Global Convergence of Stochastic Policy Gradient Methods with Momentum and Entropy

Yuhao Ding
Department of IEOR
UC, Berkeley
Berkeley, CA 94704
yuhao_ding@berkeley.edu

Junzi Zhang*
Amazon Advertising
Amazon.com, Inc.
Berkeley, CA, 94706
junziz@amazon.com

Javad Lavaei
Department of IEOR
UC, Berkeley
Berkeley, CA 94704
lavaei@berkeley.edu

Abstract

Policy gradient methods are popular and efficient for large-scale reinforcement learning due to their relative stability and incremental nature, while little is known on their global convergence and sample efficiency compared to their value-based counterparts. In recent years, the empirical success of policy gradient methods has led to the development of a theoretical foundation for these methods. In this work, we generalize this line of research by incorporating momentum terms and entropy regularization, which have been demonstrated to be efficient recipes for improving policy gradient methods. For momentum-based policy gradient 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 policy gradient methods by $\tilde{\mathcal{O}}(\epsilon^{-1.5})$ and $\tilde{\mathcal{O}}(\epsilon^{-1})$, respectively, where $\epsilon > 0$ is the target tolerance. We then extend our analysis to Maximum Entropy (MaxEnt) reinforcement learning, and show that both vanilla policy gradient methods and momentum-based policy gradient methods achieve polynomial sample complexities in ϵ, with the momentum methods achieving a better dependency on the logarithmic terms. As a by product, this work also provides the first global convergence results for stochastic entropy regularized policy gradient methods. Our analysis also formalizes a new trajectory-based entropy gradient estimator to cope with momentum terms, which may be of independent and practical interests.

1 Introduction

In this paper, we study the global convergence of stochastic policy gradient (PG) methods with momentum terms in both the classical and Maximum Entropy (MaxEnt) reinforcement learning (RL) settings. In both cases, we consider the scenario where the agent interacts with an infinite-horizon discounted Markov decision process (MDP) environment in an episodic manner, with a finite-horizon termination in each episode [19, 32, 53, 2, 47, 41, 62, 27].

Policy gradient methods can be dated back to the pioneering work [48] in 1980s, and have evolved into a rich family of RL algorithms [23, 20, 44, 39, 25, 40]. 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 [64, 43, 58, 21, 50].

On the theoretical side, the understanding of these algorithms is mostly restricted to convergence to a stationary point of the value function [46, 24, 32]. To address this issue, a recent line of research

Preprint. Under review.

*Work done prior to joining or outside of Amazon.

has focused on studying the global convergence properties of these algorithms, which started from a linear-quadratic regulator setting [13] and was later extended to generic MDPs (typically with finite state and action spaces). These works have investigated different aspects of the global convergence of policy gradient methods, including but not limited to second-order stationarity [63], global optimality with exact gradients [5, 30, 8, 6], function approximation [2, 47], exploration [1], batch sizes [62] and variance reduction [27], as well as extensions to actor-critic variants [16, 55, 51, 56, 14] and general utility functions [61, 60].

In this work, we extend this line of research to cope with *momentum terms* and *entropy regularization* with *stochastic trajectory based* policy gradient estimators.

Momentum techniques have been demonstrated as a powerful and generic recipe for accelerating stochastic gradient methods, especially for nonconvex optimization and deep learning [35, 22, 37]. Recent works have also extended momentum techniques to improve policy gradient methods both in theory and in practice [52, 59, 34, 17], yet the analyses in those works have been restricted to the classical stationary point convergence. Despite the recent advances in the global convergence theory of policy gradient methods, it remains unclear how to establish the global convergence and sample complexity of momentum-based stochastic policy gradient methods, especially when adaptive step-sizes are involved as in [17].

MaxEnt is a recent framework for reinforcement learning, which introduces a non-greedy entropy regularization term to improve the learned policies in terms of deep exploration, pretraining performance and robustness to disturbances, among others [15, 38, 10, 11]. Recently, the convergence analysis of policy gradient methods for MaxEnt RL has attracted considerable attention [30, 8]. These works have demonstrated the power of entropy regularization for both vanilla and natural policy gradient methods, but the analyses in these works rely heavily on their assumption of access to exact and deterministic policy gradients. The difficulty of extending the above results to stochastic gradients is further exacerbated when one tries to incorporate momentum terms, since the entropy regularization makes it unclear how the policy gradient estimator should be defined so that the importance sampling weights can be applied.

The notions used in this paper and a more detailed literature review can be found in Appendix A and Appendix G, respectively.

Contribution. The major contributions of this paper are summarized below:

- We obtain the first global optimality sample complexity bounds for *momentum-based* policy gradient methods. Our analysis handles both the soft-max and generic Fisher-non-degenerate policy parametrizations, and shows that adding momentum terms improves the existing sample complexity bounds in [62, 27] for these two types of parametrizations by $\tilde{\mathcal{O}}(1/\epsilon^{1.5})$ and $\tilde{\mathcal{O}}(1/\epsilon)$ (with $\epsilon > 0$ denoting the target tolerance), respectively, while allowing for constant batch sizes. As a by-product, our analysis also provides a general framework for analyzing the global convergence rates of stochastic policy gradient methods with adaptive step-sizes, which can be easily applied and extended to different policy gradient estimators.
- We obtain the first global convergence results for *stochastic* entropy regularized policy gradient methods, both with and without momentum terms. To deal with the importance sampling challenge for momentum terms, we also provide a new trajectory-based entropy gradient estimator, which may be of independent and practical interests.

2 Preliminaries

2.1 Reinforcement learning

Reinforcement learning is generally modeled as a discounted Markov decision process (MDP) defined by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, $\mathcal{P}(s'|s,a)$ 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 to the parameter $\theta \in \Theta$, where Θ is a convex constraint set in $\mathbb{R}^d$.

Let $\tau = \{s_0, a_0, s_1, \ldots\}$ denote the data of a sampled trajectory under policy π_θ with the probability distribution over trajectory as $p(\tau|\theta, \rho) = \rho(s_0)\prod_{h=1}^{\infty}\mathcal{P}(s_{h+1}|s_h, a_h)\pi_\theta(a_h|s_h)$, 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 any policy π, one can define the state-action value function $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as $Q^\pi(s,a) := \mathbb{E}_{a_h\sim\pi(\cdot|s_h), s_{h+1}\sim\mathbb{P}(\cdot|s_h,a_h)}\left(\sum_{h=0}^{\infty}\gamma^h r(s_h,a_h)\middle| s_0 = s, a_0 = a\right)$. 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 $V^\pi(s) := \mathbb{E}_{a\sim\pi(\cdot|s)}[Q^\pi(s,a)]$ and $A^\pi(s,a) := Q^\pi(s,a) - V^\pi(s)$, respectively. Then, the goal is to find an optimal policy in the policy class that maximizes the expected discounted return, namely,

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

For notional convenience, we will denote $J_\rho(\pi_\theta)$ by the shorthand notation $J_\rho(\theta)$. In practice, a truncated version of the value function is used to approximate the infinite sum of rewards in (1). Let $\tau_i^H = \{s_0^i, a_0^i, s_1^i, \ldots, s_{H-1}^i, a_{H-1}^i, s_H^i\}$ denote the truncation of the full trajectory τ_i of length H and the truncated version of the value function is defined as $J_\rho^H(\theta) := \mathbb{E}_{s_0\sim\rho, a_h\sim\pi_\theta(\cdot|s_h), s_{h+1}\sim\mathbb{P}(\cdot|s_h,a_h)}\left(\sum_{h=0}^{H-1}\gamma^h r(s_h,a_h)\middle| s_0\right)$.

2.2 Exploratory initial distribution

The discounted state visitation distribution $d_{s_0}^\pi$ of a policy π is defined as $d_{s_0}^\pi(s) := (1-\gamma)\sum_{h=0}^{\infty}\gamma^h\mathcal{P}(s_h = s|s_0, \pi)$, where $\mathcal{P}(s_h = s|s_0, \pi)$ is the state visitation probability that s_h is equal to s under the policy π starting from 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 $v_\rho^\pi(s,a) := (1-\gamma)\mathbb{E}_{s_0\sim\rho}\sum_{h=0}^{\infty}\gamma^h\mathcal{P}(s_h = s, a_h = a|s_0, \pi)$, where $\mathcal{P}(s_h = s, a_h = a|s_0, \pi)$ is the state-action visitation probability that $s_h = s$ and $a_h = a$ under π starting from state s_0. Then, the notion of the distribution mismatch coefficient is defined below:

Definition 2.1 *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 [2] that the difficulty of the exploration problem faced by policy gradient algorithms can be captured through this distribution mismatch coefficient.

2.3 Policy parameterization

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

Soft-max parameterization. For an unconstrained parameter $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $\pi_\theta(a|s)$ 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 MDPs with finite state and action spaces. It is complete in the sense that any stochastic policy can be represented by this class.

Fisher-non-degenerate parameterization. We study the 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 essentially states that $F_\rho(\theta)$ is well-behaved as a pre-conditioner in the natural PG update [19]. It is shown in [27] 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 any full-rank exponential family paramertrization with their mean parameterized by $\mu_\theta(s)$ if $\mu_\theta(s)$ is full-row rank for all

$\theta \in \mathbb{R}^d$. This parameterization can be used for general MDPs but may be restrictive in the sense that it may not contain all stochastic policies and, therefore, may not contain the optimal policy.

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

2.4 Trajectory-based policy gradient estimator

The policy gradient method is one of the standard ways to solve the optimization problem (1) [45]. 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 $\theta_{t+1} = \theta_t + \frac{\eta_t}{B}\sum_{i=1}^{B} u_t$, 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 Assumptions 2.2 and 2.3 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 the REINFORCE [48], the PGT [46] and the GPOMDP [4]. Compared with PG estimators based on the state-action visitation measure [2], the trajectory-based PG estimators are often used in practice due to their sample efficiency and amenability to use 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 PGT is given by:

$$g(\tau_i^H|\theta,\mu) = \sum_{h=0}^{H-1}\sum_{j=h}^{H-1} \nabla \log \pi_\theta(a_h^i, s_h^i)\left(\gamma^j r_j(s_j^i, a_j^i)\right). \tag{2}$$

We first make two essential assumptions for PG estimators.

Assumption 2.2 *Gradient and Hessian of the function* $\log \pi_\theta(a|s)$ *are bounded, i.e., for all* $\theta \in \Theta$, *there exist constants* $M_g, M_h > 0$ *such that* $\|\nabla \log \pi_\theta(a|s)\|_2 \le M_g$ *and* $\left\|\nabla^2 \log \pi_\theta(a|s)\right\|_2 \le M_h$.

Assumption 2.3 *The variance of the stochastic PG* $g(\tau^H|\theta,\mu)$ *is bounded for all* $\theta \in \Theta$, *i.e., there exists a constant* $\sigma > 0$ *such that* $Var(g(\tau^H|\theta,\mu)) = \mathbb{E}\left\|g(\tau^H|\theta,\mu) - \nabla J_\mu^H(\theta)\right\|_2^2 \le \sigma^2$ *for all* $\theta \in \Theta$.

For the soft-max parameterization, Assumption 2.2 is satisfied with $M_g = 2$ and $M_h = 1$ (see Lemma B.2 in appendix). Assumptions 2.2 and 2.3 have also been commonly used in the analysis of the PG [32, 54, 53, 42, 27, 17].

3 Global convergence of RL with constant batch size

3.1 Momentum-based policy gradient

Due to the high sample complexity of the vanilla PG, many receent works have turned onto the variance reduction methods for PG, including the momentum-based PG. The momentum-based policy gradient with the batch size of B and the sampled trajectory of length H is defined as

$$u_t^H = \frac{\beta_t}{B}\sum_{i=1}^{B} g(\tau_i^H|\theta_t,\mu) + (1-\beta_t)\left[u_{t-1}^H + \frac{1}{B}\sum_{i=1}^{B}\left(g(\tau_i^H|\theta_t,\mu) - w(\tau_i^H|\theta_{t-1},\theta_t)g(\tau_i^H|\theta_{t-1},\mu)\right)\right] \tag{3}$$

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

$$w(\tau^H|\theta',\theta) = \frac{p(\tau^H|\theta',\mu)}{p(\tau^H|\theta,\mu)} = \prod_{h=0}^{H-1}\frac{\pi_{\theta'}(a_h|s_t)}{\pi_\theta(a_h|s_t)}. \tag{4}$$

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 3.1 *The variance of* $w(\tau^H|\theta_1,\theta_2) = p(\tau^H|\theta_1,\mu)/p(\tau^H|\theta_2,\mu)$ *is bounded, i.e., there exists a constant* $W > 0$*, such that* $Var(w(\tau^H|\theta_1,\theta_2)) \leq W$ *for all* $\theta_1,\theta_2 \in \Theta$ *and* $\tau \sim p(\tau^H|\theta_2,\mu)$.

Assumption 3.1 have been commonly used in the analysis of some variance reduced variants of PG [32, 54, 53, 42, 27, 17]. It is worthwhile to note that the bounded importance sampling weight in Assumption 3.1 may be violated in practice. A commonly used remedy to make the algorithm more effective is to clip the importance sampling weights [17].

3.2 Soft-max parameterization with log barrier penalty

We now study the soft-max policy parameterization, 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 sending some parameters to 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

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

where $\text{KL}(p,q) := \mathbb{E}_{x\sim p}[-\log q(x)/p(x)]$ and $\text{Unif}_\mathcal{X}$ denotes the uniform distribution over a set $\mathcal{X}$. Although the optimization problem defined above is non-convex in general, [2, Theorem 5.2] has shown that the first-order stationary points of the regularized objective are approximately globally optimal solutions of the $J_\rho(\theta)$ when the regularization parameter λ is sufficiently small and the exact PG is available. Motivated by this result, when analyzing the stochastic PG, we carefully count the number of "good" iterates such that the norms of the first-order stationary points of the regularized objective are small. Then, the following result relates the global convergence under the soft-max parameterization to the convergence of the first-order stationary points of the regularized objective.

Lemma 3.2 *Consider a soft-max parameterized policy* π_θ *with the objective function* $L_{\lambda,\rho}(\cdot)$ *and* $\lambda = \frac{\epsilon(1-\gamma)}{2}$*. Let* θ^* *denote a global maximum of* $J_\rho(\theta)$*. Let* $\{\theta_t\}_{t=1}^T$ *be generated by a general update of the form* $\theta_{t+1} = \theta_t + \eta_t u_t^\lambda$ *for all* $t \in \{1,\ldots,T\}$*, where* $\{\eta_t\}_{t=1}^T$ *is a non-increasing step-size sequence and* u_t^λ *is an estimator of* $\nabla L_{\lambda,\mu}(\theta_t)$*. Then, we have*

$$\frac{1}{T}\sum_{t=1}^T\mathbb{E}\left(J_\rho(\theta^*) - J_\rho(\theta_t)\right) \leq \frac{16|\mathcal{S}|^2|\mathcal{A}|^2\sum_{t=1}^T\eta_t\mathbb{E}[\|\nabla L_{\lambda,\mu}(\theta_t)\|_2^2]}{\epsilon^2 T\eta_T(1-\gamma)^3} + \frac{T+1}{T}\left\|\frac{d_\rho^{\pi_{\theta^*}}}{\mu}\right\|_\infty\epsilon.$$

By applying the momentum-based PG with a constant batch size B for the soft-max parameterization with log barrier regularization (see Algorithm 2 in Appendix B), we arrive at the following result:

Lemma 3.3 *Consider a sequence* $\{\theta_t\}_{t=1}^T$ *generated by Algorithm 2. Let* $\bar{\lambda}$ *be an arbitrary constant such that* $\bar{\lambda} \geq \lambda$*. Let* $b^2 = L_g^2 + G^2C_w^2$*,* $C_w = \sqrt{H(2HM_g^2 + M_h)(W+1)}$*,* $L_g = M_h/(1-\gamma)^2$*,* $G = M_g/(1-\gamma)^2$*,* $c = \frac{1}{3k^3L_J^H} + 96b^2$*,* $m = \max\{2,(12L_J^Hk)^3,(\frac{ck}{12L_J^H})^3\}$*,* $k > 0$*, and* $\eta_0 = \frac{k}{m^{1/3}}$*, where* $M_g = 2$*,* $M_h = 1$ *and* $L_J^H = \frac{8}{(1-\gamma)^3} + \frac{2\bar{\lambda}}{|\mathcal{S}|}$*. It holds that*

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

where $\Gamma_1 = \left(\frac{c^2\sigma^2k^3\ln(T+2)}{36b^2} + \frac{m^{1/3}}{72b^2k}\sigma^2 + \frac{4}{9}\left(L_{\lambda,\mu}^H(\theta^*) - L_{\lambda,\mu}^H(\theta_1)\right)\right)$*,* $\Gamma_2 = 12(L_{\lambda,\mu}^H(\theta^*) - L_{\lambda,\mu}^H(\theta_1))$*, and* $\Gamma_3 = \frac{1}{k}(\frac{\sigma^2m^{1/3}}{8b^2k} + \frac{c^2\sigma^2k^3}{4b^2}\ln(2+T))$.

By combining Lemmas 3.2 and 3.3, and using a sufficiently large horizon H, we obtain the following global convergence result for Algorithm 2.

