On the Global Convergence of Momentum-based Policy Gradient

Yuhao Ding
University of California, Berkeley
yuhao_ding@berkeley.edu

Junzi Zhang[1]
Amazon Advertising
junziz@amazon.com

Javad Lavaei
University of California, Berkeley
lavaei@berkeley.edu

Abstract

Policy gradient (PG) methods are popular and efficient for large-scale reinforcement learning due to their relative stability and incremental nature. In recent years, the empirical success of PG methods has led to the development of a theoretical foundation for these methods. In this work, we generalize this line of research by studying the global convergence of stochastic PG methods with momentum terms, which have been demonstrated to be efficient recipes for improving PG methods. We study both the soft-max and the Fisher-non-degenerate policy parametrizations, and show that adding a momentum improves the global optimality sample complexity of vanilla PG methods by $\tilde{\mathcal{O}}(\epsilon^{-1.5})$ and $\tilde{\mathcal{O}}(\epsilon^{-1})$, respectively, where $\epsilon > 0$ is the target tolerance. Our work is the first one that obtains global convergence results for the momentum-based PG methods. For the generic Fisher-non-degenerate policy parametrizations, our result is the first single-loop and finite-batch PG algorithm achieving $\tilde{O}(\epsilon^{-3})$ global optimality sample complexity. Finally, as a by-product, our methods also provide general framework for analyzing the global convergence rates of stochastic PG methods, which can be easily applied and extended to different PG estimators.

1 Introduction

Policy gradient methods can be dated back to the pioneering work Williams (1992), and have evolved into a rich family of reinforcement learning (RL) algorithms (Konda and Tsitsiklis, 2000; Kakade, 2001; Silver et al., 2014; Schulman et al., 2015; Lillicrap et al., 2015; Schulman et al., 2017). In recent years, due to their amenability to function approximation and the development of deep neural networks, they have been successfully applied to a wide range of problems with significant empirical success, including robotic control, game playing, natural language processing, neural architecture search, and operations research (Zoph and Le, 2016; Silver et al., 2016; Yi et al., 2018; Khan et al., 2020; Wu et al., 2020a).

Momentum techniques have been demonstrated as a powerful and generic recipe for accelerating stochastic gradient methods, especially for nonconvex optimization and deep learning (Qian, 1999; Kingma and Ba, 2015; Reddi et al., 2019). Recent works have also extended momentum techniques to improve policy gradient methods (Xiong et al., 2020; Yuan et al., 2020; Pham et al., 2020; Huang et al., 2020). As a state-of-the-art variance reduction technique, the momentum-based PG methods have been shown to outperform non-momentum methods such as SVRPG (Papini et al., 2018), SRVR-PG (Xu et al., 2020b) and HAPG (Shen et al., 2019) in practice. In particular, Xiong et al. (2020) studies Adam-based policy gradient methods, but only achieves $O(\epsilon^{-4})$ sample complexities, which is the same as the one for the vanilla REINFORCE algorithm. Inspired by the STORM algorithm for stochastic optimization in Cutkosky and Orabona (2019), a new STORM-PG method is proposed in Yuan et al. (2020), which incorporates momentum in the updates and matches the sample complexity as the SRVR-PG method proposed in Xu et al. (2020b) (and also VRMPO) while requiring only single-loop updates and large initialization batches, whereas SRVR-PG and VRMPO require double-loop updates and large batch sizes throughout all iterations. Concurrently, Pham et al. (2020) proposes a hybrid estimator combining the momentum idea with SARAH and considers a more general setting with regularization, and achieves the same $O(\epsilon^{-3})$ sample complexity and again with single-loop updates and large initialization batches. Finally, independently inspired by the STORM algorithm in Cutkosky and Orabona (2019), Huang et al. (2020) proposes a class of momentum-based policy gradient algorithms with adaptive time-steps, single-loop updates and small batch sizes, which match the sample complexity as in Xu et al. (2020b). However, all the

[1]Work done prior to joining or outside of Amazon.

above sample complexity results for momentum-based policy gradient methods only apply to convergence to a first-order stationary point, which may have an arbitrarily poor performance, in contrast to the more desired global convergence guarantees studied in the current work. The sample complexity of momentum-based stochastic PG methods for the global convergence is still an open question.

Inspired by recent advances in the global convergence theory of PG methods (Agarwal et al., 2019; Zhang et al., 2021a,b; Liu et al., 2020), we address the above-mentioned problem in this paper. We focus on the study of STORM-based PG method introduced in Huang et al. (2020) due to its sample efficiency and the simplicity of the algorithm. Our work is the first one that obtains global convergence results for the momentum-based PG. We summarize our contributions below:

- For the soft-max policy parameterization, we show that adding momentum terms improves the existing global optimality sample complexity bounds of PG in Zhang et al. (2021a) by $\tilde{\mathcal{O}}(\epsilon^{-1.5})$.
- For generic Fisher-non-degenerate policy parametrization, adding momentum terms improves the existing sample complexity bounds of PG in Liu et al. (2020) by $\tilde{\mathcal{O}}(\epsilon^{-1})$ and matches the sample complexity bounds of SRVR-PG in Liu et al. (2020). Our result is the first single-loop and finite-batch policy gradient algorithm achieving $\tilde{O}(\epsilon^{-3})$ global optimality sample complexity.
- As a by-product, our methods also provide general frameworks (cf. Lemmas 4.2 and 4.7) for analyzing the global optimality sample complexity of stochastic policy gradient methods with the soft-max and Fisher-non-degenerate policy parameterizations, which can be easily applied and extended to different policy gradient estimators.

Comparisons with the existing global convergence results of policy gradient methods can be found in Table 1. Due to space restrictions, we provide a more detailed literature review and introduce the notations in Sections 6 and 7 of the appendix.

2 Preliminaries

2.1 Reinforcement learning

Reinforcement learning is generally modeled as a discounted Markov decision process (MDP) defined by a tuple $(\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the finite state and action spaces, $\mathbb{P}(s'|s,a)$ is the probability that the agent transits from the state s to the state s' under the action $a \in \mathcal{A}$. $r(s,a)$ is the reward function, i.e., the agent obtains the reward $r(s_h, a_h)$ after it takes the action a_h at the state s_h at time h. We also assume that the reward is bounded, i.e., $r(s,a): \mathcal{S} \times \mathcal{A} \to [0,1]$. $\gamma \in (0,1)$ is the discount factor. The policy $\pi(a|s)$ at the state s is usually represented by a conditional probability distribution $\pi_\theta(a|s)$ associated with the parameter $\theta \in \mathbb{R}^d$, where d is the dimension of the parameter space. Let $\tau = \{s_0, a_0, s_1, a_1, \ldots\}$ denote the data of a sampled trajectory under policy π_θ with the probability distribution over trajectory as

$$p(\tau|\theta, \rho) = \rho(s_0) \prod_{h=1}^{\infty} \mathbb{P}(s_{h+1}|s_h, a_h)\pi_\theta(a_h|s_h),$$

where $\rho \sim \Delta(\mathcal{S})$ is the probability distribution of the initial state s_0. Here, $\Delta(\mathcal{X})$ denotes the probability simplex over a finite set $\mathcal{X}$. For every policy π, one can define the state-action value function $Q^\pi: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as

$$Q^\pi(s,a) \coloneqq \mathbb{E}_{\substack{a_h \sim \pi(\cdot|s_h) \\ s_{h+1} \sim \mathbb{P}(\cdot|s_h, a_h)}} \left(\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \middle| s_0 = s, a_0 = a \right).$$

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

$$V^\pi(s) \coloneqq \mathbb{E}_{a \sim \pi(\cdot|s)}[Q^\pi(s,a)],$$
$$A^\pi(s,a) \coloneqq Q^\pi(s,a) - V^\pi(s).$$

Then, the goal is to find an optimal policy in the policy class that maximizes the expected discounted return, namely,

$$\max_\theta \quad J_\rho(\pi_\theta) \coloneqq \mathbb{E}_{s_0 \sim \rho}[V^{\pi_\theta}(s_0)]. \tag{1}$$

For notional convenience, we denote $J_\rho(\pi_\theta)$ by the shorthand notation $J_\rho(\theta)$ and also let θ^* denote a global maximum of $J_\rho(\theta)$. In practice, a truncated version of the value function is used to approximate the infinite sum of rewards in (1). Let

$$\tau_H = \{s_0, a_0, s_1, \ldots, s_{H-1}, a_{H-1}, s_H\}$$

denote the truncation of the full trajectory τ of length H. The truncated version of the value function is defined as

$$J_\rho^H(\theta) \coloneqq \mathbb{E}_{\substack{s_0 \sim \rho, a_h \sim \pi_\theta(\cdot|s_h) \\ s_{h+1} \sim \mathbb{P}(\cdot|s_h, a_h)}} \left(\sum_{h=0}^{H-1} \gamma^h r(s_h, a_h) \middle| s_0 \right).$$

2.2 Discounted state visitation distributions

The discounted state visitation distribution $d_{s_0}^\pi$ of a policy π is defined as

$$d_{s_0}^\pi(s) \coloneqq (1-\gamma) \sum_{h=0}^{\infty} \gamma^h \mathbb{P}(s_h = s|s_0, \pi),$$

Algorithm	Parametrization	Complexity	Single Loop	Finite Batch
PG (Wang et al., 2019)	neural	$\tilde{O}(\epsilon^{-4})$	✓	✗
PG (Liu et al., 2020)	Fisher-non-degenerate	$\tilde{O}(\epsilon^{-4})$	✓	✗
PG (Zhang et al., 2021a)	soft-max	$\tilde{O}(\epsilon^{-6})$	✓	✓
SRVR-PG (Liu et al., 2020)	Fisher-non-degenerate	$\tilde{O}(\epsilon^{-3})$	✗	✗
Ours	Fisher-non-degenerate	$\tilde{O}(\epsilon^{-3})$	✓	✓
Ours	soft-max	$\tilde{O}(\epsilon^{-4.5})$	✓	✓

Table 1: We summarize comparisons with the existing global convergence results for policy gradient methods. The (sample) complexity is defined as the number of trajectories needed to reach the global sub-optimality gap $\epsilon > 0$ plus some inherent function approximation error (if any), and we ignore logarithmic terms. Note that "single loop" only refers to the policy update step, and does not refer to the policy evaluation part (for actor-critic versions of PG).

where $\mathbb{P}(s_h = s|s_0, \pi)$ is the state visitation probability that s_h is equal to s under the policy π starting from the state s_0. Then, the discounted state visitation distribution under the initial distribution ρ is defined as $d_\rho^\pi(s) := \mathbb{E}_{s_0\sim\rho}[d_{s_0}^\pi(s)]$. Furthermore, the state-action visitation distribution induced by π and the initial state distribution ρ is defined as $v_\rho^\pi(s,a) := d_\rho^\pi(s)\pi(a|s)$, which can also be written as

$$v_\rho^\pi(s,a) := (1-\gamma)\mathbb{E}_{s_0\sim\rho}\sum_{h=0}^{\infty}\gamma^h\mathbb{P}(s_h = s, a_h = a|s_0,\pi),$$

where $\mathbb{P}(s_h = s, a_h = a|s_0, \pi)$ is the state-action visitation probability that $s_h = s$ and $a_h = a$ under π starting from the state s_0.

2.3 Policy parameterization

In this work, we consider the following two different policy classes:

Soft-max parameterization. For an unconstrained parameter $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, the policy $\pi_\theta(a|s)$ is chosen to be

$$\frac{\exp(\theta_{s,a})}{\sum_{a'\in\mathcal{A}}\exp(\theta_{s,a'})}.$$

The soft-max parameterization is generally used for Markov Decision Processes (MDPs) with finite state and action spaces. It is complete in the sense that every stochastic policy can be represented by this class.

Fisher-non-degenerate parameterization. We study the general policy class that satisfies Assumption 2.1 given below:

Assumption 2.1 *For all $\theta \in \mathbb{R}^d$, there exists some constant $\mu_F > 0$ such that the Fisher information matrix $F_\rho(\theta)$ induced by the policy π_θ and the initial state distribution ρ satisfies*

$$F_\rho(\theta) = \mathbb{E}_{(s,a)\sim v_\rho^{\pi_\theta}}[\nabla\log\pi_\theta(a|s)\nabla\log\pi_\theta(a|s)^\top] \succcurlyeq \mu_F\cdot I_d.$$

Assumption 2.1, which is also used in Liu et al. (2020), essentially states that $F_\rho(\theta)$ is well-behaved as a preconditioner in the natural PG update (Kakade and Langford, 2002). It is shown in Liu et al. (2020) that the positive definiteness of $F_\rho(\theta)$ in Assumption 2.1 can be satisfied by certain Gaussian policies, where $\pi_\theta(\cdot|s) = \mathcal{N}(\mu_\theta(s), \Sigma)$ with the parametrized mean function $\mu_\theta(s)$ and the fixed covariance matrix $\Sigma \succ 0$, provided that the Jacobian of $\mu_\theta(s)$ is full-row rank for all $\theta \in \mathbb{R}^d$. In addition, Assumption 2.1 holds more generally for every full-rank exponential family paramertrization with their mean parameterized by $\mu_\theta(s)$ if $\mu_\theta(s)$ is full-row rank for all $\theta \in \mathbb{R}^d$.

It is worth noting that Assumption 2.1 is not satisfied by the soft-max parameterization when π_θ approaches a deterministic policy, which means that the two policy parameterizations to be studied here do not overlap.

3 Trajectory-based policy gradient estimator

The policy gradient method (Sutton and Barto, 2018) is one of the standard ways to solve the optimization problem (1). Since the distribution $p(\tau|\theta)$ is unknown, $\nabla J_\mu(\theta)$ needs to be estimated from samples. Then, a stochastic PG ascent update with the exploratory initial distribution μ at time step t is given as

$$\theta_{t+1} = \theta_t + \frac{\eta_t}{B}\sum_{i=1}^{B}u_t, \tag{2}$$

where $\eta_t > 0$ is the learning rate, B is the batch size of trajectories, and u_t can be any PG estimator of $\nabla J_\mu(\theta)$. If the parameterized policy satisfies Assumption 3.1 to be stated later and the reward function is not dependent on the parameter θ, PG estimators can be obtained from a single sampled trajectory. These trajectory-based estimators include REINFORCE (Williams, 1992), PGT (Sutton et al., 1999)

and GPOMDP (Baxter and Bartlett, 2001). Compared with PG estimators based on the state-action visitation measure (Agarwal et al., 2019), the trajectory-based PG estimators are often used in practice due to their sample efficiency and amenability to using the importance sampling for variance reduction. In practice, the truncated versions of these trajectory-based PG estimators are used to approximate the infinite sum in the PG estimator. For example, the commonly used truncated REINFORCE with a constant baseline b is given by

$$g(\tau_H^i|\theta) = \left(\sum_{h=0}^{H-1} \nabla \log \pi_\theta(a_h^i, s_h^i)\right)\left(\sum_{h=0}^{H-1} \gamma^h r_h(s_h^i, a_h^i) - b\right).$$

The commonly used truncated PGT is given by:

$$g(\tau_H^i|\theta, \mu) = \sum_{h=0}^{H-1}\sum_{j=h}^{H-1} \nabla \log \pi_\theta(a_h^i, s_h^i)\left(\gamma^j r_j(s_j^i, a_j^i)\right).$$

The PGT estimator is also equivalent to the popular truncated GPOMDP estimator defined as follows

$$g(\tau_H^i|\theta) = \sum_{h=0}^{H-1}\sum_{j=0}^{h} \nabla \log \pi_\theta(a_j^i, s_j^i)\left(\gamma^h r_h(s_h^i, a_h^i) - b_h\right). \tag{3}$$

We first make the following essential assumption for PG estimators.

