1}$$

Notice that here we use the double indices s and a for notational convenience. We use $\pi(s,\cdot) \in \mathbf{R}^A$ to denote the sub-vector $(\pi(s,1),\ldots,\pi(s,A))$. We also assume that $r(s,a)$ is deterministic for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$ for simplicity, although our results hold for any r with an almost sure uniform bound. Here r can be similarly viewed as an SA-dimensional vector. Without loss of generality, we assume that $r(s,a) \in [0,1]$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, which is a common assumption [JOA10, AKLM19, MXSS20, EDM03, JAZBJ18]. We also assume that ρ is component-wise positive, as is assumed in [BR19].

Given a policy $\pi \in \Pi$, the expected cumulative reward of the MDP is defined as

$$F(\pi) = \mathbf{E}\sum\nolimits_{t=0}^{\infty} \gamma^t r(s_t, a_t), \tag{2}$$

where $s_0 \sim \rho$, $a_t \sim \pi(\cdot|s_t)$, $s_{t+1} \sim p(\cdot|s_t,a_t)$, $\forall t \geq 0$, and the goal is to find a policy π which solves the following optimization problem:

$$\text{maximize}_{\pi\in\Pi} \quad F(\pi). \tag{3}$$

Any policy $\pi^\star \in \text{argmax}_{\pi\in\Pi} F(\pi)$ is said to be optimal, and the corresponding optimal objective value is denoted as $F^\star = F(\pi^\star)$. Note that in the literature, $F(\pi)$ is also commonly written as V_ρ^π and referred to as the value function. Here we hide the dependency on ρ as it is fixed throughout the paper.

2.2 Vanilla policy gradient method and REINFORCE algorithm

When the transition probability p and reward r are fully known, problem (3) reduces to solving an MDP, in which case various classical algorithms are available, including value iteration and policy iteration [Ber17]. In this paper, we consider the episodic reinforcement learning setting in which the agent accesses p and r by interacting with the environment over successive episodes, *i.e.*, the agent access the environment in the form of a ρ-restart model [SEM19], which is commonly adopted in the policy gradient literature [K$^+$03]. In addition, we focus on the REINFORCE algorithm, a representative policy gradient method.

Policy parametrization and surrogate objectives. Here we consider parametrizing the policy with parameter $\theta \in \Theta$, *i.e.*, $\pi_\theta : \Theta \to \Pi$, and take the following (regularized) optimization problem as an approximation to (3):

$$\text{maximize}_{\theta\in\Theta} \quad L_\lambda(\theta) = F(\pi_\theta) + \lambda R(\theta), \tag{4}$$

where $\lambda \geq 0$ and $R : \Theta \to \mathbf{R}$ is a differentiable regularization term that improves convergence, to be specified later. Although our ultimate goal is still to solve the original problem (3) this regularized optimization problem is a useful surrogate and our approach will be to tackle problem (4) with progressively smaller λ regularization penalties, thereby converging to solving the actual problem we care about.

Policy gradient method. In each episode n, the policy gradient method directly performs an online stochastic gradient ascent update on a surrogate objective $L_{\lambda^n}(\theta)$, *i.e.*,

$$\theta^{n+1} = \theta^n + \alpha^n \widehat{\nabla}_\theta L_{\lambda^n}(\theta^n), \tag{5}$$

where α^n is the step-size and λ^n is the regularization parameter. Here the stochastic gradient $\widehat{\nabla}_\theta L_{\lambda^n}(\theta^n)$ is computed by sampling a *single* trajectory τ^n following policy π_{θ^n} from $\mathcal{M}$ with the initial state distribution ρ. Here $\tau^n = (s_0^n, a_0^n, r_0^n, s_1^n, a_1^n, r_1^n, \dots, s_{H^n}^k, a_{H^n}^n, r_{H^n}^n)$, where H^n is a finite (and potentially random) stopping time of the trajectory (to be specified below), $s_0^n \sim \rho$, $a_t^n \sim \pi_{\theta^n}(\cdot|s_t^n)$, $s_{t+1}^n \sim p(\cdot|s_t^n, a_t^n)$ and $r_t^n = r(s_t^n, a_t^n)$ for all $t = 0, \dots, H^n$. We summarize the generic policy gradient method (with single trajectory gradient estimates) in Algorithm 1. An extension to mini-batch scenarios will be discussed in §5. As is always (implicitly) assumed in the literature of episodic reinforcement learning (*e.g.*, *cf.* [MT01]), given the current policy, we assume that the sampled trajectory is conditionally independent of all previous policies and trajectories.

Algorithm 1 **Policy Gradient Method with Single Trajectory Estimates**

1: **Input:** initial parameter θ^0, step-sizes α^n and regularization parameters λ^n $(n \geq 0)$.
2: **for** $n = 0, 1, \dots$ **do**
3: Choose H^n, sample trajectory τ^n from $\mathcal{M}$ following policy π_{θ^n}, and compute an approximate gradient $\widehat{\nabla}_\theta L_{\lambda^n}(\theta^n)$ of L_{λ^n} using trajectory τ^n.
4: Update $\theta^{n+1} = \theta^n + \alpha^n \widehat{\nabla}_\theta L_{\lambda^n}(\theta^n)$.
5: **end for**

REINFORCE algorithm. There are several ways of choosing the stochastic gradient operator $\widehat{\nabla}_\theta$ in the policy gradient method, and the well-known REINFORCE algorithm [Wil92] corresponds to a specific family of estimators based on the policy gradient theorem [SMSM00] (*cf.* §3). Other common alternatives include zeroth order/random search [FGKM18, MPB+18] and actor-critic [KT03] approximations. One may also choose to parametrize the policy as a mapping from the parameter space to a specific action, which would then result in deterministic policy gradient approximations [SLH+14].

Although our main goal is to study the REINFORCE algorithm, our analysis indeed holds for rather generic stochastic gradient estimates. In the next section, we introduce the (mild) assumptions needed for our convergence analysis and the detailed gradient estimation procedures in the REINFORCE algorithm, and then verify that the assumptions do hold for these gradient estimations.

2.3 Phased learning and performance criteria

Phased learning. To facilitate the exposition below, we divide the optimization in Algorithm 1 into successive phases $l = 0, 1, \dots$, each with length $T_l > 0$. We then fix the regularization coefficient λ_l within each phase $l \geq 0$. In addition, a post-processing step is enforced at the end of each phase to produce the initialization of the next phase. The resulting algorithm is described in Algorithm 2. Here the trajectory is denoted as $\tau^{l,k} = (s_0^{l,k}, a_0^{l,k}, r_0^{l,k}, \dots, s_{H^{l,k}}^{l,k}, a_{H^{k,l}}^{l,k}, r_{H^{l,k}}^{l,k})$, and we will refer to $\theta^{l,k}$ as the (l,k)-th iterate hereafter. The post-processing function is required to guarantee that the resulting policy π_θ

is lower bounded by a pre-specified tolerance $\epsilon_{\rm pp} \in (0, 1/A]$ to ensure that the regularization is bounded (*cf.* Algorithm 3 for a formal description and §3.1 for an example realization).

Note that here the k-th episode in phase l corresponds to the n-th episode in the original indexing with $n = \sum_{j=0}^{l-1} T_j + k$. For notational compactness below, for $\mathcal{T} = \{T_j\}_{j=0}^{\infty}$, we define $B_{\mathcal{T}} : \mathbf{Z}_+ \times \mathbf{Z}_+ \to \mathbf{Z}_+$, where $B_{\mathcal{T}}(l,k) = \sum_{j=0}^{l-1} T_j + k$ maps the double index (l,k) to the corresponding original episode number, with $\mathbf{dom}\, B_{\mathcal{T}} = \{(l,k) \in \mathbf{Z}_+ \times \mathbf{Z}_+ \,|\, 0 \leq k \leq T_l - 1\}$. The mapping $B_{\mathcal{T}}$ is a bijection and we denote its inverse by $G_{\mathcal{T}}$.

Algorithm 2 **Phased Policy Gradient Method**

1: **Input:** initial parameter $\tilde{\theta}^{0,0}$, step-sizes $\alpha^{l,k}$, regularization parameters λ^l, phase lengths T_l $(l,\, k \geq 0)$ and post-processing tolerance $\epsilon_{\rm pp} \in (0, 1/A]$.

2: Set $\theta^{0,0} = \texttt{PostProcess}(\tilde{\theta}^{0,0}, \epsilon_{\rm pp})$.

3: **for** phase $l = 0, 1, 2, \ldots$ **do**

4:     **for** episode $k = 0, 1, \ldots, T_l - 1$ **do**

5:         Choose $H^{l,k}$, sample trajectory $\tau^{l,k}$ from $\mathcal{M}$ following policy $\pi_{\theta^{l,k}}$, and compute an approximate gradient $\widehat{\nabla}_\theta L_{\lambda^l}(\theta^{l,k})$ of L_{λ^l} using trajectory $\tau^{l,k}$.

6:         Update $\theta^{l,k+1} = \theta^{l,k} + \alpha^{l,k} \widehat{\nabla}_\theta L_{\lambda^l}(\theta^{l,k})$.

7:     **end for**

8:     Set $\theta^{l+1,0} = \texttt{PostProcess}(\theta^{l,T_l}, \epsilon_{\rm pp})$.

9: **end for**

Algorithm 3 $\texttt{PostProcess}(\theta, \epsilon_{\rm pp})$

Input: $\epsilon_{\rm pp} \in (0, 1/A]$, $\theta \in \Theta$.

Return θ' (near θ) such that $\pi_{\theta'}(s,a) \geq \epsilon_{\rm pp}$ for each $s,\, a \in \mathcal{S} \times \mathcal{A}$.

Performance criteria. The criterion we adopt to evaluate the performance of Algorithm 2 is *regret*. For any $N \geq 0$, the regret up to episode N is defined as the cumulative sub-optimality of the policy over the N episodes. Formally, we define

$$\mathbf{regret}(N) = \sum\nolimits_{\{(l,k) | B_{\mathcal{T}}(l,k) \leq N\}} F^\star - \hat{F}^{l,k}(\pi_{\theta^{l,k}}). \tag{6}$$

Here the summation is over all (l,k)-th iterates whose corresponding original episode numbers are smaller or equal to N, and

$$\hat{F}^{l,k}(\pi_{\theta^{l,k}}) = \mathbf{E}_{l,k} \sum\nolimits_{t=0}^{H^{l,k}} \gamma^t r(s_t^{l,k}, a_t^{l,k}),$$

where $s_0 \sim \rho$, $a_t^{l,k} \sim \pi_{\theta^{l,k}}(\cdot|s_t^{l,k})$, $s_{t+1}^{l,k} \sim p(\cdot|s_t^{l,k}, a_t^{l,k})$, $\forall t \geq 0$, and $\mathbf{E}_{l,k}$ denotes the conditional expectation given the (l,k)-th iteration $\theta^{l,k}$. Notice that the regret defined above takes into account the fact that the trajectories are stopped/truncated to have finite horizons $H^{l,k}$, which characterizes the actual loss caused by sampling the trajectories in line 5 of Algorithm 2. We remark that all our regret bounds remain correct up to lower order terms when we replace $\hat{F}^{l,k}$ with F or an expectation-free version.

Similarly, we also define the single phase version of regret as follows. The regret up to episode $K \in \{0, \dots, T_l - 1\}$ in phase l is defined as

$$\mathbf{regret}_l(K) = \sum\nolimits_{k=0}^{K} F^\star - \hat{F}^{l,k}(\pi_{\theta^{l,k}}). \tag{7}$$

Notice that (6) and (7) are connected via

$$\mathbf{regret}(N) = \sum_{l=0}^{l_N - 1} \mathbf{regret}_l(T_l - 1) + \mathbf{regret}_{l_N}(k_N), \tag{8}$$

where $(l_N, k_N) = G_{\mathcal{T}}(N)$.

We provide high probability regret bounds in §4. We remark that a regret bound of the form $\mathbf{regret}(N)/(N+1) \le R$ (for some $R > 0$) immediately implies that $\min_{l,k:\, B_{\mathcal{T}}(l,k) \le N} F^\star - F(\pi_{\theta^{l,k}}) \le R$, where the latter is also a commonly adopted performance criteria in the literature [AKLM19, WCYW19].

3 Assumptions and REINFORCE gradients

3.1 Assumptions

Here we list a few fundamental assumptions that we require for our analysis.

Assumption 1 (Setting)**.** *The regularization term is a log-barrier, i.e.,*

$$R(\theta) = \frac{1}{SA} \sum\nolimits_{s \in \mathcal{S}, a \in \mathcal{A}} \log(\pi_\theta(s,a)),$$

and the policy is parametrized to be a soft-max, i.e., $\pi_\theta(s,a) = \frac{\exp(\theta_{s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta_{s,a'})}$*, with* $\Theta = \mathbf{R}^{SA}$.

The first assumption concerns the form of the policy parameterization and the regularization. Notice that the regularization term here can also be seen as a relative entropy/KL regularization (with a uniform distribution policy reference) [AKLM19]. Such kind of regularization terms are also widely adopted in practice (although typically with variations) [PMA10, SCA17].