Theorem 3.4 *Under the conditions of Lemmas 3.2 and 3.3, let* $T = \widetilde{\mathcal{O}}\left(\frac{\sigma^3|\mathcal{S}|^3|\mathcal{A}|^3}{\epsilon^{\frac{9}{2}}(1-\gamma)^{\frac{21}{2}}}\right)$*,* $B = \mathcal{O}(1)$ *and* $H = \mathcal{O}\left(\log_\gamma\left(\frac{(1-\gamma)\epsilon}{|\mathcal{S}||\mathcal{A}|}\right)\right)$*. Then, for every* $\epsilon \leq \frac{2\bar{\lambda}}{1-\gamma}$*, it holds that* $J_\rho(\theta^*) - \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^T(J_\rho(\theta_t))\right] \leq \mathcal{O}(\epsilon)$.

Remark 3.5 *Theorem 3.4 improves the result of Theorem 6 in [62] from the sample complexity of* $\widetilde{\mathcal{O}}(\epsilon^{-6})$ *to* $\widetilde{\mathcal{O}}(\epsilon^{-4.5})$ *for the soft-max parameterization.*

3.3 Fisher-non-degenerate parameterization

Our analysis will leverage the notion of *compatible function approximation* in [46] defined as the regression problem:

$$\min_{w\in\mathbb{R}^d} \mathbb{E}_{(s,a)\sim v_\rho^{\pi_\theta}}\left[\left(A^{\pi_\theta}(s,a) - (1-\gamma)w^\top \nabla \log \pi_\theta(a|s)\right)^2\right]. \tag{5}$$

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 $A^{\pi_\theta}(s,a)$. It can be easily seen that $F_\rho(\theta)^{-1}\nabla J_\rho(\theta)$ is a minimizer of (5), 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 [27, 2], we assume that the policy parameterization π_θ achieves an acceptable function approximation, as measured by the *transferred compatible function approximation error*.

Assumption 3.6 *For every $\theta \in \mathbb{R}^d$, there exists a constant $\epsilon_{bias} > 0$ such that the transferred compatible function approximation error satisfies*

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

where $v_\rho^{\pi_{\theta^*}}$ is the state-action distribution induced by an optimal policy π_{θ^*} that maximizes $J_\rho(\theta)$.

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 [47]. Inspired by the global convergence analysis of PG and natural PG in [27, 2], we present a global convergence framework of stochastic PG estimator with time-varying step-sizes.

Lemma 3.7 *Consider a general Fisher-non-degenerate policy π_θ satisfying Assumptions 2.1, 3.6 and 2.2. Let θ^* denote a global maximum of $J_\rho(\theta)$. Let $\{\theta_t\}_{t=1}^T$ be generated by a general update of the form $\theta_{t+1} = \theta_t + \eta_t u_t$, where u_t is any arbitrary update direction. Then, we have*

$$\begin{aligned} J_\rho(\theta^*) - \frac{1}{T}\sum_{t=1}^T J_\rho(\theta_t) \le \frac{\sqrt{\varepsilon_{bias}}}{1-\gamma} + \frac{M_h}{T}\sum_{t=1}^T \eta_t \|u_t\|_2^2 + \frac{M_g}{\mu_F T}\sum_{t=1}^T \|\nabla J_\rho(\theta_t)\|_2 \\ + \frac{4M_g}{T}\left(1+\frac{1}{\mu_F}\right)^2 \sum_{t=1}^T \|\nabla J_\rho(\theta_t)\|_2^2 + \frac{4M_g}{T}\sum_{t=1}^T \|u_t - \nabla J_\rho(\theta_t)\|_2^2. \end{aligned} \tag{6}$$

Lemma 3.7 relates the global convergence rates of the policy gradient to the transferred compatible function approximation error, the stationary convergence rate of the update rule and the error of the estimated gradient. We also make the following mild assumption on the policy to guarantee the smoothness of the objective.

Assumption 3.8 *Assume that $\sum_{a\in\mathcal{A}} \left|\frac{d\pi_\alpha(a|s_0)}{d\alpha}\,|_{\alpha=0}\right| \le S_1$ and $\sum_{a\in\mathcal{A}} \left|\frac{d^2\pi_\alpha(a|s_0)}{(d\alpha)^2}\,|_{\alpha=0}\right| \le S_2$, where $\pi_\alpha := \pi_{\theta+\alpha u}$ and $u \in \mathbb{R}^d$ is any arbitrary vector.*

By applying the momentum-based PG with a constant batch size B under the general Fisher-non-degenerate parameterization (see Algorithm 3 in Appendix C), we arrive at the following result:

Lemma 3.9 *Under the conditions in Lemma 3.7 and Assumption 3.8, assume that the sequences $\{\theta_t\}_{t=1}^T$ and $\{u_t^H\}_{t=1}^T$ are generated by Algorithm 3. Let $e_t^H = \nabla J_\rho^H(\theta_t) - u_t^H$, $b^2 = L_g^2 + G^2 C_w^2$, $C_w = \sqrt{H(2HM_g^2 + M_h)(W+1)}$, $L_g = M_h/(1-\gamma)^2$, $G = M_g/(1-\gamma)^2$, $k > 0$, $c = \frac{1}{3k^3 L_J^H} + 96b^2$, $m = \max\{2, (12L_J^H k)^3, (\frac{ck}{12L_J^H})^3\}$ and $\eta_0 = \frac{k}{m^{1/3}}$ where $L_J^H = \frac{S_2}{(1-\gamma)^2} + \frac{2\gamma S_1^2}{(1-\gamma)^3}$. We have*

$$\sum_{t=1}^T \left[\|e_t^H\|_2^2\right] \le \frac{(m^{1/3}+T^{1/3})\Gamma_1}{kB}, \quad \frac{1}{T}\sum_{t=1}^T \mathbb{E}[\eta_t \|u_t^H\|_2^2] \le \frac{1}{T}\left(\Gamma_2 + \frac{\Gamma_3}{B}\right),$$
$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}[\|\nabla J_\rho^H(\theta_t)\|_2^2] \le \left(\frac{\Gamma_2}{k} + \frac{\Gamma_1+\Gamma_3}{kB}\right)\left(\frac{m^{1/3}}{T} + \frac{1}{T^{2/3}}\right),$$

where $\Gamma_1 = \left(\frac{c^2\sigma^2 k^3 \ln(T+2)}{36b^2} + \frac{m^{1/3}}{72b^2 k}\sigma^2 + \frac{4}{9(1-\gamma)}\right)$, $\Gamma_2 = \frac{12}{1-\gamma}$, and $\Gamma_3 = \frac{\sigma^2 m^{1/3}}{8b^2k^2} + \frac{c^2\sigma^2 k^3}{4kb^2}\ln(2+T)$.

Then, by combining Lemmas 3.7 and 3.9, we obtain the global convergence rate of Algorithm 3.

Theorem 3.10 *Under the conditions of Lemma 3.9, let* $H = \mathcal{O}(\log_\gamma((1-\gamma)\epsilon))$, $B = \mathcal{O}(1)$ *and* $T = \widetilde{\mathcal{O}}\left(\frac{\sigma}{(1-\gamma)\epsilon^3} + \frac{\sigma}{(1-\gamma)^{2.5}\epsilon^2} + \frac{\sigma^2}{(1-\gamma)^2\epsilon^{3/2}} + \frac{\sigma^2}{(1-\gamma)^5\epsilon}\right)$. *Then, it holds that* $J_\rho(\theta^*) - \frac{1}{T}\sum_{t=1}^T \mathbb{E}[J_\rho(\theta_t)] \leq \frac{\sqrt{\varepsilon_{bias}}}{1-\gamma} + \epsilon$.

Remark 3.11 *Theorem 3.10 establishes the global convergence of the momentum-based PG proposed in [17], for which only stationary convergence was previously shown. In addition, it improves the result of Theorem 4.6 in [27] 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 [27] from using a batch size of* $\mathcal{O}(\frac{1}{\epsilon})$ *and a double-loop algorithm to a constant batch size and a single-loop algorithm.*

4 Global convergence of maximum entropy RL

4.1 Momentum-based policy gradient for maximum entropy RL

In maximum entropy RL, or sometimes called entropy-regularized RL, near-deterministic policies are penalized, which is achieved by modifying the value of a policcy π to

$$\tilde{J}_\rho(\theta) = J_\rho(\theta) + \lambda\mathbb{H}(\rho,\theta), \tag{7}$$

where $\mathbb{H}(\rho,\theta) \coloneqq \mathbb{E}_{s_0\sim\rho, a_t\sim\pi(\cdot|s_t), s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)}\left[\sum_{t=0}^\infty -\gamma^t \log\pi_\theta(a_t|s_t)\right]$ is the "discounting entropy" and $\lambda \geq 0$ determines the strength of the penalty. Obviously, the value of any policy can be obtained by adding an entropy penalty to the rewards. The gradient of $\tilde{J}_\mu(\theta)$ can be written as

$$\begin{aligned}\nabla\tilde{J}_\mu(\theta) =& \mathbb{E}_{\tau\sim p(\tau|\theta)}\left[\left(\sum_{t=0}^\infty \nabla_\theta \log\pi_\theta(a_t|s_t)\right)\left(\sum_{t=0}^\infty \gamma^t\left(r(s_t,a_t) - \log\pi_\theta(a_t|s_t)\right)\right)\right] \\ &+ \mathbb{E}_{\tau\sim p(\tau|\theta)}\left[\left(\sum_{t=0}^\infty -\gamma^t \nabla\log\pi_\theta(a_t|s_t)\right)\right].\end{aligned} \tag{8}$$

On the one hand, since the discounted state visitation distribution $d_\mu^\pi(\cdot)$ is unknown, the importance sampling weight (4) cannot be computed explicitly using the PG estimator based on the state-action visitation measure [2], which impedes the use of momentum-based PG for entropy regularized RL. On the other hand, since $\log\pi_\theta(a|s)$ can be unbounded when π_θ approaches a deterministic policy, the trajectory-based PG estimator of (8) could be unbounded, which also limits the application of momentum-based PG. However, if one can guarantee that the iterates of the algorithm belong to some bounded region $\mathcal{G}^0$ with high probability, then a trajectory-based PG estimator can be used to estimate the gradient of $\tilde{J}_\mu(\theta)$ since $\log\pi_\theta(a|s)$ is bounded over $\mathcal{G}^0$. Then, the truncated PGT in (2) can be used to approximate the first term in (8) by replacing the reward with $r(s,a) - \log\pi_\theta(a|s)$:

$$g_1(\tau_i^H|\theta,\mu) = \sum_{h=0}^{H-1}\sum_{j=h}^{H-1} \nabla\log\pi_\theta(a_h^i, s_h^i)\left(\gamma^j\left(r_j(s_j^i,a_j^i) - \log\pi_\theta(a_j^i|s_j^i)\right)\right). \tag{9}$$

In addition, the second term in (8) can be approximated by the following estimator:

$$g_2(\tau_i^H|\theta,\mu) = \sum_{h=0}^{H-1} -\gamma^t\nabla\log\pi_\theta(a_h^i, s_h^i). \tag{10}$$

Then, the truncated PG estimator for (8) can be written as:

$$g(\tau_i^H|\theta,\mu) = g_1(\tau_i^H|\theta,\mu) + g_2(\tau_i^H|\theta,\mu). \tag{11}$$

We also assume that the variance of the estimators g_1 and g_2 are bounnded over $\mathcal{G}^0$.

Assumption 4.1 *The variances of* $g_1(\tau^H|\theta,\mu)$, $g_2(\tau^H|\theta,\mu)$ *are bounded for all* $\theta \in \mathcal{G}^0$, *i.e., there exist constants* σ_R, $\sigma_\pi > 0$, *such that* $Var(g_1(\tau^H|\theta,\mu)) \leq \sigma_R^2$, $Var(g_2(\tau^H|\theta,\mu)) \leq \sigma_\pi^2$ *for all* $\theta \in \mathcal{G}^0$.

Since the importance sampling weight in (4) is well-defined for the trajectory-based policy gradient estimator, momentum-based PG in (3) can now be applied to maximum entropy RL. Then, it remains to show that θ belongs to some bounded region $\mathcal{G}^0$ with high probability.

Algorithm 1 Momentum-based PG for maximum entropy RL (MBPG-MaxEnt)

Inputs: Iteration T, horizon H, batch size B, initial input θ_1, parameters $\{\eta, \beta\}$ and initial distribution μ;
Outputs: θ_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 $u_t^H = \frac{1}{B}\sum_{i=1}^B g(\tau_i^H|\theta_t, \mu)$ where $g(\tau_i^H|\theta_t, \mu)$ is given in (11);
  else
    Compute $u_t^H = \frac{\beta}{B}\sum_{i=1}^B g(\tau_i^H|\theta_t, \mu) + (1-\beta)[u_{t-1} + \frac{1}{B}\sum_{i=1}^B (g(\tau_i^H|\theta_t, \mu) - w(\tau_i^H|\theta_{t-1}, \theta_t) g(\tau_i^H|\theta_{t-1}, \mu))]$ where $g(\tau_i^H|\theta_t, \mu)$ is given in (11);
  end if
  Update $\theta_{t+1} = \theta_t + \eta u_t^H$;
end for

4.2 Global convergence analysis

We focus attention on studying tabular discounted MDPs with the soft-max parameterization, which is an important first step and a stepping stone towards demystifying the effectiveness of entropy-regularized stochastic policy optimization in more complex settings. We first notice that the soft-max parameterization is overparameterized. It can be easily verified that $\pi_\theta = \pi_{\theta'}$ if $\theta'_{s,a} = \theta_{s,a} + c$ for all state-action pair (s, a), where c is an arbitrary constant. Thus, the optimal policy is not unique and its parameters could increase to infinity. To overcome this challenge, we impose the condition $\theta_{s,a_1} = 0$ for all $s \in \mathcal{S}$, without loss of generality. The resulting soft-max parameterization is still complete, which means that the optimal policy can still be represented by the soft-max parameterization. A key result from [30] shows that the augmented value function $\tilde{J}_\mu(\theta)$ in (7) satisfies a non-uniform Łojasiewicz inequality. We first show that a similar non-uniform Łojasiewicz inequality still holds under the restricted soft-max parameterization.

Lemma 4.2 *Suppose that $\mu(s) > 0$ for all states $s \in \mathcal{S}$. Then, it holds that $\|\nabla \tilde{J}_\mu(\theta)\|_2^2 \geq C(\theta)(\tilde{J}_\rho(\theta^*) - \tilde{J}_\rho(\theta))$, where $C(\theta) = \frac{2\lambda}{|\mathcal{S}||\mathcal{A}|}\min_s \mu(s) \min_{s,a} \pi_\theta(a|s)^2 \left\| \frac{d_\rho^{\pi^*_\lambda}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-1}$.*

Next, we show that action probabilities under the restricted soft-max parameterization are still uniformly bounded away from zero if the exact PG is available.

Lemma 4.3 *Using the exact PG (Algorithm 4 in Appendix D) with $\eta \leq \frac{(6+4\log|\mathcal{A}|)}{(1-\gamma)^3}$ for the entropy regularized objective, it holds that $\inf_{t\geq 1} \min_{s,a} \pi_{\theta_t}(a|s) > 0$.*

Thus, the optimal parameter θ^* and the iterates of the algorithm with the exact policy gradient will be bounded under the restricted soft-max policy parameterization.

Lemma 4.4 *For the restricted soft-max parameterization, each globally optimal solution θ^* of $\tilde{J}_\rho(\theta)$ is bounded. In addition, let $\{\bar{\theta}_t\}_{t=1}^T$ denote the iterates of the algorithm with the exact PG (Algorithm 4 in Appendix D) with $\eta \leq \frac{(6+4\log|\mathcal{A}|)}{(1-\gamma)^3}$. Then, there exists a bounded constant $\bar{\Delta}$ such that $\|\bar{\theta}_t - \theta^*\|_2 \leq \bar{\Delta}$ for all $t = 1, \ldots, T$.*

The difficulty of analyzing the entropy regularized RL with the estimated PG is governed by the fact that Lemma 4.3 is no longer guaranteed when the estimated PG is applied. This will further result in the loss of curvature in guaranteeing the global convergence. Our strategy to overcome this challenge is to relate the algorithm using the exact PG and the algorithm using the estimated PG. For the algorithm with the exact PG, we know from Lemma 4.4 that the iterates remain bounded throughout the course of the algorithms. If the estimation error of the estimated PG is small, then one could expect the iterates of the algorithm with the estimated PG to also remain bounded (with some high probability). We achieve this by performing a careful martingale analysis. By showing that the iterates remain within the region of the finite parameters, we can guarantee that Lemma 4.3 still holds for the

estimated PG, and therefore the global convergence and the sample efficiency of maximum entropy RL can be guaranteed. To study the convergence of the stochastic PG with regularized entropy, we first introduce some notations. For every $t \in \{1, 2, \ldots, T\}$, we define $\delta_t = J_\rho(\theta^*) - J_\rho(\theta_t)$ and the set $\mathcal{G}^0 := \{\theta | \, \|\theta - \theta^*\|_2 \leq 20\bar{\Delta}\}$ corresponding to those points θ whose Euclidean distance with respect to the optimal parameter θ^* is at most $20\bar{\Delta}$. It is desirable to show that the probability of having a large optimality gap is small if the PG estimation error can be well controlled.