Assumption 3.1 *The gradient and hessian of the function* $\log \pi_\theta(a|s)$ *are bounded, i.e., there exist constants* $M_g, M_h > 0$ *such that* $\|\nabla \log \pi_\theta(a|s)\|_2 \le M_g$ *and* $\left\|\nabla^2 \log \pi_\theta(a|s)\right\|_2 \le M_h$ *for all* $\theta \in \mathbb{R}^d$.

For the soft-max parameterization, Assumption 3.1 is satisfied with $M_g = 2$ and $M_h = 1$ (see Lemma 9.1 in appendix). Assumption 3.1 has also been commonly used in the analysis of the policy gradient (Papini et al., 2018; Xu et al., 2020b,a; Shen et al., 2019; Liu et al., 2020; Huang et al., 2020) for the more general policy parameterization. Based on Assumption 3.1, we provide some useful properties of stochastic policy gradient and the value function.

Proposition 3.2 *For the truncated GPOMDP policy gradient given in* (3) *satisfying Assumptions 3.1, the following properties hold for all initial distribution* ρ *and for all* $\theta \in \mathbb{R}^d$*:*

1. $g(\tau_H|\theta)$ *is* L_g*-Lipschitz continuous, where* $L_g := M_h/(1-\gamma)^2$.
2. $\|g(\tau_H|\theta)\|_2 \le G$, *where* $G := M_g/(1-\gamma)^2$.
3. $Var(g(\tau_H|\theta, \mu)) \le \sigma^2$, *where* $\sigma := G$.
4. $J_\rho(\theta)$ *and* $J_\rho^H(\theta)$ *are* L*-smooth, namely,* $\max\{\left\|\nabla^2 J_\rho(\theta)\right\|_2, \left\|\nabla^2 J_\rho^H(\theta)\right\|_2\} \le L$, *where* $L := \frac{2M_g^2}{(1-\gamma)^3} + \frac{M_h}{(1-\gamma)^2}$.
5. *If the infinite-sum is well-defined, then*

$$g(\tau_H|\theta) = \sum_{h=0}^{\infty}\sum_{j=h}^{\infty} \nabla \log \pi_\theta(a_h^i, s_h^i)\left(\gamma^j r_j(s_j^i, a_j^i) - b_j\right).$$

is an unbiased estimate of $\nabla J_\rho(\theta)$. *Similarly, the truncated GPOMDP estimate* $g(\tau_H|\theta)$ *given by* (3) *is an unbiased estimate of* $\nabla J_\rho^H(\theta)$.

6. $\left\|\nabla J_\rho^H(\theta) - \nabla J_\rho(\theta)\right\|_2 \le M_g\left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\right)\gamma^H$.
7. $\max\{\|\nabla J_\rho(\theta)\|_2, \left\|\nabla J_\rho^H(\theta)\right\|_2\} \le G$.

The first two properties are shown in Proposition 4.2 in Xu et al. (2020b). The third and forth properties follow from Lemma 4.2 and Lemma 4.3 in Yuan et al. (2021), respectively. The last three properties follow directly from Lemma B.1 in Liu et al. (2020).

3.1 Momentum-based policy gradient

Due to the high sample complexity of the vanilla PG, many recent works have turned onto variance reduction methods for PG, including the momentum-based policy gradient (Huang et al., 2020; Yuan et al., 2020). The momentum-based policy gradient with the batch size of B and the sampled trajectory of length H is defined as

$$u_t^H = \frac{\beta_t}{B}\sum_{i=1}^{B} g(\tau_H^i|\theta_t, \mu) + (1-\beta_t)\Big[u_{t-1}^H + \frac{1}{B}\sum_{i=1}^{B}\left(g(\tau_H^i|\theta_t, \mu) - w(\tau_H^i|\theta_{t-1}, \theta_t) g(\tau_H^i|\theta_{t-1}, \mu)\right)\Big] \tag{4}$$

for all $t \in \{2, \ldots, T\}$, where $g(\tau_H^i|\theta_t, \mu)$ is the vanilla PG estimator such as (3), $\beta_t \in [0, 1]$, and the importance sampling weight is defined as

$$w(\tau_H|\theta', \theta) = \frac{p(\tau_H|\theta', \mu)}{p(\tau_H|\theta, \mu)} = \prod_{h=0}^{H-1} \frac{\pi_{\theta'}(a_h|s_t)}{\pi_\theta(a_h|s_t)}. \tag{5}$$

This importance sampling weight guarantees that

$$\mathbb{E}_{\tau_H \sim p(\cdot|\theta,\mu)}[g(\tau_H|\theta, \mu) - w(\tau_H|\theta', \theta) g(\tau_H|\theta', \mu)] = \nabla J_\mu^H(\theta) - \nabla J_\mu^H(\theta').$$

Then, by carefully choosing η_t and β_t, the accumulated policy gradient estimation error $u_t^H - \nabla J_\mu^H(\theta_t)$ can be well controlled. To guarantee the convergence of the momentum-based policy gradient, we require the following assumption:

Assumption 3.3 *For every θ_t and θ_{t+1} satisfying (2), the variance of $w(\tau_H|\theta_t,\theta_{t+1})$ is bounded, i.e., there exists a constant $W>0$ such that $Var(w(\tau_H|\theta_t,\theta_{t+1}))\leq W$ for all $\tau_H\sim p(\cdot|\theta_{t+1},\mu)$.*

Assumption 3.3 has been commonly used in the analysis of some variance reduced variants of PG (Papini et al., 2018; Xu et al., 2020b,a; Shen et al., 2019; Liu et al., 2020; Huang et al., 2020). It is worthwhile to note that the bounded importance sampling weight in Assumption 3.3 may be violated in practice. A commonly used remedy to make the algorithm more effective is to clip the importance sampling weights (Huang et al., 2020). In addition, when π_θ is the soft-max parameterization, the importance sampling weight $w(\tau_H|\theta_t,\theta_{t+1})$ in Assumption 3.3 has a bounded variance by using the truncated policy gradient (see Lemma 5.6 in Zhang et al. (2021b)).

The key reason that the momentum can improve the convergence rates is that the momentum can help reduce the variance of the estimated stochastic gradient. To build some intuition for this, let $e_t=u_t^H-\nabla J_\rho^H(\theta_t)$. It can be verified that

$$\mathbb{E}[e_t]=(1-\beta_t)\mathbb{E}[e_{t-1}]$$

and

$$\begin{aligned}\mathbb{E}\left[||e_t||^2\right]\leq&(1-\beta_t)^2\mathbb{E}\left[||e_{t-1}||^2\right]+2\beta_t^2\mathbb{E}\left[||T_1||^2\right]\\&+2(1-\beta_t)^2\mathbb{E}\left[||T_2||^2\right],\end{aligned}$$

where

$$\begin{aligned}\mathbb{E}\left[||T_1||^2\right]&=\mathrm{Var}(g(\tau_H|\theta_t))\leq\sigma^2,\\\mathcal{O}\left(||T_2||^2\right)&=\mathcal{O}(||\theta_t-\theta_{t-1}||^2)=\mathcal{O}(\eta_t^2||u_t||^2).\end{aligned}$$

Then, the variance of the stochastic gradient u_t can be reduced with the appropriate choices of η_t and β_t.

4 Global convergence of momentum-based policy gradient

As mentioned in the previous sections, the global convergence of policy gradient depends on the parameterization of the policy. In this section, we will study the global convergence and the sample complexity of the momentum-based policy gradient for both soft-max parameterization with a log barrier penalty and the more general parameterization satisfying the fisher-non-degenerate assumption.

4.1 Soft-max parameterization with log barrier penalty

4.1.1 Preliminary tools

We first study the global convergence of momentum-based policy gradient for the soft-max policy parameterization (Algorithm 1), where $\pi_\theta(a|s)=\frac{\exp(\theta_{s,a})}{\sum_{a'\in\mathcal{A}}\exp(\theta_{s,a'})}$ for all $\theta\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. Optimization over the soft-max parameterization is problematic since the optimal policy—that is usually deterministic—is obtained by letting some parameters grow towards infinity. To prevent the parameters from becoming too large and to ensure adequate exploration, a log-barrier regularization term that penalizes the policy for becoming deterministic is commonly used. The regularized objective is defined as

$$\begin{aligned}L_{\lambda,\rho}(\theta)=&J_\rho(\theta)-\lambda\mathbb{E}_{s\sim\mathrm{Unif}_\mathcal{S}}\left[\mathrm{KL}(\mathrm{Unif}_\mathcal{A},\pi_\theta(\cdot|s))\right]\\=&J_\rho(\theta)+\frac{\lambda}{|\mathcal{A}||\mathcal{S}|}\sum_{s,a}\log\pi_\theta(a|s)+\lambda\log|\mathcal{A}|,\end{aligned}$$

where $\mathrm{KL}(p,q):=\mathbb{E}_{x\sim p}[-\log q(x)/p(x)]$ and $\mathrm{Unif}_\mathcal{X}$ denotes the uniform distribution over a set $\mathcal{X}$.

Algorithm 1 Momentum-based PG with soft-max parameterization (MBPG-S)

Inputs: Iteration T, horizon H, batch size B, initial input θ_1, parameters $\{k,m,c\}$, initial distribution μ;
Outputs: θ_ξ chosen uniformly random from $\{\theta_t\}_{t=1}^T$;
for $t=1,2,\ldots,T-1$ **do**
  Sample B trajectories $\{\tau_i^H\}_{i=1}^B$ from $p(\cdot|\theta_t,\mu)$;
  if $t=1$ **then**
    Compute $u_1^H=\frac{1}{B}\sum_{i=1}^B g(\tau^i|\theta_1,\mu)$;
  else
    Compute u_t^H based on (4);
  end if
  Compute $\eta_t=\frac{k}{(m+t)^{1/3}}$;
  Update $\theta_{t+1}=\theta_t+\eta_t(u_t^H+\frac{\lambda}{|\mathcal{A}||\mathcal{S}|}\sum_{s,a}\nabla\log\pi_\theta(a|s))$;
  Update $\beta_{t+1}=c\eta_t^2$;
end for

In addition, while we are interested in the value $J_\rho(\theta)$ under the performance measure ρ, it may be helpful to optimize under the initial distribution μ, i.e., the policy gradient is taken with respect to the optimization measure μ, where μ is usually chosen as an exploratory initial distribution that adequately covers the state distribution of some optimal policy. Then, the notion of the distribution mismatch coefficient is defined below:

Definition 4.1 *Given a policy π and measures $\rho,\mu\in\Delta(\mathcal{S})$, the distribution mismatch coefficient of π under*

ρ relative to μ is defined as $\left\|\frac{d_\rho^\pi}{\mu}\right\|_\infty$, where $\frac{d_\rho^\pi}{\mu}$ denotes componentwise division.

It is shown in Agarwal et al. (2019) that the difficulty of the exploration problem faced by policy gradient algorithms can be captured through this distribution mismatch coefficient.

Although the optimization problem defined above is non-convex in general, Theorem 5.3 in Agarwal et al. (2019) has shown that the first-order stationary points of the regularized objective are approximately globally optimal solutions of $J_\rho(\theta)$ when the regularization parameter λ is sufficiently small and the exact PG is available.

Lemma 4.1 (Agarwal et al. (2019)) *Suppose that θ satisfies the inequality $\|\nabla L_{\lambda,\mu}(\theta)\| \le \epsilon_{opt}$ with $\epsilon_{opt} \le \frac{\lambda}{2|\mathcal{S}||\mathcal{A}|}$. Then, for every initial distribution ρ, we have:*

$$J_\rho(\theta^*) - J_\rho(\theta) \le \frac{2\lambda}{1-\gamma}\left\|\frac{d_\rho^{\pi_{\theta^*}}}{\mu}\right\|_\infty.$$

4.1.2 Theoretical results

Motivated by the above result and the proof idea in Zhang et al. (2021a), we can relate the global convergence to the convergence of the first-order stationary points of the regularized objective.

Lemma 4.2 *Consider a soft-max parameterization π_θ. Given a fixed constant $\epsilon > 0$, let $\lambda = \frac{\epsilon(1-\gamma)}{4\left\|\frac{d_\rho^{\pi_{\theta^*}}}{\mu}\right\|_\infty}$. For every initial distribution ρ, every step-size sequence $\{\eta_t\}_{t=1}^T$, and every sequence $\{\theta_t\}_{t=1}^T$, we have:*

$$J_\rho(\theta^*) - \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[J_\rho(\theta_t)\right] \le \left\|\frac{d_\rho^{\pi_{\theta^*}}}{\mu}\right\|_\infty^2 \frac{64|\mathcal{S}|^2|\mathcal{A}|^2\sum_{t=1}^T \eta_t \mathbb{E}[\|\nabla L_{\lambda,\mu}(\theta_t)\|_2^2]}{\epsilon^2 T\eta_T(1-\gamma)^3} + \frac{\epsilon}{2}.$$

It is worth noting that the bound in Lemma 4.2 is agnostic to the algorithms. To prove Lemma 4.2, we first define the following set of "bad" iterates:

$$I^+ = \left\{t \in \{1,\dots,T\} \,\middle|\, \|\nabla_\theta L_{\lambda,\rho}(\theta_t)\| \ge \frac{\lambda}{2|\mathcal{S}||\mathcal{A}|}\right\}.$$

which counts the number of iterates such that the norm of the first-order stationary points of the entropy-regularized objective is large. We then carefully upper-bound the number of iterates in the set I^+ using the accumulated gradient norm. One can show that for every $\epsilon > 0$ and $\lambda = \frac{\epsilon(1-\gamma)}{4\left\|\frac{d_\rho^{\pi_{\theta^*}}}{\mu}\right\|_\infty}$, we have that

$$J_\rho(\theta^*) - J_\rho(\theta) \le \frac{\epsilon}{2}, \quad \forall k \in \{0,\dots,K\}/I^+,$$
$$J_\rho(\theta^*) - J_\rho(\theta) \le 1/(1-\gamma), \quad \forall k \in I^+,$$

where the second inequality is due to the assumption that the rewards are between 0 and 1. Finally, by combining with the first result, we obtain the desired bound. For the details of the proof, we refer the reader to the appendix in Section 9.1.

With Lemma 4.2, it remains to bound the accumulated stationary convergence of the stochastic policy gradient. Let

$$L_{\lambda,\mu}^H(\theta_t) := J_\mu^H(\theta) + \frac{\lambda}{|\mathcal{A}||\mathcal{S}|}\sum_{s,a}\log\pi_\theta(a|s) + \lambda\log|\mathcal{A}|$$

be the regularized value function with the truncated horizon H. By applying the momentum-based PG, we arrive at the following result:

Lemma 4.3 *Under the conditions in Proposition 3.2, Lemma 4.2, and Assumption 3.3, suppose that the sequence $\{\theta_t\}_{t=1}^T$ is generated by Algorithm 1 with $k > 0$, $\lambda > 0$, $c = \frac{1}{3k^3 L_\lambda} + 96b^2$, $m = \max\{2, (2L_\lambda k)^3, (\frac{ck}{2L_\lambda})^3\}$ and $\eta_0 = \frac{k}{m^{1/3}}$, where $b^2 = L_g^2 + G^2C_w^2$, $L_\lambda = L + \lambda$, and $C_w = \sqrt{H(2HM_g^2 + M_h)(W+1)}$. Then, it holds that*

$$\sum_{t=1}^T \mathbb{E}\left[\eta_t \left\|\nabla L_{\lambda,\mu}^H(\theta_t)\right\|_2^2\right] \le \Gamma_2 + \frac{\Gamma_1 + \Gamma_3}{B}, \tag{6}$$

where $\Gamma_1 = \frac{c^2\sigma^2k^3\ln(T+2)}{44b^2} + \frac{m^{1/3}\sigma^2}{88b^2k} + \frac{1}{22}\left(L_{\lambda,\mu}^H(\theta^*) - L_{\lambda,\mu}^H(\theta_1)\right)$, $\Gamma_2 = \frac{48}{11}\left(L_{\lambda,\mu}^H(\theta^*) - L_{\lambda,\mu}^H(\theta_1)\right)$ and $\Gamma_3 = (\frac{\sigma^2m^{1/3}}{44b^2k} + \frac{c^2\sigma^2k^3}{22b^2}\ln(2+T))$.