With Assumption 1, the post-processing function in Algorithm 3 can be for example realized by first calculating $\hat{\pi} = \epsilon_{\mathrm{pp}} \mathbf{1} + (1 - A\epsilon_{\mathrm{pp}})\pi_\theta$, and then return θ' with $\theta'_{s,a} = \log \hat{\pi}_{s,a} + c_s$. Here $\mathbf{1}$ is an all-one vector and $c_s \in \mathbf{R}$ ($s = 1, \dots, S$) are arbitrary real numbers.

Assumption 2 (Technical)**.** *There exist constants* C, C_1, C_2, M_1, $M_2 > 0$*, such that for all* l, $k \ge 0$*, we have* $\|\widehat{\nabla}_\theta L_{\lambda^l}(\theta^{l,k})\|_2 \le C_1$ *almost surely and that*

$$\nabla_\theta L_{\lambda^l}(\theta^{l,k})^T \mathbf{E}_{l,k} \widehat{\nabla}_\theta L_{\lambda^l}(\theta^{l,k}) \ge C_2 \|\nabla_\theta L_{\lambda^l}(\theta^{l,k})\|_2^2 - \delta_{l,k}, \tag{9}$$

$$\mathbf{E}_{l,k} \|\widehat{\nabla}_\theta L_{\lambda^k}(\theta^{l,k})\|_2^2 \le M_1 + M_2 \|\nabla_\theta L_{\lambda^l}(\theta^{l,k})\|_2^2, \tag{10}$$

where $\sum_{k=0}^{T_l - 1} \delta_{l,k}^2 \le C$, $\forall\, l \ge 0$*. In addition,* $H^{l,k} \ge \log_{1/\gamma}(k+1)$, $\forall\, l$, $k \ge 0$.

The second assumption requires that the gradient estimates are almost surely bounded, nearly unbiased and satisfy a bounded second-order moment growth condition. This is a slight generalization of standard assumptions in the stochastic gradient descent literature [BCN18]. Additionally, we also require that the trajectory lengths $H^{l,k}$ are at least logarithmically growing in k to control the loss of rewards due to truncation. For notational simplicity, hereafter we omit to mention the trajectory sampling (*i.e.*, $s_0 \sim \rho$, $a_t^{l,k} \sim \pi_{\theta^{l,k}}(\cdot|s_t^{l,k})$, $s_{t+1}^{l,k} \sim p(\cdot|s_t^{l,k}, a_t^{l,k})$, $\forall t \geq 0$) when we write down $\mathbf{E}_{l,k}$.

Notice that Assumption 2 immediately holds if $\widehat{\nabla}_\theta L_{\lambda^l}(\theta^{l,k})$ is unbiased and has a bounded second-order moment. We have implicitly assumed that L_λ is differentiable, which we can do due to the following lemma:

Proposition 1 ([AKLM19, Lemma E.4]). *Under Assumption 1, L_λ is strongly smooth with parameter $\beta_\lambda = \frac{8}{(1-\gamma)^3} + \frac{2\lambda}{S}$, i.e., $\|\nabla_\theta L_\lambda(\theta) - \nabla_\theta L_\lambda(\theta')\|_2 \leq \beta_\lambda \|\theta - \theta'\|_2$ for any θ, $\theta' \in \mathbf{R}^{SA}$.*

3.2 REINFORCE gradient estimations

Now we introduce REINFORCE gradient estimation with baselines, and specify the hyper-parameters under which the technical Assumption 2 holds, when operating under the setting Assumption 1.

REINFORCE gradient estimation with log-regularization takes the following form:

$$\begin{aligned}\widehat{\nabla}_\theta L_{\lambda^l}(\theta^{l,k}) = &\sum_{t=0}^{\lfloor \beta H^{l,k} \rfloor} \gamma^t (\widehat{Q}^{l,k}(s_t^{l,k}, a_t^{l,k}) - b(s_t^{l,k})) \nabla_\theta \log \pi_{\theta^{l,k}}(a_t^{l,k}|s_t^{l,k}) \\ &+ \frac{\lambda^l}{SA} \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \nabla_\theta \log \pi_{\theta^{l,k}}(a|s),\end{aligned} \tag{11}$$

where $\beta \in (0,1)$, $\widehat{Q}^{l,k}(s_t^{l,k}, a_t^{l,k}) = \sum_{t'=t}^{H^{l,k}} \gamma^{t'-t} r_{t'}^{l,k}$, and the second term above corresponds to the gradient of the regularization $R(\theta)$. Notice that here the outer summation is only up to $\lfloor \beta H^{l,k} \rfloor$, which ensures that $\hat{Q}^{l,k}(s_t^{l,k}, a_t^{l,k})$ is sufficiently accurate. Here $b : \mathcal{S} \to \mathbf{R}$ is called the baseline, and is required to be independent of the trajectory $\tau^{l,k}$ [AJK19, §4.1]. The purpose of subtracting b from the approximate Q-values is to (potentially) reduce the variance of the "plain" REINFORCE gradient estimation, which corresponds to the case when $b = 0$.

With this we have the following result, the proof of which can be found in the Appendix.

Lemma 2. *Suppose that Assumption 1 holds, $\beta \in (0,1)$, and that for all l, $k \geq 0$, $\lambda^l \leq \bar{\lambda}$,*

$$H^{l,k} \geq \frac{2 \log_{1/\gamma} \left(\frac{8(k+1)}{(1-\gamma)^3} \right)}{3 \min\{\beta, 1-\beta\}} (= \Theta(\log(k+1))). \tag{12}$$

Assume in addition that $|b(s)| \leq B$ for any $s \in \mathcal{S}$, where $B > 0$ is a constant. Then for the gradient estimation (11)*, Assumption 2 holds with*

$$C = 16 \left(\frac{1}{(1-\gamma)^2} + \bar{\lambda} \right)^2, \quad C_1 = \frac{2(1 + B(1-\gamma))}{(1-\gamma)^2} + 2\bar{\lambda},$$
$$C_2 = 1, \quad M_1 = \frac{32}{(1-\gamma)^4} + \bar{V}_b, \quad M_2 = 2.$$

and $\delta_{l,k} = \left(\frac{2}{(1-\gamma)^2} + 2\bar{\lambda}\right)(k+1)^{-\frac{2}{3}}$, $\forall\ l,\ k \geq 0$. *Here* $\bar{V}_b \in \left[0, 4\left(\frac{1+B(1-\gamma)}{(1-\gamma)^2} + \bar{\lambda}\right)^2\right]$ *is the uniform upper bound on the variances of policy gradient estimations in the form of* (11).

This result extends without modification to non-stationary baselines $b_t^{l,k}$, as long as each $b_t^{l,k}$ is independent of trajectory $\tau^{l,k}$ and $|b_t^{l,k}(s)| \leq B$ for any $t,\ l,\ k \geq 0$. Note that the explicit upper bound on $\bar{V}_b$ is pessimistic, and in practice $\bar{V}_b$ is usually much smaller than $\bar{V}_0$ with appropriate choices of baselines (*e.g.*, the adaptive reinforcement baselines [Wil92, ZHNS11]), although the latter has a smaller upper bound as stated in Lemma 2.

4 Main convergence results

4.1 Preliminary tools

We first present some preliminary tools required for our analysis.

Non-convexity and control of "bad" episodes. One of the key difficulties in applying policy gradient methods to solve an MDP problem towards global optimality is that problem (3) is in general non-convex [AKLM19]. Fortunately, we have the following result, which connects the gradient of the surrogate objective L_λ with the global optimality gap of the original optimization problem (3).

Proposition 3 ([AKLM19, Theorem 5.3])**.** *Under Assumption 1, for any* $\epsilon > 0$, *suppose that we have* $\|\nabla_\theta L_\lambda(\theta)\|_2 \leq \epsilon$ *and that* $\epsilon \leq \lambda/(2SA)$. *Then* $F^\star - F(\pi_\theta) \leq \frac{2\lambda}{1-\gamma}\left\|\frac{d_\rho^{\pi^\star}}{\rho}\right\|_\infty$.

Here for any policy $\pi \in \Pi$, $d_\rho^\pi = (1-\gamma)\sum_{t=0}^\infty \gamma^t \mathbf{Prob}_\pi(s_t = s | s_0 \sim \rho)$ is the discounted state visitation distribution, where $\mathbf{Prob}_\pi(s_t = s|s_0 \sim \rho)$ is the probability of arriving at s in step t starting from $s_0 \sim \rho$ following policy π in $\mathcal{M}$. In addition, the division in $d_\rho^{\pi^\star}/\rho$ is component-wise.

Now motivated by Proposition 3, when analyzing the regret up to episode K in phase l, we define the following set of "bad episodes":

$$I^+ = \{k \in \{0,\dots,K\} \mid \|\nabla_\theta L_{\lambda^l}(\theta^{l,k})\|_2 \geq \lambda^l/(2SA)\}. \tag{13}$$

Then one can show that for any $\epsilon > 0$, if we choose $\lambda^l = \frac{\epsilon(1-\gamma)}{2\|d_\rho^{\pi^\star}/\rho\|_\infty}$, we have that $F^\star - F(\pi_{\theta^{l,k}}) \leq \epsilon$ for any $k \in \{0,\dots,K\}\backslash I^+$, while $F^\star - F(\pi_{\theta^{l,k}}) \leq 1/(1-\gamma)$ holds trivially for $k \in I^+$ due to the assumption that the rewards are between 0 and 1. We then establish a sub-linear (in K) bound the size of I^+, which serves as the key stepping stone for the overall sub-linear regret bound. The details of these arguments can be found in the Appendix.

Doubling trick. The second tool is a classical doubling trick that is commonly adopted in the design of online learning algorithms [BK18, BGH20], which can be used to stitch together the regret over multiple learning phases in Algorithm 2.

Notice that Proposition 3 suggests that for any pre-specified tolerance ϵ, one can select λ proportional to ϵ and then run (stochastic) gradient ascent to drive $F^\star - F(\pi_\theta)$ below the tolerance. To obtain the eventual convergence and regret bound in the long run we apply the doubling trick, which specifies a growing phase length sequence with $T_{l+1} \approx 2T_l$ in Algorithm 2 and a suitably decaying sequence of regularization parameters $\{\lambda^l\}_{l=0}^{\infty}$.

From high probability to almost sure convergence. The last tool is an observation that an arbitrary anytime sub-linear high probability regret bound with logarithmic dependency on $1/\delta$ immediately leads to almost sure convergence of the average regret with a corresponding asymptotic rate. Although such an observation seems to be informally well-known in the theoretical computer science community, we provide a compact formal discussion below for self-contained-ness.

Lemma 4. *Suppose that* $\forall\ \delta > 0$, *with probability at least* $1 - \delta$, $\forall\ N \geq 0$, *we have*

$$\mathbf{regret}(N) \leq d_1(N + c)^{d_2}(\log(N/\delta))^{d_3} + d_4(\log N)^{d_5} \tag{14}$$

for some constants c, d_1, d_3, d_4, $d_5 \geq 0$ *and* $d_2 \in [0, 1)$. *Then we also have*

$$\mathbf{Prob}\left(\exists\ \bar{N} \in \mathbf{Z}_+,\ \textit{such that}\ \forall\ N \geq \bar{N},\ A_N\ \textit{holds}\right) = 1,$$

where the events $A_N = \{\mathbf{regret}(N)/(N + 1) \leq (*)\}$, *and*

$$(*) = d_1 N^{-(1-d_2)} \left(1 + \frac{c}{N}\right)^{d_2} (3 \log N)^{d_3} + \frac{d_4(\log N)^{d_5}}{N}.$$

To put it another way, we have

$$\lim_{N\to\infty} \mathbf{regret}(N)/(N + 1) = 0 \quad \textit{almost surely}$$

with an asymptotic rate of $(*)$.

Proof. The proof is a direct application of the well-known Borel-Cantelli lemma [Kle13]. Let $\delta_N = 1/N^2$ and define the events $\{\bar{A}_N\}_{N\geq 0}$ as

$$\bar{A}_N = \{\mathbf{regret}(N) > d_1(N + c)^{d_2}(\log(N/\delta_N))^{d_3} + d_4(\log N)^{d_5}\}.$$

Then $\mathbf{Prob}(\bar{A}_N) \leq \delta_N$, and hence $\sum_{N=1}^{\infty} \mathbf{Prob}(\bar{A}_N) \leq \sum_{N=1}^{\infty} 1/N^2 < \infty$. Hence by Borel-Cantelli lemma, we have

$$\mathbf{Prob}(\bar{A}_N \text{ occurs infinitely often}) = 0.$$

Finally, by noticing that the complement of $\bar{A}_N$ is a subset of A_N, the proof is complete. $\square$

Notice that here we restrict the right-hand side of (14) to a rather specific form simply because our bounds below are all of this form. However, similar results hold for much more general forms of bounds, as can be seen from the simple proof above.

4.2 Regret analysis

In this section, we establish the regret bound of Algorithm 2, when used with the REINFORCE gradient estimator from §3.2. We begin by bounding the regret of a single phase and then use the doubling trick to combine these into the overall regret bound.

Big-O notation. Before we proceed with our main results, we first clarify the precise definitions of the Big-O notation used in our statements below. Let $f,\, g : \mathbf{Z}_+ \times \mathbf{R}^d \to \mathbf{R}_+$ be functions of the total number of episodes N and some problem and algorithm dependent quantities (written jointly as a vector) $U \in \mathbf{R}^d$. Then we write $g(N;U) = O(f(N;U))$ to indicate that there exist constants $W > 0$ and $N_0 \in \mathbf{Z}_+$ (independent of N and U), such that $g(N;U) \leq Wf(N;U)$ for all $N \geq N_0$.