Lemma 4.5 *Let $e_i = \nabla \tilde{J}_\mu(\theta_i) - u_i$, where u_i is an unbiased estimator of $\nabla \tilde{J}_\mu(\theta_i)$. Let $\eta \leq \min\{\frac{L_J}{12}, \frac{8}{C^0}\}$, where $L_J = \frac{(12+8\log|\mathcal{A}|)}{(1-\gamma)^3}$ and $C^0 = \min_{\theta \in \mathcal{G}^0} C(\theta)$. For every $\epsilon > 0$, it holds that*

$$\mathbb{P}(\delta_T \geq \epsilon) \leq \frac{1}{2}^{\frac{\eta C^0 T}{8}} \frac{\delta_0}{\epsilon} + \sum_{i=0}^{T} \frac{5\eta}{8\epsilon} \mathbb{E}[\|e_i\|_2^2 \mathbf{1}_{\theta_i \in \mathcal{G}^0}] + \frac{(1+\eta L_J)^{T-1} \sum_{i=0}^{T} \eta \mathbb{E}[\|e_i\|_2 \mathbf{1}_{\theta_i \in \mathcal{G}^0}] + \bar{\Delta}}{20\bar{\Delta}}$$

By applying the momentum-based PG in Algorithm 1, we arrive at the following result:

Lemma 4.6 *Consider sequences $\{\theta_t\}_{t=1}^T$ and $\{u_t^H\}_{t=1}^T$ generated by Algorithm 1, and let $e_t^H = \nabla \tilde{J}_\mu^H(\theta_t) - u_t^H$, $b^2 = L_g^2 + G^2 C_w^2$, $C_w = \sqrt{H(2HM_g^2 + M_h)(W+1)}$, $L_g = M_h(1+R)/(1-\gamma)^2 + M_h/(1-\gamma)$, $G = M_g(1+R)/(1-\gamma)^2 + M_g/(1-\gamma)$, $\beta = 48b^2\eta^2$, $\eta \leq \frac{L_J}{12}$, where $\sigma_e^2 = 2(\sigma_R^2 + \sigma_\pi^2)$, $M_g = 2$, $M_h = 1$, $R := \max_{\theta \in \mathcal{G}^0} - \log \pi_\theta(a, s)$ and $L_J = \frac{(12+8\log|\mathcal{A}|)}{(1-\gamma)^3}$. We have*

$$\sum_{t=1}^{T} \mathbb{E}\left[\eta \left\|e_t^H\right\|_2^2 \mathbf{1}_{\theta_t \in \mathcal{G}^0}\right] \leq \frac{Tc^2\sigma_e^2\eta^3}{12b^2B} + \frac{\sigma_e^2}{24b^2\eta B} + \frac{4\eta}{3B} \frac{1 + \lambda \log(|\mathcal{A}|)}{1-\gamma}.$$

Finally, we show that Algorithm 1 achieves a polynomial sample complexity in ϵ.

Theorem 4.7 *Under the conditions in Lemmas 4.5 and 4.6, let θ_T be generated by Algorithm 1. Let $\eta = \frac{L_J}{12T^{\frac{1}{4}}}$ and $T = \left(\frac{96}{L_J C^0} \log_{\frac{1}{2}}(\frac{\epsilon}{60\delta_0})\right)^{\frac{4}{3}}$, where $L_J = \frac{(12+8\log|\mathcal{A}|)}{(1-\gamma)^3}$. Let $H = \mathcal{O}\left(\log_\gamma\left(\frac{\epsilon(1-\gamma)}{\log_{\frac{1}{2}}(\epsilon)}\right)\right)$ and $B = \mathcal{O}\left(\max\left\{\left(\log_{\frac{1}{2}}(\epsilon)\right)^{\frac{1}{3}}\epsilon^{-1}, \left(\log_{\frac{1}{2}}(\epsilon)\right)^{\frac{4}{3}}\epsilon^{-\frac{16L_J}{C^0 \ln 2}}\right\}\right)$. Then, for all $\epsilon \leq \frac{60\delta_0}{2^{\left(\frac{L_J C^0}{96}\right)^4}}$, we have $\mathbb{P}(\delta_T \leq \epsilon) \geq \frac{9}{10}$. In total, it requires $\mathcal{O}\left(\log_\gamma\left(\frac{\epsilon(1-\gamma)}{\log_{\frac{1}{2}}(\epsilon)}\right) \cdot \max\left\{\left(\log_{\frac{1}{2}}(\epsilon)\right)^{\frac{5}{3}}\epsilon^{-1}, \left(\log_{\frac{1}{2}}(\epsilon)\right)^{\frac{8}{3}}\epsilon^{-\frac{16L_J}{C^0 \ln 2}}\right\}\right)$ samples to achieve an ϵ-optimal policy.*

Remark 4.8 *Algorithm 1 is guaranteed to return an ϵ-accurate solution with the constant probability $\frac{9}{10}$. This probability bound of $\frac{9}{10}$ in itself can be sharpened by a slightly more refined analysis with different constants. This constant probability guarantee has also been used for the optimality guarantees of linear quadratic systems [28]. In addition, this result provides the last-iterate convergence, which is theoretically stronger and more appealing than the time-average convergence. Finally, an any-time bound can be easily obtained by resorting to the classical doubling technique.*

In Appendix D.3.3, we also show that the convergence can be established for vanilla entropy regularized PG but with a worse sample complexity in terms of the order of the logarithmic term.

5 Conclusion

In this work, we studied the global convergence and the sample complexity of momentum-based stochastic policy gradient methods for both classical RL and MaxEnt RL. We showed that adding the 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. We then extended our analysis to MaxEnt RL, and formalized a new trajectory-based entropy gradient estimator to cope with momentum terms. Our results show that both vanilla policy gradient methods and momentum-based policy gradient methods achieve a polynomial sample complexity in the target tolerance. This work provides the first global convergence results for stochastic entropy regularized policy gradient methods. It remains an open question whether the sample complexity of MaxEnt RL with the stochastic entropy gradient estimator can be further improved.

References

[1] Alekh Agarwal, Mikael Henaff, Sham Kakade, and Wen Sun. Pc-pg: Policy cover directed exploration for provable policy gradient learning. *Advances in Neural Information Processing Systems*, 2020.

[2] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *arXiv preprint arXiv:1908.00261*, 2019.

[3] Zeyuan Allen-Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. In *International conference on machine learning*, pages 699–707. PMLR, 2016.

[4] Jonathan Baxter and Peter L Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

[5] Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

[6] Jalaj Bhandari and Daniel Russo. On the linear convergence of policy gradient methods for finite mdps. In *International Conference on Artificial Intelligence and Statistics*, pages 2386–2394. PMLR, 2021.

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

[8] Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *arXiv preprint arXiv:2007.06558*, 2020.

[9] Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *Nips*, volume 10, pages 442–450. Citeseer, 2010.

[10] Benjamin Eysenbach and Sergey Levine. If MaxEnt RL is the answer, what is the question? *arXiv preprint arXiv:1910.01913*, 2019.

[11] Benjamin Eysenbach and Sergey Levine. Maximum entropy rl (provably) solves some robust rl problems. *arXiv preprint arXiv:2103.06257*, 2021.

[12] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path integrated differential estimator. *arXiv preprint arXiv:1807.01695*, 2018.

[13] Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476. PMLR, 2018.

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

[15] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, pages 1352–1361. PMLR, 2017.

[16] Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *arXiv preprint arXiv:2007.05170*, 2020.

[17] Feihu Huang, Shangqian Gao, Jian Pei, and Heng Huang. Momentum-based policy gradient methods. In *International Conference on Machine Learning*, pages 4422–4433. PMLR, 2020.

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

[19] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

[20] Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

[21] Arbaaz Khan, Ekaterina Tolstaya, Alejandro Ribeiro, and Vijay Kumar. Graph policy gradients for large scale robot control. In *Conference on Robot Learning*, pages 823–834. PMLR, 2020.

[22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[23] Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in neural information processing systems*, pages 1008–1014. Citeseer, 2000.

[24] Vijay R Konda and John N Tsitsiklis. On actor-critic algorithms. *SIAM journal on Control and Optimization*, 42(4):1143–1166, 2003.

[25] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[26] Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural trust region/proximal policy optimization attains globally optimal policy. *Advances in Neural Information Processing Systems*, 32:10565–10576, 2019.

[27] Yanli Liu, Kaiqing Zhang, Tamer Basar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33, 2020.

[28] Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter Bartlett, and Martin Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2916–2925. PMLR, 16–18 Apr 2019.

[29] Jincheng Mei, Yue Gao, Bo Dai, Csaba Szepesvari, and Dale Schuurmans. Leveraging non-uniformity in first-order non-convex optimization. *International Conference on Machine Learning*, 2021.

[30] Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pages 6820–6829. PMLR, 2020.

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

[32] Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. In *International Conference on Machine Learning*, pages 4026–4035. PMLR, 2018.

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

[34] Nhan Pham, Lam Nguyen, Dzung Phan, Phuong Ha Nguyen, Marten Dijk, and Quoc Tran-Dinh. A hybrid stochastic policy gradient algorithm for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 374–385. PMLR, 2020.

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

[36] Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pages 314–323. PMLR, 2016.

[37] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

[38] John Schulman, Pieter Abbeel, and Xi Chen. Equivalence between policy gradients and soft q-learning. *CoRR*, abs/1704.06440, 2017.

[39] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

[40] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[41] Lior Shani, Yonathan Efroni, and Shie Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5668–5675, 2020.

[42] Zebang Shen, Alejandro Ribeiro, Hamed Hassani, Hui Qian, and Chao Mi. Hessian aided policy gradient. In *International Conference on Machine Learning*, pages 5729–5738. PMLR, 2019.

[43] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[44] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International conference on machine learning*, pages 387–395. PMLR, 2014.

[45] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[46] Richard S Sutton, David A McAllester, Satinder P Singh, Yishay Mansour, et al. Policy gradient methods for reinforcement learning with function approximation. In *NIPs*, volume 99, pages 1057–1063. Citeseer, 1999.

[47] Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019.

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

[49] Cathy Wu, Aravind Rajeswaran, Yan Duan, Vikash Kumar, Alexandre M Bayen, Sham Kakade, Igor Mordatch, and Pieter Abbeel. Variance reduction for policy gradient with action-dependent factorized baselines. *International Conference on Learning Representations*, 2018.

[50] Tianhao Wu, Mingzhi Jiang, and Lin Zhang. Cooperative multiagent deep deterministic policy gradient (comaddpg) for intelligent connected transportation with unsignalized intersection. *Mathematical Problems in Engineering*, 2020, 2020.

[51] Yue Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite time analysis of two time-scale actor critic methods. *Advances in Neural Information Processing Systems*, 2020.

[52] Huaqing Xiong, Tengyu Xu, Yingbin Liang, and Wei Zhang. Non-asymptotic convergence of adam-type reinforcement learning algorithms under markovian sampling. *arXiv preprint arXiv:2002.06286*, 2020.

[53] Pan Xu, Felicia Gao, and Quanquan Gu. An improved convergence analysis of stochastic variance-reduced policy gradient. In *Uncertainty in Artificial Intelligence*, pages 541–551. PMLR, 2020.

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

[55] Tengyu Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for (natural) actor-critic algorithms. *Advances in Neural Information Processing Systems*, 33, 2020.

[56] Tengyu Xu, Zhe Wang, and Yingbin Liang. Non-asymptotic convergence analysis of two time-scale (natural) actor-critic algorithms. *arXiv preprint arXiv:2005.03557*, 2020.

[57] Tianbing Xu, Qiang Liu, and Jian Peng. Stochastic variance reduction for policy gradient estimation. *arXiv preprint arXiv:1710.06034*, 2017.

[58] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *arXiv preprint arXiv:1810.02338*, 2018.

[59] Huizhuo Yuan, Xiangru Lian, Ji Liu, and Yuren Zhou. Stochastic recursive momentum for policy gradient methods. *arXiv preprint arXiv:2003.04302*, 2020.

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

[61] Junyu Zhang, Chengzhuo Ni, Zheng Yu, Csaba Szepesvari, and Mengdi Wang. On the convergence and sample efficiency of variance-reduced policy gradient method. *arXiv preprint arXiv:2102.08607*, 2021.

[62] Junzi Zhang, Jongho Kim, Brendan O'Donoghue, and Stephen Boyd. Sample efficient reinforcement learning with REINFORCE. *35th AAAI Conference on Artificial Intelligence*, 2021.

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

[64] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.

A 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 $\mathbf{1}_{\{X\}}$ to denote the indicator function of the event X. The notion I_d represents the identity matrix in $\mathbb{R}^{d\times d}$. 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 scalars $a, b \in \mathbb{R}$, we use $a \cdot b$ to denote the scalar product. 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. Lastly, 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.

B Proof of results in Section 3.2

Algorithm 2 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 $u_t^H = \frac{\beta_t}{B}\sum_{i=1}^B g(\tau_i^H|\theta_t, \mu) + (1-\beta_t)[u_{t-1}^H + \frac{1}{B}\sum_{i=1}^B (g(\tau_i^H|\theta_t, \mu) - w(\tau_i^H|\theta_{t-1}, \theta_t) g(\tau_i^H|\theta_{t-1}, \mu))]$;
  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

B.1 Proof of Lemma 3.2

Proposition B.1 (Theorem 5.3 in [2]) *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 starting distribution ρ, we have:*

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

B.1.1 Proof of Lemma 3.2

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

$$I^+ = \left\{t \in \{1, \ldots, T\} \,|\, \|\nabla_\theta L_{\lambda,\mu}(\theta_t)\|_2 \ge \frac{\lambda}{2|\mathcal{S}||\mathcal{A}|}\right\}, \tag{12}$$

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 $\lambda = \frac{\epsilon(1-\gamma)}{2}$, we have that $J_\rho(\theta^*) - J_\rho(\theta) \le \left\|\frac{d_\rho^{\pi^*}}{\mu}\right\|_2 \epsilon$ for all $k \in \{0, \ldots, K\}/I^+$, while $J_\rho(\theta^*) - J_\rho(\theta) \le 1/(1-\gamma)$ 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

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

Thus,

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

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 Proposition B.1 and the choice of $\lambda = \frac{\epsilon(1-\gamma)}{2}$, it holds that

$$J_\rho(\theta^*) - J_\rho(\theta_t) \leq \left\| \frac{d_\rho^{\pi_{\theta^*}}}{\mu} \right\|_\infty \epsilon, \quad \forall t \notin I^+.$$

Therefore,

$$\begin{aligned}\sum_{t=1}^{T} \left(J_\rho(\theta^*) - J_\rho(\theta_t)\right) &= \sum_{t\in I^+} \left(J_\rho(\theta^*) - J_\rho(\theta_t)\right) + \sum_{t\notin I^+} \left(J_\rho(\theta^*) - J_\rho(\theta_t)\right) \\ &\leq |I^+| \frac{1}{1-\gamma} + \left(T + 1 - |I^+|\right) \left\| \frac{d_\rho^{\pi_{\theta^*}}}{\mu} \right\|_\infty \epsilon \\ &\leq \frac{4|\mathcal{S}|^2|\mathcal{A}|^2 \sum_{t=1}^{T} \eta_t \left\|\nabla L_{\lambda,\rho}(\theta_t)\right\|_2^2}{\lambda^2 \eta_T (1-\gamma)} + (T+1) \left\| \frac{d_\rho^{\pi_{\theta^*}}}{\mu} \right\|_\infty \epsilon \qquad (13)\end{aligned}$$

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

$$\sum_{t=1}^{T} \left(J_\rho(\theta^*) - J_\rho(\theta_t)\right) \leq (T+1) \left\| \frac{d_\rho^{\pi_{\theta^*}}}{\mu} \right\|_\infty \epsilon,$$

and hence (13) always holds. This completes the proof. □

B.1.2 Proof of Lemma 3.3

We first notice that Assumption 2.2 is satisfied by the soft-max parameterization with $M_g = 2$ and $M_h = 1$.

Lemma B.2 *For the soft-max parameterization, Assumption 2.2 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, $\left\|\nabla_\theta \log \pi_\theta(a|s)\right\|_2 \leq 2$ 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 [30], 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 B.3 *Suppose that the stochastic policy gradient u_t^H is generated by Algorithm 2 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]\\
\leq& \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}\left[\|(\nabla J_\rho^H(\theta_t) - \frac{1}{B}\sum_{i=1}^B g(\tau_i^H|\theta_t,\mu))\|_2^2\right]\\
&+ 2\eta_{t-1}^{-1}(1-\beta_t)^2\mathbb{E}\left[\|\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\right]\\
=& \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]\\
\leq& \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 Assumption 2.3 and $\mathbb{E}\|\xi - \mathbb{E}[\xi]\|_2^2 \leq \mathbb{E}\|\xi\|_2^2$.