Lemma 4.3 shows that $\sum_{t=1}^T \mathbb{E}\left[\eta_t \left\|\nabla L_{\lambda,\mu}^H(\theta_t)\right\|_2^2\right]$ is upper-bounded by a constant up to some logarithmic terms. To prove this result, we first show that $\mathbb{E}[\eta_t\|\nabla L_{\lambda,\mu}^H(\theta_t)\|_2^2]$ can be bounded by two terms, namely the successive iteration differences $\mathbb{E}[\eta_t^{-1}\|\theta_t - \theta_{t-1}\|_2^2]$ and the policy gradient estimation errors $\mathbb{E}[\eta_t\|u_t^H - \nabla J_\mu^H(\theta_t)\|_2^2]$. The proof then reduces to bounding each of these two terms. In light of the momentum term and the carefully chosen hyper-parameters β_t and η_t, a recursive inequality on the policy gradient estimation errors can be used to make the accumulated policy gradient estimation errors small even with a constant batch size. Finally, the successive iteration differences can be upper-bounded by the smoothness of the regularized value function and the construction of a non-trivial Lyapunov function. We

refer the reader to the appendix in Section 9.2 for the details.

Finally, by combining Lemmas 4.2 and 4.3, and using a sufficiently large horizon H, we obtain the following global convergence and sample complexity result for Algorithm 1.

Theorem 4.4 *Under the conditions of Lemmas 4.2 and 4.3, let* $T = \widetilde{\mathcal{O}}\left(\frac{c_\infty^{\frac{3}{2}}}{\epsilon^{\frac{9}{2}}(1-\gamma)^{\frac{33}{2}}} + \frac{c_\infty W}{\epsilon^3(1-\gamma)^{12}}\right)$, $B = \mathcal{O}(1)$ *and* $H = \mathcal{O}\left(\log_\gamma\left(\frac{(1-\gamma)\epsilon}{|\mathcal{S}||\mathcal{A}|}\left\|\frac{d_\rho^{\pi_{\theta^*}}}{\mu}\right\|_\infty^{-1}\right)\right)$, *where* $c_\infty = \left\|\frac{d_\rho^{\pi_{\theta^*}}}{\mu}\right\|_\infty^2 |\mathcal{S}|^2|\mathcal{A}|^2(1+W)$. *Then, it holds that*

$$J_\rho(\theta^*) - \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}(J_\rho(\theta_t))\right] \leq \epsilon.$$

In total, it requires $\widetilde{O}(\epsilon^{-4.5})$ *samples to achieve an* ϵ*-optimal policy.*

The detailed proof of Theorem 4.4 can be found in Section 9.3 in the appendix.

Remark 4.5 *Theorem 4.4 improves the result of Theorem 6 in Zhang et al. (2021a) from the sample complexity of* $\widetilde{\mathcal{O}}(\epsilon^{-6})$ *to* $\widetilde{\mathcal{O}}(\epsilon^{-4.5})$ *for the soft-max parameterization with a log barrier penalty while only using a constant number of batch size* B*.*

4.2 Fisher-non-degenerate parameterization

4.2.1 Preliminary tools

Algorithm 2 Momentum-based PG with Fisher-non-degenerate parameterization (MBPG-F)

1: **Inputs**: Iteration T, Horizon H, batch size B, initial input θ_1, parameters $\{k, m, c\}$ and initial distribution ρ;
2: **Outputs**: θ_ξ chosen uniformly random from $\{\theta_t\}_{t=1}^T$;
3: **for** $t = 1, 2, \ldots, T-1$ **do**
4: Sample B trajectories $\{\tau_i^H\}_{i=1}^B$ from $p(\cdot|\theta_t, \rho)$;
5: **if** $t = 1$ **then**
6: Compute $u_1^H = \frac{1}{B}\sum_{i=1}^B g(\tau_i^H|\theta_1, \rho)$;
7: **else**
8: Compute u_t^H based on (4);
9: **end if**
10: Compute $\eta_t = \frac{k}{(m+t)^{1/3}}$;
11: Update $\theta_{t+1} = \theta_t + \eta_t u_t^H$;
12: Update $\beta_{t+1} = c\eta_t^2$;
13: **end for**

We now study the global convergence of momentum-based policy gradient for the general parameterization satisfying the fisher-non-degenerate assumption in Assumption 2.1 (Algorithm 1). Since this parameterization can be used for general MDPs, it may be restrictive in the sense that it may not contain all stochastic policies and, therefore, may not contain the optimal policy. Thus, there may be some approximation errors. Our analysis will leverage the notion of *compatible function approximation* in Sutton et al. (1999) defined as the regression problem:

$$\min_{w\in\mathbb{R}^d} \mathbb{E}_{(s,a)\sim v_\rho^{\pi_\theta}}\left[\left(A^{\pi_\theta}(s,a) - (1-\gamma)w^\top \nabla \log \pi_\theta(a|s)\right)^2\right]. \tag{7}$$

The notion of *compatible function approximation* measures the ability of using the score function $\nabla \log \pi_\theta(a|s)$ as the features to approximate the advantage function $A^{\pi_\theta}(s,a)$. It can be seen that $F_\rho(\theta)^{-1}\nabla J_\rho(\theta)$ is a minimizer of (7), due to the first-order optimality conditions. Since even the best linear fit using $\nabla \log \pi_\theta(a|s)$ as the features may not perfectly match $A^{\pi_\theta}(s,a)$, the *compatible function approximation error* may not be 0 in practice. Following the assumptions in Liu et al. (2020) and Agarwal et al. (2019), we assume that the policy parameterization π_θ achieves an acceptable function approximation, as measured by the *compatible function approximation error* under a shifted distribution $v_\rho^{\pi_{\theta^*}}$.

Assumption 4.6 *For every* $\theta \in \mathbb{R}^d$*, there exists a constant* $\epsilon_{bias} > 0$ *such that the transferred compatible function approximation error satisfies*

$$\mathbb{E}_{(s,a)\sim v_\rho^{\pi_{\theta^*}}}\left[\left(A^{\pi_\theta}(s,a) - (1-\gamma)u^{*\top}\nabla \log \pi_\theta(a|s)\right)^2\right] \leq \epsilon_{bias},$$

where $v_\rho^{\pi_{\theta^*}}$ *is the state-action distribution induced by an optimal policy* π_{θ^*} *that maximizes* $J_\rho(\theta)$ *and* $u^* := F_\rho(\theta)^{-1}\nabla J_\rho(\theta)$ *is the solution of* (7) *.*

Assumption 4.6, which is also used in Liu et al. (2020), means that the parameterization of π_θ makes the advantage function $A^{\pi_\theta}(s,a)$ be able to nearly approximated by using the score function $\nabla \log \pi_\theta(a|s)$ as the features. When π_θ is a soft-max parameterization, ϵ_{bias} is 0. When π_θ is a rich neural parameterization, ϵ_{bias} is very small (Wang et al., 2019).

4.2.2 Theoretical results

Inspired by the global convergence analysis of PG and natural PG in Liu et al. (2020) and Agarwal et al. (2019), we present a simpler and more general global convergence framework for stochastic PG estimator.

Lemma 4.7 *Consider a general Fisher-non-degenerate policy* π_θ *satisfying Assumptions 2.1,*

3.1 and 4.6. Then, we have

$$J_\rho(\theta^*) - J_\rho(\theta) \leq \frac{\sqrt{\epsilon_{bias}}}{(1-\gamma)} + \frac{M_g}{\mu_F} \|\nabla J_\rho(\theta)\|. \quad (8)$$

Lemma 4.7 relates the global convergence rates of the policy gradient to the transferred compatible function approximation error and the first-order stationary convergence. It can be regarded as the gradient domination condition for Fisher-non-degenerate parametrizations. Compared with Proposition 4.5 in Liu et al. (2020), our result is more general. In particular, the bound in (8) does not depend on the update rule for θ and does not require a Lipschitz continuity assumption for the score function $\nabla \log \pi_\theta(a|s)$.

To prove Lemma 4.7, we first relate the global convergence optimality gap with the stationary convergence of the natural policy gradient $F_\rho(\theta)^{-1}\nabla J_\rho(\theta)$ and the *transferred compatible function approximation error*. This is achieved by the following two observations: (1) the advantage function $A^{\pi_\theta}(\cdot,\cdot)$ appears in both the performance difference lemma and the definition of the *transferred compatible function approximation error*; (2) the natural policy gradient update is connected with the *transferred compatible function approximation error*. In light of Assumption 2.1, one can relate the first-order stationary convergence of natural policy gradient with the first-order stationary convergence of policy gradient. For the details, we refer the reader to the appendix in Section 10.1.

By applying the momentum-based PG with a constant batch size B under the general Fisher-non-degenerate parameterization (see Algorithm 2), we arrive at the following result:

Lemma 4.8 *Under the conditions in Proposition 3.2, Lemma 4.7, and Assumption 3.3, suppose that the sequences $\{\theta_t\}_{t=1}^T$ and $\{u_t^H\}_{t=1}^T$ are generated by Algorithm 2. Let $k > 0$, $c = \frac{1}{3k^3L} + 96b^2$, $m = \max\{2, (2Lk)^3, (\frac{ck}{2L})^3\}$ and $\eta_0 = \frac{k}{m^{1/3}}$, where $b^2 = L_g^2 + G^2C_w^2$ and $C_w = \sqrt{H(2HM_g^2 + M_h)(W+1)}$. Then, it holds that*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\|\nabla J_\rho^H(\theta_t)\|_2] \leq \sqrt{\frac{\Gamma_2}{k} + \frac{\Gamma_1+\Gamma_3}{kB}}\left(\frac{m^{1/6}}{\sqrt{T}} + \frac{1}{T^{1/3}}\right), \quad (9)$$

where $\Gamma_1 = \left(\frac{c^2\sigma^2k^3\ln(T+2)}{48b^2} + \frac{m^{1/3}}{96b^2k}\sigma^2 + \frac{1}{22(1-\gamma)}\right)$, $\Gamma_2 = \frac{48}{11(1-\gamma)}$, and $\Gamma_3 = \frac{\sigma^2m^{1/3}}{44b^2k^2} + \frac{c^2\sigma^2k^3}{22kb^2}\ln(2+T)$.

The proof sketch for Lemma 4.8 is similar to that of Lemma 4.3 and the detailed proof is provided in Appendix 10.2. Finally, by combining Lemmas 4.7 and 4.8, we obtain the global convergence and sample complexity of Algorithm 2.

Theorem 4.9 *Under the conditions of Lemma 4.8, let $H = \mathcal{O}(\log_\gamma((1-\gamma)\mu_F\epsilon))$, $B = \mathcal{O}(1)$ and $T = \tilde{\mathcal{O}}\left(\frac{(1+W)^{\frac{3}{2}}}{\epsilon^3\mu_F^3(1-\gamma)^{12}} + \frac{1+W}{\epsilon^2\mu_F^2(1-\gamma)^{11}} + \frac{W(1+W)}{\epsilon^2\mu_F^2(1-\gamma)^9}\right)$. Then, it holds that*

$$J_\rho(\theta^*) - \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[J_\rho(\theta_t)] \leq \frac{\sqrt{\varepsilon_{bias}}}{1-\gamma} + \epsilon.$$

In total, it requires $\widetilde{O}(\epsilon^{-3})$ samples to achieve an $(\frac{\sqrt{\varepsilon_{bias}}}{1-\gamma} + \epsilon)$-optimal policy.

The detailed proof of Theorem 4.9 can be found in Section 10.3 in the appendix.

Remark 4.10 *Theorem 4.9 establishes the global convergence of the momentum-based PG proposed in Huang et al. (2020), for which only stationary convergence was previously shown. The results in Theorem 4.9 do not hold for the soft-max parameterization. The reason is that the soft-max parameterization lacks the exploration and thus Assumption 2.1 is not satisfied. In addition, it improves the result of Theorem 4.6 in Liu et al. (2020) from the sample complexity of $\mathcal{O}(\frac{1}{\epsilon^4})$ to $\widetilde{\mathcal{O}}(\frac{1}{\epsilon^3})$. It also improves Theorem 4.11 in Liu et al. (2020) from using a batch size of $\mathcal{O}(\epsilon^{-1})$ to a constant batch size and from using a double-loop algorithm to a single-loop algorithm, where the later improvement is due to the momentum introduced in Huang et al. (2020); Yuan et al. (2020).*

5 Conclusion

In this work, we studied the global convergence and the sample complexity of momentum-based stochastic policy gradient methods for both soft-max parameterization and more general parameterization satisfying the fisher-non-degenerate assumption. We showed that adding a momentum improves the global optimally sample complexity of vanilla policy gradient methods in both soft-max and Fisher-non-degenerate policy parametereizations with a constant batch size. This work provides the first global convergence results for momentum-based policy gradient methods.

There are also several open problems that may be addressed by combining the techniques introduced in this paper with the existing results in the literature. First, it remains as an open question whether the momentum-based policy gradient can be combined with the natural policy gradient (Kakade, 2001) to achieve or exceed the state-of-the-art sample complexity. In addition, it is desirable to generalize our soft-max setting to the general class of log-linear policies and remove the "exploration" assumption about the initial state distribution being component-wise positive. This may be achieved by combining our results with the COPOE method in Zanette et al. (2021).

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. (2019). On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *arXiv preprint arXiv:1908.00261*.

Allen-Zhu, Z. and Hazan, E. (2016). Variance reduction for faster non-convex optimization. In *International conference on machine learning*, pages 699–707. PMLR.

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

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

Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M., and Lee, M. (2009). Natural actor–critic algorithms. *Automatica*, 45(11):2471–2482.

Cortes, C., Mansour, Y., and Mohri, M. (2010). Learning bounds for importance weighting. In *Nips*, volume 10, pages 442–450. Citeseer.

Cutkosky, A. and Orabona, F. (2019). Momentum-based variance reduction in non-convex sgd. *arXiv preprint arXiv:1905.10018*.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. (2018). Spider: Near-optimal non-convex optimization via stochastic path integrated differential estimator. *arXiv preprint arXiv:1807.01695*.

Fazel, M., Ge, R., Kakade, S., and Mesbahi, M. (2018). Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476. PMLR.

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

Hong, M., Wai, H.-T., Wang, Z., and Yang, Z. (2020). A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *arXiv preprint arXiv:2007.05170*.

Huang, F., Gao, S., Pei, J., and Huang, H. (2020). Momentum-based policy gradient methods. In *International Conference on Machine Learning*, pages 4422–4433. PMLR.

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26:315–323.

Kakade, S. and Langford, J. (2002). Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer.

Kakade, S. M. (2001). A natural policy gradient. *Advances in neural information processing systems*, 14.

Khan, A., Tolstaya, E., Ribeiro, A., and Kumar, V. (2020). Graph policy gradients for large scale robot control. In *Conference on Robot Learning*, pages 823–834. PMLR.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Konda, V. R. and Tsitsiklis, J. N. (2000). Actor-critic algorithms. In *Advances in neural information processing systems*, pages 1008–1014. Citeseer.

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

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Liu, B., Cai, Q., Yang, Z., and Wang, Z. (2019). Neural trust region/proximal policy optimization attains globally optimal policy. *Advances in Neural Information Processing Systems*, 32:10565–10576.

Liu, Y., Zhang, K., Basar, T., and Yin, W. (2020). An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33.