Single phase analysis. We begin by bounding the regret defined in (7) of each phase in Algorithm 2. Note that a single phase in Algorithm 2 is exactly Algorithm 1 terminated in episode T_l, with $\lambda^n = \lambda^l$ for all $n \geq 0$ and $\theta^0 = \theta^{l,0}$. Also notice that for a given phase $l \geq 0$, in order for Theorem 5 below to hold, we actually only need the conditions in Assumption 2 to be satisfied for this specific l.

Theorem 5. *Under Assumptions 1 and 2, for phase $l \geq 0$ suppose that we choose $\alpha^{l,k} = C_{l,\alpha}\frac{1}{\sqrt{k+3}\log_2(k+3)}$ for some $C_{l,\alpha} \in (0, C_2/(M_2\beta_{\lambda^l})]$. Then for any $\epsilon > 0$, if we choose $\lambda^l = \frac{\epsilon(1-\gamma)}{2\left\|d_\rho^{\pi^\star}/\rho\right\|_\infty}$, then $\forall\ \delta \in (0,1)$, with probability at least $1-\delta$, for any $K \in \{0,\dots,T_l - 1\}$, we have*

$$\mathbf{regret}_l(K) \leq U_1 \frac{\sqrt{K+1}\log_2(K+3)\sqrt{\log(2/\delta)}}{\epsilon^2} + (K+1)\epsilon + \frac{2\gamma}{1-\gamma}\log(K+3). \tag{15}$$

Here the constant U_1 only depends on the underlying MDP $\mathcal{M}$, phase initialization $\theta^{l,0}$ and the constants C, C_1, C_2, M_1, $C_{l,\alpha}$, λ^l.

The proof as well as a more formal statement of Theorem 5 with details of the constants (*cf.* Theorem 10) are deferred to the Appendix. Here the constant β_λ is the smoothness constant from Proposition 1. We remark that when ϵ is fixed, the regret bound (15) can be seen as a sub-linear (in K as $K \to \infty$) regret term plus an error term $(K+1)\epsilon + \frac{2\gamma}{1-\gamma}\log(K+3)$. Alternatively, one can interpret it as follows:

$$\mathbf{regret}_l(K)/(K+1) \leq U_1 \frac{\log_2(K+3)\sqrt{\log(2/\delta)}}{\sqrt{K+1}\epsilon^2} + \frac{2\gamma}{1-\gamma}\frac{\log(K+3)}{K+1} + \epsilon.$$

Namely, the average regret in episode l converges to the pre-specified tolerance ϵ at a sub-linear rate (as $K \to \infty$).

Overall regret bound. Now we stitch together the single phase regret bounds established above to obtain the overall regret bound of Algorithm 2, with the help of the doubling trick. This leads to the following theorem.

Theorem 6 (Regret for REINFORCE)**.** *Under Assumption 1, suppose that for each $l \geq 0$, we choose $\alpha^{l,k} = C_{l,\alpha}\frac{1}{\sqrt{k+3}\log_2(k+3)}$, with $C_{l,\alpha} \in [1/(2\beta_{\bar{\lambda}}), 1/(2\beta_{\lambda^l})]$ and $\bar{\lambda} = \frac{1-\gamma}{2\|d_\rho^{\pi^\star}/\rho\|_\infty}$, and choose $T_l = 2^l$, $\epsilon^l = T_l^{-1/6} = 2^{-l/6}$, $\lambda^l = \frac{\epsilon^l(1-\gamma)}{2\left\|d_\rho^{\pi^\star}/\rho\right\|_\infty}$ and $\epsilon_{\rm pp} = 1/(2A)$. In addition, suppose that (11) is adopted to evaluate $\widehat{\nabla}_\theta L_{\lambda^l}(\theta^{l,k})$, with $\beta \in (0,1)$, $|b(s)| \leq B$ for any $s \in \mathcal{S}$ (where*

$B > 0$ is a constant), and that (12) holds for $H^{l,k}$ for all l, $k \geq 0$. Then we have for any $\delta \in (0,1)$, with probability at least $1-\delta$, for any $N \geq 0$, we have

$$\mathbf{regret}(N) = O\left(\frac{S^2A^2}{(1-\gamma)^3}\left\|\frac{d_\rho^{\pi^\star}}{\rho}\right\|_\infty^2 N^{\frac{5}{6}}(\log(N/\delta))^{\frac{5}{2}}\right). \tag{16}$$

In addition, we have

$$\lim_{N\to\infty} \mathbf{regret}(N)/(N+1) = 0 \quad \textit{almost surely} \tag{17}$$

with an asymptotic rate of $O\left(\frac{S^2A^2}{(1-\gamma)^3}\left\|d_\rho^{\pi^\star}/\rho\right\|_\infty^2 N^{-\frac{1}{6}}(\log N)^{\frac{5}{2}}\right)$.

Note that the almost sure convergence (17) is immediately implied by the high probability bound (16) via Lemma 4. Here for clarity, we have restricted the statement to the case when we use the REINFORCE gradient estimation from §3.2. A more general counterpart result can be found in Appendix B.2, from which Theorem 6 is immediately implied. See also Appendix C for a more formal statement of the regret bound (*cf.* Corollary 12) for REINFORCE with detailed constants.

Notice that compared to the single phase regret bound in (15), the overall regret bound in (16) now gets rid of the dependency on a pre-specified tolerance $\epsilon > 0$. This should be attributed to the adaptivity in the regularization parameter sequence. Also notice that here we have followed the convention of the reinforcement learning literature to make all the problem dependent quantities (*e.g.*, γ, S, A, etc.) explicit in the big-O notation.

One crucial difference between our regret bound and those in the existing literature of model-free policy gradient methods in the stochastic MDP settings (which are sometimes not stated in the form of regret, but can be easily deduced from their proofs in those cases) is that the previous results contain a non-vanishing $\Theta(1/M^p)$ term, with M being the mini-batch size (of the trajectories) and $p > 0$ being some exponent (with a typical value of $1/2$). As a result, these regret bounds are linear in the total number of episodes N. By removing such non-vanishing terms, we obtain the first sub-linear regret bound.

5 Extension to mini-batch updates

We now consider extending our previous results to mini-batch settings, by modifying Algorithm 2 as follows. Firstly, in each inner iteration, instead of sampling only one trajectory in line 5, we sample $M \geq 1$ independent trajectories $\tau_1^{l,k}, \ldots, \tau_M^{l,k}$ from $\mathcal{M}$ following policy $\pi_{\theta^{l,k}}$ and then compute an approximate gradient $\widehat{\nabla}_\theta^{(i)} L_{\lambda^l}(\theta^{l,k})$ ($i = 1, \ldots, M$) using each of these M trajectories. We then modify the update in line 6 as

$$\theta^{l,k+1} = \theta^{l,k} + \alpha^{l,k}\frac{1}{M}\sum_{i=1}^M \widehat{\nabla}_\theta^{(i)} L_{\lambda^l}(\theta^{l,k}).$$

See Algorithm 4 in Appendix D for a formal description of the modified algorithm.

Then we have the following lemma, which transfers guarantees on $\widehat{\nabla}_\theta^{(i)} L_{\lambda^l}(\theta^{l,k})$ ($i = 1,\dots,M$) to the averaged gradient estimation $\frac{1}{M}\sum_{i=1}^M \widehat{\nabla}_\theta^{(i)} L_{\lambda^l}(\theta^{l,k})$. The proof follows directly from the fact that the variance of the sum of independent random variables is the sum of the variances, and is thus omitted.

Lemma 7. *Suppose that each $\widehat{\nabla}_\theta^{(i)} L_{\lambda^l}(\theta^{l,k})$ ($i = 1,\dots,M$) is computed using* (11) *with the corresponding trajectory, and that the same assumptions as in Lemma 2 hold. Then Assumption 2 also holds for $\widehat{\nabla}_\theta L_{\lambda^l}(\theta^{l,k}) = \frac{1}{M}\sum_{i=1}^M \widehat{\nabla}_\theta^{(i)} L_{\lambda^l}(\theta^{l,k})$ with the same constants $C, C_1, C_2, M_2, \delta_{l,k}$ and $\bar{V}_b$ as in Lemma 2, while $M_1 = \frac{32}{(1-\gamma)^4} + \frac{\bar{V}_b}{M}$.*

Regret with mini-batches. Notice that since each inner iteration (in Algorithm 4) now consists of M episodes, we need to slightly modify the definition of the regret up to episode N ($N \geq 0$) as follows:

$$\begin{aligned}\mathbf{regret}(N;M) = &\sum\nolimits_{\left\{(l,k)\mid B_{\mathcal{T}}(l,k)\leq \lfloor \frac{N}{M}\rfloor - 1\right\}} M(F^\star - \hat{F}^{l,k}(\pi_{\theta^{l,k}})) \\ &+ \left(N - M\left\lfloor \frac{N}{M}\right\rfloor\right)(F^\star - \hat{F}^{l,k}(\pi_{\theta^{l_{N,M},k_{N,M}}})),\end{aligned} \tag{18}$$

where $(l_{N,M}, k_{N,M}) = G_{\mathcal{T}}(\lfloor N/M \rfloor)$ and $\hat{F}^{l,k}(\pi_{\theta^{l,k}})$ is the same as in (6). The above definition accounts for the fact that each of the M episodes in an inner iteration/step (l,k) corresponds to the same iterate $\theta^{l,k}$ and hence has the same contribution to the regret. The second term on the right-hand side accounts for the contribution of the (remaining) $N - M\lfloor N/M \rfloor$ episodes (among a total of M episodes) in inner iteration/step $(l_{N,M}, k_{N,M})$.

Then the following regret bound can be established.

Corollary 8 (Regret for mini-batch REINFORCE)**.** *Under Assumption 1, suppose that for each $l \geq 0$, we choose $\alpha^{l,k} = C_{l,\alpha}\frac{1}{\sqrt{k+3}\log_2(k+3)}$, with $C_{l,\alpha} \in [1/(2\beta_{\bar{\lambda}}), 1/(2\beta_{\lambda^l})]$ and $\bar{\lambda} = \frac{1-\gamma}{2\|d_\rho^{\pi^\star}/\rho\|_\infty}$, and choose $T_l = 2^l$, $\epsilon^l = T_l^{-1/6} = 2^{-l/6}$, $\lambda^l = \frac{\epsilon^l(1-\gamma)}{2\left\|d_\rho^{\pi^\star}/\rho\right\|_\infty}$ and $\epsilon_{\rm pp} = 1/(2A)$. In addition, suppose that the assumptions in Lemma 7 hold (note that Assumption 1 and $\lambda^l \leq \bar{\lambda}$ already automatically hold by the other assumptions). Then we have for any $\delta \in (0,1)$, with probability at least $1-\delta$, jointly for all episodes N, we have (for the mini-batch Algorithm 4)*

$$\mathbf{regret}(N;M) = O\left(\frac{S^2A^2}{(1-\gamma)^3}\left\|\frac{d_\rho^{\pi^\star}}{\rho}\right\|_\infty^2 (M^{\frac{1}{6}} + M^{-\frac{5}{6}})(N+M)^{\frac{5}{6}}(\log(N/\delta))^{\frac{5}{2}} + \frac{M(\log N)^2}{1-\gamma}\right).$$

In addition, we also have

$$\lim_{N\to\infty} \mathbf{regret}(N;M)/(N+1) = 0 \quad \textit{almost surely}$$

with an asymptotic rate of

$$O\left(\frac{S^2A^2}{(1-\gamma)^3}\left\|\frac{d_\rho^{\pi^\star}}{\rho}\right\|_\infty^2 (M^{\frac{1}{6}} + M^{-\frac{5}{6}})N^{-\frac{1}{6}}\left(1+\frac{M}{N}\right)^{\frac{5}{6}}(\log N)^{\frac{5}{2}} + \frac{M(\log N)^2}{(1-\gamma)N}\right).$$

Again, we note that the almost sure convergence above is directly implied by the high probability bound via Lemma 4. The proof and a more formal statement of this corollary (*cf.* Corollary 13) can be found in Appendix D. In particular, when $M = 1$, the bound above reduces to (16). In addition, we can see that there might be a trade-off between the terms $M^{1/6}$ and $M^{-5/6}$. The intuition behind this is a trade-off between lower variance with larger batch sizes and more frequent updates with smaller batch sizes.

6 Conclusion and open problems

In this work, we establish the global convergence rates of practical policy gradient algorithms with a fixed size mini-batch of trajectories combined with REINFORCE gradient estimation.

Although in §4 and §5, we only instantiate the bounds for the REINFORCE gradient estimators, we note that our general results (in particular, Theorem 11 in Appendix B.2) can be easily applied to other gradient estimators (*e.g.*, actor-critic and state-action visitation measure based estimators) as well, as long as one can verify the existence of the constants in Assumption 2 in a similar way to Lemma 2. In addition, one can also easily derive sample complexity results as by-products of our analysis. In fact, our proof of Theorem 5 immediately implies a $\tilde{O}(1/\epsilon^4)$ sample complexity bound (for Algorithm 1 with REINFORCE gradient estimators and a constant regularization parameter) for any pre-specified tolerance $\epsilon > 0$, where we use $\tilde{O}$ to indicate the big-O notation with logarithmic terms suppressed. We have focused only on regret in this paper mainly for clarity purposes.