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 E.1, and the third inequality is due to Proposition E.2, and the last inequality holds by Proposition E.3. 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. □

Lemma B.4 *Under the conditions of Lemma B.3, consider a constant* $\bar{\lambda} \in [\lambda, \infty)$*, and let* $k > 0$*,* $c = \frac{1}{3k^3L_J^H} + 96B^2$*,* $m = \max\{2, (12L_J^Hk)^3, (\frac{ck}{12L_J^H})^3\}$ *and* $\eta_0 = \frac{k}{m^{1/3}}$ *where* $L_J^H = \frac{8}{(1-\gamma)^3} + \frac{2\bar{\lambda}}{|\mathcal{S}|}$*. Then,*

$$
\sum_{t=1}^{T}\mathbb{E}\left[\eta_t\left\|e_t^H\right\|_2^2\right] \le \frac{\Gamma_1}{B},
$$

where $\Gamma_1 = \left(\frac{c^2\sigma^2k^3\ln(T+2)}{36b^2} + \frac{m^{1/3}}{72b^2k}\sigma^2 + \frac{4}{9}\left(L_{\lambda,\rho}^H(\theta^*) - L_{\lambda,\rho}^H(\theta_1)\right)\right)$.

Proof. From Lemma F.2, we know that $L_{\lambda,\mu}(\theta)$ is $\frac{8}{(1-\gamma)^3} + \frac{2\lambda}{|\mathcal{S}|}$-smooth. Since $\bar{\lambda} \ge \lambda$, $L_{\lambda,\mu}(\theta)$ is also L_J^H-smooth. Due to $m \ge (12L_J^Hk)^3$, we have $\eta_t \le \eta_0 = \frac{k}{m^{1/3}} \le \frac{1}{12L_J^H}$. Since $\eta_t \le \frac{1}{12L_J^H}$, we obtain that $\beta_{t+1} = c\eta_t^2 \le \frac{c\eta_t}{12L_J^H} \le \frac{ck}{12L_J^Hm^{1/3}} \le 1$. Now, it results from Lemma B.3 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

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

where the second inequality holds by $m \ge 2$, and the fifth inequality uses the property $0 < \eta \le \frac{1}{2L_J^H}$. Then, it holds that

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

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

$$
\mathbb{E}[\eta_t^{-1}\|e_{t+1}^H\|_2^2 - \eta_{t-1}^{-1}\|e_t^H\|_2^2] \le \mathbb{E}\left[-96b^2\eta_t\|e_t^H\|_2^2 + \frac{2\eta_t^{-1}\beta_{t+1}^2\sigma^2}{B} + \frac{4b^2\eta_t^{-1}}{B}\|\theta_{t+1} - \theta_t\|_2^2\right]. \tag{14}
$$

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

$$\frac{1}{96b^2}\left(\mathbb{E}\left[\frac{\left\|e_{T+1}^H\right\|_2^2}{\eta_T}-\frac{\left\|e_1^H\right\|_2^2}{\eta_0}\right]\right)\leq\sum_{t=1}^{T}\mathbb{E}\left[\frac{c^2\eta_t^3\sigma^2}{48b^2B}-\eta_t\left\|e_t^H\right\|_2^2+\frac{1}{24\eta_tB}\left\|\theta_{t+1}-\theta_t\right\|_2^2\right] \tag{15}$$

Then, it follows from Lemma E.4 that

$$\begin{aligned}
\frac{1}{96b^2}\left(\mathbb{E}\left[\frac{\left\|e_{T+1}^H\right\|_2^2}{\eta_T}-\frac{\left\|e_1^H\right\|_2^2}{\eta_0}\right]\right)\leq&\sum_{\tau=1}^{T}\mathbb{E}\left[\frac{1}{48b^2B}c^2\eta_t^3\sigma^2-\frac{(12B-3)\eta_t}{12B}\left\|e_t^H\right\|_2^2+\frac{1}{3B}\left(L_{\lambda,\rho}^H(\theta_{t+1})-L_{\lambda,\rho}^H(\theta_t)\right)\right]\\
\leq&\sum_{t=1}^{T}\left(\frac{1}{48b^2B}c^2\eta_t^3\sigma^2-\mathbb{E}\left[\frac{3\eta_t}{4}\left\|e_t^H\right\|_2^2\right]\right)+\frac{1}{3B}\left(L_{\lambda,\rho}^H(\theta^*)-L_{\lambda,\rho}^H(\theta_1)\right)\\
\leq&\frac{c^2\sigma^2k^3}{48b^2B}\sum_{t=1}^{T}\frac{1}{m+t}-\sum_{t=1}^{T}\mathbb{E}\left[\frac{3\eta_t}{4}\left\|e_t^H\right\|_2^2\right]+\frac{1}{3B}\left(L_{\lambda,\rho}^H(\theta^*)-L_{\lambda,\rho}^H(\theta_1)\right)\\
\leq&\frac{c^2\sigma^2k^3}{48b^2B}\sum_{t=1}^{T}\frac{1}{2+t}-\sum_{t=1}^{T}\mathbb{E}\left[\frac{3\eta_t}{4}\left\|e_t^H\right\|_2^2\right]+\frac{1}{3B}\left(L_{\lambda,\rho}^H(\theta^*)-L_{\lambda,\rho}^H(\theta_1)\right)\\
\leq&\frac{c^2\sigma^2k^3\ln(T+2)}{48b^2B}-\sum_{t=1}^{T}\mathbb{E}\left[\frac{3\eta_t}{4}\left\|e_t^H\right\|_2^2\right]+\frac{1}{3B}\left(L_{\lambda,\rho}^H(\theta^*)-L_{\lambda,\rho}^H(\theta_1)\right).
\end{aligned} \tag{16}$$

By rearranging the above inequality, we have

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}\left[\frac{3\eta_t}{4}\left\|e_t^H\right\|_2^2\right]\leq&\frac{c^2\sigma^2k^3\ln(T+2)}{48b^2B}+\frac{1}{96b^2\eta_0}\mathbb{E}\left[\left\|e_1^H\right\|_2^2\right]+\frac{1}{3B}\left(L_{\lambda,\rho}^H(\theta^*)-L_{\lambda,\rho}^H(\theta_1)\right)\\
\leq&\frac{1}{B}\left(\frac{c^2\sigma^2k^3\ln(T+2)}{48b^2}+\frac{m^{1/3}\sigma^2}{96b^2k}+\frac{1}{3}\left(L_{\lambda,\rho}^H(\theta^*)-L_{\lambda,\rho}^H(\theta_1)\right)\right).
\end{aligned}$$

Multiplying both sides by $\frac{4}{3}$ yields that

$$\sum_{t=1}^{T}\mathbb{E}\left[\eta_t\left\|e_t^H\right\|_2^2\right]\leq\frac{1}{B}\left(\frac{c^2\sigma^2k^3\ln(T+2)}{36b^2}+\frac{m^{1/3}\sigma^2}{72b^2k}+\frac{4}{9}\left(L_{\lambda,\rho}^H(\theta^*)-L_{\lambda,\rho}^H(\theta_1)\right)\right).$$

□

Lemma B.5 *Under the same conditions of Lemma B.4, we have*

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

where $\Gamma_1=\left(\frac{c^2\sigma^2k^3\ln(T+2)}{36b^2}+\frac{m^{1/3}}{72b^2k}\sigma^2+\frac{4}{9}\left(L_{\lambda,\rho}^H(\theta^*)-L_{\lambda,\rho}^H(\theta_1)\right)\right)$, $\Gamma_2=12(L_{\lambda,\rho}^H(\theta^*)-L_{\lambda,\rho}^H(\theta_1))$, *and* $\Gamma_3=\frac{1}{k}(\frac{\sigma^2m^{1/3}}{8b^2k}+\frac{c^2\sigma^2k^3}{4b^2}\ln(2+T))$.

Proof. We define a Lyapunov function $\Phi_t(\theta_t)=L_{\lambda,\rho}^H(\theta_t)-\frac{1}{96b^2\eta_{t-1}}\left\|e_t^H\right\|_2^2$ for all $t\geq1$. Then,

$$\begin{aligned}
\mathbb{E}[\Phi_{t+1}-\Phi_t]=&\mathbb{E}\left[L_{\lambda,\rho}^H(\theta_{t+1})-L_{\lambda,\rho}^H(\theta_t)-\frac{1}{96b^2\eta_t}\left\|e_{t+1}^H\right\|_2^2+\frac{1}{96b^2\eta_{t-1}}\left\|e_t^H\right\|_2^2\right]\\
\geq&\mathbb{E}\left[-\frac{3\eta_t}{4}\left\|e_t^H\right\|_2^2+\frac{1}{8\eta_t}\left\|\theta_{t+1}-\theta_t\right\|_2^2-\frac{1}{96b^2\eta_t}\left\|e_{t+1}^H\right\|_2^2+\frac{1}{96b^2\eta_{t-1}}\left\|e_t^H\right\|_2^2\right]\\
\geq&\mathbb{E}\left[\frac{1}{8\eta_t}\left\|\theta_{t+1}-\theta_t\right\|_2^2-\frac{\beta_t^2\sigma^2}{48b^2B\eta_t}-\frac{1}{24B\eta_t}\|\theta_{t+1}-\theta_t\|^2\right]\\
\geq&\mathbb{E}\left[\frac{1}{12\eta_t}\left\|\theta_{t+1}-\theta_t\right\|_2^2-\frac{c^2\eta_t^3\sigma^2}{48b^2B}\right]
\end{aligned}$$

where the first inequality holds by Lemma E.4 and the second inequality holds due to (14). Summing the above inequality over t from 1 to T, we obtain

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}\left[\eta_t^{-1}\left\|\theta_{t+1}-\theta_t\right\|_2^2\right] \leq & \mathbb{E}\left[12(\Phi_{T+1}-\Phi_1)+\sum_{t=1}^{T}\frac{c^2\eta_t^3\sigma^2}{4b^2B}\right] \\
\leq & \mathbb{E}\left[12\left(L_{\lambda,\rho}^H(\theta^*)-L_{\lambda,\rho}^H(\theta_1)\right)+\frac{1}{8b^2\eta_0}\mathbb{E}\|e_1^H\|_2^2+\frac{c^2\sigma^2k^3}{4b^2B}\sum_{t=1}^{T}\frac{1}{m+t}\right] \\
\leq & \mathbb{E}\left[12\left(L_{\lambda,\rho}^H(\theta^*)-L_{\lambda,\rho}^H(\theta_1)\right)+\frac{1}{8b^2\eta_0}\mathbb{E}\|e_1^H\|_2^2+\frac{c^2\sigma^2k^3}{4b^2B}\sum_{t=1}^{T}\frac{1}{2+t}\right] \\
\leq & \mathbb{E}\left[12\left(L_{\lambda,\rho}^H(\theta^*)-L_{\lambda,\rho}^H(\theta_1)\right)+\frac{\sigma^2m^{1/3}}{8b^2kB}+\frac{c^2\sigma^2k^3}{4b^2B}\ln(2+T)\right] \\
\leq & \Gamma_2+\frac{\Gamma_3}{B},
\end{aligned}$$

where $\Gamma_2 = 12\left(L_{\lambda,\rho}^H(\theta^*)-L_{\lambda,\rho}^H(\theta_1)\right)$ and $\Gamma_3 = (\frac{\sigma^2m^{1/3}}{8b^2k}+\frac{c^2\sigma^2k^3}{4b^2}\ln(2+T))$. Then, it follows from Lemma B.4 that

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}\left[\eta_t\left\|\nabla L_{\lambda,\rho}^H(\theta_t)\right\|_2^2\right] \leq & \sum_{t=1}^{T}\mathbb{E}\left[\eta_t\left\|u_t^H+\frac{\lambda}{|\mathcal{A}||\mathcal{S}|}\sum_{s,a}\nabla\log\pi_\theta(a|s)\right\|_2^2\right]+\sum_{t=1}^{T}\mathbb{E}\left[\eta_t\left\|e_t^H\right\|_2^2\right] \\
= & \sum_{t=1}^{T}\mathbb{E}\left[\eta_t^{-1}\left\|\theta_{t+1}-\theta_t\right\|_2^2\right]+\sum_{t=1}^{T}\mathbb{E}\left[\eta_t\left\|e_t^H\right\|_2^2\right] \\
\leq & \Gamma_2+\frac{\Gamma_1+\Gamma_3}{B}.
\end{aligned}$$

□

B.2 Proof of Theorem 3.4

Proof. From Lemma E.1, we have

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

In light of Lemma 3.2, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\left(J_\rho(\theta^*)-J_\rho(\theta_t)\right)\right] \leq & \frac{4|\mathcal{S}|^2|\mathcal{A}|^2\sum_{t=1}^{T}\eta_t\mathbb{E}\left[\left\|\nabla L_{\lambda,\rho}(\theta_t)\right\|_2^2\right]}{\lambda^2\eta_T(1-\gamma)}+(T+1)\left\|\frac{d_\rho^{\pi_{\theta^*}}}{\rho}\right\|_\infty\epsilon \\
\leq & \frac{4|\mathcal{S}|^2|\mathcal{A}|^2\sum_{t=1}^{T}\eta_t\mathbb{E}\left[\left\|\nabla L_{\lambda,\rho}^H(\theta_t)\right\|_2^2\right]}{\lambda^2\eta_T(1-\gamma)}+(T+1)\left\|\frac{d_\rho^{\pi_{\theta^*}}}{\rho}\right\|_\infty\epsilon \\
& +\frac{4|\mathcal{S}|^2|\mathcal{A}|^2\sum_{t=1}^{T}\eta_t}{\lambda^2\eta_T(1-\gamma)}\left(M_g\left(\frac{H+1}{1-\gamma}+\frac{\gamma}{(1-\gamma)^2}\right)\gamma^H\right)^2
\end{aligned}$$

By choosing $\eta_t = \frac{k}{(m+t)^{1/3}}$ and $\epsilon \leq \frac{2\bar{\lambda}}{1-\gamma}$, it results from Lemma 3.3 that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\left(J_\rho(\theta^*) - J_\rho(\theta_t)\right)\right] \leq & \frac{4|\mathcal{S}|^2|\mathcal{A}|^2}{\lambda^2\eta_T(1-\gamma)}\left(\Gamma_2 + \frac{\Gamma_1+\Gamma_3}{B}\right) + (T+1)\left\|\frac{d_\rho^{\pi_{\theta^*}}}{\rho}\right\|_\infty \epsilon \\
& + \frac{4|\mathcal{S}|^2|\mathcal{A}|^2\sum_{t=1}^{T}\eta_t}{\lambda^2\eta_T(1-\gamma)}\left(M_g\left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\right)\gamma^H\right)^2 \\
\leq & \frac{4|\mathcal{S}|^2|\mathcal{A}|^2(m+T)^{1/3}}{\lambda^2 k(1-\gamma)}\left(\Gamma_2 + \frac{\Gamma_1+\Gamma_3}{B}\right) + (T+1)\left\|\frac{d_\rho^{\pi_{\theta^*}}}{\rho}\right\|_\infty \epsilon \\
& + \frac{6|\mathcal{S}|^2|\mathcal{A}|^2(m+T)}{\lambda^2(1-\gamma)}\left(M_g\left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\right)\gamma^H\right)^2
\end{aligned}$$

By substituting $\lambda = \epsilon(1-\gamma)/2$ arrive at

$$\begin{aligned}
\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\left(J_\rho(\theta^*) - J_\rho(\theta_t)\right)\right] \leq & \frac{16|\mathcal{S}|^2|\mathcal{A}|^2(m+T)^{1/3}}{\epsilon^2(1-\gamma)^3 kT}\left(\Gamma_2 + \frac{\Gamma_1+\Gamma_3}{B}\right) + \frac{T+1}{T}\left\|\frac{d_\rho^{\pi_{\theta^*}}}{\rho}\right\|_\infty \epsilon \\
& + \frac{24|\mathcal{S}|^2|\mathcal{A}|^2(m+T)}{\epsilon^2(1-\gamma)^3 T}\left(M_g\left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\right)\gamma^H\right)^2 \\
\leq & \frac{16|\mathcal{S}|^2|\mathcal{A}|^2(m+T)^{1/3}}{\epsilon^2(1-\gamma)^3 kT}\left(\Gamma_2 + \frac{\Gamma_1+\Gamma_3}{B}\right) + 2\left\|\frac{d_\rho^{\pi_{\theta^*}}}{\rho}\right\|_\infty \epsilon \\
& + \frac{24|\mathcal{S}|^2|\mathcal{A}|^2(m+T)}{\epsilon^2(1-\gamma)^3 T}\left(M_g\left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\right)\gamma^H\right)^2 \\
\leq & \tilde{\mathcal{O}}\left(\frac{\sigma^2|\mathcal{S}|^2|\mathcal{A}|^2}{\epsilon^2(1-\gamma)^7 T^{2/3}}\right) + 2\left\|\frac{d_\rho^{\pi_{\theta^*}}}{\rho}\right\|_\infty \epsilon \\
& + \frac{24|\mathcal{S}|^2|\mathcal{A}|^2(m+T)}{\epsilon^2(1-\gamma)^3 T}\left(M_g\left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\right)\gamma^H\right)^2.
\end{aligned}$$

By taking

$$H = \mathcal{O}\left(\log_\gamma\left(\frac{(1-\gamma)\epsilon}{|\mathcal{S}||\mathcal{A}|}\right)\right)$$

we obtain that

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\left(J_\rho(\theta^*) - J_\rho(\theta_t)\right)\right] \leq \tilde{\mathcal{O}}\left(\frac{\sigma^2|\mathcal{S}|^2|\mathcal{A}|^2}{\epsilon^2(1-\gamma)^7 T^{2/3}} + \left\|\frac{d_\rho^{\pi_{\theta^*}}}{\rho}\right\|_\infty \epsilon\right)$$