Mei, J., Gao, Y., Dai, B., Szepesvari, C., and Schuurmans, D. (2021). Leveraging non-uniformity in first-order non-convex optimization. *International Conference on Machine Learning*.

Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. (2020). On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pages 6820–6829. PMLR.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. (2017). Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621. PMLR.

Papini, M., Binaghi, D., Canonaco, G., Pirotta, M., and Restelli, M. (2018). Stochastic variance-reduced policy gradient. In *International Conference on Machine Learning*, pages 4026–4035. PMLR.

Peters, J. and Schaal, S. (2008). Natural actor-critic. *Neurocomputing*, 71(7-9):1180–1190.

Pham, N., Nguyen, L., Phan, D., Nguyen, P. H., Dijk, M., and Tran-Dinh, Q. (2020). A hybrid stochastic policy gradient algorithm for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 374–385. PMLR.

Qian, N. (1999). On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151.

Reddi, S. J., Hefny, A., Sra, S., Poczos, B., and Smola, A. (2016). Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pages 314–323. PMLR.

Reddi, S. J., Kale, S., and Kumar, S. (2019). On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. (2015). Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shen, Z., Ribeiro, A., Hassani, H., Qian, H., and Mi, C. (2019). Hessian aided policy gradient. In *International Conference on Machine Learning*, pages 5729–5738. PMLR.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. (2014). Deterministic policy gradient algorithms. In *International conference on machine learning*, pages 387–395. PMLR.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Sutton, R. S., McAllester, D. A., Singh, S. P., Mansour, Y., et al. (1999). Policy gradient methods for reinforcement learning with function approximation. In *NIPs*, volume 99, pages 1057–1063. Citeseer.

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

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

Wu, C., Rajeswaran, A., Duan, Y., Kumar, V., Bayen, A. M., Kakade, S., Mordatch, I., and Abbeel, P. (2018). Variance reduction for policy gradient with action-dependent factorized baselines. *International Conference on Learning Representations*.

Wu, T., Jiang, M., and Zhang, L. (2020a). Cooperative multiagent deep deterministic policy gradient (comaddpg) for intelligent connected transportation with unsignalized intersection. *Mathematical Problems in Engineering*, 2020.

Wu, Y., Zhang, W., Xu, P., and Gu, Q. (2020b). A finite time analysis of two time-scale actor critic methods. *Advances in Neural Information Processing Systems*.

Xiong, H., Xu, T., Liang, Y., and Zhang, W. (2020). Non-asymptotic convergence of adam-type reinforcement learning algorithms under markovian sampling. *arXiv preprint arXiv:2002.06286*.

Xu, P., Gao, F., and Gu, Q. (2020a). An improved convergence analysis of stochastic variance-reduced policy gradient. In *Uncertainty in Artificial Intelligence*, pages 541–551. PMLR.

Xu, P., Gao, F., and Gu, Q. (2020b). Sample efficient policy gradient methods with recursive variance reduction. *International Conference on Learning Representations*.

Xu, T., Liu, Q., and Peng, J. (2017). Stochastic variance reduction for policy gradient estimation. *arXiv preprint arXiv:1710.06034*.

Xu, T., Wang, Z., and Liang, Y. (2020c). Improving sample complexity bounds for (natural) actor-critic algorithms. *Advances in Neural Information Processing Systems*, 33.

Xu, T., Wang, Z., and Liang, Y. (2020d). Non-asymptotic convergence analysis of two time-scale (natural) actor-critic algorithms. *arXiv preprint arXiv:2005.03557*.

Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., and Tenenbaum, J. B. (2018). Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *arXiv preprint arXiv:1810.02338*.

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

Yuan, R., Gower, R. M., and Lazaric, A. (2021). A general sample complexity analysis of vanilla policy gradient. *arXiv preprint arXiv:2107.11433*.

Zanette, A., Cheng, C.-A., and Agarwal, A. (2021). Cautiously optimistic policy optimization and exploration with linear function approximation. *arXiv preprint arXiv:2103.12923*.

Zhang, J., Kim, J., O'Donoghue, B., and Boyd, S. (2021a). Sample efficient reinforcement learning with REINFORCE. *35th AAAI Conference on Artificial Intelligence*.

Zhang, J., Koppel, A., Bedi, A. S., Szepesvari, C., and Wang, M. (2020). Variational policy gradient method for reinforcement learning with general utilities. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4572–4583. Curran Associates, Inc.

Zhang, J., Ni, C., Yu, Z., Szepesvari, C., and Wang, M. (2021b). On the convergence and sample efficiency of variance-reduced policy gradient method. *arXiv preprint arXiv:2102.08607*.

Zoph, B. and Le, Q. V. (2016). Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*.

Supplementary Materials

6 Related work.

Momentum-based policy gradient. Conventional approaches to reducing the high variance in PG methods include adding the baselines (Sutton et al., 1999; Wu et al., 2018) and using the actor-critic algorithms (Konda and Tsitsiklis, 2000; Bhatnagar et al., 2009; Peters and Schaal, 2008). The idea of variance reduction, inspired by its successes in the stochastic nonconvex optimization (Johnson and Zhang, 2013; Allen-Zhu and Hazan, 2016; Reddi et al., 2016; Fang et al., 2018; Nguyen et al., 2017), is also incorporated to improve the PG methods (Xu et al., 2017; Papini et al., 2018; Xu et al., 2020b). In addition, momentum techniques, which are demonstrated as a powerful and generic recipe for accelerating stochastic gradient methods for nonconvex optimization (Qian, 1999; Kingma and Ba, 2015; Reddi et al., 2019), have also been extended to improve PG methods both in theory and in practice (Xiong et al., 2020; Yuan et al., 2020; Pham et al., 2020; Huang et al., 2020). Xiong et al. (2020) studies Adam-based policy gradient methods but only achieved $O(\epsilon^{-4})$ sample complexities, which is the same as the vanilla REINFORCE algorithm. A new STORM-PG method is proposed in Yuan et al. (2020), which incorporates momentum in the updates and matches the sample complexity of the SRVR-PG method proposed in Xu et al. (2020b) (and also VRMPO) while requiring only single-loop updates and large initialization batches, whereas SRVR-PG and VRMPO require double-loop updates and large batch sizes throughout all iterations. Concurrently, Pham et al. (2020) proposes a hybrid estimator combining the momentum idea with SARAH and considers a more general setting with regularization, and achieves the same $O(\epsilon^{-3})$ sample complexity and again with single-loop updates and large initialization batches. Finally, independently inspired by STORM algorithm for stochastic optimization in Cutkosky and Orabona (2019), Huang et al. (2020) proposes a class of momentum-based policy gradient algorithms, with adaptive time-steps, single-loop updates and small batch sizes, which matches the sample complexity in Xu et al. (2020b).

Global convergence of (stochastic) policy gradient. The understanding of the PG methods is mostly restricted to their convergence to stationary points of the value function (Sutton et al., 1999; Konda and Tsitsiklis, 2003; Papini et al., 2018). It was not until very recently that a series of works emerged to establish the global convergence properties of these algorithms. Fazel et al. (2018) shows that the linear quadratic regulator problem satisfies a gradient domination condition although it has a nonconvex landscape, implying that the PG methods could converge to the globally optimal policy. Bhandari and Russo (2019) generalizes the results in Fazel et al. (2018) from the linear quadratic regulator problem to several control tasks by relating the objective for policy gradient to the objective associated with the Bellman operator. For the soft-max parameterization, Mei et al. (2020, 2021) show that the value function satisfies a non-uniform Łojasiewicz inequality and a fast global convergence rate can be achieved if the exact PG is available. In addition, Agarwal et al. (2019) provides a fairly general characterization of global convergence for the PG methods and a sample complexity result for sample-based natural PG updates. By incorporating the variance reduction techniques in the PG methods, an improved sample complexity for the global convergence is established in Liu et al. (2020) for both PG and natural PG methods. When overparameterized neural networks are used for function approximation, the global convergence is proved for the (natural) PG methods (Wang et al., 2019) and for the trust-region policy optimization (Liu et al., 2019). Very recently, a series of non-asymptotic global convergence results (Hong et al., 2020; Xu et al., 2020c; Wu et al., 2020b; Xu et al., 2020d; Fu et al., 2020) have also been established for actor-critic algorithms with (natural) PG or proximal policy optimization used in the actor step. Apart from RL systems with a cumulative sum of rewards, the global convergence results of PG methods for RL systems whose objectives are a general utility function of the state-action occupancy measure are studied in Zhang et al. (2020, 2021b).

7 Notation

The set of real numbers is shown as $\mathbb{R}$. $u \sim \mathcal{U}$ means that u is a random vector sampled from the distribution $\mathcal{U}$. We use $|\mathcal{X}|$ to denote the number of elements in a finite set X. The notions $\mathbb{E}_\xi[\cdot]$ and $\mathbb{E}[\cdot]$ refer to the expectation over the random variable ξ and over all of the randomness. The notion $\mathrm{Var}[\cdot]$ refers to the variance. For vectors $x, y \in \mathbb{R}^d$, let $\|x\|_1$, $\|x\|_2$ and $\|x\|_\infty$ denote the ℓ_1-norm, ℓ_2-norm and ℓ_∞-norm. We use $\langle x, y\rangle$ to denote the inner product. For a matrix A, $A \succcurlyeq 0$ means that A is positive semi-definite. Given a variable x, the notation $a = \mathcal{O}(b(x))$ means that $a \le C \cdot b(x)$ for some constant $C > 0$ that is independent of x. Similarly, $a = \widetilde{\mathcal{O}}(b(x))$ indicates that the previous inequality may also depend on the function $\log(x)$, where $C > 0$ is again independent of x. We use $\mathcal{N}(\mu, \sigma^2)$ to denote a Gaussian distribution with the mean μ and the variance σ^2.

8 Supporting results

Proposition 8.1 (Lemma 1 in Cortes et al. (2010)) *Let $w(x) = P(x)/Q(x)$ be the importance weight for two distributions P and Q. The following identities hold for the expectation, second moment, and variance of $w(x)$:*

$$\mathbb{E}[w(x)] = 1,\ \mathbb{E}[w^2(x)] = d_2(P\|Q),\ \ Var[w(x)] = d_2(P\|Q) - 1,$$

where $d_2(P\|Q) = 2^{D(P\|Q)}$ and $D(P\|Q)$ is the Rényi divergence between the distributions P and Q.

Proposition 8.2 (Lemma 6.1 in Xu et al. (2020b)) *Under Assumptions 3.1 and 3.3, let $w(\tau|\theta_{t-1}, \theta_t) = p(\tau|\theta_{t-1}, \mu)/p(\tau|\theta_t, \mu)$. We have*

$$Var[w(\tau|\theta_{t-1}, \theta_t)] \le C_w^2 \left\|\theta_t - \theta_{t-1}\right\|_2^2,$$

for any state distribution μ, where $C_w = \sqrt{H(2HM_g^2 + M_h)(W+1)}$.

Lemma 8.3 *Suppose that $f(x)$ is L-smooth. Given $0 < \eta_t \le \frac{1}{2L}$ for all $t \ge 1$, let $\{x_t\}_{t=1}^T$ be generated by a general update of the form $x_{t+1} = x_t + \eta_t u_t$ and let $e_t = u_t - \nabla f(x_t)$. We have*

$$f(x_{t+1}) \ge f(x_t) + \frac{\eta_t}{4}\left\|u_t\right\|_2^2 - \frac{\eta_t}{2}\left\|e_t\right\|_2^2.$$

Proof. Since $f(f)$ is L-smooth, one can write

$$\begin{aligned}
&f(x_{t+1}) - f(x_t) - \langle u_t, x_{t+1} - x_t\rangle \\
=&f(x_{t+1}) - f(x_t) - \langle \nabla f(x_t), x_{t+1} - x_t\rangle + \langle \sqrt{\eta_t}(\nabla f(x_t) - u_t), \frac{1}{\sqrt{\eta_t}}(x_{t+1} - x_t)\rangle \\
\ge& -\frac{L}{2}\left\|x_{t+1} - x_t\right\|^2 - \frac{b\eta_t}{2}\left\|\nabla f(x_t) - u_t\right\|_2^2 - \frac{1}{2b\eta_t}\left\|x_{t+1} - x_t\right\|_2^2 \\
=&(-\frac{1}{2b\eta_t} - \frac{L}{2})\left\|x_{t+1} - x_t\right\|_2^2 - \frac{b\eta_t}{2}\left\|e_t\right\|_2^2,
\end{aligned}$$

where the constant $b > 0$ is to be determined later. By the above inequality and the definition of x_{t+1}, we have

$$\begin{aligned}
f(x_{t+1}) \ge& f(x_t) + \langle u_t, x_{t+1} - x_t\rangle - (\frac{1}{2b\eta_t} + \frac{L}{2})\left\|x_{t+1} - x_t\right\|_2^2 - \frac{b\eta_t}{2}\left\|e_t\right\|_2^2 \\
=& f(x_t) + \eta_t\left\|u_t\right\|^2 - (\frac{\eta_t}{2b} + \frac{L\eta_t^2}{2})\left\|u_t\right\|_2^2 - \frac{b\eta_t}{2}\left\|e_t\right\|_2^2
\end{aligned}$$

By choosing $b = 1$ and using the fact that $0 < \eta_t \le \frac{1}{2L}$, it holds that

$$\begin{aligned}
f(x_{t+1}) \ge& f(x_t) + (\frac{\eta_t}{2} - \frac{L\eta_t^2}{2})\left\|u_t\right\|_2^2 - \frac{\eta_t}{2}\left\|e_t\right\|_2^2 \\
\ge& f(x_t) + \frac{\eta_t}{4}\left\|u_t\right\|_2^2 - \frac{\eta_t}{2}\left\|e_t\right\|_2^2.
\end{aligned}$$

This completes the proof. □

9 Proofs of results in Section 4.1

9.1 Proof of Lemma 4.2

Proof. We first define the following set of "bad" iterates:

$$I^+ = \left\{ t \in \{1, \dots, T\} \,\middle|\, \|\nabla_\theta L_{\lambda,\mu}(\theta_t)\|_2 \geq \frac{\lambda}{2|\mathcal{S}||\mathcal{A}|} \right\}, \tag{10}$$

which counts the number of iterates such that the norms of the first-order stationary points of the entropy-regularized objective are large. Then, one can show that for every $\epsilon > 0$ and $\lambda = \frac{\epsilon(1-\gamma)}{4\left\|\frac{d_\rho^{\pi_{\theta^*}}}{\mu}\right\|_\infty}$, we have that $J_\rho(\theta^*) - J_\rho(\theta) \leq \frac{\epsilon}{2}$ for all $k \in \{0, \dots, K\}/I^+$, while $J_\rho(\theta^*) - J_\rho(\theta) \leq 1/(1-\gamma)$ holds trivially for all $k \in I^+$ due to the assumption that the rewards are between 0 and 1. Then, by controlling the number of "bad" iterates, we obtain the desired optimality guarantee. For simplicity, assume for now that $|I^+| > 0$. Since η_t is non-increasing in t, we have

$$\begin{aligned}
\sum_{t=1}^{T} \eta_t \|\nabla L_{\lambda,\mu}(\theta_t)\|_2^2 &\geq \sum_{t \in I^+} \eta_t \|\nabla L_{\lambda,\mu}(\theta_t)\|_2^2 \\
&\geq \frac{\lambda^2}{4|\mathcal{S}|^2|\mathcal{A}|^2} \sum_{t \in I^+} \eta_t \\
&\geq \frac{\lambda^2}{4|\mathcal{S}|^2|\mathcal{A}|^2} \sum_{t=T-|I^+|+1}^{T} \eta_t \\
&\geq \frac{\lambda^2 |I^+| \eta_T}{4|\mathcal{S}|^2|\mathcal{A}|^2}.
\end{aligned}$$

Thus,

$$|I^+| \leq \frac{4|\mathcal{S}|^2|\mathcal{A}|^2 \sum_{t=1}^{T} \eta_t \|\nabla L_{\lambda,\mu}(\theta_t)\|_2^2}{\lambda^2 \eta_T}.$$

Since $J_\rho(\theta) \in [0, \frac{1}{1-\gamma}]$ for every θ, it holds that $J_\rho(\theta^*) - J_\rho(\theta_t) \leq \frac{1}{1-\gamma}$ for all $t \in I^+$. In addition, by Lemma 4.1 and the choice of $\lambda = \frac{\epsilon(1-\gamma)}{4\left\|\frac{d_\rho^{\pi_{\theta^*}}}{\mu}\right\|_\infty}$, it holds that

$$J_\rho(\theta^*) - J_\rho(\theta_t) \leq \frac{\epsilon}{2}, \quad \forall t \notin I^+.$$

Therefore,

$$\begin{aligned}
\sum_{t=1}^{T} (J_\rho(\theta^*) - J_\rho(\theta_t)) &= \sum_{t \in I^+} (J_\rho(\theta^*) - J_\rho(\theta_t)) + \sum_{t \notin I^+} (J_\rho(\theta^*) - J_\rho(\theta_t)) \\
&\leq |I^+| \frac{1}{1-\gamma} + (T - |I^+|) \frac{\epsilon}{2} \\
&\leq \frac{4|\mathcal{S}|^2|\mathcal{A}|^2 \sum_{t=1}^{T} \eta_t \|\nabla L_{\lambda,\mu}(\theta_t)\|_2^2}{\lambda^2 \eta_T (1-\gamma)} + \frac{T\epsilon}{2} \\
&\leq \frac{4|\mathcal{S}|^2|\mathcal{A}|^2 \sum_{t=1}^{T} \eta_t \|\nabla L_{\lambda,\mu}(\theta_t)\|_2^2}{\lambda^2 \eta_T (1-\gamma)} + \frac{T\epsilon}{2}.
\end{aligned} \tag{11}$$

Now if $|I^+| = 0$,

$$\sum_{t=1}^{T} (J_\rho(\theta^*) - J_\rho(\theta_t)) \leq \frac{T\epsilon}{2}$$

and hence (11) always holds. This completes the proof. □

9.2 Proof of Lemma 4.3

We first notice that Assumption 3.1 is satisfied by the soft-max parameterization with $M_g = 2$ and $M_h = 1$.