It is also not difficult to extend our results to finite horizon non-stationary settings, in which the soft-max policy parametrization will have a dimension of SAH and different policy gradient estimators can be adopted (without trajectory truncation), with H being the horizon of each episode. The results are almost identical to those in this work apart from a smaller constant D_l (due to the unbiased estimators, which implies that $\delta^{l,k} = 0$) and replacing each $1/(1-\gamma)$ term with an H term. In this case, it's also easy to rewrite the regret bound as a function of the total number of time steps $T \leq HN$, where N is the total number of episodes. Other straightforward extensions include refined convergence to stationary points (in both almost sure and high probability sense and with no requirement on large batch sizes), and inexact convergence results when $\delta^{l,k}$ (*cf.* Assumption 2) is not square summable (*e.g.*, when $H^{l,k}$ is fixed or not growing sufficiently fast).

There are also several open problems that may be resolved by combining the techniques introduced in this paper with existing results in the literature. Firstly, it would be desirable to remove the "exploration" assumption that the initial distribution ρ is component-wise positive. This may be achieved by combining our results with the policy cover technique in [AHKS20] or the optimistic bonus tricks in [CYJW19, ESRM20]. Secondly, the bounds in our paper are likely far from optimal (*i.e.*, sharp). Hence it would be desirable to either refine our analysis or apply our techniques to accelerated policy gradient methods (*e.g.*, IS-MBPG [HGPH20]) to obtain better global convergence rates and/or last-iterate convergence. Thirdly, it would be very interesting to see if global convergence results still hold for REINFORCE when the relative entropy regularization term used in this paper is replaced with the practically adopted entropy regularization term in the literature. The answer is affirmative when exact gradient estimations are available [MXSS20, CCC+20], but it remains unknown how these results might be generalized to the stochastic settings in our paper. We conjecture that entropy regularization leads to better global convergence rates and can help us remove the necessity of the `PostProcess` steps in Algorithm 2 as they are uniformly bounded. Finally, one may also consider relaxing the uniform bound assumption on the rewards r to instead being sub-Gaussian, introducing function approximation, and extending our results to natural policy gradient and actor-critic methods as well as more modern policy gradient methods like DPG, PPO and TRPO.

Acknowledgment

We would like to thank Anran Hu for pointing out a mistake in the proof of an early draft of this paper. We also thank Guillermo Angeris, Shane Barratt, Haotian Gu, Xin Guo, Yusuke Kikuchi, Bennet Meyers, Xinyue Shen, Mahan Tajrobehkar, Jonathan Tuck, Jiaming Wang and Xiaoli Wei for their feedback on some preliminary results in this paper. We would like to appreciate the support from Samsung Scholarship.

References

[AHKS20] A. Agarwal, M. Henaff, S. Kakade, and W. Sun. PC-PG: Policy cover directed exploration for provable policy gradient learning. *arXiv preprint arXiv:2007.08459*, 2020.

[AJK19] A. Agarwal, N. Jiang, and S. Kakade. Reinforcement Learning: Theory and Algorithms. Technical report, Department of Computer Science, University of Washington, 2019.

[AKLM19] A. Agarwal, S. Kakade, J. Lee, and G. Mahajan. Optimality and approximation with policy gradient methods in Markov decision processes. *arXiv preprint arXiv:1908.00261*, 2019.

[AYBB+19] Y. Abbasi-Yadkori, P. Bartlett, K. Bhatia, N. Lazic, C. Szepesvari, and G. Weisz. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pages 3692–3702, 2019.

[BB01] J. Baxter and P. Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

[BCN18] L. Bottou, F. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[Ber17] D. Bertsekas. *Dynamic programming and optimal control*, volume II. Athena scientific Belmont, MA, 4th edition, 2017.

[BGH20] M. Basei, X. Guo, and A. Hu. Linear quadratic reinforcement learning: Sublinear regret in the episodic continuous-time framework. *arXiv preprint arXiv:2006.15316*, 2020.

[BK18] L. Besson and E. Kaufmann. What doubling tricks can and can't do for multi-armed bandits. *arXiv preprint arXiv:1803.06971*, 2018.

[BR19] J. Bhandari and D. Russo. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

[BR20] J. Bhandari and D. Russo. A note on the linear convergence of policy gradient methods. *arXiv preprint arXiv:2007.11120*, 2020.

[CCC+20] S. Cen, C. Cheng, Y. Chen, Y. Wei, and Y. Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *arXiv preprint arXiv:2007.06558*, 2020.

[CLT19] R. Carmona, M. Laurière, and Z. Tan. Linear-quadratic mean-field reinforcement learning: convergence of policy gradient methods. *arXiv preprint arXiv:1910.04295*, 2019.

[CYJW19] Q. Cai, Z. Yang, C. Jin, and Z. Wang. Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830*, 2019.

[EDM03] E. Even-Dar and Y. Mansour. Learning rates for Q-learning. *Journal of machine learning Research*, 5(Dec):1–25, 2003.

[ESRM20] Y. Efroni, L. Shani, A. Rosenberg, and S. Mannor. Optimistic policy optimization with bandit feedback. *arXiv preprint arXiv:2002.08243*, 2020.

[FGKM18] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. *arXiv preprint arXiv:1801.05039*, 2018.

[FYCW19] Z. Fu, Z. Yang, Y. Chen, and Z. Wang. Actor-critic provably finds Nash equilibria of linear-quadratic mean-field games. *arXiv preprint arXiv:1910.07498*, 2019.

[FYW20] Z. Fu, Z. Yang, and Z. Wang. Single-timescale actor-critic provably finds globally optimal policy. *arXiv preprint arXiv:2008.00483*, 2020.

[FYWX20] Y. Fei, Z. Yang, Z. Wang, and Q. Xie. Dynamic regret of policy optimization in non-stationary environments. *arXiv preprint arXiv:2007.00148*, 2020.

[GHXZ20] X. Guo, A. Hu, R. Xu, and J. Zhang. A general framework for learning mean-field games. *arXiv preprint arXiv:2003.06069*, 2020.

[GLSM15] S. Gu, S. Levine, I. Sutskever, and A. Mnih. MuProp: Unbiased backpropagation for stochastic neural networks. *arXiv preprint arXiv:1511.05176*, 2015.

[Gly86] P. Glynn. Stochastic approximation for Monte Carlo optimization. In *Proceedings of the 18th conference on Winter simulation*, pages 356–365, 1986.

[HGPH20] F. Huang, S. Gao, J. Pei, and H. Huang. Momentum-based policy gradient methods. *arXiv preprint arXiv:2007.06680*, 2020.

[JAZBJ18] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

[JOA10] T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

[JSW20] Z. Jin, J. Schmitt, and Z. Wen. On the analysis of model-free methods for the linear quadratic regulator. *arXiv preprint arXiv:2007.03861*, 2020.

[K$^+$03] S. Kakade et al. *On the sample complexity of reinforcement learning*. PhD thesis, University of London London, England, 2003.

[Kle13] A. Klenke. *Probability Theory: A Comprehensive Course*. Springer Science & Business Media, 2013.

[KT03] V. Konda and J. Tsitsiklis. On actor-critic algorithms. *SIAM journal on Control and Optimization*, 42(4):1143–1166, 2003.

[KvHW18] W. Kool, H. van Hoof, and M. Welling. Attention, learn to solve routing problems! *arXiv preprint arXiv:1803.08475*, 2018.

[KvHW20] W. Kool, H. van Hoof, and M. Welling. Estimating gradients for discrete random variables by sampling without replacement. *arXiv preprint arXiv:2002.06043*, 2020.

[LCYW19] B. Liu, Q. Cai, Z. Yang, and Z. Wang. Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv preprint arXiv:1906.10306*, 2019.

[MBG+16] V. Mnih, M. Badia, A.and Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937, 2016.

[MG14] A. Mnih and K. Gregor. Neural variational inference and learning in belief networks. *arXiv preprint arXiv:1402.0030*, 2014.

[MPB+18] D. Malik, A. Pananjady, K. Bhatia, K. Khamaru, P. Bartlett, and M. Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. *arXiv preprint arXiv:1812.08305*, 2018.

[MRJS19] E. Mazumdar, L. Ratliff, M. Jordan, and S. Sastry. Policy-gradient algorithms have no guarantees of convergence in linear quadratic games. *arXiv preprint arXiv:1907.03712*, 2019.

[MT01] P. Marbach and J. Tsitsiklis. Simulation-based optimization of Markov reward processes. *IEEE Transactions on Automatic Control*, 46(2):191–209, 2001.

[MXSS20] J. Mei, C. Xiao, C. Szepesvari, and D. Schuurmans. On the global convergence rates of softmax policy gradient methods. *arXiv preprint arXiv:2005.06392*, 2020.

[NJG17] G. Neu, A. Jonsson, and V. Gómez. A unified view of entropy-regularized Markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

[PBC+18] M. Papini, D. Binaghi, G. Canonaco, M. Pirotta, and M. Restelli. Stochastic variance-reduced policy gradient. *arXiv preprint arXiv:1806.05618*, 2018.

[PMA10] J. Peters, K. Mülling, and Y. Altun. Relative entropy policy search. In *AAAI*, volume 10, pages 1607–1612. Atlanta, 2010.

[PNP+20] N. Pham, L. Nguyen, D. Phan, P. Nguyen, M. van Dijk, and Q. Tran-Dinh. A hybrid stochastic policy gradient algorithm for reinforcement learning. *arXiv preprint arXiv:2003.00430*, 2020.

[RB16] E. Ryu and S. Boyd. Primer on monotone operator methods. *Appl. Comput. Math*, 15(1):3–43, 2016.

[RMM+17] S. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024, 2017.

[SB18] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.

[SCA17] J. Schulman, X. Chen, and P. Abbeel. Equivalence between policy gradients and soft Q-learning. *arXiv preprint arXiv:1704.06440*, 2017.

[SEM19] L. Shani, Y. Efroni, and S. Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized MDPs. *arXiv preprint arXiv:1909.02769*, 2019.

[SLA+15] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897, 2015.

[SLH+14] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller. Deterministic policy gradient algorithms. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32(1), pages 387–395, 2014.

[SMSM00] R. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

[SRH+19] Z. Shen, A. Ribeiro, H. Hassani, H. Qian, and C. Mi. Hessian aided policy gradient. In *International Conference on Machine Learning*, pages 5729–5738, 2019.

[SWD+17] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[WCYW19] L. Wang, Q. Cai, Z. Yang, and Z. Wang. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019.

[Wil92] R. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[XXLZ20] H. Xiong, T. Xu, Y. Liang, and W. Zhang. Non-asymptotic convergence of Adam-type reinforcement learning algorithms under Markovian sampling. *arXiv preprint arXiv:2002.06286*, 2020.

[YLLZ20] H. Yuan, X. Lian, J. Liu, and Y. Zhou. Stochastic recursive momentum for policy gradient methods. *arXiv preprint arXiv:2003.04302*, 2020.

[ZHNS11] T. Zhao, H. Hachiya, G. Niu, and M. Sugiyama. Analysis and improvement of policy gradient estimation. In *Advances in Neural Information Processing Systems*, pages 262–270, 2011.

[ZKB+20] J. Zhang, A. Koppel, A. Bedi, C. Szepesvari, and M. Wang. Variational policy gradient method for reinforcement learning with general utilities. *arXiv preprint arXiv:2007.02151*, 2020.

[ZKZB19] K. Zhang, A. Koppel, H. Zhu, and T. Başar. Global convergence of policy gradient methods to (almost) locally optimal policies. *arXiv preprint arXiv:1906.08383*, 2019.

[ZLW19] M. Zhao, Y. Li, and Z. Wen. A stochastic trust-region framework for policy optimization. *arXiv preprint arXiv:1911.11640*, 2019.

[ZYB19] K. Zhang, Z. Yang, and T. Basar. Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games. In *Advances in Neural Information Processing Systems*, pages 11602–11614, 2019.

Appendix

In this appendix, we provide detailed proofs and formal statements of the results in the main text. For notational simplicity, we sometimes abbreviate "almost sure" as "a.s." or even omit "a.s." whenever it is clear from the context. Also notice that as is always implicitly assumed in the literature of episodic reinforcement learning (*e.g.*, *cf.* [MT01]), given the current policy, the sampled trajectory is conditionally independent of all previous policies and trajectories.

A Proofs for REINFORCE gradient estimations

A.1 Proof of Lemma 2 (with $b = 0$)

Proof. We validate the three groups conditions in Assumption 2 in order. For the simplicity of exposition, we first restrict to the case when $b = 0$, *i.e.*, no baseline is incorporated.