This completes the proof. □

C Proof of results in Section 3.3

Algorithm 3 Momentum-based PG with Fisher-non-degenerate parameterization (MBPG-F)

Inputs: Iteration T, Horizon H, batch size B, initial input θ_1, parameters $\{k, m, c\}$ and initial distribution ρ;
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 $p(\cdot|\theta_t, \rho)$;
  if $t = 1$ **then**
    Compute $u_1^H = \frac{1}{B}\sum_{i=1}^B g(\tau_i^H|\theta_1, \rho)$;
  else
    Compute $u_t^H = \frac{\beta_t}{B}\sum_{i=1}^B g(\tau_i^H|\theta_t, \rho) + (1-\beta_t)[u_{t-1}^H + \frac{1}{B}\sum_{i=1}^B (g(\tau_i^H|\theta_t, \rho) - w(\tau_i^H|\theta_{t-1}, \theta_t) g(\tau_i^H|\theta_{t-1}, \rho))]$;
  end if
  Compute $\eta_t = \frac{k}{(m+t)^{1/3}}$;
  Update $\theta_{t+1} = \theta_t + \eta_t u_t^H$;
  Update $\beta_{t+1} = c\eta_t^2$;
end for

C.1 Proof of Lemma 3.7

Proposition C.1 *[Proposition 4.5 in [27]] Let $\{\theta_t\}_{t=1}^T$ be generated by a general update of the form $\theta_{t+1} = \theta_t + \eta_t u_t$ and let $u_t^* = F_\rho^{-1}(\theta_t)\nabla J_\rho(\theta_t)$. Then*

$$J_\rho(\theta^*) - \frac{1}{T}\sum_{t=1}^T J_\rho(\theta_t) \le \frac{\sqrt{\varepsilon_{bias}}}{1-\gamma} + \frac{M_g}{T}\sum_{t=1}^T \|u_t - u_t^\star\|_2 + \frac{M_h}{2T}\sum_{t=1}^T \eta_t \|u_t\|_2^2 + \frac{1}{T}\sum_{k=1}^{T-1}\frac{1}{\eta_t}\mathbb{E}_{s\sim d_\rho^{\pi_{\theta_*}}}\mathbb{E}_{a\sim\pi_{\theta^*}(\cdot|s)}\left[\log\frac{\pi_{\theta_{t+1}}(a\mid s)}{\pi_{\theta_t}(a\mid s)}\right].$$

C.1.1 Proof of Lemma 3.7

Proof. By the M_h-smoothness of the score function in Assumption 2.2, we have

$$\begin{aligned}
&\mathbb{E}_{s\sim d_\rho^{\pi_{\theta_*}}}\mathbb{E}_{a\sim\pi_{\theta^*}(\cdot|s)}\left[\log\frac{\pi_{\theta_{t+1}}(a\mid s)}{\pi_{\theta_t}(a\mid s)}\right] \\
&\le \mathbb{E}_{s\sim d_\rho^{\pi^*}}\mathbb{E}_{a\sim\pi_{\theta^*}(\cdot|s)}\left[\nabla_\theta \log\pi_{\theta_t}(a\mid s)\cdot(\theta_{t+1}-\theta_t)\right] + \frac{M_h}{2}\|\theta_{t+1}-\theta_t\|_2^2 \\
&\le M_g\|\theta_{t+1}-\theta_t\|_2 + \frac{M_h}{2}\|\theta_{t+1}-\theta_t\|_2^2 \\
&= M_g\eta_t\|u_t\|_2 + \frac{M_h}{2}\eta_t^2\|u_t\|_2^2 \\
&\le M_g\eta_t\|u_t - u_t^*\|_2 + M_g\eta_t\|u_t^*\|_2 + \frac{M_h}{2}\eta_t^2\|u_t\|_2^2.
\end{aligned}$$

Combining it with Proposition C.1 yields that

$$J_\rho(\theta^*) - \frac{1}{T}\sum_{t=1}^T J_\rho(\theta_t) \le \frac{\sqrt{\varepsilon_{\text{bias}}}}{1-\gamma} + \frac{M_h}{T}\sum_{t=1}^T \eta_t\|u_t\|_2^2 + \frac{M_g}{T}\sum_{t=1}^T\|u_t^*\|_2 + \frac{2M_g}{T}\sum_{t=1}^T\|u_t - u_t^\star\|_2. \quad (17)$$

By Assumption 2.1, we have

$$\begin{aligned}
\|u_t - u_t^\star\|_2^2 &= \left\|u_t - \nabla J_\rho(\theta_t) + \nabla J_\rho(\theta_t) - F^{-1}(\theta_t)\nabla J_\rho(\theta_t)\right\|_2^2 \\
&\le 2\|u_t - \nabla J_\rho(\theta_t)\|_2^2 + 2\left\|\nabla J_\rho(\theta_t) - F^{-1}(\theta_t)\nabla J_\rho(\theta_t)\right\|_2^2 \\
&\le 2\|u_t - \nabla J_\rho(\theta_t)\|_2^2 + 2(1+\frac{1}{\mu_F})_2^2\|\nabla J_\rho(\theta_t)\|_2^2.
\end{aligned}$$

and $\|u_t^*\|_2 \leq \frac{1}{\mu_F}\|\nabla J_\rho(\theta_t)\|_2$. Now, it can be concluded from (17) that

$$J_\rho(\theta^*) - \frac{1}{T}\sum_{t=1}^{T} J_\rho(\theta_t) \leq \frac{\sqrt{\varepsilon_{\text{bias}}}}{1-\gamma} + \frac{M_h}{T}\sum_{t=1}^{T}\eta_t\|u_t\|_2^2 + \frac{M_g}{\mu_F T}\sum_{t=1}^{T}\|\nabla J_\rho(\theta_t)\|_2 + \frac{2M_g}{T}\sum_{t=1}^{T}\left(2\|u_t - \nabla J_\rho(\theta_t)\|_2^2 + 2(1+\frac{1}{\mu_F})^2\|\nabla J_\rho(\theta_t)\|_2^2\right).$$

This completes the proof. □

C.2 Proof of Lemma 3.9

Lemma C.2 *Under Assumption 2.2, suppose that the stochastic policy gradient u_t is generated by Algorithm 3 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 B.3 with M_g and M_h defined in Assumption 2.2. The details are omitted for brevity. □

C.2.1 Proof of Lemma 3.9

Proof. The function $J_\rho^H(\theta)$ is L_J^H-smooth due to Assumption 3.8 and Lemma F.1. Moreover, becasue of $m \geq (12L_J^H k)^3$, it holds that $\eta_t \leq \eta_0 = \frac{k}{m^{1/3}} \leq \frac{1}{12L_J^H}$. Since $\eta_t \leq \frac{1}{12L_J^H}$, we have $\beta_{t+1} = c\eta_t^2 \leq \frac{c\eta_t}{12L_J^H} \leq \frac{ck}{12L_J^H m^{1/3}} \leq 1$. It follows from Lemma C.2 that

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

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

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

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

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

where the last equality holds by $c = \frac{1}{3k^3L_J^H} + 96b^2$. 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

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

Then, Lemma E.4 can be used to obtain

$$
\begin{aligned}
&\frac{1}{96b^2}\left(\mathbb{E}\left[\frac{\left\|e_{T+1}^H\right\|_2^2}{\eta_T}-\frac{\left\|e_1^H\right\|_2^2}{\eta_0}\right]\right)\\
&\leq\sum_{\tau=1}^{T}\mathbb{E}\left[\frac{1}{48b^2B}c^2\eta_t^3\sigma^2-\frac{(12B-3)\eta_t}{12B}\left\|e_t^H\right\|_2^2+\frac{1}{3B}\left(J_\rho^H(\theta_{t+1})-J_\rho^H(\theta_t)\right)\right]\\
&\leq\sum_{t=1}^{T}\left(\frac{1}{48b^2B}c^2\eta_t^3\sigma^2-\mathbb{E}\left[\frac{3\eta_t}{4}\left\|e_t^H\right\|_2^2\right]\right)+\frac{1}{3B}\left(J_\rho^H(\theta_{T+1})-J_\rho^H(\theta_1)\right)\\
&\leq\frac{c^2\sigma^2k^3}{48b^2B}\sum_{t=1}^{T}\frac{1}{m+t}-\sum_{t=1}^{T}\mathbb{E}\left[\frac{3\eta_t}{4}\left\|e_t^H\right\|_2^2\right]+\frac{1}{3B}\left(J_\rho^H(\theta^*)-J_\rho^H(\theta_1)\right)\\
&\leq\frac{c^2\sigma^2k^3}{48b^2B}\sum_{t=1}^{T}\frac{1}{2+t}-\sum_{t=1}^{T}\mathbb{E}\left[\frac{3\eta_t}{4}\left\|e_t^H\right\|_2^2\right]+\frac{1}{3B}\left(J_\rho^H(\theta^*)-J_\rho^H(\theta_1)\right)\\
&\leq\frac{c^2\sigma^2k^3\ln(T+2)}{48b^2B}-\sum_{t=1}^{T}\mathbb{E}\left[\frac{3\eta_t}{4}\left\|e_t^H\right\|_2^2\right]+\frac{1}{3B}\left(J_\rho^H(\theta^*)-J_\rho^H(\theta_1)\right).
\end{aligned}
\tag{19}
$$

Rearranging the above inequality gives rise to

$$
\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}\left[\frac{3\eta_t}{4}\left\|e_t^H\right\|_2^2\right]\leq&\frac{c^2\sigma^2k^3\ln(T+2)}{48b^2B}+\frac{1}{96b^2\eta_0}\mathbb{E}\left[\left\|e_1^H\right\|_2^2\right]+\frac{1}{3B}\left(J_\rho^H(\theta^*)-J_\rho^H(\theta_1)\right)\\
\leq&\frac{1}{B}\left(\frac{c^2\sigma^2k^3\ln(T+2)}{48b^2}+\frac{m^{1/3}\sigma^2}{96b^2k}+\frac{1}{3}\left(J_\rho^H(\theta^*)-J_\rho^H(\theta_1)\right)\right).
\end{aligned}
$$

Multiplying both sides by $\frac{4}{3}$ yields that

$$
\sum_{t=1}^{T}\mathbb{E}\left[\eta_t\left\|e_t^H\right\|_2^2\right]\leq\frac{1}{B}\left(\frac{c^2\sigma^2k^3\ln(T+2)}{36b^2}+\frac{m^{1/3}\sigma^2}{72b^2k}+\frac{4}{9}\left(J_\rho^H(\theta^*)-J_\rho^H(\theta_1)\right)\right).
\tag{20}
$$

Since η_t is decreasing, we have $\eta_t\geq\eta_T$ for all $t=1,\ldots,T$. Therefore,

$$
\begin{aligned}
\sum_{t=1}^{T}\left[\left\|e_t^H\right\|_2^2\right]\leq&\frac{1}{\eta_T}\sum_{t=1}^{T}\mathbb{E}\left[\eta_t\left\|e_t^H\right\|_2^2\right]\\
\leq&\frac{(m+T)^{1/3}}{kB}\left(\frac{c^2\sigma^2k^3\ln(T+2)}{36b^2}+\frac{m^{1/3}\sigma^2}{72b^2k}+\frac{4}{9}\left(J_\rho^H(\theta^*)-J_\rho^H(\theta_1)\right)\right)\\
\leq&\frac{(m^{1/3}+T^{1/3})}{kB}\left(\frac{c^2\sigma^2k^3\ln(T+2)}{36b^2}+\frac{m^{1/3}\sigma^2}{72b^2k}+\frac{4}{9(1-\gamma)}\right).
\end{aligned}
$$

where the last inequality holds due to $(a+b)^{1/3}\leq a^{1/3}+b^{1/3}$ for all $a>0,b>0$ as well as $J_\rho^H(\theta^*)-J_\rho^H(\theta_1)\leq\frac{1}{1-\gamma}$.

Next, we define a Lyapunov function $\Phi_t(\theta_t)=J_\rho^H(\theta_t)-\frac{1}{96b^2\eta_{t-1}}\left\|e_t^H\right\|^2$ for all $t\geq1$. One can write

$$
\begin{aligned}
\mathbb{E}[\Phi_{t+1}-\Phi_t]=&\mathbb{E}\left[J_\rho^H(\theta_{t+1})-J_\rho^H(\theta_t)-\frac{1}{96b^2\eta_t}\left\|e_{t+1}^H\right\|_2^2+\frac{1}{96b^2\eta_{t-1}}\left\|e_t^H\right\|_2^2\right]\\
\geq&\mathbb{E}\left[-\frac{3\eta_t}{4}\left\|e_t^H\right\|_2^2+\frac{1}{8\eta_t}\left\|\theta_{t+1}-\theta_t\right\|_2-\frac{1}{96b^2\eta_t}\left\|e_{t+1}^H\right\|_2^2+\frac{1}{96b^2\eta_{t-1}}\left\|e_t^H\right\|_2^2\right]\\
\geq&\mathbb{E}\left[\frac{1}{8\eta_t}\left\|\theta_{t+1}-\theta_t\right\|_2-\frac{\beta_t^2\sigma^2}{48b^2B\eta_t}-\frac{1}{24B\eta_t}\|\theta_{t+1}-\theta_t\|_2^2\right]\\
\geq&\mathbb{E}\left[\frac{1}{12\eta_t}\left\|\theta_{t+1}-\theta_t\right\|_2-\frac{c^2\eta_t^3\sigma^2}{48b^2B}\right]
\end{aligned}
$$

where the first inequality holds by Lemma E.4 and the second inequality follows from (18). Summing the above inequality over t from 1 to T yields that

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}\left[\eta_t\left\|u_t^H\right\|_2^2\right]=&\sum_{t=1}^{T}\mathbb{E}\left[\eta_t^{-1}\left\|\theta_{t+1}-\theta_t\right\|_2^2\right]\\
\leq&\mathbb{E}\left[12(\Phi_{T+1}-\Phi_1)+\sum_{t=1}^{T}\frac{c^2\eta_t^3\sigma^2}{4b^2B}\right]\\
\leq&\mathbb{E}\left[12(J_\rho^H(\theta_{T+1})-J_\rho^H(\theta_1))+\frac{1}{8b^2\eta_0}\mathbb{E}\left[\|e_1^H\|_2^2\right]+\frac{c^2\sigma^2k^3}{4b^2B}\sum_{t=1}^{T}\frac{1}{m+t}\right]\\
\leq&\mathbb{E}\left[12(J_\rho^H(\theta_{T+1})-J_\rho^H(\theta_1))+\frac{1}{8b^2\eta_0}\mathbb{E}\left[\|e_1^H\|_2^2\right]+\frac{c^2\sigma^2k^3}{4b^2B}\sum_{t=1}^{T}\frac{1}{2+t}\right]\\
\leq&12\left(J_\rho^H(\theta^*)-J_\rho^H(\theta_1)\right)+\frac{\sigma^2m^{1/3}}{8b^2kB}+\frac{c^2\sigma^2k^3}{4b^2B}\ln(2+T)\\
\leq&\frac{12}{1-\gamma}+\frac{\sigma^2m^{1/3}}{8b^2kB}+\frac{c^2\sigma^2k^3}{4b^2B}\ln(2+T)\\
=&\Gamma_2+\frac{\Gamma_3}{B}.
\end{aligned}$$

where the last inequality is due to $J_\rho^H(\theta^*)-J_\rho^H(\theta_1)\leq\frac{1}{1-\gamma}$. It results from (20) that

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

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

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

□

C.3 Proof of Theorem 3.10

Proof. Let $e_t^H=u_t^H-\nabla J_\rho^H(\theta_t)$. By Lemma E.1, we have

$$\begin{aligned}
J_\rho(\theta^*)-\frac{1}{T}\sum_{t=1}^{T}J_\rho(\theta_t)\leq&\frac{\sqrt{\varepsilon_{\text{bias}}}}{1-\gamma}+\frac{M_h}{T}\sum_{t=1}^{T}\eta_t\left\|u_t^H\right\|_2^2+\frac{M_g}{\mu_F T}\sum_{t=1}^{T}\left\|\nabla J_\rho^H(\theta_t)\right\|_2+\frac{4M_g}{T}\sum_{t=1}^{T}\left\|e_t^H\right\|_2^2\\
&+\frac{4M_g}{T}(1+\frac{1}{\mu_F})^2\sum_{t=1}^{T}\left\|\nabla J_\rho^H(\theta_t)\right\|_2^2+\frac{M_g^2}{\mu_F}\left(\frac{H+1}{1-\gamma}+\frac{\gamma}{(1-\gamma)^2}\right)\gamma^H\\
&+\left(4M_g^2(1+\frac{1}{\mu_F})^2+4M_g^2\right)\left(\frac{H+1}{1-\gamma}+\frac{\gamma}{(1-\gamma)^2}\right)\gamma^H.
\end{aligned}\tag{21}$$