Lemma 9.1 *For the soft-max parameterization, Assumption 3.1 is satisfied with $M_g = 2$ and $M_h = 1$.*

Proof. For the soft-max parameterization, we have

$$\frac{\alpha \log \pi_\theta(a|s)}{\alpha\theta(s,\cdot)} = \mathbf{1}_a - \mathbb{E}_{a'\sim\pi_\theta(\cdot|s)}\mathbf{1}_{a'},$$

where $\mathbf{1}_a \in \mathbb{R}^{|\mathcal{A}|}$ is a vector with zero entries except one nonzero entry corresponding to the action a. In addition, $\frac{\alpha \log \pi_\theta(a|s)}{\alpha\theta(s',\cdot)} = \mathbf{0}$ for all $s \neq s'$. Hence, $\|\nabla_\theta \log \pi_\theta(a|s)\|_2 \leq 2$ for every $(s,a) \in \mathcal{S}\times\mathcal{A}$.

Similarly, we have

$$\frac{\alpha^2 \log \pi_\theta(a|s)}{\alpha\theta(s,\cdot)^2} = \left(\frac{d\pi_\theta(\cdot|s)}{d\theta(s,\cdot)}\right)^\top = \mathrm{diag}(\pi(\cdot|s)) - \pi(\cdot|s)\pi(\cdot|s)^\top.$$

From Lemma 22 of Mei et al. (2020), we know that the largest eigenvalue of the matrix $\mathrm{diag}(\pi(\cdot|s)) - \pi(\cdot|s)\pi(\cdot|s)^\top$ is less than 1. Thus, $\left\|\nabla_\theta^2 \log \pi_\theta(a|s)\right\|_2 \leq 1$. □

Lemma 9.2 *Suppose that the stochastic policy gradient u_t^H is generated by Algorithm 1 with the soft-max parameterization. Let $e_t^H = u_t^H + \frac{\lambda}{|\mathcal{A}||\mathcal{S}|}\sum_{s,a}\nabla \log \pi_\theta(a|s) - \nabla L_{\lambda,\mu}^H(\theta_t)$. It holds that*

$$\mathbb{E}\left[\eta_{t-1}^{-1}\|e_t^H\|_2^2\right] \leq \mathbb{E}\left[\eta_{t-1}^{-1}(1-\beta_t)^2\|e_{t-1}^H\|_2^2 + \frac{2\eta_{t-1}^{-1}\beta_t^2\sigma^2}{B} + \frac{4b^2\eta_{t-1}^{-1}(1-\beta_t)^2}{B}\|\theta_t - \theta_{t-1}\|_2^2\right],$$

where $b^2 = L_g^2 + G^2C_w^2$, $L_g = M_h/(1-\gamma)^2$, $G = M_g/(1-\gamma)^2$ and $C_w = \sqrt{H(2HM_g^2 + M_h)(W+1)}$.

Proof. First note that $e_t^H = u_t^H - \nabla J_\mu^H(\theta)$. Then, by the definition of u_t^H, we have

$$u_t^H - u_{t-1}^H = -\beta_t u_{t-1}^H + \frac{\beta_t}{B}\sum_{i=1}^B g(\tau_i^H|\theta_t,\mu) + \frac{(1-\beta_t)}{B}\sum_{i=1}^B (g(\tau_i^H|\theta_t,\mu) - w(\tau_i^H|\theta_{t-1},\theta_t)g(\tau_i^H|\theta_{t-1},\mu)).$$

As a result,

$$
\begin{aligned}
\mathbb{E}\left[\eta_{t-1}^{-1}\|e_t^H\|_2^2\right] =& \mathbb{E}\left[\eta_{t-1}^{-1}\|\nabla J_\rho^H(\theta_{t-1}) - u_{t-1} + \nabla J_\rho^H(\theta_t) - \nabla J_\rho^H(\theta_{t-1}) - (u_t^H - u_{t-1}))\|_2^2\right] \\
=& \mathbb{E}\Big[\eta_{t-1}^{-1}\|\nabla J_\rho^H(\theta_{t-1}) - u_{t-1} + \nabla J_\rho^H(\theta_t) - \nabla J_\rho^H(\theta_{t-1}) \\
&+\beta_t u_{t-1} - \frac{\beta_t}{B}\sum_{i=1}^B g(\tau_i^H|\theta_t,\mu) - \frac{(1-\beta_t)}{B}\sum_{i=1}^B (g(\tau_i^H|\theta_t,\mu) - w(\tau_i^H|\theta_{t-1},\theta_t)g(\tau_i^H|\theta_{t-1},\mu)))\|_2^2\Big] \\
=& \mathbb{E}\Bigg[\eta_{t-1}^{-1}\|(1-\beta_t)(\nabla J_\rho^H(\theta_{t-1}) - u_{t-1}) + \beta_t(\nabla J_\rho^H(\theta_t) - \frac{1}{B}\sum_{i=1}^B g(\tau_i^H|\theta_t,\mu)) \\
&-\frac{(1-\beta_t)}{B}\sum_{i=1}^B (g(\tau_i^H|\theta_t,\mu) - w(\tau_i^H|\theta_{t-1},\theta_t)g(\tau_i^H|\theta_{t-1},\mu)) - (\nabla J_\rho^H(\theta_t) - \nabla J_\rho^H(\theta_{t-1})))\|_2^2\Bigg] \\
=& \eta_{t-1}^{-1}(1-\beta_t)^2\mathbb{E}\left[\|\nabla J_\rho^H(\theta_{t-1}) - u_{t-1}\|_2^2\right] + \eta_{t-1}^{-1}\mathbb{E}\Bigg[\|\beta_t(\nabla J_\rho^H(\theta_t) - \frac{1}{B}\sum_{i=1}^B g(\tau_i^H|\theta_t,\mu)) \\
&-\frac{(1-\beta_t)}{B}\sum_{i=1}^B (g(\tau_i^H|\theta_t,\mu) - w(\tau_i^H|\theta_{t-1},\theta_t)g(\tau_i^H|\theta_{t-1},\mu)) - (\nabla J_\rho^H(\theta_t) - \nabla J_\rho^H(\theta_{t-1})))\|_2^2\Bigg] \\
\le& \eta_{t-1}^{-1}(1-\beta_t)^2\mathbb{E}\left[\|\nabla J_\rho^H(\theta_{t-1}) - u_{t-1}\|_2^2\right] + 2\eta_{t-1}^{-1}\beta_t^2\mathbb{E}\Bigg[\|(\nabla J_\rho^H(\theta_t) - \frac{1}{B}\sum_{i=1}^B g(\tau_i^H|\theta_t,\mu))\|_2^2\Bigg] \\
&+ 2\eta_{t-1}^{-1}(1-\beta_t)^2\mathbb{E}\Bigg[\|\frac{1}{B}\sum_{i=1}^B (g(\tau_i^H|\theta_t,\mu) - w(\tau_i^H|\theta_{t-1},\theta_t)g(\tau_i^H|\theta_{t-1},\mu)) - (\nabla J_\rho^H(\theta_t) - \nabla J_\rho^H(\theta_{t-1})))\|_2^2\Bigg] \\
=& \eta_{t-1}^{-1}(1-\beta_t)^2\mathbb{E}\left[\|e_{t-1}^H\|_2^2\right] + 2\eta_{t-1}^{-1}\beta_t^2\frac{1}{B}\mathbb{E}\left[\|g(\tau_i^H|\theta_t,\mu) - \nabla J_\rho^H(\theta_t)\|_2^2\right] \\
&+ 2\eta_{t-1}^{-1}(1-\beta_t)^2\frac{1}{B}\mathbb{E}\left[\|g(\tau_i^H|\theta_t,\mu) - w(\tau_i^H|\theta_{t-1},\theta_t)g(\tau_i^H|\theta_{t-1},\mu) - (\nabla J_\rho^H(\theta_t) - \nabla J_\rho^H(\theta_{t-1}))\|_2^2\right] \\
\le& \eta_{t-1}^{-1}(1-\beta_t)^2\mathbb{E}\left[\|e_{t-1}^H\|_2^2\right] + \frac{2\eta_{t-1}^{-1}\beta_t^2\sigma^2}{B} \\
&+ 2\eta_{t-1}^{-1}(1-\beta_t)^2\frac{1}{B}\mathbb{E}\left[\|g(\tau_i^H|\theta_t,\mu) - w(\tau_i^H|\theta_{t-1},\theta_t)g(\tau_i^H|\theta_{t-1},\mu)\|_2^2\right],
\end{aligned}
$$

where the fourth equality is due to $\mathbb{E}_{\tau_i^H}[g(\tau_i^H|\theta_t,\mu)] = \nabla J_\rho^H(\theta_t)$ and $\mathbb{E}_{\tau_i^H}[g(\tau_i^H|\theta_t,\mu) - w(\tau_i^H|\theta_{t-1},\theta_t)g(\tau_i^H|\theta_{t-1},\mu)] = \nabla J_\rho^H(\theta_t) - \nabla J_\rho^H(\theta_{t-1})$, the first inequality follows from Young's inequality, the second inequality holds by $\mathbb{E}\|\frac{1}{B}\sum_{i=1}^B \xi_i - \mathbb{E}[\xi_i]\|_2^2 = \frac{1}{B}\mathbb{E}\|\xi_i - \mathbb{E}[\xi_i]\|_2^2$ for the i.i.d. samples of $\{\xi_i\}_{i=1}^B$, and the last inequality is due to the bounded variance of stochastic policy gradient under the soft-max parameterization and $\mathbb{E}\|\xi - \mathbb{E}[\xi]\|_2^2 \le \mathbb{E}\|\xi\|_2^2$.

In addition,

$$
\begin{aligned}
&\mathbb{E}\left[\|g(\tau_i^H|\theta_t,\mu) - w(\tau_i^H|\theta_{t-1},\theta_t)g(\tau_i^H|\theta_{t-1},\mu)\|_2^2\right] \\
&= \mathbb{E}\left[\|g(\tau_i^H|\theta_t,\mu) - g(\tau_i^H|\theta_{t-1},\mu) + g(\tau_i^H|\theta_{t-1},\mu) - w(\tau_i^H|\theta_{t-1},\theta_t)g(\tau_i^H|\theta_{t-1},\mu)\|_2^2\right] \\
&\le 2\mathbb{E}\left[\|g(\tau_i^H|\theta_t,\mu) - g(\tau_i^H|\theta_{t-1},\mu)\|_2^2 + 2\mathbb{E}[\|g(\tau_i^H|\theta_{t-1},\mu) - w(\tau_i^H|\theta_{t-1},\theta_t)g(\tau_i^H|\theta_{t-1},\mu)\|_2^2\right] \\
&\le 2L_g^2\mathbb{E}\left[\|\theta_t - \theta_{t-1})\|_2^2\right] + 2G^2\mathbb{E}\left[\|1 - w(\tau_i^H|\theta_{t-1},\theta_t)\|_2^2\right] \\
&\le 2L_g^2\mathbb{E}\left[\|\theta_t - \theta_{t-1})\|_2^2\right] + 2G^2\mathrm{Var}(w(\tau_i^H|\theta_{t-1},\theta_t)) \\
&\le 2(L_g^2 + G^2C_w^2)\mathbb{E}\left[\|\theta_t - \theta_{t-1})\|_2^2\right],
\end{aligned}
$$

where the second inequality follows from Lemma 3.2, and the third inequality is due to Proposition 8.1, and the last inequality holds by Proposition 8.2. By selecting $b^2 = L_g^2 + G^2C_w^2$, we have

$$
\mathbb{E}\left[\eta_{t-1}^{-1}\|e_t^H\|_2^2\right] \le \mathbb{E}\left[\eta_{t-1}^{-1}(1-\beta_t)^2\|e_{t-1}^H\|_2^2 + \frac{2\eta_{t-1}^{-1}\beta_t^2\sigma^2}{B} + \frac{4b^2\eta_{t-1}^{-1}(1-\beta_t)^2}{B}\|\theta_t - \theta_{t-1}\|_2^2\right],
$$

which completes the proof. □

9.2.1 Proof of Lemma 4.3

Proof. From Proposition 3.2, we know that $J_\mu(\theta)$ is L-smooth. Since $\left\|\nabla^2 \log \pi_\theta(a|s)\right\|_2 \le 1$ for the soft-max parameterization, it holds that $L_{\lambda,\mu}(\theta)$ is L_λ-smooth, where $L_\lambda \coloneqq L + \lambda$.