Gradient estimation boundedness. Firstly, notice that since $r(s, a) \in [0, 1]$, we have $\hat{Q}^{l,k}(s_t^{l,k}, a_t^{l,k}) \leq 1/(1 - \gamma)$. And by the soft-max parametrization in Assumption 1, we have

$$\nabla_\theta \log \pi_{\theta^{l,k}}(a|s) = \mathbf{1}_{s,a} - \mathbf{E}_{a' \sim \pi_{\theta^{l,k}}(\cdot|s)} \mathbf{1}_{s,a'},$$

where the vector $\mathbf{1}_{s,a} \in \mathbf{R}^{SA}$ has all zero entries apart from the one corresponding to the state-action pair (s, a). Hence $\|\nabla_\theta \log \pi_{\theta^{l,k}}(a|s)\|_2 \leq 2$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, and we see that

$$\begin{aligned} \left\| \widehat{\nabla}_\theta L_{\lambda^l}(\theta^{l,k}) \right\|_2 &\leq \frac{1}{1 - \gamma} \sum_{t=0}^{\infty} \gamma^t \|\nabla_\theta \log \pi_{\theta^{l,k}}(a_t^{l,k}|s_t^{l,k})\|_2 + \frac{\bar{\lambda}}{SA} \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \|\nabla_\theta \log \pi_{\theta^{l,k}}(a|s)\|_2 \\ &\leq 2/(1 - \gamma)^2 + 2\bar{\lambda}, \quad \text{a.s.} \end{aligned} \tag{19}$$

Hence we can take $C_1 = 2/(1 - \gamma)^2 + 2\bar{\lambda}$.

Validation of nearly unbiasedness. Secondly, notice that

$$\begin{aligned} \mathbf{E}_{l,k} \widehat{\nabla}_\theta L_{\lambda^l}(\theta^{l,k}) = \mathbf{E}_{l,k} &\left(\sum_{t=0}^{\lfloor \beta H^{l,k} \rfloor} \gamma^t \mathbf{E}_{l,k} \left(\hat{Q}^{l,k}(s_t^{l,k}, a_t^{l,k}) \middle| s_t^{l,k}, a_t^{l,k} \right) \nabla_\theta \log \pi_{\theta^{l,k}}(a_t^{l,k}|s_t^{l,k}) \right) \\ &+ \frac{\lambda^l}{SA} \sum\nolimits_{s \in \mathcal{S}, a \in \mathcal{A}} \nabla_\theta \log \pi_{\theta^{l,k}}(a|s) = J_1 + J_2 + J_3, \end{aligned}$$

where

$$\begin{aligned} J_1 = \mathbf{E}_{l,k} &\left(\sum_{t=0}^{\infty} \gamma^t \mathbf{E}_{l,k} \left(\sum\nolimits_{t'=t}^{\infty} \gamma^{t'-t} r_{t'}^{l,k} \middle| s_t^{l,k}, a_t^{l,k} \right) \nabla_\theta \log \pi_{\theta^{l,k}}(a_t^{l,k}|s_t^{l,k}) \right) \\ &+ \frac{\lambda^l}{SA} \sum\nolimits_{s \in \mathcal{S}, a \in \mathcal{A}} \nabla_\theta \log \pi_{\theta^{l,k}}(a|s), \end{aligned}$$

$$J_2 = -\mathbf{E}_{l,k}\left(\sum_{t=\lfloor \beta H^{l,k}\rfloor+1}^{\infty}\gamma^t \mathbf{E}_{l,k}\left(\sum\nolimits_{t'=t}^{\infty}\gamma^{t'-t}r_{t'}^{l,k}\middle|s_t^{l,k},a_t^{l,k}\right)\nabla_\theta \log \pi_{\theta^{l,k}}(a_t^{l,k}|s_t^{l,k})\right),$$

$$J_3 = -\mathbf{E}_{l,k}\left(\sum_{t=0}^{\lfloor \beta H^{l,k}\rfloor}\gamma^t \mathbf{E}_{l,k}\left(\sum\nolimits_{t'=H^{l,k}+1}^{\infty}\gamma^{t'-t}r_{t'}^{l,k}\middle|s_t^{l,k},a_t^{l,k}\right)\nabla_\theta \log \pi_{\theta^{l,k}}(a_t^{l,k}|s_t^{l,k})\right).$$

By [AJK19, Theorem 4.6], we have

$$\begin{aligned} J_1 = {} & \mathbf{E}_{l,k}\left(\sum_{t=0}^{\infty}\gamma^t Q^{\pi_{\theta^{l,k}}}(s_t^{l,k},a_t^{l,k})\nabla_\theta \log \pi_{\theta^{l,k}}(a_t^{l,k}|s_t^{l,k})\right) \\ & + \frac{\lambda^l}{SA}\sum_{s\in\mathcal{S},a\in\mathcal{A}}\nabla_\theta \log \pi_{\theta^k}(a|s) \\ = {} & \nabla_\theta L_{\lambda^l}(\theta^{l,k}). \end{aligned} \tag{20}$$

Here for any $\pi \in \Pi$,

$$Q^\pi(s,a) = \mathbf{E}\left(\sum\nolimits_{t=0}^{\infty}\gamma^t r(s_t,a_t)\middle|s_0=s,\ a_0=a\right),$$

with $a_t \sim \pi(s_t,\cdot),\ s_{t+1}\sim p(\cdot|s_t,a_t),\ \forall t>0$.

And since $r(s,a)\in[0,1]$, we have

$$\begin{aligned} \|J_2\|_2 &\le \frac{1}{1-\gamma}\sum_{t=\lfloor \beta H^{l,k}\rfloor+1}^{\infty}\gamma^t \|\nabla_\theta \log \pi_{\theta^{l,k}}(a_t^{l,k}|s_t^{l,k})\|_2 \\ &\le 2\gamma^{\beta H^{l,k}}/(1-\gamma)^2, \end{aligned}$$

and similarly

$$\begin{aligned} \|J_3\|_2 &\le \sum_{t=0}^{\lfloor \beta H^{l,k}\rfloor}\gamma^t \sum_{t'=H^{l,k}+1}^{\infty}\gamma^{t'-t}\|\nabla_\theta \log \pi_{\theta^k}(a_t^{l,k}|s_t^{l,k})\|_2 \\ &\le \sum_{t=0}^{\lfloor \beta H^{l,k}\rfloor}\gamma^t \times 2\gamma^{(1-\beta)H^{l,k}}/(1-\gamma) \\ &\le 2\gamma^{(1-\beta)H^{l,k}}/(1-\gamma)^2. \end{aligned}$$

Hence for any $\eta_0>0$, by taking

$$H^{l,k} \ge \frac{1+\eta_0}{(2+\eta_0)\min\{\beta,1-\beta\}}\log_{1/\gamma}\left(\frac{4^{(2+\eta_0)/(1+\eta_0)}(k+1)}{(1-\gamma)^{(4+2\eta_0)/(1+\eta_0)}}\right)(=\Theta(\log(k+1))),$$

we have $H^{l,k}\ge \log_{1/\gamma}(k+1)$, and that

$$\left\|\mathbf{E}_{l,k}\widehat{\nabla}_\theta L_{\lambda^l}(\theta^{l,k}) - \nabla_\theta L_{\lambda^l}(\theta^{l,k})\right\|_2 \le \frac{4\gamma^{\min\{\beta,1-\beta\}H^{l,k}}}{(1-\gamma)^2} \le (k+1)^{-\frac{1+\eta_0}{2+\eta_0}}, \tag{21}$$

which implies that

$$
\begin{aligned}
&\nabla_\theta L_{\lambda^l}(\theta^{l,k})^T \mathbf{E}_{l,k} \widehat{\nabla}_\theta L_{\lambda^l}(\theta^{l,k}) \\
&\quad = \|\nabla_\theta L_{\lambda^l}(\theta^{l,k})\|_2^2 + \left(\mathbf{E}_{l,k} \widehat{\nabla}_\theta L_{\lambda^l}(\theta^{l,k}) - \nabla_\theta L_{\lambda^l}(\theta^{l,k})\right)^T \nabla_\theta L_{\lambda^l}(\theta^{l,k}) \\
&\quad \geq \|\nabla_\theta L_{\lambda^l}(\theta^{l,k})\|_2^2 - \|\nabla_\theta L_{\lambda^l}(\theta^{l,k})\|_2 (k+1)^{-\frac{1+\eta_0}{2+\eta_0}} \\
&\quad \geq \|\nabla_\theta L_{\lambda^l}(\theta^{l,k})\|_2^2 - \left(\frac{2}{(1-\gamma)^2} + 2\bar{\lambda}\right)(k+1)^{-\frac{1+\eta_0}{2+\eta_0}},
\end{aligned}
$$

where the last two steps used Cauchy inequality, (21) and the fact that by (20),

$$
\begin{aligned}
&\|\nabla_\theta L_{\lambda^l}(\theta^{l,k})\|_2 \\
&\quad \leq \sum_{t=0}^{\infty} \gamma^t \mathbf{E}_{l,k} \left(Q^{\pi_\theta^{l,k}}(s_t^{l,k}, a_t^{l,k}) \left\|\nabla_\theta \log \pi_{\theta^{l,k}}(a_t^{l,k}|s_t^{l,k})\right\|_2\right) + \frac{\bar{\lambda}}{SA} \sum_{s\in\mathcal{S}, a\in\mathcal{A}} \|\nabla_\theta \log \pi_{\theta^{l,k}}(a|s)\|_2 \\
&\quad \leq 2/(1-\gamma)^2 + 2\bar{\lambda}.
\end{aligned}
$$

Hence we can take $C_2 = 1$ and $\delta_{l,k} = \left(\frac{2}{(1-\gamma)^2} + 2\bar{\lambda}\right)(k+1)^{-\frac{1+\eta_0}{2+\eta_0}}$. Thus we have

$$
\begin{aligned}
\sum_{k=0}^{T_l-1} \delta_{l,k}^2 &= \left(\frac{2}{(1-\gamma)^2} + 2\bar{\lambda}\right)^2 \sum_{k=0}^{\infty} (k+1)^{-\frac{2+2\eta_0}{2+\eta_0}} \\
&\leq 8\left(\frac{1}{(1-\gamma)^2} + \bar{\lambda}\right)^2 \left(1 + \frac{1}{\eta_0}\right),
\end{aligned}
$$

and hence we can take $C = 8\left(\frac{1}{(1-\gamma)^2} + \bar{\lambda}\right)^2 \left(1 + \frac{1}{\eta_0}\right)$. Notice that for notational simplicity, we have taken $\eta_0 = 1$ in the statement of the proposition.

Validation of bounded second-order moment growth. Finally, we bound the second-order moment of the policy gradient. In the following, for a random vector $X = (X_1, \dots, X_n) \in \mathbf{R}^n$, we define $\mathbf{Var}\, X = \sum_{i=1}^n \mathbf{var}\, X_i$, and similarly $\mathbf{Var}_{l,k}\, X = \sum_{i=1}^n \mathbf{var}_{l,k}\, X_i$, where $\mathbf{var}_{l,k}$ denotes the conditional variance given the (l,k)-th iteration $\theta^{l,k}$. Now define the constant $\bar{V}$ as the uniform upper bound on the variance of the policy gradient vector, *i.e.*,

$$
\bar{V} = \sup_{H\geq 0,\, \theta\in\Theta,\, \lambda\in[0,\bar{\lambda}]} \mathbf{Var}\left(\sum_{t=0}^{\lfloor \beta H \rfloor} \gamma^t \widehat{Q}(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t) + \frac{\lambda}{SA} \sum_{s\in\mathcal{S}, a\in\mathcal{A}} \nabla_\theta \log \pi_\theta(a|s)\right),
$$

where $\tau = (s_0, a_0, r_0, \dots, s_H, a_H, r_H)$ is sampled from $\mathcal{M}$ following policy π_θ, and $\hat{Q}(s_t, a_t) = \sum_{t'=t}^{H} \gamma^{t'-t} r_{t'}$.

Then we have $\mathbf{Var}_{l,k}\, \widehat{\nabla}_\theta L_{\lambda^l}(\theta^{l,k}) \leq \bar{V}$ for any $l,\, k \geq 0$ by definition. In addition, since for any random vector $X \in \mathbf{R}^n$,

$$
\mathbf{var}\, X \leq \sum_{i=1}^n \mathbf{E} X_i^2 = \mathbf{E}\|X\|_2^2,
$$

we have by the same argument as (19) that

$$
\begin{aligned}
\bar{V} &\leq \mathbf{E}\left\|\sum_{t=0}^{\lfloor\beta H\rfloor}\gamma^t\widehat{Q}(s_t,a_t)\nabla_\theta\log\pi_\theta(a_t|s_t)+\frac{\lambda}{SA}\sum_{s\in\mathcal{S},a\in\mathcal{A}}\nabla_\theta\log\pi_\theta(a|s)\right\|_2^2\\
&\leq \frac{4(1+\bar{\lambda}(1-\gamma)^2)^2}{(1-\gamma)^4}.
\end{aligned}
$$

Finally, since for any random vector $X\in\mathbf{R}^n$,

$$
\mathbf{E}_{l,k}\|X\|^2=\mathbf{E}_{l,k}\sum\nolimits_{i=1}^n X_i^2=\sum\nolimits_{i=1}^n(\mathbf{E}_{l,k}X_i^2+\mathbf{var}_{l,k}X_i)=\|\mathbf{E}_{l,k}X\|_2^2+\mathbf{Var}_{l,k}X,
$$

we have

$$
\begin{aligned}
\mathbf{E}_{l,k}\left\|\widehat{\nabla}_\theta L_{\lambda^l}(\theta^{l,k})\right\|_2^2 &\leq \|J_1+J_2+J_3\|_2^2+\bar{V}\\
&\leq 2\|J_1\|_2^2+2(\|J_2\|_2+\|J_3\|_2)^2+\bar{V}\\
&\leq 2\|\nabla_\theta L_{\lambda^l}(\theta^{l,k})\|_2^2+\frac{32\gamma^{2\min\{\beta,1-\beta\}H^k}}{(1-\gamma)^4}+\bar{V}\\
&\leq 2\|\nabla_\theta L_{\lambda^l}(\theta^{l,k})\|_2^2+\frac{32}{(1-\gamma)^4}+\bar{V},
\end{aligned}
$$

and hence we can take $M_2=2$ and $M_1=32/(1-\gamma)^4+\bar{V}$. This completes our proof. □

A.2 Proof of Lemma 2

Proof. The proof is nearly identical to the case when $b=0$ above. Hence we only outline the proof while highlighting the differences.