Using Lemma 3.9, one can write

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

where $\Gamma_1 = \left(\frac{c^2\sigma^2k^3\ln(T+2)}{36b^2} + \frac{m^{1/3}}{72b^2k}\sigma^2 + \frac{4}{9(1-\gamma)}\right)$, $\Gamma_2 = \frac{12}{1-\gamma}$, and $\Gamma_3 = \frac{1}{k}(\frac{\sigma^2m^{1/3}}{8b^2k} + \frac{c^2\sigma^2k^3}{4b^2}\ln(2+T))$. Now one can use Jensen's inequality to conclude that

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

where the last inequality follows from the inequality $(a+b)^{1/2} \leq a^{1/2} + b^{1/2}$ for all $a, b > 0$. Now, it results from (21) that

$$\begin{aligned}J_\rho(\theta^*) - \frac{1}{T}\sum_{t=1}^{T}J_\rho(\theta_t) \leq& \frac{\sqrt{\varepsilon_{\text{bias}}}}{1-\gamma} + \frac{M_h}{T}\left(\Gamma_2 + \frac{\Gamma_3}{B}\right) + \frac{4M_g(m^{1/3}+T^{1/3})\Gamma_1}{kBT}\\ &+ 4M_g\left(1+\frac{1}{\mu_F}\right)^2(\frac{\Gamma_2}{k} + \frac{\Gamma_1+\Gamma_3}{kB})(\frac{m^{1/3}}{T} + \frac{1}{T^{2/3}})\\ &+ \frac{M_g}{\mu_F}\sqrt{\frac{\Gamma_2}{k} + \frac{\Gamma_1+\Gamma_3}{kB}}(\frac{m^{1/6}}{\sqrt{T}} + \frac{1}{T^{1/3}}) + \frac{M_g^2}{\mu_F}\left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\right)\gamma^H\\ &+ \left(4M_g^2(1+\frac{1}{\mu_F})^2 + 4M_g^2\right)\left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\right)\gamma^H\\ =& \frac{\sqrt{\varepsilon_{\text{bias}}}}{1-\gamma} + \tilde{\mathcal{O}}\left(\frac{\sigma}{(1-\gamma)T^{1/3}} + \frac{\sigma}{(1-\gamma)^{2.5}T^{1/2}} + \frac{\sigma^2}{(1-\gamma)^2T^{2/3}} + \frac{\sigma^2}{(1-\gamma)^5T}\right)\\ &+ \frac{M_g^2}{\mu_F}\left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\right)\gamma^H\\ &+ \left(4M_g^2(1+\frac{1}{\mu_F})^2 + 4M_g^2\right)\left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\right)\gamma^H.\end{aligned}$$

By taking

$$H = \mathcal{O}(\log_\gamma((1-\gamma)\epsilon)),$$

and

$$T = \tilde{\mathcal{O}}\left(\frac{\sigma}{(1-\gamma)\epsilon^3} + \frac{\sigma}{(1-\gamma)^{2.5}\epsilon^2} + \frac{\sigma^2}{(1-\gamma)^2\epsilon^{3/2}} + \frac{\sigma^2}{(1-\gamma)^5\epsilon}\right)$$

we obtain that

$$J_\rho(\theta^*) - \frac{1}{T}\sum_{t=1}^{T}J_\rho(\theta_t) \leq \frac{\sqrt{\varepsilon_{\text{bias}}}}{1-\gamma} + \mathcal{O}(\epsilon).$$

This completes the proof.

□

D Proof of results in Section 4

D.1 Proof of Lemmas 4.2, 4.3 and 4.4

First, given any distribution $\pi(\cdot|s)$ over $\mathcal{A}$, let $H(\pi(\cdot|s)) := \text{diag}(\pi(\cdot|s)) - \pi(\cdot|s)\pi(\cdot|s)^\top$, where $\text{diag}(x) \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{A}|}$ is the diagonal matrix whose diagonal entries are the elements of x. Let $X_{i:j}$ be the sub-matrix of X which contains i^{th} row to j^{th} row of the matrix of X. We define $H^-(\pi(\cdot|s)) \in \mathbb{R}^{(|\mathcal{A}|-1)\times|\mathcal{A}|}$ as

$$H^-(\pi(\cdot|s)) = H(\pi(\cdot|s))_{2:|\mathcal{A}|}.$$

It can be shown that for the overparameterized softmax parameterization, where the variables are $\theta_{s,a}, \forall a \in \mathcal{A}, s \in \mathcal{S}$, we have $\left(\frac{d\pi_\theta(\cdot|s)}{d\theta(s,\cdot)}\right)^\top = H(\pi_\theta(\cdot|s))$. Since the restricted softmax parameterization is studied, where $\theta_{s,a_1} = 0$ for all s, we have $\left(\frac{d\pi_\theta(\cdot|s)}{d\theta(s,\cdot)}\right)^\top = H^-(\pi_\theta(\cdot|s))$. We first study the relationship between H^- and H.

Lemma D.1 *For every vector $x \in \mathbb{R}^{|\mathcal{A}|}$, it holds*

$$\|H(\pi(\cdot|s))x\|_2 \leq \sqrt{|\mathcal{A}|}\,\|H^-(\pi(\cdot|s))x\|_2\,,$$

Proof. Let H_i denotes the i^{th} row of the matrix H. Since $\sum_{i=1}^{|\mathcal{A}|} H_i = \mathbf{0}$, one can write

$$(H_1 x)^2 = \left(-\sum_{i=2}^{|\mathcal{A}|} H_i x\right)^2 \leq (|\mathcal{A}|-1)\sum_{i=2}^{|\mathcal{A}|}(H_i x)^2,$$

where the last inequality follows from the Cauchy-Schwarz inequality. Hence,

$$\left(\sum_{i=1}^{|\mathcal{A}|} H_i x\right)^2 \leq |\mathcal{A}|\sum_{i=2}^{|\mathcal{A}|}(H_i x)^2.$$

This completes the proof. □

Now, we study the the gradient of the entropy regularized objective under the restricted soft-max policy parameterization:

Lemma D.2 *For the restricted softmax policy parameterization, the following equations hold true for all $s \in \mathcal{S}$ and $a \in \mathcal{A}/a_1$:*

$$\frac{\alpha\tilde{J}_\mu(\theta)}{\alpha\theta(s,\cdot)} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s) H^-(\pi_\theta(\cdot|s))\left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \lambda\log\pi_\theta(\cdot|s)\right],$$
$$\frac{\alpha\tilde{J}_\mu(\theta)}{\alpha\theta(s,a)} = \frac{1}{1-\gamma}\cdot d_\mu^{\pi_\theta}(s)\cdot\pi_\theta(a|s)\cdot\tilde{A}^{\pi_\theta}(s,a),$$

where $\tilde{A}^{\pi_\theta}(s,a)$ is the soft advantage function defined as

$$\tilde{A}^{\pi_\theta}(s,a) \coloneqq \tilde{Q}^{\pi_\theta}(s,a) - \lambda\log\pi_\theta(a|s) - \tilde{V}^{\pi_\theta}(s),$$
$$\tilde{Q}^{\pi_\theta}(s,a) \coloneqq r(s,a) + \gamma\sum_{s'}\mathcal{P}(s'|s,a)\tilde{V}^{\pi_\theta}(s'),$$
$$\tilde{V}^{\pi_\theta}(s) \coloneqq V^{\pi_\theta}(s) + \lambda\mathbb{H}(s,\theta).$$

Proof. According to the definition of $\tilde{J}_\rho(\theta)$,

$$\tilde{J}_\mu(\theta) = \mathbb{E}_{s\sim\mu}\sum_a \pi_\theta(a|s)\left[\tilde{Q}^{\pi_\theta}(a|s) - \lambda\log\pi_\theta(a|s)\right].$$

Taking derivative with respect to θ yields that

$$\begin{aligned}
\frac{\alpha\tilde{J}_\mu(\theta)}{\alpha\theta} =& \mathbb{E}_{s\sim\mu}\sum_a \frac{\alpha\pi_\theta(a|s)}{\alpha\theta}\left[\tilde{Q}^{\pi_\theta}(s,a) - \lambda\log\pi_\theta(a|s)\right] + \mathbb{E}_{s\sim\mu}\sum_a \pi_\theta(a|s)\left[\frac{\tilde{Q}^{\pi_\theta}(s,a)}{\alpha\theta} - \lambda\frac{1}{\pi_\theta(a|s)}\frac{\alpha\pi_\theta(a|s)}{\alpha\theta}\right]\\
=& \mathbb{E}_{s\sim\mu}\sum_a \frac{\alpha\pi_\theta(a|s)}{\alpha\theta}\left[\tilde{Q}^{\pi_\theta}(s,a) - \lambda\log\pi_\theta(a|s)\right] + \mathbb{E}_{s\sim\mu}\sum_a \pi_\theta(a|s)\frac{\tilde{Q}^{\pi_\theta}(s,a)}{\alpha\theta}\\
=& \mathbb{E}_{s\sim\mu}\sum_a \frac{\alpha\pi_\theta(a|s)}{\alpha\theta}\left[\tilde{Q}^{\pi_\theta}(s,a) - \lambda\log\pi_\theta(a|s)\right] + \lambda\mathbb{E}_{s\sim\mu}\sum_a \pi_\theta(a|s)\sum_{s'}\mathcal{P}(s'|s,a)\frac{\tilde{V}^{\pi_\theta}(s)}{\alpha\theta}\\
=& \frac{1}{1-\gamma}\sum_s d_\mu^{\pi_\theta}(s)\sum_a \frac{\alpha\pi_\theta(a|s)}{\alpha\theta}\left[\tilde{Q}^{\pi_\theta}(s,a) - \lambda\log\pi_\theta(a|s)\right],
\end{aligned}$$

where the second equation is due to

$$\sum_a \pi_\theta(a|s)\frac{1}{\pi_\theta(a|s)}\frac{\alpha\pi_\theta(a|s)}{\alpha\theta} = 0.$$

Using the fact that $\frac{\pi_\theta(a|s)}{\alpha\theta(s',\cdot)} = \mathbf{0}$ for $s \neq s'$, one can write

$$\begin{aligned}
\frac{\alpha \tilde{J}_\mu(\theta)}{\alpha\theta(s,\cdot)} =& \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s) \sum_a \frac{\alpha\pi_\theta(a|s)}{\alpha\theta(s,\cdot)} \left[\tilde{Q}^{\pi_\theta}(s,a) - \lambda \log \pi_\theta(a|s)\right] \\
=& \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s) H^{-}(\pi_\theta(\cdot|s)) \left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \lambda \log \pi_\theta(\cdot|s)\right] \\
=& \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s) H^{-}(\pi_\theta(\cdot|s)) \left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \lambda \log \pi_\theta(\cdot|s)\right] \\
=& \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s) H^{-}(\pi_\theta(\cdot|s)) \left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \lambda \log \pi_\theta(\cdot|s) + \lambda \log \sum_a \exp(\theta(s,a)) \cdot \mathbf{1}\right]
\end{aligned}$$

Since $H^{-}(\pi_\theta(\cdot|s))\mathbf{1} = 0$, we have

$$\frac{\alpha \tilde{J}_\mu(\theta)}{\alpha\theta(s,\cdot)} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s) H^{-}(\pi_\theta(\cdot|s)) \left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \lambda \log \pi_\theta(\cdot|s)\right]$$

It follows from Lemma 10 in [30] that,

$$\frac{\alpha \tilde{J}_\mu(\theta)}{\alpha\theta(s,a)} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot \tilde{A}^{\pi_\theta}(s,a), \quad \forall a \in \mathcal{A}/a_1.$$

□

D.1.1 Proof of Lemma 4.2

Proof. It results from the definition of soft value functions that

$$\begin{aligned}
&\tilde{J}_\rho(\pi_\lambda^*) - \tilde{J}_\rho(\pi_\theta) \\
&= \mathbb{E}_{s_0\sim\rho, a_t\sim\pi_\lambda^*(\cdot|s_t), s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)} \left[\sum_{t=0}^{\infty} \gamma^t (r(s_t,a_t) - \tau \log \pi_\lambda^*(a_t|s_t))\right] - \tilde{J}_\rho(\theta) \\
&= \mathbb{E}_{s_0\sim\rho, a_t\sim\pi_\lambda^*(\cdot|s_t), s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)} \left[\sum_{t=0}^{\infty} \gamma^t (r(s_t,a_t) - \tau \log \pi_\lambda^*(a_t|s_t)) + \tilde{V}^{\pi_\theta}(s_t) - \tilde{V}^{\pi_\theta}(s_t)\right] - \tilde{J}_\rho(\theta) \\
&= \mathbb{E}_{s_0\sim\rho, a_t\sim\pi_\lambda^*(\cdot|s_t), s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)} \left[\sum_{t=0}^{\infty} \gamma^t (r(s_t,a_t) - \tau \log \pi_\lambda^*(a_t|s_t)) + \gamma\tilde{V}^{\pi_\theta}(s_{t+1}) - \tilde{V}^{\pi_\theta}(s_t)\right] \\
&= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\lambda^*}(s) \left[\sum_a \pi_\lambda^*(a|s)(r(s,a) - \tau \log \pi_\lambda^*(a_t|s_t)) + \gamma \sum_{s'} \mathcal{P}(s'|s,a)\tilde{V}^{\pi_\theta}(s') - \tilde{V}^{\pi_\theta}(s_t)\right] \\
&= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\lambda^*}(s) \left[\sum_a \pi_\lambda^*(a|s)(\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\lambda^*(a_t|s_t)) - \tilde{V}^{\pi_\theta}(s_t)\right].
\end{aligned}$$

Next, we define the "soft greedy policy" $\bar{\pi}_\theta(\cdot|s) = \text{softmax}(\tilde{Q}^{\pi_\theta})(s,\cdot)/\lambda$ as follows:

$$\bar{\pi}_\theta(\cdot|s) = \frac{\exp\left(\tilde{Q}^{\pi_\theta}(s,a)/\tau\right)}{\sum_{a'} \exp\left(\tilde{Q}^{\pi_\theta}(s,a')/\tau\right)}, \quad \forall a \in \mathcal{A}, \forall s \in \mathcal{S}.$$

Therefore,

$$\begin{aligned}
\sum_a \pi_\tau^*(a|s)\left[\tilde{Q}^{\pi_\theta}(s,a) - \lambda \log \pi_\lambda^*(a|s)\right] \leq& \max_{\pi(\cdot|s)} \sum_a \pi(a|s)\left[\tilde{Q}^{\pi_\theta}(s,a) - \lambda \log \pi(a|s)\right] \\
=& \sum_a \bar{\pi}_\theta(a|s)\left[\tilde{Q}^{\pi_\theta}(s,a) - \lambda \log \pi(a|s)\right] \\
=& \lambda \log \sum_a \exp\left(\tilde{Q}^{\pi_\theta}(s,a)/\lambda\right).
\end{aligned}$$

In addition, it holds that

$$\begin{aligned}\tilde{V}^{\pi_\theta}(s) &= \sum_a \pi_\theta(a|s)[\tilde{Q}^{\pi_\theta}(s,a) - \lambda \log \pi_\theta(a|s)] \\ &= \sum_a \pi_\theta(a|s)[\tilde{Q}^{\pi_\theta}(s,a) - \lambda \log \bar{\pi}_\theta(a|s) + \lambda \log \bar{\pi}_\theta(a|s) - \lambda \log \pi_\theta(a|s)] \\ &= \sum_a \pi_\theta(a|s)[\tilde{Q}^{\pi_\theta}(s,a) - \lambda \log \bar{\pi}_\theta(a|s)] - \lambda \mathrm{KL}(\pi_\theta(\cdot|s), \bar{\pi}_\theta(\cdot|s)) \\ &= \lambda \log \sum_a \exp\left(\tilde{Q}^{\pi_\theta}(s,a)/\lambda\right) - \lambda \mathrm{KL}(\pi_\theta(\cdot|s), \bar{\pi}_\theta(\cdot|s)).\end{aligned}$$

Combining the above, we have

$$\begin{aligned}&\tilde{J}_\rho(\pi_\lambda^*) - \tilde{J}_\rho(\pi_\theta) \\ &= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\lambda^*} \left[\sum_a \pi_\lambda^*(a|s)[\tilde{Q}^{\pi_\theta}(s,a) - \lambda \log \pi_\lambda^*(a|s)] - \tilde{V}^{\pi_\theta}(s) \right] \\ &\le \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\lambda^*} \left[\lambda \log \sum_a \exp\left(\tilde{Q}^{\pi_\theta}(s,a)/\lambda\right) - \tilde{V}^{\pi_\theta}(s) \right] \\ &= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\lambda^*} \cdot \lambda \cdot \mathrm{KL}(\pi_\theta(\cdot|s), \bar{\pi}_\theta(\cdot|s))\end{aligned}$$

The KL-logit inequality in Lemma 27 of [30] can be leveraged to write

$$\begin{aligned}\tilde{J}_\rho(\pi_\lambda^*) - \tilde{J}_\rho(\pi_\theta) &\le \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\lambda^*} \cdot \frac{\lambda}{2} \cdot \left\| \frac{\tilde{Q}^{\pi_\theta}(s,\cdot)}{\lambda} - \theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot)/\lambda - \theta(s,\cdot))^\top \mathbf{1}}{|\mathcal{A}|} \cdot \mathbf{1} \right\|_\infty^2 \\ &= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\lambda^*} \cdot \frac{1}{2\lambda} \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \lambda\theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot) - \lambda\theta(s,\cdot))^\top \mathbf{1}}{|\mathcal{A}|} \cdot \mathbf{1} \right\|_\infty^2.\end{aligned}$$