Due to $m \ge (2L_\lambda k)^3$, we have $\eta_t \le \eta_0 = \frac{k}{m^{1/3}} \le \frac{1}{2L_\lambda}$. Since $\eta_t \le \frac{1}{2L_\lambda}$, we obtain that $\beta_{t+1} = c\eta_t^2 \le \frac{c\eta_t}{2L_\lambda} \le \frac{ck}{2L_\lambda m^{1/3}} \le 1$. Now, it results from Lemma 9.2 that

$$\begin{aligned}
&\mathbb{E}\left[\eta_t^{-1}\|e_{t+1}^H\|_2^2 - \eta_{t-1}^{-1}\|e_t^H\|_2^2\right] \\
&\le \mathbb{E}\left[(\eta_t^{-1}(1-\beta_{t+1})^2 - \eta_{t-1}^{-1})\|e_t^H\|_2^2 + \frac{2\eta_t^{-1}\beta_{t+1}^2\sigma^2}{B} + \frac{4b^2\eta_t^{-1}(1-\beta_{t+1})^2}{B}\|\theta_{t+1}-\theta_t\|_2^2\right] \\
&\le \mathbb{E}\left[(\eta_t^{-1}(1-\beta_{t+1}) - \eta_{t-1}^{-1})\|e_t^H\|_2^2 + \frac{2\eta_t^{-1}\beta_{t+1}^2\sigma^2}{B} + \frac{4b^2\eta_t^{-1}}{B}\|\theta_{t+1}-\theta_t\|_2^2\right],
\end{aligned}$$

where the last inequality holds by $0 < \beta_{t+1} \le 1$. Since the function $x^{1/3}$ is concave, we have $(x+y)^{1/3} \le x^{1/3} + yx^{-2/3}/3$. Then, we have

$$\begin{aligned}
\eta_t^{-1} - \eta_{t-1}^{-1} =& \frac{1}{k}((m+t)^{1/3} - (m+t-1)^{1/3}) \le \frac{1}{3k(m+t-1)^{2/3}} \\
\le& \frac{1}{3k(m/2+t)^{2/3}} \le \frac{2^{2/3}}{3k^3}\eta_t^2 \le \frac{2^{2/3}}{6k^3L}\eta_t \le \frac{1}{3k^3L}\eta_t,
\end{aligned}$$

where the second inequality holds by $m \ge 2$, and the forth inequality uses the property $0 < \eta_t \le \frac{1}{2L_\lambda}$. Then, it holds that

$$(\eta_t^{-1}(1-\beta_{t+1}) - \eta_{t-1}^{-1})\|e_t^H\|_2^2 = \left(\frac{1}{3k^3L} - c\right)\eta_t\left\|e_t^H\right\|_2^2 = -96b^2\eta_t\left\|e_t^H\right\|_2^2,$$

where the last equality is based on the relation $c = \frac{1}{3k^3L_\lambda} + 96b^2$. Combining the above results yields that

$$\mathbb{E}[\eta_t^{-1}\|e_{t+1}^H\|_2^2 - \eta_{t-1}^{-1}\|e_t^H\|_2^2] \le \mathbb{E}\left[-96b^2\eta_t\|e_t^H\|_2^2 + \frac{2\eta_t^{-1}\beta_{t+1}^2\sigma^2}{B} + \frac{4b^2\eta_t^{-1}}{B}\|\theta_{t+1}-\theta_t\|_2^2\right]. \tag{12}$$

To streamline the presentation, we denote $u_{t,\lambda}^H \coloneqq \frac{1}{\eta_t}(\theta_{t+1} - \theta_t)$. By summing up the above inequality and dividing the both sides by $96b^2$, we obtain

$$\frac{1}{96b^2}\left(\mathbb{E}\left[\frac{\left\|e_{T+1}^H\right\|_2^2}{\eta_T} - \frac{\left\|e_1^H\right\|_2^2}{\eta_0}\right]\right) \le \sum_{t=1}^T \mathbb{E}\left[\frac{c^2\eta_t^3\sigma^2}{48b^2B} - \eta_t\left\|e_t^H\right\|_2^2 + \frac{\eta_t}{24B}\left\|u_{t,\lambda}^H\right\|_2^2\right] \tag{13}$$

For $\eta_t\left\|u_{t,\lambda}^H\right\|_2^2$, it follows from Lemma 8.3 that

$$\frac{\eta_t}{4}\left\|u_{t,\lambda}^H\right\|_2^2 \le L_{\lambda,\mu}^H(\theta_{t+1}) - L_{\lambda,\mu}^H(\theta_t) + \frac{\eta_t}{2}\left\|e_t^H\right\|^2.$$

Then, it holds that

$$\begin{aligned}
\frac{1}{96b^2}\left(\mathbb{E}\left[\frac{\left\|e_{T+1}^H\right\|_2^2}{\eta_T} - \frac{\left\|e_1^H\right\|_2^2}{\eta_0}\right]\right) &\le \sum_{\tau=1}^T \mathbb{E}\left[\frac{c^2\eta_t^3\sigma^2}{48b^2B} - \frac{(12B-1)\eta_t}{12B}\left\|e_t^H\right\|_2^2 + \frac{1}{6B}\left(L_{\lambda,\mu}^H(\theta_{t+1}) - L_{\lambda,\mu}^H(\theta_t)\right)\right] \\
&\le \sum_{t=1}^T\left(\frac{c^2\eta_t^3\sigma^2}{48b^2B} - \mathbb{E}\left[\frac{11\eta_t}{12}\left\|e_t^H\right\|_2^2\right]\right) + \frac{1}{6B}\left(L_{\lambda,\mu}^H(\theta^*) - L_{\lambda,\mu}^H(\theta_1)\right) \\
&\le \frac{c^2\sigma^2k^3}{48b^2B}\sum_{t=1}^T\frac{1}{m+t} - \sum_{t=1}^T\mathbb{E}\left[\frac{11\eta_t}{12}\left\|e_t^H\right\|_2^2\right] + \frac{1}{6B}\left(L_{\lambda,\mu}^H(\theta^*) - L_{\lambda,\mu}^H(\theta_1)\right) \\
&\le \frac{c^2\sigma^2k^3}{48b^2B}\sum_{t=1}^T\frac{1}{2+t} - \sum_{t=1}^T\mathbb{E}\left[\frac{11\eta_t}{12}\left\|e_t^H\right\|_2^2\right] + \frac{1}{6B}\left(L_{\lambda,\mu}^H(\theta^*) - L_{\lambda,\mu}^H(\theta_1)\right) \\
&\le \frac{c^2\sigma^2k^3\ln(T+2)}{48b^2B} - \sum_{t=1}^T\mathbb{E}\left[\frac{11\eta_t}{12}\left\|e_t^H\right\|_2^2\right] + \frac{1}{6B}\left(L_{\lambda,\mu}^H(\theta^*) - L_{\lambda,\mu}^H(\theta_1)\right).
\end{aligned} \tag{14}$$

By rearranging the above inequality, we have

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}\left[\frac{11\eta_t}{12}\left\|e_t^H\right\|_2^2\right]\leq&\frac{c^2\sigma^2k^3\ln(T+2)}{48b^2B}+\frac{1}{96b^2\eta_0}\mathbb{E}\left[\left\|e_1^H\right\|_2^2\right]+\frac{1}{6B}\left(L_{\lambda,\mu}^H(\theta^*)-L_{\lambda,\mu}^H(\theta_1)\right)\\ \leq&\frac{1}{B}\left(\frac{c^2\sigma^2k^3\ln(T+2)}{48b^2}+\frac{m^{1/3}\sigma^2}{96b^2k}+\frac{1}{6}\left(L_{\lambda,\mu}^H(\theta^*)-L_{\lambda,\mu}^H(\theta_1)\right)\right).\end{aligned}$$

Multiplying both sides by $\frac{12}{11}$ yields that

$$\sum_{t=1}^{T}\mathbb{E}\left[\eta_t\left\|e_t^H\right\|_2^2\right]\leq\frac{1}{B}\left(\frac{c^2\sigma^2k^3\ln(T+2)}{44b^2}+\frac{m^{1/3}\sigma^2}{88b^2k}+\frac{1}{22}\left(L_{\lambda,\mu}^H(\theta^*)-L_{\lambda,\mu}^H(\theta_1)\right)\right).$$

To bound $\sum_{t=1}^{T}\mathbb{E}\left[\eta_t\left\|u_{t,\lambda}^H\right\|_2^2\right]$, we define a Lyapunov function $\Phi_t(\theta_t)=L_{\lambda,\mu}^H(\theta_t)-\frac{1}{192b^2\eta_{t-1}}\left\|e_t^H\right\|_2^2$ for all $t\geq 1$. Then,

$$\begin{aligned}\mathbb{E}[\Phi_{t+1}-\Phi_t]=&\mathbb{E}\left[L_{\lambda,\mu}^H(\theta_{t+1})-L_{\lambda,\mu}^H(\theta_t)-\frac{1}{192b^2\eta_t}\left\|e_{t+1}^H\right\|_2^2+\frac{1}{192b^2\eta_{t-1}}\left\|e_t^H\right\|_2^2\right]\\ \geq&\mathbb{E}\left[-\frac{\eta_t}{2}\left\|e_t^H\right\|_2^2+\frac{\eta_t}{4}\left\|u_{t,\lambda}^H\right\|_2^2-\frac{1}{192b^2\eta_t}\left\|e_{t+1}^H\right\|_2^2+\frac{1}{192b^2\eta_{t-1}}\left\|e_t^H\right\|_2^2\right]\\ \geq&\mathbb{E}\left[\frac{\eta_t}{4}\left\|u_{t,\lambda}^H\right\|_2^2-\frac{\beta_t^2\sigma^2}{96b^2B\eta_t}-\frac{\eta_t}{48B}\|u_{t,\lambda}^H\|^2\right]\\ \geq&\mathbb{E}\left[\frac{11\eta_t}{48}\left\|u_{t,\lambda}^H\right\|_2^2-\frac{c^2\eta_t^3\sigma^2}{96b^2B}\right],\end{aligned}$$

where the first inequality holds by Lemma 8.3 and the second inequality holds due to (12). Summing the above inequality over t from 1 to T, we obtain

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}\left[\eta_t\left\|u_{t,\lambda}^H\right\|_2^2\right]\leq&\mathbb{E}\left[\frac{48}{11}(\Phi_{T+1}-\Phi_1)+\sum_{t=1}^{T}\frac{c^2\eta_t^3\sigma^2}{22b^2B}\right]\\ \leq&\mathbb{E}\left[\frac{48}{11}\left(L_{\lambda,\mu}^H(\theta^*)-L_{\lambda,\mu}^H(\theta_1)\right)+\frac{1}{44b^2\eta_0}\mathbb{E}\|e_1^H\|_2^2+\frac{c^2\sigma^2k^3}{22b^2B}\sum_{t=1}^{T}\frac{1}{m+t}\right]\\ \leq&\mathbb{E}\left[\frac{48}{11}\left(L_{\lambda,\mu}^H(\theta^*)-L_{\lambda,\mu}^H(\theta_1)\right)+\frac{1}{44b^2\eta_0}\mathbb{E}\|e_1^H\|_2^2+\frac{c^2\sigma^2k^3}{22b^2B}\sum_{t=1}^{T}\frac{1}{2+t}\right]\\ \leq&\mathbb{E}\left[\frac{48}{11}\left(L_{\lambda,\mu}^H(\theta^*)-L_{\lambda,\mu}^H(\theta_1)\right)+\frac{\sigma^2m^{1/3}}{44b^2kB}+\frac{c^2\sigma^2k^3}{22b^2B}\ln(2+T)\right]\\ \leq&\Gamma_2+\frac{\Gamma_3}{B},\end{aligned}$$

where $\Gamma_2=\frac{48}{11}\left(L_{\lambda,\mu}^H(\theta^*)-L_{\lambda,\mu}^H(\theta_1)\right)$ and $\Gamma_3=(\frac{\sigma^2m^{1/3}}{44b^2k}+\frac{c^2\sigma^2k^3}{22b^2}\ln(2+T))$. Finally, by the triangle inequality, we have

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}\left[\eta_t\left\|\nabla L_{\lambda,\mu}^H(\theta_t)\right\|_2^2\right]\leq&\sum_{t=1}^{T}\mathbb{E}\left[\eta_t\left\|u_{t,\lambda}^H\right\|_2^2\right]+\sum_{t=1}^{T}\mathbb{E}\left[\eta_t\left\|e_t^H\right\|_2^2\right]\\ \leq&\Gamma_2+\frac{\Gamma_1+\Gamma_3}{B}.\end{aligned}$$

This completes the proof. □

9.3 Proof of Theorem 4.4

Proof. From Proposition 3.2, we have

$$\begin{aligned}\sum_{t=1}^{T}\eta_t\mathbb{E}\left[\|\nabla L_{\lambda,\mu}(\theta_t)\|_2^2\right] &= \sum_{t=1}^{T}\eta_t\mathbb{E}\left[\left\|\nabla L_{\lambda,\mu}^H(\theta_t)\right\|_2^2+\left\|\nabla L_{\lambda,\mu}^H(\theta_t)-\nabla L_{\lambda,\mu}(\theta_t)\right\|_2^2\right]\\ &\le \sum_{t=1}^{T}\eta_t\mathbb{E}\left[\left\|\nabla L_{\lambda,\mu}^H(\theta_t)\right\|_2^2+\left\|\nabla J_{\mu}^H(\theta_t)-\nabla J_{\mu}(\theta_t)\right\|_2^2\right]\\ &\le \sum_{t=1}^{T}\eta_t\mathbb{E}\left[\left\|\nabla L_{\lambda,\mu}^H(\theta_t)\right\|_2^2\right]+\left(M_g\left(\frac{H+1}{1-\gamma}+\frac{\gamma}{(1-\gamma)^2}\right)\gamma^H\right)^2\sum_{t=1}^{T}\eta_t.\end{aligned}$$

In light of Lemma 4.2, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\left(J_\rho(\theta^*)-J_\rho(\theta_t)\right)\right] &\le \left\|\frac{d_\rho^{\pi_{\theta^*}}}{\mu}\right\|_\infty^2\frac{64|\mathcal{S}|^2|\mathcal{A}|^2\sum_{t=1}^{T}\eta_t\mathbb{E}[\|\nabla L_{\lambda,\mu}(\theta_t)\|_2^2]}{\epsilon^2\eta_T(1-\gamma)^3}+\frac{T\epsilon}{2}\\ &\le \left\|\frac{d_\rho^{\pi_{\theta^*}}}{\mu}\right\|_\infty^2\frac{64|\mathcal{S}|^2|\mathcal{A}|^2\sum_{t=1}^{T}\eta_t\mathbb{E}[\left\|\nabla L_{\lambda,\mu}^H(\theta_t)\right\|_2^2]}{\epsilon^2\eta_T(1-\gamma)^3}+\frac{T\epsilon}{2}\\ &\quad+\left\|\frac{d_\rho^{\pi_{\theta^*}}}{\mu}\right\|_\infty^2\frac{64|\mathcal{S}|^2|\mathcal{A}|^2\sum_{t=1}^{T}\eta_t}{\epsilon^2\eta_T(1-\gamma)^3}\left(M_g\left(\frac{H+1}{1-\gamma}+\frac{\gamma}{(1-\gamma)^2}\right)\gamma^H\right)^2.\end{aligned}$$

By choosing $\eta_t=\frac{k}{(m+t)^{1/3}}$, it results from Lemma 4.3 that

$$\begin{aligned}\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\left(J_\rho(\theta^*)-J_\rho(\theta_t)\right)\right] &\le \left\|\frac{d_\rho^{\pi_{\theta^*}}}{\mu}\right\|_\infty^2\frac{64|\mathcal{S}|^2|\mathcal{A}|^2(m+T)^{1/3}}{\epsilon^2k(1-\gamma)^3T}\left(\Gamma_2+\frac{\Gamma_1+\Gamma_3}{B}\right)+\frac{\epsilon}{2}\\ &\quad+\left\|\frac{d_\rho^{\pi_{\theta^*}}}{\mu}\right\|_\infty^2\frac{96|\mathcal{S}|^2|\mathcal{A}|^2(m+T)}{\epsilon^2(1-\gamma)^3T}\left(M_g\left(\frac{H+1}{1-\gamma}+\frac{\gamma}{(1-\gamma)^2}\right)\gamma^H\right)^2.\end{aligned}$$