Firstly, similar to (19), we have

$$
\left\|\widehat{\nabla}_\theta L_{\lambda^l}(\theta^{l,k})\right\|_2\leq\left(\frac{1}{1-\gamma}+B\right)\frac{2}{1-\gamma}+2\bar{\lambda}\quad\text{a.s.},
$$

and hence we can take $C_1=\frac{2+2B(1-\gamma)}{(1-\gamma)^2}+2\bar{\lambda}$.

Secondly, by the proof of [AJK19, Lemma 4.10], we have

$$
\mathbf{E}_{l,k}\left(\sum_{t=0}^{\lfloor\beta H^{l,k}\rfloor}\gamma^t b(s_t^{l,k})\nabla_\theta\log\pi_{\theta^{l,k}}(a_t^{l,k}|s_t^{l,k})\right)=0. \tag{22}
$$

Hence $\mathbf{E}_{l,k}\widehat{\nabla}_\theta L_{\lambda^l}(\theta^{l,k})$ is the same as in the proof when $b=0$ above, and hence we can take

$$
\begin{aligned}
&C_2=1,\quad C=16\left(\frac{1}{(1-\gamma)^2}+\bar{\lambda}\right)^2,\\
&\delta_{l,k}=\left(\frac{2}{(1-\gamma)^2}+2\bar{\lambda}\right)(k+1)^{-\frac{2}{3}},\quad H^{l,k}\geq\frac{3\log_{1/\gamma}\left(\frac{8(k+1)}{(1-\gamma)^3}\right)}{2\min\{\beta,1-\beta\}}.
\end{aligned}
$$

Finally, by definition, $\bar{V}_b$ can be written explicitly as

$$\bar{V}_b = \sup_{H\geq 0,\, \theta\in\Theta,\, \lambda\in[0,\bar{\lambda}]} \mathbf{Var}\left(\sum_{t=0}^{\lfloor \beta H \rfloor} \gamma^t(\widehat{Q}(s_t,a_t) - b(s_t))\nabla_\theta \log \pi_\theta(a_t|s_t) + \frac{\lambda}{SA}\sum_{s\in\mathcal{S},a\in\mathcal{A}} \nabla_\theta \log \pi_\theta(a|s)\right),$$

where $\tau = (s_0, a_0, r_0, \dots, s_H, a_H, r_H)$ is sampled from $\mathcal{M}$ following policy π_θ, and $\hat{Q}(s_t,a_t) = \sum_{t'=t}^{H} \gamma^{t'-t} r_{t'}$.

Hence similar to $\bar{V}$ in then proof when $b = 0$ above, we have $\bar{V}_b \leq 4\left(\frac{1+B(1-\gamma)}{(1-\gamma)^2} + \bar{\lambda}\right)^2$, $M_2 = 2$ and $M_1 = 32/(1-\gamma)^4 + \bar{V}_b$. □

B Proofs for convergence analysis

Firstly, we state a simple result from elementary analysis, which will be used repeatedly in our proof below.

Lemma 9. *Let $x_k = \frac{1}{\sqrt{k+3}\log_2(k+3)}$ for $k \geq 0$. Then we have*

$$\sum_{k=K_1}^{K_2} x_k \geq \frac{K_2 - K_1 + 1}{\sqrt{K_2+3}\log_2(K_2+3)}, \quad \sum_{k=0}^{\infty} x_k^4 \leq \sum_{k=0}^{\infty} x_k^2 \leq 1.$$

Proof. The first inequality immediately comes from the fact that x_k is monotonically decreasing in k. The second inequality can be derived by noticing that $x_k < 1$ for any $k \geq 0$, and that

$$\sum_{k=0}^{\infty} x_k^2 \leq \int_0^{\infty} \frac{1}{(x+2)(\log_2(x+2))^2} dx = 1.$$

This completes the proof. □

B.1 Proof of Theorem 5

Theorem 10 (Formal statement of Theorem 5)**.** *Under Assumptions 1 and 2, for phase $l \geq 0$ suppose that we choose $\alpha^{l,k} = C_{l,\alpha}\frac{1}{\sqrt{k+3}\log_2(k+3)}$ for some $C_{l,\alpha} \in (0, C_2/(M_2\beta_{\lambda^l})]$. Then for any $\epsilon > 0$, if we choose $\lambda^l = \frac{\epsilon(1-\gamma)}{2\left\|d_\rho^{\pi^\star}/\rho\right\|_\infty}$, then for any $\delta \in (0,1)$, with probability at least $1-\delta$, for any $K \in \{0,\dots,T_l-1\}$, we have*

$$\begin{aligned}&\mathbf{regret}_l(K)\\ &\leq \frac{4(D_l + \sqrt{2C_l\log(2/\delta)})}{(1-\gamma)C_2E_l\epsilon^2}\sqrt{K+1}\log_2(K+3) + \epsilon(K+1) + \frac{\gamma+\gamma\log(K+1)}{1-\gamma}.\end{aligned} \tag{23}$$

Here $E_l = \frac{C_{l,\alpha}(1-\gamma)^2}{16S^2A^2\left\|d_\rho^{\pi^\star}/\rho\right\|_\infty^2}$, *and*

$$C_l = 32C_1^2C_{l,\alpha}^2\left(\frac{1}{(1-\gamma)^2}+\lambda^l\right)^2 + \frac{\beta_{\lambda^l}^2C_1^4C_{l,\alpha}^4}{2},$$
$$D_l = CC_{l,\alpha}^2 + \beta_{\lambda^l}M_1C_{l,\alpha}^2 + F^\star - L_{\lambda^l}(\theta^{l,0}).$$

Proof. The proof consists of two parts. In the first part, we establish an upper bound on the weighted gradient norms sum of previous iterates in the current phase. The second part then utilizes this bound to establish an upper bound on the phase regret.

Bounding the weighted gradient norms sum. By Proposition 1 and an equivalent definition of strongly smoothness (*cf.* [RB16, Appendix]), we have

$$\begin{aligned}-L_{\lambda^l}(\theta^{l,k+1}) - (-L_{\lambda^l}(\theta^{l,k})) &\le -\nabla_\theta L_{\lambda^l}(\theta^{l,k})^T(\theta^{l,k+1}-\theta^{l,k}) + \frac{\beta_{\lambda^l}}{2}\|\theta^{l,k+1}-\theta^{l,k}\|_2^2\\ &= \underbrace{-\alpha^{l,k}\nabla_\theta L_{\lambda^l}(\theta^{l,k})^T\widehat{\nabla}_\theta L_{\lambda^l}(\theta^{l,k}) + \frac{\beta_{\lambda^l}(\alpha^{l,k})^2}{2}\|\widehat{\nabla}_\theta L_{\lambda^l}(\theta^{l,k})\|_2^2}_{Y_{l,k}}.\end{aligned}$$

Let $Z_{l,k} = Y_{l,k} - \mathbf{E}_{l,k}[Y_{l,k}]$. Then the above inequality implies that

$$\begin{aligned}&L_{\lambda^l}(\theta^{l,k}) - L_{\lambda^l}(\theta^{l,k+1})\\ &\le -\alpha^{l,k}L_{\lambda^l}(\theta^{l,k})^T\mathbf{E}_{l,k}\widehat{\nabla}_\theta L_{\lambda^l}(\theta^{l,k}) + \frac{\beta_{\lambda^l}(\alpha^{l,k})^2}{2}\mathbf{E}_{l,k}\|\widehat{\nabla}_\theta L_{\lambda^l}(\theta^{l,k})\|_2^2 + Z_{l,k}\\ &\le -\alpha^{l,k}\left(C_2\|\nabla_\theta L_{\lambda^l}(\theta^{l,k})\|_2^2 - \delta_{l,k}\right) + \frac{\beta_{\lambda^l}(\alpha^{l,k})^2}{2}\left(M_1 + M_2\|\nabla_\theta L_{\lambda^l}(\theta^{l,k})\|_2^2\right) + Z_{l,k}\\ &= -\alpha^{l,k}(C_2 - M_2\beta_{\lambda^l}\alpha^{l,k}/2)\|\nabla_\theta L_{\lambda^l}(\theta^{l,k})\|_2^2 + \alpha^{l,k}\delta_{l,k} + \frac{\beta_{\lambda^l}M_1(\alpha^{l,k})^2}{2} + Z_{l,k}\\ &\le -\frac{C_2\alpha^{l,k}}{2}\|\nabla_\theta L_{\lambda^l}(\theta^{l,k})\|_2^2 + \alpha^{l,k}\delta_{l,k} + \frac{\beta_{\lambda^l}M_1(\alpha^{l,k})^2}{2} + Z_{l,k}.\end{aligned} \tag{24}$$

Now define $X_{l,K} = \sum_{k=0}^{K-1} Z_{l,k}$ (with $X_{l,0}=0$), then

$$\mathbf{E}(X_{l,K+1}|\mathcal{F}_{l,K}) = \sum_{k=0}^{K-1} Z_{l,k} + \mathbf{E}(Y_{l,K} - \mathbf{E}_{l,K}Y_{l,K}|\mathcal{F}_{l,K}) = X_{l,K}. \tag{25}$$

Here $\mathcal{F}_{l,K}$ is the filtration up to episode K in phase l, *i.e.*, the σ-algebra generated by all iterations $\{\theta^{0,0},\dots,\theta^{0,T_0},\dots,\theta^{l,0},\dots,\theta^{l,K}\}$ up to the (l,K)-th one. Notice that the second equality makes use of the fact that given the current policy, the correspondingly sampled trajectory is conditionally independent of all previous policies and trajectories.

In addition, for any $K \ge 1$,

$$\begin{aligned}|X_{l,K} - X_{l,K-1}| = &|Z_{l,K-1}| \le \alpha^{l,K-1}\|\nabla_\theta L_{\lambda^l}(\theta^{l,K-1})\|_2\|\mathbf{E}_{l,K-1}\widehat{\nabla}_\theta L_{\lambda^l}(\theta^{l,K-1}) - \widehat{\nabla}_\theta L_{\lambda^l}(\theta^{l,K-1})\|_2\\ &+ \frac{\beta_{\lambda^l}(\alpha^{l,K-1})^2}{2}\left|\mathbf{E}_{l,K-1}\|\widehat{\nabla}_\theta L_{\lambda^l}(\theta^{l,K-1})\|_2^2 - \|\widehat{\nabla}_\theta L_{\lambda^l}(\theta^{l,K-1})\|_2^2\right|\\ \le &\underbrace{2C_1\left(\frac{2}{(1-\gamma)^2}+2\lambda^l\right)\alpha^{l,K-1} + \frac{\beta_{\lambda^l}}{2}C_1^2(\alpha^{l,K-1})^2}_{c_{l,K}}.\end{aligned}$$

Here we use the fact that

$$\|\nabla_\theta L_{\lambda^l}(\theta^{l,K-1})\|_2 \leq 2/(1-\gamma)^2 + 2\lambda^l,$$

which follows from the same argument as (19). The above inequality on $|X_{l,K} - X_{l,K-1}|$ also implies that $\mathbf{E}|X_{l,K}| < \infty$, which, together with (25), implies that $X_{l,K}$ is a martingale. Notice that although $X_{l,K}$ is only defined for $K = 0, \dots, T_l$, we can augment the sequence by setting $X_{l,K} = X_{l,T_l}$ and $\mathcal{F}_{l,K} = \mathcal{F}_{l,T_l}$ for all $K > T_l$, and it's obvious that (25) and $\mathbf{E}|X_{l,K}| < \infty$ also hold for $K \geq T_l$. And by saying that $X_{l,K}$ is a martingale, we are indeed referring to this (infinite length) augmented sequence of random variables.