Then, by $\|x\|_2^2 \le \|x\|_1^2$, one can write

$$\begin{aligned}&\tilde{J}_\rho(\pi_\lambda^*) - \tilde{J}_\rho(\pi_\theta) \\ &= \frac{1}{1-\gamma} \sum_s \left(\sqrt{d_\rho^{\pi_\lambda^*}} \cdot \sqrt{\frac{1}{2\lambda}} \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \lambda\theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot) - \lambda\theta(s,\cdot))^\top \mathbf{1}}{|\mathcal{A}|} \cdot \mathbf{1} \right\|_\infty \right)^2 \\ &\le \frac{1}{1-\gamma} \left(\sum_s \sqrt{d_\rho^{\pi_\lambda^*}} \cdot \sqrt{\frac{1}{2\lambda}} \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \lambda\theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot) - \lambda\theta(s,\cdot))^\top \mathbf{1}}{|\mathcal{A}|} \cdot \mathbf{1} \right\|_\infty \right)^2 \\ &\le \frac{1}{1-\gamma} \frac{1}{2\lambda} \left\| \frac{d_\rho^{\pi_\lambda^*}}{d_\mu^{\pi_\theta}} \right\|_\infty \left(\sum_s \sqrt{d_\rho^{\pi_\lambda^*}} \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \lambda\theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot) - \lambda\theta(s,\cdot))^\top \mathbf{1}}{|\mathcal{A}|} \cdot \mathbf{1} \right\|_\infty \right)^2. \quad (22)\end{aligned}$$

On the other hand, the entropy regularized policy gradient norm is lower bounded as

$$\left\| \frac{\alpha \tilde{J}_\mu(\theta)}{\alpha\theta} \right\|_2^2 = \sum_{s,a} \left(\frac{\alpha \tilde{J}_\mu(\theta)}{\alpha\theta(s,a)} \right)^2 = \sum_{s,a} \left\| \frac{\alpha \tilde{J}_\mu(\theta)}{\alpha\theta(s,\cdot)} \right\|_2^2 \ge \frac{1}{\sqrt{|\mathcal{S}|}} \sum_s \left\| \frac{\alpha \tilde{J}_\mu(\theta)}{\alpha\theta(s,\cdot)} \right\|_2^2,$$

where the last inequality is due to the Cauchy-Schwarz inequality. In light of Lemma D.2, we have

$$\begin{aligned}\left\| \frac{\alpha \tilde{J}_\mu(\theta)}{\alpha\theta} \right\|_2^2 &\ge \frac{1}{|\mathcal{S}|} \frac{1}{(1-\gamma)^2} \sum_s (d_\mu^{\pi_\theta}(s))^2 \left\| H^-(\pi_\theta(s,\cdot)[\tilde{Q}^{\pi_\theta}(s,\cdot) - \lambda\theta(s,\cdot)] \right\|_2^2 \\ &= \frac{1}{|\mathcal{S}|} \frac{1}{(1-\gamma)^2} \sum_s (d_\mu^{\pi_\theta}(s))^2 \left\| H^-(\pi_\theta(s,\cdot)[\tilde{Q}^{\pi_\theta}(s,\cdot) - \lambda\theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot) - \lambda\theta(s,\cdot))^\top \mathbf{1}}{|\mathcal{A}|} \cdot \mathbf{1}] \right\|_2^2,\end{aligned}$$

where the last equality follows from $H^- \cdot \mathbf{1} = 0$. It can be concluded from Lemma D.1 that

$$\left\| \frac{\alpha \tilde{J}_\mu(\theta)}{\alpha\theta} \right\|_2^2 \ge \frac{1}{|\mathcal{S}|} \frac{1}{(1-\gamma)^2} \sum_s d_\mu^{\pi_\theta}(s) \frac{1}{|\mathcal{A}|} \left\| H(\pi_\theta(s,\cdot)[\tilde{Q}^{\pi_\theta}(s,\cdot) - \lambda\theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot) - \lambda\theta(s,\cdot))^\top \mathbf{1}}{|\mathcal{A}|} \cdot \mathbf{1}] \right\|_2^2.$$

Now, it follows from Lemma 23 of [30] that

$$\begin{aligned}
&\left\|\frac{\alpha\tilde{J}_\mu(\theta)}{\alpha\theta}\right\|_2^2 \\
&\geq \frac{1}{|\mathcal{S}||\mathcal{A}|}\frac{1}{(1-\gamma)^2}\sum_s (d_\mu^{\pi_\theta}(s))^2(\min_a \pi_\theta(a|s))^2\left\|\tilde{Q}^{\pi_\theta}(s,\cdot)-\lambda\theta(s,\cdot)-\frac{(\tilde{Q}^{\pi_\theta}(s,\cdot)-\lambda\theta(s,\cdot))^\top\mathbf{1}}{|\mathcal{A}|}\cdot\mathbf{1}\right\|_2^2 \\
&\geq \frac{1}{|\mathcal{S}||\mathcal{A}|}\frac{1}{(1-\gamma)^2}\sum_s (d_\mu^{\pi_\theta}(s))^2(\min_a \pi_\theta(a|s))^2\left\|\tilde{Q}^{\pi_\theta}(s,\cdot)-\lambda\theta(s,\cdot)-\frac{(\tilde{Q}^{\pi_\theta}(s,\cdot)-\lambda\theta(s,\cdot))^\top\mathbf{1}}{|\mathcal{A}|}\cdot\mathbf{1}\right\|_\infty^2.
\end{aligned}$$

By denoting $\xi_\theta(s)=\tilde{Q}^{\pi_\theta}(s,\cdot)-\lambda\theta(s,\cdot)-\frac{(\tilde{Q}^{\pi_\theta}(s,\cdot)-\lambda\theta(s,\cdot))^\top\mathbf{1}}{|\mathcal{A}|}\cdot\mathbf{1}$, we have

$$\begin{aligned}
&\left\|\frac{\alpha\tilde{J}_\mu(\theta)}{\alpha\theta}\right\|_2^2 \\
&\geq \frac{1}{|\mathcal{S}||\mathcal{A}|}\frac{1}{(1-\gamma)^2}\sum_s (d_\mu^{\pi_\theta}(s))^2(\min_a \pi_\theta(a|s))^2\,\|\xi_\theta(s)\|_\infty^2 \\
&\geq \frac{1}{|\mathcal{S}||\mathcal{A}|}\frac{1}{(1-\gamma)}\min_s d_\mu^{\pi_\theta}(s)(\min_{s,a}\pi_\theta(a|s))^2\cdot 2\lambda\cdot\left\|\frac{d_\rho^{\pi_\lambda^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1}\cdot\left(\frac{1}{1-\gamma}\frac{1}{2\lambda}\left\|\frac{d_\rho^{\pi_\lambda^*}}{d_\mu^{\pi_\theta}}\right\|_\infty\sum_s d_\mu^{\pi_\theta}(s)\cdot\|\xi_\theta(s)\|_\infty\right)^2
\end{aligned}$$

It results from (22) that

$$\begin{aligned}
\left\|\frac{\alpha\tilde{J}_\mu(\theta)}{\alpha\theta}\right\|_2^2 &\geq \frac{1}{|\mathcal{S}||\mathcal{A}|}\frac{1}{(1-\gamma)}\min_s d_\mu^{\pi_\theta}(s)(\min_{s,a}\pi_\theta(a|s))^2\cdot 2\lambda\cdot\left\|\frac{d_\rho^{\pi_\lambda^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1}\cdot\left(\tilde{J}_\rho(\pi_\lambda^*)-\tilde{J}_\rho(\pi_\theta)\right) \\
&\geq \frac{2\lambda}{|\mathcal{S}||\mathcal{A}|}\min_s\mu(s)\cdot\min_{s,a}\pi_\theta(a|s)^2\cdot\left\|\frac{d_\rho^{\pi_\lambda^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1}\cdot\left(\tilde{J}_\rho(\pi_\lambda^*)-\tilde{J}_\rho(\pi_\theta)\right),
\end{aligned}$$

where the last inequality is due to $d_\mu^{\pi_\theta}(s)\geq(1-\gamma)\cdot\mu(s)$. This completes the proof. □

D.1.2 Proof of Lemma 4.3

Next, it can be shown that the action probabilities are uniformly bounded away from zero if the exact PG in Algorithm 4 is used. The proof is similar to Lemma 16 in [30].

Algorithm 4 Policy Gradient Method

1: **Inputs**: Learning rate η, initial input θ_1 and initial distribution μ;
2: **Outputs**: θ_T;
3: **for** $t=1,2,\ldots,T-1$ **do**
4: $\quad\theta_{t+1}=\theta_t+\eta\nabla\tilde{J}_\mu(\theta_t)$
5: **end for**

Proof. The augmented value function $\tilde{J}_\rho(\theta_t)$ is monotonically increasing following gradient update due to smoothness when $\eta\leq\frac{L_{\tilde{J}}}{2}$ where $L_{\tilde{J}}=\frac{(12+8\log|\mathcal{A}|)}{(1-\gamma)^3}$. Therefore,

$$\begin{aligned}
\tilde{J}_\rho(\theta_t) =&\mathbb{E}_{s_0\sim\rho,a_t\sim\pi_{\theta_t}(\cdot|s_t),s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)}\left[\sum_{t=0}^\infty\gamma^t(r(s_t,a_t)-\lambda\log\pi_{\theta_t}(a_t|s_t))\right] \\
=&\frac{1}{1-\gamma}\sum_s d_\rho^{\pi_{\theta_t}}(s)\left[\sum_a\pi_{\theta_t}(a|s)(r(s_t,a_t)-\lambda\log\pi_{\theta_t}(a_t|s_t))\right] \\
\leq&\frac{1}{1-\gamma}\sum_s d_\rho^{\pi_{\theta_t}}(s)(1+\lambda\log|\mathcal{A}|) \\
\leq&\frac{1+\lambda\log|\mathcal{A}|}{1-\gamma},
\end{aligned}\tag{23}$$

where the first inequality is due to $r(s,a) \le 1$ and $-\sum_a \pi_{\theta_t}(a|s) \cdot \log \pi_{\theta_t}(a|s) \le \log|\mathcal{A}|$. According to the monotone convergence theorem, $\tilde{J}_\rho(\theta_t)$ converges to a finite value. Suppose that $\pi_{\theta_t}(a|s) \to \pi_{\theta_\infty}(a|s)$. For each state $s \in \mathcal{S}$, define the sets

$$\mathcal{A}_0(s) = \{a : \pi_{\theta_\infty}(a|s) = 0\},$$
$$\mathcal{A}_+(s) = \{a : \pi_{\theta_\infty}(a|s) > 0\}.$$

Note that $\mathcal{A} = \mathcal{A}_0(s) \cup \mathcal{A}_+(s)$ since $\pi_\infty(a|s) \ge 0, \forall a \in \mathcal{A}$. We prove by contradiction that, under the restricted soft-max parameterization, $\mathcal{A}_0(s)$ is empty for any state $s \in \mathcal{S}$. Suppose that there exists a state $s \in \mathcal{S}$, such that $\mathcal{A}_0(s)$ is non-empty. For every $a' \in \mathcal{A}_0(s)$, we have $\pi_{\theta_t}(a_0|s) \to \pi_{\theta_\infty}(a'|s)$, which implies that $-\log \pi_{\theta_t}(a'|s) \to \infty$. There exists $t_0 \ge 1$ such that,

$$-\log \pi_{\theta_t}(a'|s) \ge \frac{1 + \lambda \log|\mathcal{A}|}{\lambda(1-\gamma)}, \quad \forall t \ge t_0.$$

According to Lemma D.2, if $a' \in \mathcal{A}/a_1$, the following holds for all $t \ge t_0$,

$$\begin{aligned}
\frac{\alpha \tilde{J}_\mu(\theta_t)}{\alpha \theta_t(s,a')} =& \frac{1}{1-\gamma} \cdot d_\mu^{\pi_{\theta_t}}(s) \cdot \pi_{\theta_t}(a'|s) \cdot \tilde{A}^{\pi_{\theta_t}}(s,a') \\
=& \frac{1}{1-\gamma} \cdot d_\mu^{\pi_{\theta_t}}(s) \cdot \pi_{\theta_t}(a'|s) \cdot [\tilde{Q}^{\pi_{\theta_t}}(s,a') - \lambda \log \pi_{\theta_t}(a_0|s) - \tilde{V}^{\pi_{\theta_t}}(s)] \\
=& \frac{1}{1-\gamma} \cdot d_\mu^{\pi_{\theta_t}}(s) \cdot \pi_{\theta_t}(a'|s) \cdot \left[0 - \lambda \log \pi_{\theta_t}(a_0|s) - \frac{1 + \lambda \log|\mathcal{A}|}{1-\gamma}\right] \\
=& \frac{1}{1-\gamma} \cdot d_\mu^{\pi_{\theta_t}}(s) \cdot \pi_{\theta_t}(a'|s) \cdot \left[0 + \frac{1 + \lambda \log|\mathcal{A}|}{(1-\gamma)} - \frac{1 + \lambda \log|\mathcal{A}|}{1-\gamma}\right] = 0,
\end{aligned}$$

where the first inequality is due to $\tilde{Q}^{\pi_{\theta_t}}(s,a') = r(s,a') + \gamma \sum_{s'} \mathcal{P}(s'|s,a') \tilde{V}^{\pi_{\theta_t}}(s') \ge 0$. If $a' = a_1$, we simply have $\frac{\alpha \tilde{J}_\mu(\theta_t)}{\alpha \theta_t(s,a')} = 0$. Thus, $\theta_t(s,a')$ is non-decreasing for all $t \ge t_0$, which in turn implies that $\theta_\infty(s,a')$ is lower bounded by a constant c, i.e., $\theta_\infty(s,a') \ge c$, and thus $\exp(\theta_\infty(a'|s)) \ge \exp(c) > 0$. Since

$$\pi_{\theta_\infty}(a'|s) = \frac{\exp(\theta_\infty(a'|s))}{\sum_a \exp(\theta_\infty(a|s))} = 0,$$

we have $\sum_a \exp(\theta_\infty(a|s)) = \infty$. On the other hand, for every $a_+ \in \mathcal{A}_+(s)$, it follows from

$$\pi_{\theta_\infty}(a_+|s) = \frac{\exp(\theta_\infty(a_+|s))}{\sum_a \exp(\theta_\infty)(a|s)} > 0$$

that

$$\exp(\theta_\infty(a_+|s)) = \infty, \forall a_+ \in \mathcal{A}_+(s),$$

which further implies that

$$\sum_{a_+ \in \mathcal{A}_+(s)} \theta_\infty(a_+|s) = \infty \tag{24}$$

and $a_0 \notin \mathcal{A}_+(s)$. Note that

$$\sum_a \frac{\alpha \tilde{J}_\mu(\theta_t)}{\alpha \theta_t(s,a)} = \sum_{a' \in \mathcal{A}_0(s)} \frac{\alpha \tilde{J}_\mu(\theta_t)}{\alpha \theta_t(s,a')} + \sum_{a_+ \in \mathcal{A}_+(s)} \frac{\alpha \tilde{J}_\mu(\theta_t)}{\alpha \theta_t(s,a_+)} = 0.$$

Since $\sum_{a' \in \mathcal{A}_0(s)} \frac{\alpha \tilde{J}_\mu(\theta_t)}{\alpha \theta_t(s,a')} \ge 0$ for every $t \ge t_0$, it can be concluded that

$$\sum_{a_+ \in \mathcal{A}_+(s)} \frac{\alpha \tilde{J}_\mu(\theta_t)}{\alpha \theta_t(s,a_+)} \le 0, \quad \forall t \ge t_0$$

which implies that $\sum_{a_+ \in \mathcal{A}_+(s)} \theta_\infty(a_+|s)$ will not increase when $t \ge t_0$. This contradicts (24).

So far, we have shown that $\mathcal{A}_0(s) = \varnothing$ for all state $s \in \mathcal{S}$, i.e., $\pi_{\theta_t}(\cdot|s)$ will converge to the interior of the probabilistic simplex $\Delta(\mathcal{A})$. Since $\pi_{\theta_t} \to \pi_{\theta_\infty}(a|s)$, there exists $t_0 \ge 1$ such that

$$0.9 \cdot \pi_{\theta_\infty}(a|s) \le \pi_{\theta_t}(a|s) \le 1.1 \cdot \pi_{\theta_\infty}(a|s), \forall s \in \mathcal{S}, \quad \forall a \in \mathcal{A}, \quad \forall t \ge t_0.$$

This implies that

$$\inf_{t \ge 1} \min_{s,a} \pi_{\theta_t}(a|s) = \min\left\{\min_{1 \le t \le t_0} \min_{s,a} \pi_{\theta_t}(a|s), \inf_{t \ge t_0} \min_{s,a} \pi_{\theta_t}(a|s)\right\} > 0.$$

□

D.1.3 Proof of Lemma 4.4

Proof. It follows from Lemma 4.3 that $\pi_{\bar{\theta}_t}$ generated by the exact PG in Algorithm 4 will converge to an interior point of the probabilistic simplex $\Delta(\mathcal{A})$ and thus $\inf_{t\geq 1}\min_{s,a}\pi_{\bar{\theta}_t}(a|s) > 0$. Then, by the gradient domination condition in Lemma 4.2, we know that $\pi_{\bar{\theta}_t}$ will converge to π_{θ^*}, where $\theta^* \in \operatorname{argmax}_\theta \tilde{J}_\rho(\theta)$. Hence, $\min_{s,a}\pi_{\theta^*}(a|s) > 0$.