Notice that, in order to guarantee

$$\left\|\frac{d_\rho^{\pi_{\theta^*}}}{\mu}\right\|_\infty^2\frac{96|\mathcal{S}|^2|\mathcal{A}|^2(m+T)}{\epsilon^2(1-\gamma)^3T}\left(M_g\left(\frac{H+1}{1-\gamma}+\frac{\gamma}{(1-\gamma)^2}\right)\gamma^H\right)^2\le\frac{\epsilon}{4},$$

it suffices to have

$$H=\mathcal{O}\left(\log_\gamma\left(\frac{(1-\gamma)\epsilon}{|\mathcal{S}||\mathcal{A}|\left\|\frac{d_\rho^{\pi_{\theta^*}}}{\mu}\right\|_\infty}\right)\right).$$

Recall that $\Gamma_1=\frac{c^2\sigma^2k^3\ln(T+2)}{44b^2}+\frac{m^{1/3}\sigma^2}{88b^2k}+\frac{1}{22}\left(L_{\lambda,\mu}^H(\theta^*)-L_{\lambda,\mu}^H(\theta_1)\right)$, $\Gamma_2=\frac{48}{11}\left(L_{\lambda,\mu}^H(\theta^*)-L_{\lambda,\mu}^H(\theta_1)\right)$ and $\Gamma_3=(\frac{\sigma^2m^{1/3}}{44b^2k}+\frac{c^2\sigma^2k^3}{22b^2}\ln(2+T))$. By only considering the dependencies on the parameters c,σ,b,λ and m, we have

$$\begin{aligned}\Gamma_2+\frac{\Gamma_1+\Gamma_3}{B} &=\widetilde{\mathcal{O}}\left(\frac{c^2\sigma^2}{b^2}+\frac{m^{1/3}\sigma^2}{b^2}+\frac{\sigma^2m^{1/3}}{b^2}+\frac{c^2\sigma^2}{b^2}+L_{\lambda,\mu}^H(\theta^*)-L_{\lambda,\mu}^H(\theta_1)\right)\\ &=\widetilde{\mathcal{O}}\left(\frac{\sigma^2}{b^2}\left(c^2+m^{1/3}\right)+L_{\lambda,\mu}^H(\theta^*)-L_{\lambda,\mu}^H(\theta_1)\right).\end{aligned}\tag{15}$$

In addition, by the definitions $b^2=L_g^2+G^2C_w^2$, $L_g=M_h/(1-\gamma)^2$, $G=M_g/(1-\gamma)^2$, $C_w=\sqrt{H(2HM_g^2+M_h)(W+1)}$, $c=\frac{1}{3k^3L_\lambda}+96b^2$, $m=\max\{2,(2L_\lambda k)^3,(\frac{ck}{2L_\lambda})^3\}$, and $\lambda=\frac{\epsilon(1-\gamma)}{4\left\|\frac{d_\rho^{\pi_{\theta^*}}}{\mu}\right\|_\infty}$, we know that

$$b^2=\mathcal{O}\left(\frac{1+W}{(1-\gamma)^4}\right),\ \sigma^2=\mathcal{O}\left(\frac{1}{(1-\gamma)^4}\right),\ c=\mathcal{O}\left(\frac{1+W}{(1-\gamma)^4}\right),\ m^{1/3}=\mathcal{O}\left(\frac{\epsilon(1-\gamma)}{\left\|\frac{d_\rho^{\pi_{\theta^*}}}{\mu}\right\|_\infty}+\frac{1}{(1-\gamma)^3}+\frac{W}{1-\gamma}\right).$$

In addition,

$$L^H_{\lambda,\mu}(\theta^*) - L^H_{\lambda,\mu}(\theta_1) \leq \mathcal{O}\left(\frac{1}{1-\gamma} + \frac{\epsilon(1-\gamma)}{\left\|\frac{d_\rho^{\pi_{\theta^*}}}{\mu}\right\|_\infty}\right).$$

Substituting the above results into (15) gives rise to

$$\Gamma_2 + \frac{\Gamma_1+\Gamma_3}{B} = \widetilde{\mathcal{O}}\left(\frac{1+W}{(1-\gamma)^8} + \frac{\epsilon(1+W)(1-\gamma)}{\left\|\frac{d_\rho^{\pi_{\theta^*}}}{\mu}\right\|_\infty}\right).$$

Thus, we obtain

$$\begin{aligned}
&\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\left(J_\rho(\theta^*) - J_\rho(\theta_t)\right)\right] \\
&\qquad \leq \widetilde{\mathcal{O}}\left(\left\|\frac{d_\rho^{\pi_{\theta^*}}}{\mu}\right\|_\infty^2 \frac{|\mathcal{S}|^2|\mathcal{A}|^2(m+T)^{1/3}(1+W)}{\epsilon^2(1-\gamma)^{11}T}\right) + \frac{3\epsilon}{4} \\
&\qquad \leq \left\|\frac{d_\rho^{\pi_{\theta^*}}}{\mu}\right\|_\infty^2 |\mathcal{S}|^2|\mathcal{A}|^2(1+W)\cdot\widetilde{\mathcal{O}}\left(\frac{m^{1/3}}{\epsilon^2(1-\gamma)^{11}T} + \frac{1}{\epsilon^2(1-\gamma)^{11}T^{2/3}}\right) + \frac{3\epsilon}{4}.
\end{aligned}$$

Since $m^{1/3} = \mathcal{O}\left(\frac{\epsilon(1-\gamma)}{\left\|\frac{d_\rho^{\pi_{\theta^*}}}{\mu}\right\|_\infty} + \frac{1}{(1-\gamma)^3} + \frac{W}{1-\gamma}\right)$, one can write

$$\begin{aligned}
&\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\left(J_\rho(\theta^*) - J_\rho(\theta_t)\right)\right] \\
&\qquad \leq \left\|\frac{d_\rho^{\pi_{\theta^*}}}{\mu}\right\|_\infty^2 |\mathcal{S}|^2|\mathcal{A}|^2(1+W)\cdot\widetilde{\mathcal{O}}\left(\frac{1}{\epsilon^2(1-\gamma)^{14}T} + \frac{W}{\epsilon^2(1-\gamma)^{12}T} + \frac{1}{\epsilon^2(1-\gamma)^{11}T^{2/3}}\right) + \frac{3\epsilon}{4}.
\end{aligned}$$

Finally, to guarantee $\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\left(J_\rho(\theta^*) - J_\rho(\theta_t)\right)\right] \leq \epsilon$, it suffices to have

$$T = \widetilde{\mathcal{O}}\left(\frac{\left\|\frac{d_\rho^{\pi_{\theta^*}}}{\mu}\right\|_\infty^3 |\mathcal{S}|^3|\mathcal{A}|^3(1+W)^{\frac{3}{2}}}{\epsilon^{\frac{9}{2}}(1-\gamma)^{\frac{33}{2}}} + \frac{\left\|\frac{d_\rho^{\pi_{\theta^*}}}{\mu}\right\|_\infty^2 |\mathcal{S}|^2|\mathcal{A}|^2(1+W)}{\epsilon^3}\cdot\frac{W}{(1-\gamma)^{12}}\right).$$

This completes the proof. □

10 Proofs of results in Section 4.2

10.1 Proof of Lemma 4.7

Proof. By the performance difference lemma Kakade and Langford (2002), we know that

$$\mathbb{E}_{(s,a)\sim v_\rho^{\pi_{\theta^*}}}\left[A^{\pi_{\theta_t}}(s,a)\right] = (1-\gamma)\left(J_\rho(\theta^*) - J_\rho(\theta_t)\right). \tag{16}$$

In addition, by Assumption 4.6, we know that the advantage function is also related to the defined *transferred compatible function approximation error* that measures the richness of the policy parameterization:

$$\begin{aligned}
\epsilon_{\text{bias}} \geq& \mathbb{E}_{(s,a)\sim v_\rho^{\pi_{\theta^*}}}\left[\left(A^{\pi_{\theta_t}}(s,a) - (1-\gamma)u_t^{*\top}\nabla\log\pi_{\theta_t}(a|s)\right)^2\right] \\
\geq& \left(\mathbb{E}_{(s,a)\sim v_\rho^{\pi_{\theta^*}}}\left[A^{\pi_{\theta_t}}(s,a) - (1-\gamma)u_t^{*\top}\nabla\log\pi_{\theta_t}(a|s)\right]\right)^2
\end{aligned} \tag{17}$$

where the second inequality uses the Jensen's inequality. Then, by combining (16) and (17), we have

$$\begin{aligned}\sqrt{\epsilon_{\text{bias}}} \geq & \mathbb{E}_{(s,a)\sim v_\rho^{\pi_{\theta^*}}}\left[A^{\pi_{\theta_t}}(s,a) - (1-\gamma)u_t^{*\top}\nabla \log \pi_{\theta_t}(a|s)\right] \\ = & (1-\gamma)\left(J_\rho(\theta^\star) - J_\rho(\theta_t)\right) - \mathbb{E}_{(s,a)\sim v_\rho^{\pi_{\theta^*}}}\left[(1-\gamma)u_t^{*\top}\nabla \log \pi_{\theta_t}(a|s)\right].\end{aligned}$$

The rearrangement of the above inequality gives

$$\begin{aligned}\left(J_\rho(\theta^\star) - J_\rho(\theta_t)\right) \leq & \frac{1}{(1-\gamma)}\sqrt{\epsilon_{\text{bias}}} + \mathbb{E}_{(s,a)\sim v_\rho^{\pi_{\theta^*}}}\left[u_t^{*\top}\nabla \log \pi_{\theta_t}(a|s)\right] \\ \leq & \frac{1}{(1-\gamma)}\sqrt{\epsilon_{\text{bias}}} + \mathbb{E}_{(s,a)\sim v_\rho^{\pi_{\theta^*}}}\left[\|u_t^*\|\,\|\nabla \log \pi_{\theta_t}(a|s)\|\right] \\ \leq & \frac{1}{(1-\gamma)}\sqrt{\epsilon_{\text{bias}}} + M_g\|u_t^*\|.\end{aligned}$$

In addition, by the definition of u_t^*, we have

$$\begin{aligned}J_\rho(\theta^\star) - J_\rho(\theta_t) \leq & \frac{1}{(1-\gamma)}\sqrt{\epsilon_{\text{bias}}} + M_g\left\|F^{-1}(\theta_t)\nabla J_\rho(\theta_t)\right\| \\ \leq & \frac{1}{(1-\gamma)}\sqrt{\epsilon_{\text{bias}}} + \frac{M_g}{\mu_F}\|\nabla J_\rho(\theta_t)\|,\end{aligned}$$

where the second inequality follows from Assumption 2.1. This completes the proof.

□

10.2 Proof of Lemma 4.8

Lemma 10.1 *Under Assumption 3.1, suppose that the stochastic policy gradient u_t^H is generated by Algorithm 2 with the restricted parameterization. Let $e_t^H = \nabla J_\rho^H(\theta_t) - u_t^H$. Then*

$$\mathbb{E}\left[\eta_{t-1}^{-1}\|e_t^H\|_2^2\right] \leq \mathbb{E}\left[\eta_{t-1}^{-1}(1-\beta_t)^2\|e_{t-1}^H\|_2^2 + \frac{2\eta_{t-1}^{-1}\beta_t^2\sigma^2}{B} + \frac{4b^2\eta_{t-1}^{-1}(1-\beta_t)^2}{B}\|\theta_t - \theta_{t-1}\|_2^2\right],$$

where $b^2 = L_g^2 + G^2C_w^2$, $L_g = M_h/(1-\gamma)^2$, $G = M_g/(1-\gamma)^2$ and $C_w = \sqrt{H(2HM_g^2 + M_h)(W+1)}$.

Proof. This proof is similar to the proof of Lemma 9.2 with M_g and M_h defined in Assumption 3.1. The details are omitted for brevity. □

10.2.1 Proof of Lemma 4.8

Proof. Let $e_t^H = \nabla J_\rho^H(\theta_t) - u_t^H$. The function $J_\rho^H(\theta)$ is L-smooth due to Lemma 3.2. Moreover, because of $m \geq (2Lk)^3$, it holds that $\eta_t \leq \eta_0 = \frac{k}{m^{1/3}} \leq \frac{1}{2L}$. Since $\eta_t \leq \frac{1}{2L}$, we have $\beta_{t+1} = c\eta_t^2 \leq \frac{c\eta_t}{2L} \leq \frac{ck}{2Lm^{1/3}} \leq 1$. It follows from Lemma 10.1 that

$$\begin{aligned}&\mathbb{E}\left[\eta_t^{-1}\|e_{t+1}^H\|_2^2 - \eta_{t-1}^{-1}\|e_t^H\|_2^2\right] \\ &\leq \mathbb{E}\left[(\eta_t^{-1}(1-\beta_{t+1})^2 - \eta_{t-1}^{-1})\|e_t^H\|_2^2 + \frac{2\eta_t^{-1}\beta_{t+1}^2\sigma^2}{B} + \frac{4b^2\eta_t^{-1}(1-\beta_{t+1})^2}{B}\|\theta_{t+1} - \theta_t\|_2^2\right] \\ &\leq \mathbb{E}\left[(\eta_t^{-1}(1-\beta_{t+1}) - \eta_{t-1}^{-1})\|e_t^H\|_2^2 + \frac{2\eta_t^{-1}\beta_{t+1}^2\sigma^2}{B} + \frac{4b^2\eta_t^{-1}}{B}\|\theta_{t+1} - \theta_t\|_2^2\right],\end{aligned}$$

where the last inequality holds by $0 < \beta_{t+1} \leq 1$. Since the function $x^{1/3}$ is concave, we have $(x+y)^{1/3} \leq x^{1/3} + yx^{-2/3}/3$. As a result

$$\begin{aligned}\eta_t^{-1} - \eta_{t-1}^{-1} = & \frac{1}{k}\left((m+t)^{1/3} - (m+t-1)^{1/3}\right) \leq \frac{1}{3k(m+t-1)^{2/3}} \\ \leq & \frac{1}{3k(m/2+t)^{2/3}} \leq \frac{2^{2/3}}{3k^3}\eta_t^2 \leq \frac{2^{2/3}}{6k^3L}\eta_t \leq \frac{1}{3k^3L}\eta_t,\end{aligned}$$

where the second inequality is due to $m \geq 2$, and the fifth inequality holds by $0 < \eta \leq \frac{1}{2L}$. Then, it holds that

$$\left(\eta_t^{-1}(1-\beta_{t+1}) - \eta_{t-1}^{-1}\right)\|e_t^H\|_2^2 = \left(\frac{1}{3k^3L} - c\right)\eta_t \left\|e_t^H\right\|_2^2 = -96b^2\eta_t \left\|e_t^H\right\|_2^2,$$

where the last equality holds by $c = \frac{1}{3k^3L} + 96b^2$. Combining the above results leads to

$$\mathbb{E}\left[\eta_t^{-1}\|e_{t+1}^H\|_2^2 - \eta_{t-1}^{-1}\|e_t^H\|_2^2\right] \leq \mathbb{E}\left[-96b^2\eta_t\|e_t^H\|_2^2 + \frac{2\eta_t^{-1}\beta_{t+1}^2\sigma^2}{B} + \frac{4b^2\eta_t^{-1}}{B}\|\theta_{t+1}-\theta_t\|_2^2\right]. \tag{18}$$

By summing up the above inequality and dividing both sides by $96b^2$, we obtain

$$\begin{aligned}\frac{1}{96b^2}\left(\mathbb{E}\left[\frac{\left\|e_{T+1}^H\right\|_2^2}{\eta_T} - \frac{\left\|e_1^H\right\|_2^2}{\eta_0}\right]\right) &\leq \sum_{t=1}^{T}\mathbb{E}\left[\frac{c^2\eta_t^3\sigma^2}{48b^2B} - \eta_t\left\|e_t^H\right\|_2^2 + \frac{1}{24\eta_t B}\left\|\theta_{t+1}-\theta_t\right\|_2^2\right] \\ &\leq \sum_{t=1}^{T}\mathbb{E}\left[\frac{c^2\eta_t^3\sigma^2}{48b^2B} - \eta_t\left\|e_t^H\right\|_2^2 + \frac{\eta_t}{24B}\left\|u_t^H\right\|_2^2\right].\end{aligned}$$