Now by the definition of $\alpha^{l,k}$, it's easy to see that $\sum_{K=1}^{T_l} c_{l,K}^2 \leq C_l < \infty$, where

$$C_l = 32 C_1^2 C_{l,\alpha}^2 \left(\frac{1}{(1-\gamma)^2} + \lambda^l\right)^2 + \frac{\beta_{\lambda^l}^2 C_1^4 C_{l,\alpha}^4}{2}. \tag{26}$$

Hence by Azuma-Hoeffding inequality, for any $c > 0$,

$$\mathbf{Prob}(|X_{l,T_l}| \geq c) \leq 2e^{-c^2/(2C_l)}. \tag{27}$$

Then by summing up the inequalities (24) from $k = 0$ to K, we obtain that

$$\begin{aligned}
\frac{C_2}{2}\sum_{k=0}^{K} \alpha^{l,k}\|\nabla_\theta L_{\lambda^l}(\theta^{l,k})\|_2^2 &\leq \frac{C_2}{2}\sum_{k=0}^{T_l-1} \alpha^{l,k}\|\nabla_\theta L_{\lambda^l}(\theta^{l,k})\|_2^2 \\
&\leq \sum_{k=0}^{\infty} \alpha^{l,k}\delta_{l,k} + \frac{\beta_{\lambda^l} M_1 \sum_{k=0}^{\infty}(\alpha^{l,k})^2}{2} + \sum_{k=0}^{T_l-1} Z_{l,k} + \sup_{\theta\in\Theta} L_{\lambda^l}(\theta) - L_{\lambda^l}(\theta^{l,0}) \\
&\leq \sum_{k=0}^{\infty}(\alpha^{l,k})^2 \sum_{k=0}^{\infty}\delta_{l,k}^2 + \frac{\beta_{\lambda^l} M_1}{2}\sum_{k=0}^{\infty}(\alpha^{l,k})^2 + X_{l,T_l} + F^\star - L_{\lambda^l}(\theta^{l,0}) \\
&\leq \underbrace{C C_{l,\alpha}^2 + \beta_{\lambda^l} M_1 C_{l,\alpha}^2 + F^\star - L_{\lambda^l}(\theta^{l,0})}_{D_l} + X_{l,T_l},
\end{aligned} \tag{28}$$

where we use the fact that the regularization term $R(\theta) \leq 0$ for all $\theta \in \Theta$.

Hence we have

$$\sum_{k=0}^{K} \alpha^{l,k}\|\nabla_\theta L_{\lambda^l}(\theta^{l,k})\|_2^2 \leq \frac{2(D_l + X_{l,T_l})}{C_2}. \tag{29}$$

Bounding the phase regret. We now establish the regret bound in phase l using (29).

Fix $l \geq 0$ and $K \in \{0, \dots, T_l - 1\}$. Let

$$I^+ = \{k \in \{0, \dots, K\} \mid \|\nabla_\theta L_{\lambda^l}(\theta^{l,k})\|_2 \geq \lambda^l/(2SA)\}.$$

For simplicity, assume for now that $|I^+| > 0$. Then since $\alpha^{l,k}$ is decreasing in k, we have

$$\begin{aligned}
2(D_l + X_{l,T_l})/C_2 &\geq \frac{(\lambda^l)^2}{4S^2A^2} \sum_{k=K-|I^+|+1}^{K} \alpha^{l,k} \\
&= \epsilon^2 \underbrace{\frac{C_{l,\alpha}(1-\gamma)^2}{16S^2A^2 \left\|d_\rho^{\pi^\star}/\rho\right\|_\infty^2}}_{E_l} \sum_{k=K-|I^+|+1}^{K} \frac{1}{\sqrt{k+3}\log_2(k+3)} \\
(\text{By Lemma 9}) &\geq E_l\epsilon^2 \frac{|I^+|}{\sqrt{K+3}\log_2(K+3)}.
\end{aligned}$$

Hence we have (by the simple fact that $\sqrt{K+3} \leq 2\sqrt{K+1}$ for any $K \geq 0$)

$$|I^+| \leq \frac{4(D_l + X_{l,T_l})}{C_2 E_l \epsilon^2} \sqrt{K+1} \log_2(K+3) \tag{30}$$

Now by Proposition 3 and the choice of λ^l, we have that for any $k \notin I^+$,

$$F^\star - F(\pi_{\theta^{l,k}}) \leq \epsilon.$$

Since for any $\pi \in \Pi$, $F(\pi) \in [0, 1/(1-\gamma)]$, we have $F^\star - F(\pi) \leq 1/(1-\gamma)$. Hence by (30), we have

$$\begin{aligned}
&\textstyle\sum_{k\leq K} F^\star - F(\pi_{\theta^{l,k}}) \\
&\quad = \sum_{k\in I^+} F^\star - F(\pi_{\theta^{l,k}}) + \sum_{k\notin I^+} F^\star - F(\pi_{\theta^{l,k}}) \\
&\quad \leq |I^+|/(1-\gamma) + \big(K+1-|I^+|\big)\,\epsilon \\
&\quad \leq \frac{4(D_l + X_{l,T_l})}{(1-\gamma)C_2 E_l \epsilon^2} \sqrt{K+1} \log_2(K+3) + (K+1)\epsilon.
\end{aligned}$$

This immediately implies that

$$\begin{aligned}
\mathbf{regret}_l(K) &= \textstyle\sum_{k\leq K} F^\star - F(\pi_{\theta^{l,k}}) + \sum_{k\leq K} F(\pi_{\theta^{l,k}}) - \hat{F}^{l,k}(\pi_{\theta^{l,k}}) \\
&\leq \textstyle\sum_{k\leq K} F^\star - F(\pi_{\theta^{l,k}}) + \sum_{k\leq K} \mathbf{E}_{l,k} \sum_{t=H^{l,k}+1}^{\infty} \gamma^t r(s_t^{l,k}, a_t^{l,k}) \\
&\leq \textstyle\sum_{k\leq K} F^\star - F(\pi_{\theta^{l,k}}) + \displaystyle\sum_{k\leq K} \frac{\gamma/(k+1)}{1-\gamma} \\
&\leq \frac{4(D_l + X_{l,T_l})}{(1-\gamma)C_2 E_l \epsilon^2} \sqrt{K+1} \log_2(K+3) + (K+1)\epsilon + \frac{\gamma + \gamma\log(K+1)}{1-\gamma}.
\end{aligned} \tag{31}$$

Now if $|I^+| = 0$, then we immediately have that

$$\mathbf{regret}_l(K) \leq (K+1)\epsilon + \frac{\gamma + \gamma\log(K+1)}{1-\gamma},$$

and hence (31) always holds.

Finally, by (27), we have that with probability at least $1-\delta$, for all $K \in \{0, \ldots, T_l - 1\}$,

$$\mathbf{regret}_l(K) \leq \frac{4(D_l + \sqrt{2C_l \log(2/\delta)})}{(1-\gamma)C_2 E_l \epsilon^2} \sqrt{K+1} \log_2(K+3) + \epsilon(K+1) + \frac{\gamma + \gamma \log(K+1)}{1-\gamma}.$$

This completes our proof. $\square$

B.2 Overall regret bound for general policy gradient estimators

In this section, we state and prove the overall regret bound for general policy gradient estimators, which generalizes Theorem 6 for REINFORCE gradient estimators.

Theorem 11 (General regret bound)**.** *Under Assumptions 1 and 2, suppose that for each $l \geq 0$, we choose $\alpha^{l,k} = C_{l,\alpha} \frac{1}{\sqrt{k+3} \log_2(k+3)}$ for some $C_{l,\alpha} \in [\underline{C}^{\alpha}, C_2/(M_2 \beta_{\lambda^l})]$, with $\underline{C}^{\alpha} \in (0, C_2/(M_2 \beta_{\bar{\lambda}})]$ and $\bar{\lambda} = \frac{1-\gamma}{2\|d_\rho^{\pi^\star}/\rho\|_\infty}$. In addition, suppose that we specify $T_0 \geq 1$, choose $T_l = 2^l T_0$, $\epsilon^l = T_l^{-1/6}$ and $\lambda^l = \frac{\epsilon^l(1-\gamma)}{2\left\|d_\rho^{\pi^\star}/\rho\right\|_\infty}$ for each $l \geq 0$. Then we have for any $\delta \in (0,1)$, with probability at least $1-\delta$, for any $N \geq 0$, we have*

$$\mathbf{regret}(N) \leq \bar{R}_1(N) + \bar{R}_2(N) = O(N^{5/6}(\log(N/\delta))^{5/2}), \tag{32}$$

where

$$\begin{aligned}
\bar{R}_1(N) &= \left(\frac{4(\bar{D} + \sqrt{2\bar{C}((\log_2(N+1)+2)\log 2 + \log(1/\delta))})}{(1-\gamma)C_2\underline{E}} + 1\right) \\
&\quad \times (N+T_0)^{\frac{5}{6}} (\log_2(2N+2T_0+1))^2, \\
\bar{R}_2(N) &= \frac{(\log_2(N+1)+1)(\gamma + \gamma\log(N+T_0))}{1-\gamma}.
\end{aligned} \tag{33}$$

Here the constants $\underline{E} = \frac{\underline{C}^{\alpha}(1-\gamma)^2}{16 S^2 A^2 \left\|d_\rho^{\pi^\star}/\rho\right\|_\infty^2}$,

$$\bar{D} = C\bar{C}_\alpha^2 + \beta_{\bar{\lambda}} M_1 \bar{C}_\alpha^2 + \frac{1}{1-\gamma} + \log(1/\epsilon_{\mathrm{pp}}),$$

$$\bar{C} = 32 C_1^2 \bar{C}_\alpha^2 \left(\frac{1}{(1-\gamma)^2} + \bar{\lambda}\right)^2 + \frac{\beta_{\bar{\lambda}}^2 C_1^4 \bar{C}_\alpha^4}{2},$$

with $\bar{C}_\alpha = \frac{C_2(1-\gamma)^3}{8M_2}$.

In addition, we also have

$$\lim_{N\to\infty} \mathbf{regret}(N)/(N+1) = 0 \quad \textit{almost surely}$$

with an asymptotic rate of $O(N^{-1/6}(\log N)^{5/2})$.

Remark 1. *Notice that the constant $\underline{E}$ is a uniform lower bound of E_l ($l \geq 0$), while the constants $\bar{D}$ and $\bar{C}$ are uniform upper bounds of D_l and C_l ($l \geq 0$), respectively. Here the constants E_l, D_l, C_l are those defined in Theorem 10.*

Remark 2. *In the big-O notation above, we have (temporarily) hidden the problem dependent quantities, which will be made explicit when we specialize the results to the REINFORCE gradient estimation below.*

Proof. We first prove the high probability result. By (23) and the choices of ϵ^l and λ^l, we have that for any phase $l \geq 0$, with probability at least $1-\delta/2^{l+1}$, for all $K \in \{0, \ldots, T_l - 1\}$,

$$\mathbf{regret}_l(K) \leq \left(\frac{4(\bar{D} + \sqrt{2\bar{C}((l+2)\log 2 + \log(1/\delta))})}{(1-\gamma)C_2\underline{E}} + 1 \right) T_l^{5/6} \log_2(T_l + 2) + \frac{\gamma + \gamma \log T_l}{1-\gamma}.$$

where $\underline{E} = \frac{\underline{C}^\alpha (1-\gamma)^2}{16 S^2 A^2 \left\| d_\rho^{\pi^\star}/\rho \right\|_\infty^2}$,

$$\bar{D} = C\bar{C}_\alpha^2 + \beta_{\bar{\lambda}} M_1 \bar{C}_\alpha^2 + \frac{1}{1-\gamma} + \log(1/\epsilon_{\text{pp}}),$$

$$\bar{C} = 32 C_1^2 \bar{C}_\alpha^2 \left(\frac{1}{(1-\gamma)^2} + \bar{\lambda} \right)^2 + \frac{\beta_{\bar{\lambda}}^2 C_1^4 \bar{C}_\alpha^4}{2},$$

with $\bar{C}_\alpha = \frac{C_2(1-\gamma)^3}{8M_2}$ and $\bar{\lambda} = \frac{1-\gamma}{2\|d_\rho^{\pi^\star}/\rho\|_\infty}$. Here we used the fact that $\epsilon^l \leq 1$, which then implies that $\lambda^l \leq \bar{\lambda}$ and

$$\frac{8}{(1-\gamma)^3} \leq \beta_{\lambda^l} \leq \beta_{\bar{\lambda}} = \frac{8}{(1-\gamma)^3} + \frac{2\bar{\lambda}}{S}.$$

We also used the fact that $F^\star - F(\pi) \leq 1/(1-\gamma)$ for any $\pi \in \Pi$ and that by the definition of `PostProcess`, $R_{\lambda^l}(\pi_{\theta^{l,0}}) \geq \log \epsilon_{\text{pp}}$.

Now recall that for any $N \geq 0$, we have

$$\begin{aligned}
\mathbf{regret}(N) &= \sum_{l=0}^{l_N - 1} \mathbf{regret}_l(T_l - 1) + \mathbf{regret}_{l_N}(k_N) \\
&\leq \sum_{l=0}^{l_N} \mathbf{regret}_l(T_l - 1),
\end{aligned}$$

where $(l_N, k_N) = G_{\mathcal{T}}(N)$. In addition, by the choices of T_l, we have that for any $0 \leq k \leq T_l - 1$,

$$\begin{aligned}
B_{\mathcal{T}}(l, k) &= \sum_{j=0}^{l-1} T_j + k \\
&= (2^l - 1)T_0 + k \\
&\geq (2^l - 1)T_0.
\end{aligned}$$

Hence for any $N \geq 0$, we have $l_N \leq \log_2 \left(\frac{N}{T_0} + 1 \right) \leq \log_2(N+1)$.