For $\eta_t \left\|u_t^H\right\|_2^2$, it follows from Lemma 8.3 that

$$\frac{\eta_t}{4}\left\|u_{t,\lambda}^H\right\|_2^2 \leq J_\rho^H(\theta_{t+1}) - J_\rho^H(\theta_t) + \frac{\eta_t}{2}\left\|e_t^H\right\|^2.$$

Then, it holds that

Then, Lemma 8.3 can be used to obtain

$$\begin{aligned}&\frac{1}{96b^2}\left(\mathbb{E}\left[\frac{\left\|e_{T+1}^H\right\|_2^2}{\eta_T} - \frac{\left\|e_1^H\right\|_2^2}{\eta_0}\right]\right) \\ &\leq \sum_{\tau=1}^{T}\mathbb{E}\left[\frac{1}{48b^2B}c^2\eta_t^3\sigma^2 - \frac{(12B-1)\eta_t}{12B}\left\|e_t^H\right\|_2^2 + \frac{1}{6B}\left(J_\rho^H(\theta_{t+1}) - J_\rho^H(\theta_t)\right)\right] \\ &\leq \sum_{t=1}^{T}\left(\frac{1}{48b^2B}c^2\eta_t^3\sigma^2 - \mathbb{E}\left[\frac{11\eta_t}{12}\left\|e_t^H\right\|_2^2\right]\right) + \frac{1}{6B}\left(J_\rho^H(\theta_{T+1}) - J_\rho^H(\theta_1)\right) \\ &\leq \frac{c^2\sigma^2k^3}{48b^2B}\sum_{t=1}^{T}\frac{1}{m+t} - \sum_{t=1}^{T}\mathbb{E}\left[\frac{11\eta_t}{12}\left\|e_t^H\right\|_2^2\right] + \frac{1}{6B}\left(J_\rho^H(\theta^*) - J_\rho^H(\theta_1)\right) \\ &\leq \frac{c^2\sigma^2k^3}{48b^2B}\sum_{t=1}^{T}\frac{1}{2+t} - \sum_{t=1}^{T}\mathbb{E}\left[\frac{11\eta_t}{12}\left\|e_t^H\right\|_2^2\right] + \frac{1}{6B}\left(J_\rho^H(\theta^*) - J_\rho^H(\theta_1)\right) \\ &\leq \frac{c^2\sigma^2k^3\ln(T+2)}{48b^2B} - \sum_{t=1}^{T}\mathbb{E}\left[\frac{11\eta_t}{12}\left\|e_t^H\right\|_2^2\right] + \frac{1}{6B}\left(J_\rho^H(\theta^*) - J_\rho^H(\theta_1)\right).\end{aligned} \tag{19}$$

Rearranging the above inequality gives rise to

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}\left[\frac{11\eta_t}{12}\left\|e_t^H\right\|_2^2\right] &\leq \frac{c^2\sigma^2k^3\ln(T+2)}{48b^2B} + \frac{1}{96b^2\eta_0}\mathbb{E}\left[\left\|e_1^H\right\|_2^2\right] + \frac{1}{6B}\left(J_\rho^H(\theta^*) - J_\rho^H(\theta_1)\right) \\ &\leq \frac{1}{B}\left(\frac{c^2\sigma^2k^3\ln(T+2)}{48b^2} + \frac{m^{1/3}\sigma^2}{96b^2k} + \frac{1}{6}\left(J_\rho^H(\theta^*) - J_\rho^H(\theta_1)\right)\right).\end{aligned}$$

Multiplying both sides by $\frac{12}{11}$ yields that

$$\sum_{t=1}^{T}\mathbb{E}\left[\eta_t\left\|e_t^H\right\|_2^2\right] \leq \frac{1}{B}\left(\frac{c^2\sigma^2k^3\ln(T+2)}{44b^2} + \frac{m^{1/3}\sigma^2}{88b^2k} + \frac{1}{22}\left(J_\rho^H(\theta^*) - J_\rho^H(\theta_1)\right)\right) \tag{20}$$

$$\leq \frac{1}{B}\left(\frac{c^2\sigma^2k^3\ln(T+2)}{44b^2} + \frac{m^{1/3}\sigma^2}{88b^2k} + \frac{1}{22(1-\gamma)}\right), \tag{21}$$

where the last inequality holds due to $J_\rho^H(\theta^*) - J_\rho^H(\theta_1) \le \frac{1}{1-\gamma}$.

Next, we define a Lyapunov function $\Phi_t(\theta_t) = J_\rho^H(\theta_t) - \frac{1}{192b^2\eta_{t-1}} \left\|e_t^H\right\|^2$ for all $t \ge 1$. One can write

$$\begin{aligned}
\mathbb{E}[\Phi_{t+1} - \Phi_t] =& \mathbb{E}\left[J_\rho^H(\theta_{t+1}) - J_\rho^H(\theta_t) - \frac{1}{192b^2\eta_t}\left\|e_{t+1}^H\right\|_2^2 + \frac{1}{192b^2\eta_{t-1}}\left\|e_t^H\right\|_2^2\right] \\
\ge& \mathbb{E}\left[-\frac{\eta_t}{2}\left\|e_t^H\right\|_2^2 + \frac{\eta_t}{4}\left\|u_t^H\right\|_2 - \frac{1}{192b^2\eta_t}\left\|e_{t+1}^H\right\|_2^2 + \frac{1}{192b^2\eta_{t-1}}\left\|e_t^H\right\|_2^2\right] \\
\ge& \mathbb{E}\left[\frac{\eta_t}{4}\left\|u_t^H\right\|_2 - \frac{\beta_t^2\sigma^2}{96b^2B\eta_t} - \frac{\eta_t}{48B}\|u_t^H\|_2^2\right] \\
\ge& \mathbb{E}\left[\frac{11\eta_t}{48}\left\|u_t^H\right\|_2 - \frac{c^2\eta_t^3\sigma^2}{96b^2B}\right]
\end{aligned}$$

where the first inequality holds by Lemma 8.3 and the second inequality follows from (18). Summing the above inequality over t from 1 to T yields that

$$\begin{aligned}
\sum_{t=1}^T \mathbb{E}\left[\eta_t\left\|u_t^H\right\|_2^2\right] \le& \mathbb{E}\left[\frac{48}{11}(\Phi_{T+1} - \Phi_1) + \sum_{t=1}^T \frac{c^2\eta_t^3\sigma^2}{22b^2B}\right] \\
\le& \mathbb{E}\left[\frac{48}{11}(J_\rho^H(\theta_{T+1}) - J_\rho^H(\theta_1)) + \frac{1}{44b^2\eta_0}\mathbb{E}\left[\|e_1^H\|_2^2\right] + \frac{c^2\sigma^2k^3}{22b^2B}\sum_{t=1}^T \frac{1}{m+t}\right] \\
\le& \mathbb{E}\left[\frac{48}{11}(J_\rho^H(\theta_{T+1}) - J_\rho^H(\theta_1)) + \frac{1}{44b^2\eta_0}\mathbb{E}\left[\|e_1^H\|_2^2\right] + \frac{c^2\sigma^2k^3}{22b^2B}\sum_{t=1}^T \frac{1}{2+t}\right] \\
\le& \frac{48}{11}\left(J_\rho^H(\theta^*) - J_\rho^H(\theta_1)\right) + \frac{\sigma^2m^{1/3}}{44b^2kB} + \frac{c^2\sigma^2k^3}{22b^2B}\ln(2+T) \\
\le& \frac{1}{1-\gamma}\frac{48}{11} + \frac{\sigma^2m^{1/3}}{44b^2kB} + \frac{c^2\sigma^2k^3}{22b^2B}\ln(2+T) \\
=& \Gamma_2 + \frac{\Gamma_3}{B}.
\end{aligned}$$

where the last inequality is due to $J_\rho^H(\theta^*) - J_\rho^H(\theta_1) \le \frac{1}{1-\gamma}$. This completes the proof.

It results from (20) that

$$\sum_{t=1}^T \mathbb{E}\left[\eta_t\left\|\nabla J_\rho^H(\theta_t)\right\|_2^2\right] \le \sum_{t=1}^T \mathbb{E}\left[\eta_t\left\|u_t^H\right\|_2^2\right] + \sum_{t=1}^T \mathbb{E}\left[\eta_t\left\|e_t^H\right\|_2^2\right] \le \Gamma_2 + \frac{\Gamma_1+\Gamma_3}{B}.$$

Since $\eta_t = \frac{k}{(m+t)^{1/3}}$ is decreasing, we have

$$\begin{aligned}
\sum_{t=1}^T \mathbb{E}\left[\left\|\nabla J_\rho^H(\theta_t)\right\|_2^2\right] \le& 1/\eta_T \sum_{t=1}^T \mathbb{E}\left[\eta_t\left\|\nabla J_\rho^H(\theta_t)\right\|_2^2\right] \\
=& \left(\frac{\Gamma_2}{k} + \frac{\Gamma_1+\Gamma_3}{kB}\right)(m+T)^{1/3}.
\end{aligned}$$

Finally, one can use Jensen's inequality to conclude that

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\left\|\nabla J_\rho^H(\theta_t)\right\|_2\right] \le& \left(\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left\|\nabla J_\rho^H(\theta_t)\right\|_2^2\right)^{1/2} \\
\le& \sqrt{\frac{\Gamma_2}{k} + \frac{\Gamma_1+\Gamma_3}{kB}}\left(\frac{m^{1/6}}{\sqrt{T}} + \frac{1}{T^{1/3}}\right),
\end{aligned}$$

where the last inequality follows from the inequality $(a+b)^{1/2} \le a^{1/2} + b^{1/2}$ for all $a, b > 0$.

□

10.3 Proof of Theorem 4.9

Proof. By Lemma 4.7 and the triangle inequality, we have

$$J_\rho(\theta^*) - \frac{1}{T}\sum_{t=1}^{T} J_\rho(\theta_t) \le \frac{\sqrt{\varepsilon_{\text{bias}}}}{1-\gamma} + \frac{M_g}{\mu_F}\frac{1}{T}\sum_{t=1}^{T}\left\|\nabla J_\rho^H(\theta_t)\right\| + \frac{M_g}{\mu_F}\max_{t=1,\dots,T}\left\{\left\|\nabla J_\rho^H(\theta_t) - \nabla J_\rho(\theta_t)\right\|\right\}$$
$$\le \frac{\sqrt{\varepsilon_{\text{bias}}}}{1-\gamma} + \frac{M_g}{\mu_F}\frac{1}{T}\sum_{t=1}^{T}\left\|\nabla J_\rho^H(\theta_t)\right\| + \frac{M_g^2}{\mu_F}\left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\right)\gamma^H, \tag{22}$$

where the last inequality follows from Proposition 3.2. Then, due to Lemma 4.8, we know that

$$J_\rho(\theta^*) - \frac{1}{T}\sum_{t=1}^{T} J_\rho(\theta_t) \le \frac{\sqrt{\varepsilon_{\text{bias}}}}{1-\gamma} + \frac{M_g}{\mu_F}\sqrt{\frac{\Gamma_2}{k} + \frac{\Gamma_1+\Gamma_3}{kB}}\left(\frac{m^{1/6}}{\sqrt{T}} + \frac{1}{T^{1/3}}\right) + \frac{M_g^2}{\mu_F}\left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\right)\gamma^H,$$

Notice that, in order to guarantee

$$\frac{M_g^2}{\mu_F}\left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\right)\gamma^H \le \frac{\epsilon}{2},$$

it suffices to have

$$H = \mathcal{O}\left(\log_\gamma\left(\mu_F(1-\gamma)\epsilon\right)\right).$$

Recall that $\Gamma_1 = \left(\frac{c^2\sigma^2k^3\ln(T+2)}{48b^2} + \frac{m^{1/3}}{96b^2k}\sigma^2 + \frac{1}{22(1-\gamma)}\right)$, $\Gamma_2 = \frac{48}{11(1-\gamma)}$, and $\Gamma_3 = \frac{\sigma^2m^{1/3}}{44b^2k^2} + \frac{c^2\sigma^2k^3}{22kb^2}\ln(2+T)$. By only considering the dependencies on the parameters c, σ, b, γ and m, we have

$$\frac{\Gamma_2}{k} + \frac{\Gamma_1+\Gamma_3}{kB} = \widetilde{\mathcal{O}}\left(\frac{c^2\sigma^2}{b^2} + \frac{m^{1/3}\sigma^2}{b^2} + \frac{\sigma^2m^{1/3}}{b^2} + \frac{c^2\sigma^2}{b^2} + \frac{1}{1-\gamma}\right)$$
$$= \widetilde{\mathcal{O}}\left(\frac{\sigma^2}{b^2}\left(c^2 + m^{1/3}\right) + \frac{1}{1-\gamma}\right). \tag{23}$$

In addition, by the definitions $b^2 = L_g^2 + G^2C_w^2$, $L_g = M_h/(1-\gamma)^2$, $G = M_g/(1-\gamma)^2$, $C_w = \sqrt{H(2HM_g^2 + M_h)(W+1)}$, $c = \frac{1}{3k^3L} + 96b^2$, and $m = \max\{2, (2Lk)^3, (\frac{ck}{2L})^3\}$, we know that

$$b^2 = \mathcal{O}\left(\frac{1+W}{(1-\gamma)^4}\right),\ \sigma^2 = \mathcal{O}\left(\frac{1}{(1-\gamma)^4}\right),\ c = \mathcal{O}\left(\frac{1+W}{(1-\gamma)^4}\right),\ m^{1/3} = \mathcal{O}\left(\frac{1}{(1-\gamma)^3} + \frac{W}{1-\gamma}\right).$$

Substituting the above results into (23) gives rise to

$$\frac{\Gamma_2}{k} + \frac{\Gamma_1+\Gamma_3}{kB} = \widetilde{\mathcal{O}}\left(\frac{1+W}{(1-\gamma)^8}\right).$$

Thus, we obtain

$$J_\rho(\theta^*) - \frac{1}{T}\sum_{t=1}^{T} J_\rho(\theta_t) \le \frac{\sqrt{\varepsilon_{\text{bias}}}}{1-\gamma} + \frac{M_g}{\mu_F}\left(\frac{m^{1/6}}{\sqrt{T}} + \frac{1}{T^{1/3}}\right)\cdot\widetilde{\mathcal{O}}\left(\frac{\sqrt{1+W}}{(1-\gamma)^4}\right) + \frac{\epsilon}{2}$$
$$\le \frac{\sqrt{\varepsilon_{\text{bias}}}}{1-\gamma} + \widetilde{\mathcal{O}}\left(\left(\frac{1}{\sqrt{(1-\gamma)^3T}} + \frac{\sqrt{W}}{\sqrt{(1-\gamma)T}} + \frac{1}{T^{1/3}}\right)\frac{\sqrt{1+W}}{(1-\gamma)^4\mu_F}\right) + \frac{\epsilon}{2}$$

In order to guarantee

$$\widetilde{\mathcal{O}}\left(\left(\frac{1}{\sqrt{(1-\gamma)^3T}} + \frac{\sqrt{W}}{\sqrt{(1-\gamma)T}} + \frac{1}{T^{1/3}}\right)\frac{\sqrt{1+W}}{(1-\gamma)^4\mu_F}\right) \le \frac{\epsilon}{2},$$

it suffices to take

$$T = \tilde{\mathcal{O}}\left(\frac{(1+W)^{\frac{3}{2}}}{\epsilon^3\mu_F^3(1-\gamma)^{12}} + \frac{1+W}{\epsilon^2\mu_F^2(1-\gamma)^{11}} + \frac{W(1+W)}{\epsilon^2\mu_F^2(1-\gamma)^9}\right).$$

This completes the proof. □