Thus we have that with probability at least $1-\sum_{l=0}^{l_N}\delta/2^{l+1}\geq 1-\delta$, for any $N\geq 0$,

$$\mathbf{regret}(N)\leq (l_N+1)(\hat{R}_1(N)+\hat{R}_2(N)),$$

where

$$\begin{aligned}\hat{R}_1(N)&=\left(\frac{4(\bar{D}+\sqrt{2\bar{C}((l_N+2)\log 2+\log(1/\delta))})}{(1-\gamma)C_2\underline{E}}+1\right)(N+T_0)^{\frac{5}{6}}\log_2(N+T_0+2),\\ \hat{R}_2(N)&=\frac{\gamma+\gamma\log(N+T_0)}{1-\gamma}.\end{aligned}$$

Finally, noticing that $l_N+1\leq\log_2(N+1)+1\leq\log_2(2N+2T_0+1)$, we have

$$\begin{aligned}(l_N+1)\hat{R}_1(N)&\leq\left(\frac{4(\bar{D}+\sqrt{2\bar{C}((\log_2(N+1)+2)\log 2+\log(1/\delta))})}{(1-\gamma)C_2\underline{E}}+1\right)\\ &\quad\times(N+T_0)^{\frac{5}{6}}(\log_2(2N+2T_0+1))^2,\\ (l_N+1)\hat{R}_2(N)&\leq\frac{(\log_2(N+1)+1)(\gamma+\gamma\log(N+T_0))}{1-\gamma},\end{aligned}$$

which immediately imply (32) and (33). Notice that here we used the fact that $\log_2(N+1)+1\leq\log_2(2N+2T_0+1)$ (since $T_0\geq 1$), and that $T_l\leq N+1\leq N+T_0$ for all $l=0,\dots,l_N-1$ and $T_{l_N}=2^{l_N}T_0\leq N+T_0$.

Finally, by invoking Lemma 4, we immediately obtain the almost sure convergence result. This completes our proof. □

C Formal statement of REINFORCE regret bound

Here we provide a slightly more formal restatement of Theorem 6, with details about the constants in the big-O notation in the main text. Recall that our goal there is specialize (and slightly simplify) the regret bound in Theorem 11 to the case when the REINFORCE gradient estimation in §3.2 is adopted to evaluate $\widehat{\nabla}_\theta L_{\lambda^l}(\theta^{l,k})$. In particular, we have the following corollary. The proof is done by simply combining Lemma 2 (with $\lambda^l\leq\bar{\lambda}$ by their definitions in Theorem 6 or Corollary 12 below) and Theorem 11, together with the specific choices of the hyper-parameters as well as the constants in Lemma 2 plugged in and some elementary algebraic simplifications, and is hence omitted.

Corollary 12 (Formal statement of Theorem 6)**.** *Under Assumption 1, suppose that for each $l\geq 0$, we choose $\alpha^{l,k}=C_{l,\alpha}\frac{1}{\sqrt{k+3}\log_2(k+3)}$, with $C_{l,\alpha}\in[\underline{C}^\alpha,1/(2\beta_{\lambda^l})]$, $\underline{C}^\alpha\in(0,1/(2\beta_{\bar{\lambda}})]$ and $\bar{\lambda}=\frac{1-\gamma}{2\|d_\rho^{\pi^\star}/\rho\|_\infty}$, and choose $T_l=2^l$, $\epsilon^l=T_l^{-1/6}=2^{-l/6}$, $\lambda^l=\frac{\epsilon^l(1-\gamma)}{2\left\|d_\rho^{\pi^\star}/\rho\right\|_\infty}$ and $\epsilon_{\rm pp}=1/(2A)$. In addition, suppose that (11) is adopted to evaluate $\widehat{\nabla}_\theta L_{\lambda^l}(\theta^{l,k})$, with $\beta\in(0,1)$, $|b(s)|\leq B$ for any $s\in\mathcal{S}$ (where $B>0$ is a constant), and that (12) holds for $H^{l,k}$ for all $l,k\geq 0$. Then we have for any $\delta\in(0,1)$, with probability at least $1-\delta$, for any $N\geq 0$, we have*

$$\mathbf{regret}(N)\leq\tilde{R}_1(N)+\tilde{R}_2(N),$$

where

$$\tilde{R}_1(N) = \left(\frac{4(\tilde{D} + \sqrt{2\tilde{C}((\log_2(N+1)+2)\log 2 + \log(1/\delta))})}{(1-\gamma)\underline{E}} + 1\right)(N+1)^{\frac{5}{6}}(\log_2(2N+3))^2,$$

$$\tilde{R}_2(N) = \frac{\gamma(\log_2(N+1)+1)^2}{1-\gamma}.$$

Here the constants are $\underline{E} = \frac{\underline{C}^\alpha(1-\gamma)^2}{16S^2A^2\left\|d_\rho^{\pi^\star}/\rho\right\|_\infty^2}$, *and*

$$\tilde{D} = (1-\gamma)^6\left(\frac{1}{(1-\gamma)^2} + \bar{\lambda}\right)^2 + \frac{1}{256}(1-\gamma)^6\beta_{\bar{\lambda}}\left(\frac{32}{(1-\gamma)^4} + \bar{V}_b\right) + \frac{1}{1-\gamma} + \log(2A),$$

$$\tilde{C} = \frac{\beta_{\bar{\lambda}}^2(1-\gamma)^{12}\left(\frac{(1+B(1-\gamma))}{(1-\gamma)^2} + \bar{\lambda}\right)^4}{8192} + \frac{1}{2}(1-\gamma)^6\left(\frac{(1+B(1-\gamma))}{(1-\gamma)^2} + \bar{\lambda}\right)^4.$$

Here $\bar{V}_b$ *is the variance bound defined in Lemma 2.*

Suppose in addition that we specify $\underline{C}^\alpha = 1/(2\beta_{\bar{\lambda}})$, *then we can simplify the regret bound into the following simple form:*

$$\mathbf{regret}(N) = O\left(\frac{S^2A^2}{(1-\gamma)^3}\left\|\frac{d_\rho^{\pi^\star}}{\rho}\right\|_\infty^2 N^{\frac{5}{6}}(\log(N/\delta))^{\frac{5}{2}}\right).$$

In addition, we also have

$$\lim_{N\to\infty}\mathbf{regret}(N)/(N+1) = 0 \quad \textit{almost surely}$$

with an asymptotic rate of $O\left(\frac{S^2A^2}{(1-\gamma)^3}\left\|d_\rho^{\pi^\star}/\rho\right\|_\infty^2 N^{-\frac{1}{6}}(\log N)^{\frac{5}{2}}\right)$.

Remark 3. *Notice that here (and below), with the specific choices of algorithm hyper-parameters and gradient estimators we are finally able to make all the problem dependent quantities (e.g.,* γ*,* S*,* A*, etc.) explicit in the big-O notation, which is consistent with the convention of the reinforcement learning literature. Here the only hidden quantities are some absolute constants.*

D Mini-batch phased policy gradient method

Here we formalize the mini-batch version of Algorithm 2 described at the beginning of §5 as Algorithm 4, and provide a formal statement as well as a proof for Corollary 8.

Corollary 13 (Formal statement of Corollary 8)**.** *Under Assumption 1, suppose that for each* $l \geq 0$*, we choose* $\alpha^{l,k} = C_{l,\alpha}\frac{1}{\sqrt{k+3}\log_2(k+3)}$*, with* $C_{l,\alpha} \in [\underline{C}^\alpha, 1/(2\beta_{\lambda^l})]$*,* $\underline{C}^\alpha \in (0, 1/(2\beta_{\bar{\lambda}})]$ *and* $\bar{\lambda} = \frac{1-\gamma}{2\|d_\rho^{\pi^\star}/\rho\|_\infty}$*, and choose* $T_l = 2^l$*,* $\epsilon^l = T_l^{-1/6} = 2^{-l/6}$*,* $\lambda^l = \frac{\epsilon^l(1-\gamma)}{2\left\|d_\rho^{\pi^\star}/\rho\right\|_\infty}$ *and* $\epsilon_{\mathrm{pp}} = 1/(2A)$*. In* *addition, suppose that the assumptions in Lemma 7 hold (note that Assumption 1 and $\lambda^l \leq \bar{\lambda}$ already automatically hold by the other assumptions). Then we have for any $\delta \in (0,1)$, with probability at least $1-\delta$, for any $N \geq M \geq 1$, we have that (for the mini-batch Algorithm 4)*

$$\mathbf{regret}(N-M;M) \leq \tilde{R}_1(N;M) + \tilde{R}_2(N;M),$$

where

$$\begin{aligned}\tilde{R}_1(N;M) =& \left(\frac{4(\tilde{D}_M + \sqrt{2\tilde{C}((\log_2(N/M)+2)\log 2 + \log(1/\delta))})}{(1-\gamma)\underline{E}} + 1\right) \\ & \times M^{\frac{1}{6}} N^{\frac{5}{6}} (\log_2(2(N/M)+1))^2, \\ \tilde{R}_2(N;M) =& \frac{\gamma M (\log_2(N/M)+1)^2}{1-\gamma}.\end{aligned}$$

Here the constants $\underline{E}$ and $\tilde{C}$ are the same as in Corollary 12, while

$$\tilde{D}_M = (1-\gamma)^6 \left(\frac{1}{(1-\gamma)^2} + \bar{\lambda}\right)^2 + \frac{1}{256}(1-\gamma)^6 \beta_{\bar{\lambda}} \left(\frac{32}{(1-\gamma)^4} + \frac{\bar{V}_b}{M}\right) + \frac{1}{1-\gamma} + \log(2A).$$

Here $\bar{V}_b$ is the variance bound defined in Lemma 2.

Suppose in addition that we specify $\underline{C}^\alpha = 1/(2\beta_{\bar{\lambda}})$. Then we can simplify the regret bound into the following simple form:

$$\mathbf{regret}(N;M) = O\left(\frac{S^2A^2}{(1-\gamma)^3}\left\|\frac{d_\rho^{\pi^\star}}{\rho}\right\|_\infty^2 (M^{\frac{1}{6}} + M^{-\frac{5}{6}})(N+M)^{\frac{5}{6}}(\log(N/\delta))^{\frac{5}{2}} + \frac{M(\log N)^2}{1-\gamma}\right).$$

In addition, we also have

$$\lim_{N\to\infty} \mathbf{regret}(N;M)/(N+1) = 0 \quad \textit{almost surely}$$

with an asymptotic rate of

$$O\left(\frac{S^2A^2}{(1-\gamma)^3}\left\|\frac{d_\rho^{\pi^\star}}{\rho}\right\|_\infty^2 (M^{\frac{1}{6}} + M^{-\frac{5}{6}})N^{-\frac{1}{6}}\left(1+\frac{M}{N}\right)^{\frac{5}{6}}(\log N)^{\frac{5}{2}} + \frac{M(\log N)^2}{(1-\gamma)N}\right).$$

Proof of Corollary 8. By the definition of $\mathbf{regret}(N;M)$, we immediately see that

$$\mathbf{regret}(N;M) \leq M\mathbf{regret}(\lfloor N/M \rfloor), \tag{34}$$

where $\mathbf{regret}(J)$ ($J \geq 0$) is the original regret (6) in the mini-batch setting, which is defined only for the total number of inner iterations/steps (instead of episodes, so not magnified with a factor of M). More precisely, we have that for any $J \geq 0$,

$$\mathbf{regret}(J) = \sum\nolimits_{\{(l,k)|B_\mathcal{T}(l,k)\leq J\}} F^\star - F(\pi_{\theta^{l,k}}).$$

Now by Lemma 7 and Theorem 11, and following the same simplification as is done in Corollary 12, we have that for any $J \geq 0$,

$$
\begin{aligned}
&\mathbf{regret}(J) \\
&\quad \leq \left(\frac{4(\tilde{D}_M + \sqrt{2\tilde{C}((\log_2(J+1)+2)\log 2 + \log(1/\delta))})}{(1-\gamma)\underline{E}} + 1 \right) (J+1)^{\frac{5}{6}} (\log_2(2J+3))^2 \\
&\qquad + \frac{\gamma(\log_2(J+1)+1)^2}{1-\gamma},
\end{aligned}
$$

where the constants are as stated in the Corollary claims.

The proof is then complete by plugging the bound of $\mathbf{regret}(J)$ above into (34) and invoking Lemma 4. □

Algorithm 4 Mini-Batch Phased Policy Gradient Method

Input: initial parameter $\tilde{\theta}^{0,0}$, step-sizes $\alpha^{l,k}$, regularization parameters λ^l, phase lengths T_l $(l,\, k \geq 0)$, post-processing tolerance ϵ_{pp} and batch size $M > 0$.

Set $\theta^{0,0} = \texttt{PostProcess}(\tilde{\theta}^{0,0}, \epsilon_{\text{pp}})$.

for phase $l = 0, 1, 2, \ldots$ **do**

    for step $k = 0, 1, \ldots, T_l - 1$ **do**

        Choose $H^{l,k}$, sample IID trajectories $\{\tau_i^{l,k}\}_{i=1}^M$ (each with horizon $H^{l,k}$) from $\mathcal{M}$ following policy $\pi_{\theta^{l,k}}$, and compute an approximate gradient $\widehat{\nabla}_\theta^{(i)} L_{\lambda^l}(\theta^{l,k})$ of L_{λ^l} for each trajectory $\tau_i^{l,k}$ $(i = 1, \ldots, M)$.

        Update $\theta^{l,k+1} = \theta^{l,k} + \alpha^{l,k} \frac{1}{M} \sum_{i=1}^M \widehat{\nabla}_\theta^{(i)} L_{\lambda^l}(\theta^{l,k})$.

    end for

    Set $\theta^{l+1,0} = \texttt{PostProcess}(\theta^{l,T_l}, \epsilon_{\text{pp}})$.

end for