We now prove by contradiction that θ^* under the restricted soft-max parameterization must be bounded. Suppose that there exists an action $a' \in \mathcal{A}/a_1$ such that $\theta^*_{s,a'}$ is unbounded. We study two cases:

- Case 1: $\theta^*_{s,a'} = +\infty$. One can write

$$\pi_{\theta^*}(a_1|s) = \frac{\exp(\theta^*(s,a_1))}{\sum_a \exp(\theta^*(s,a))} \leq \frac{\exp(\theta^*(s,a_1))}{\exp(\theta^*(s,a'))} = 0.$$

  This contradicts the inequality $\min_{s,a}\pi_{\theta^*}(a|s) > 0$.
- Case 2: $\theta^*_{s,a'} = -\infty$. One can write

$$\pi_{\theta^*}(a'|s) = \frac{\exp(\theta^*(s,a'))}{\sum_a \exp(\theta^*(s,a))} \leq \frac{\exp(\theta^*(s,a'))}{\exp(\theta^*(s,a_1))} = 0.$$

  This contradicts the inequality $\min_{s,a}\pi_{\theta^*}(a|s) > 0$.

Thus, θ^* must be bounded. In addition, the boundedness of $\{\bar{\theta}_t\}_{t=1}^T$ generated by the exact PG in Algorithm 4 can be established similarly. Therefore, there must exist a bounded constant $\bar{\Delta}$ such that $\left\|\bar{\theta}_t - \theta^*\right\| \leq \bar{\Delta}$ for all $t = 1, \ldots, T$. □

D.2 Proof of Lemma 4.5

To show the relationship between the probability of having a large optimality gap and the PG estimation error, we first define the stopping time $\tau \coloneqq \min\{t| \left\|\theta_t - \theta^*\right\|_2 > 20\bar{\Delta}\}$, where $\bar{\Delta}$ is defined in Lemma 4.4. In other words, the time τ is the index of the first iterate that exits the bounded region $\mathcal{G}^0$.

Lemma D.3 *Let $e_t = \nabla\tilde{J}_\mu(\theta_t) - u_t$ where u_t is an unbiased estimator of $\nabla\tilde{J}_\mu(\theta_t)$ and $\eta \leq \frac{(3+2\log|\mathcal{A}|)}{3(1-\gamma)^3}$. We have*

$$\mathbb{E}[\delta_T \mathbf{1}_{\tau>T}] \leq \left(1 - \frac{\eta C^0}{8}\right)^{T-1}\delta_1 + \sum_{i=1}^{T-1}\frac{5\eta}{8}\mathbb{E}\left[\left\|e_i\right\|_2^2 \mathbf{1}_{\theta_i\in\mathcal{G}^0}\right].$$

Proof. Let $\mathcal{F}_t$ denote the sigma field generated by the randomness up to iteration t. We define $\mathbb{E}^t \coloneqq \mathbb{E}[\cdot|\mathcal{F}_t]$ as the expectation operator conditioned on the sigma field $\mathcal{F}_t$. Since $\tilde{J}_\mu(\theta)$ is $\frac{(12+8\log|\mathcal{A}|)}{(1-\gamma)^3}$-smooth from Lemma F.3. Thus, from Lemma E.4, for every $\eta \leq \frac{(3+2\log|\mathcal{A}|)}{3(1-\gamma)^3}$, we have

$$\begin{aligned}
\mathbb{E}^t[\delta_{t+1} - \delta_t] =& \mathbb{E}^t\left[\tilde{J}_\mu(\theta_t) - \tilde{J}_\mu(\theta_{t+1})\right] \\
\leq& \mathbb{E}^t\left[-\frac{\eta}{8}\left\|u_t\right\|_2^2 + \frac{3\eta}{4}\left\|e_t\right\|_2^2\right] \\
\leq& \mathbb{E}^t\left[-\frac{\eta}{8}\left\|u_t - \nabla\tilde{J}_\mu(\theta_t) + \nabla\tilde{J}_\mu(\theta_t)\right\|_2^2 + \frac{3\eta}{4}\left\|e_t\right\|_2^2\right] \\
=& \mathbb{E}^t\left[-\frac{\eta}{8}\left\|u_t - \nabla\tilde{J}_\mu(\theta_t)\right\|_2^2 - \frac{\eta}{8}\left\|\nabla\tilde{J}_\mu(\theta_t)\right\|_2^2 + \frac{3\eta}{4}\left\|e_t\right\|_2^2\right] \\
=& \mathbb{E}^t\left[-\frac{\eta}{8}\left\|\nabla\tilde{J}_\mu(\theta_t)\right\|_2^2 + \frac{5\eta}{8}\left\|e_t\right\|_2^2\right] \\
\leq& \mathbb{E}^t\left[-\frac{\eta C(\theta_t)}{8}\delta_t + \frac{5\eta}{8}\left\|e_t\right\|_2^2\right],
\end{aligned}$$

where the second inequality uses the fact that u_t is an unbiased estimator of $\nabla \tilde{J}_\mu(\theta_t)$ and the last inequality is due to Lemma 4.2. Since $\theta_t \in \mathcal{G}^0$, we have $C(\theta_t) \geq C^0$. Therefore,

$$\mathbb{E}^t[\delta_{t+1}] \leq \left(1 - \frac{\eta C^0}{8}\right)\delta_t + \frac{5\eta}{8}\mathbb{E}^t\left[\|e_t\|_2^2\right].$$

In addition, conditioning on this $\mathcal{F}_t$ yields that

$$\mathbb{E}[\delta_{t+1}\mathbf{1}_{\tau>t+1}|\mathcal{F}_t] \leq \mathbb{E}[\delta_{t+1}\mathbf{1}_{\tau>t}|\mathcal{F}_t] = \mathbb{E}[\delta_{t+1}|\mathcal{F}_t]\mathbf{1}_{\tau>t},$$

where the last equality uses the fact that τ is a stopping time and the random variable $\mathbf{1}_{\tau>t}$ is determined completely by the sigma-field $\mathcal{F}_t$.

We now consider two cases.

- Case 1: Assume that $\tau > t$, which implies that $x_t \in \mathcal{G}^0$. Then,

$$\mathbb{E}[\delta_{t+1}|\mathcal{F}_t] \leq \left(1 - \frac{\eta C^0}{8}\right)\delta_t + \frac{5\eta}{8}\mathbb{E}\left[\|e_t\|_2^2|\mathcal{F}_t\right].$$

- Case 2: Assume that $\tau \leq t$, leading to

$$\mathbb{E}[\delta_{t+1}|\mathcal{F}_t]\mathbf{1}_{\tau>t} = 0.$$

Now combining the above two cases yields the inequality

$$\begin{aligned}\mathbb{E}[\delta_{t+1}|\mathcal{F}_t]\mathbf{1}_{\tau>t} &\leq \left\{(1 - \frac{\eta C^0}{8})\delta_t + \frac{5\eta}{8}\mathbb{E}\left[\|e_t\|_2^2|\mathcal{F}_t\right]\right\}\mathbf{1}_{\tau>t} \\ &\leq \left(1 - \frac{\eta C^0}{8}\right)\delta_t\mathbf{1}_{\tau>t} + \frac{5\eta}{8}\mathbb{E}\left[\|e_t\|_2^2\mathbf{1}_{\theta_t\in\mathcal{G}^0}|\mathcal{F}_t\right].\end{aligned}$$

Taking the expectations over the sigma-field $\mathcal{F}_t$ and then arguing inductively yields that

$$\begin{aligned}\mathbb{E}[\delta_{t+1}\mathbf{1}_{\tau>t+1}] &\leq \prod_{i=0}^{t}\left(1 - \frac{\eta C^0}{8}\right)\delta_1 + \sum_{i=0}^{t}\left(1 - \frac{\eta C^0}{8}\right)^i\frac{5\eta}{8}\mathbb{E}\left[\|e_i\|_2^2\mathbf{1}_{\theta_i\in\mathcal{G}^0}\right] \\ &\leq \left(1 - \frac{\eta C^0}{8}\right)^t\delta_1 + \sum_{i=1}^{t}\frac{5\eta}{8}\mathbb{E}\left[\|e_i\|_2^2\mathbf{1}_{\theta_i\in\mathcal{G}^0}\right].\end{aligned}$$

By setting $t + 1 = T$, we obtain that

$$\mathbb{E}[\delta_T\mathbf{1}_{\tau>T}] \leq \left(1 - \frac{\eta C^0}{8}\right)^{T-1}\delta_1 + \sum_{i=1}^{T-1}\frac{5\eta}{8}\mathbb{E}\left[\|e_i\|_2^2\mathbf{1}_{\theta_i\in\mathcal{G}^0}\right].$$

This completes the proof. □

Lemma D.4 *By letting $e_t = \nabla \tilde{J}_\mu(\theta_t) - u_t$ where u_t is an unbiased estimator of $\nabla \tilde{J}_\mu(\theta_t)$, it holds that*

$$\mathbb{P}(\tau \leq T) \leq \frac{(1+\eta L)^{T-1}\eta\sum_{i=0}^{T-1}\mathbb{E}[\|e_i\|_2\mathbf{1}_{\theta_i\in\mathcal{G}^0}] + \bar{\Delta}}{20\bar{\Delta}}.$$

Proof. We first define $\Delta_t = \|\theta_t - \theta^*\|_2$. By the triangle inequality and the fact that the iterations of the algorithm with the exact policy gradient are bounded by $\bar{\Delta}$, we have

$$\Delta_t \leq \left\|\theta_t - \bar{\theta}_t\right\|_2 + \left\|\theta^* - \bar{\theta}_t\right\|_2\left\|\theta_t - \bar{\theta}_t\right\|_2 + \bar{\Delta}.$$

Using the update rule of the algorithm with the exact PG $\nabla \tilde{J}_\mu(\bar{\theta}_i)$ and the stochastic PG $u_t(\theta_i)$, one can write

$$\begin{aligned}
\Delta_t &= \left\| (\theta_1 + \sum_{i=1}^{t-1} \eta u_t(\theta_i)) - (\theta_1 + \sum_{i=1}^{t-1} \eta \nabla \tilde{J}_\mu(\bar{\theta}_i)) \right\|_2 + \bar{\Delta} \\
&\leq \sum_{i=1}^{t-1} \eta \left\| u_t(\theta_i) - \nabla \tilde{J}_\mu(\bar{\theta}_i) \right\|_2 + \bar{\Delta} \\
&= \sum_{i=1}^{t-1} \eta \left\| u_t(\theta_i) - \nabla \tilde{J}_\mu(\theta_i) + \nabla \tilde{J}_\mu(\theta_i) - \nabla \tilde{J}_\mu(\bar{\theta}_i) \right\|_2 + \bar{\Delta} \\
&\leq \sum_{i=1}^{t-1} \eta \left\| e_i \right\|_2 + \sum_{i=1}^{t-1} \eta L \left\| \theta_i - \bar{\theta}_i \right\|_2 + \bar{\Delta}.
\end{aligned}$$

By expanding $\left\| \theta_i - \bar{\theta}_i \right\|_2$ recursively, it can be concluded that

$$\begin{aligned}
\Delta_t \leq & \sum_{i=1}^{t-1} \eta \left\| e_i \right\|_2 + \eta L \left\| \theta_{t-1} - \bar{\theta}_{t-1} \right\|_2 + \sum_{i=1}^{t-2} \eta L \left\| \theta_i - \bar{\theta}_i \right\|_2 + \bar{\Delta} \\
\leq & \eta \sum_{i=1}^{t-1} \left\| e_i \right\|_2 + \eta^2 L \sum_{i=1}^{t-2} \left\| e_i \right\|_2 + (\eta L)^2 \sum_{i=1}^{t-2} \left\| \theta_i - \bar{\theta}_i \right\|_2 + \eta L \sum_{i=1}^{t-2} \left\| \theta_i - \bar{\theta}_i \right\|_2 + \bar{\Delta} \\
= & \eta \sum_{i=1}^{t-1} \left\| e_i \right\|_2 + \eta^2 L \sum_{i=1}^{t-2} \left\| e_i \right\|_2 + \left(\eta L + (\eta L)^2 \right) \sum_{i=1}^{t-2} \left\| \theta_i - \bar{\theta}_i \right\|_2 + \bar{\Delta} \\
= & \eta \sum_{i=1}^{t-1} \left\| e_i \right\|_2 + \eta^2 L \sum_{i=1}^{t-2} \left\| e_i \right\|_2 + \eta \left(\eta L + (\eta L)^2 \right) \sum_{i=1}^{t-3} \left\| e_i \right\|_2 + \left(\eta L + 2(\eta L)^2 + (\eta L)^3 \right) \sum_{i=1}^{t-3} \left\| \theta_i - \bar{\theta}_i \right\|_2 + \bar{\Delta} \\
= & \eta \sum_{i=1}^{t-1} \left((1 + \eta L)^{t-1-i} \left\| e_i \right\|_2 \right) + \bar{\Delta} \\
\leq & \eta (1 + \eta L)^{t-1} \sum_{i=1}^{t-1} \left\| e_i \right\|_2 + \bar{\Delta}.
\end{aligned}$$

Define $\tilde{\Delta}_t := \bar{\Delta} + \eta \sum_{i=0}^{t-1} (1+\eta L)^{T-1-i} \left\| e_i \right\|_2$. With the stopping time τ, we define the stopped process as

$$Y_t := \tilde{\Delta}_{\tau \wedge t} + \eta \sum_{i=t}^{T-1} (1 + \eta L)^{T-1-i} \mathbb{E}[\left\| \tilde{e}_i \right\|_2 \mathbf{1}_{\theta_i \in \mathcal{G}^0} | \mathcal{F}_t].$$

By construction, each random variable Y_t is non-negative and almost surely bounded by the bounded variance of $\tilde{e}_i$ for $\theta_i \in \mathcal{G}^0$.

We claim that $\{Y_t\}_{t=1}^T$ is a martingale. In order to prove this claim, we first write

$$\begin{aligned}
\mathbb{E}[Y_{t+1} | \mathcal{F}_t] = & \mathbb{E}[\tilde{\Delta}_{\tau \wedge (t+1)} \mathbf{1}_{\tau \leq t} | \mathcal{F}_t] + \mathbb{E}[\tilde{\Delta}_{\tau \wedge (t+1)} \mathbf{1}_{\tau > t} | \mathcal{F}_t] \\
& + \eta \sum_{i=t+1}^{T-1} (1 + \eta L)^{T-1-i} \mathbb{E}[\mathbb{E}[\left\| \tilde{e}_i \right\|_2 \mathbf{1}_{\theta_i \in \mathcal{G}^0} | \mathcal{F}_{t+1}] | \mathcal{F}_t] \\
= & \mathbb{E}[\tilde{\Delta}_{\tau \wedge (t+1)} \mathbf{1}_{\tau \leq t} | \mathcal{F}_t] + \mathbb{E}[\tilde{\Delta}_{\tau \wedge (t+1)} \mathbf{1}_{\tau > t} | \mathcal{F}_t] \\
& + \eta \sum_{i=t+1}^{T-1} (1 + \eta L)^{T-1-i} \mathbb{E}[\left\| \tilde{e}_i \right\|_2 \mathbf{1}_{\theta_i \in \mathcal{G}^0} | \mathcal{F}_t].
\end{aligned} \tag{25}$$

The first term on the right-hand side can be written as

$$\mathbb{E}[\tilde{\Delta}_{\tau \wedge (t+1)} \mathbf{1}_{\tau \leq t} | \mathcal{F}_t] = \mathbb{E} \tilde{\Delta}_{\tau \wedge t} \mathbf{1}_{\tau \leq t} | \mathcal{F}_t] = \tilde{\Delta}_{\tau \wedge t} \mathbf{1}_{\tau \leq t}. \tag{26}$$

As for the second term, it holds that

$$\begin{aligned}
\mathbb{E}[\tilde{\Delta}_{\tau\wedge(t+1)}\mathbf{1}_{\tau>t}|\mathcal{F}_t] =&\mathbb{E}[\tilde{\Delta}_{t+1}|\mathcal{F}_t]\mathbf{1}_{\tau>t}\\
=&\mathbb{E}[\bar{\Delta}_t+\eta\sum_{i=0}^{t-1}(1+\eta L)^{T-1-i}\left\|e_i\right\|_2+\eta(1+\eta L)^{T-1-t}\left\|e_t\right\|_2|\mathcal{F}_t]\mathbf{1}_{\tau>t}\\
=&\mathbb{E}[\bar{\Delta}_t+\eta\sum_{i=0}^{t-1}(1+\eta L)^{T-1-i}\left\|e_i\right\|_2|\mathcal{F}_t]\mathbf{1}_{\tau>t}+\eta(1+\eta L)^{T-1-t}\mathbb{E}[\left\|e_t\right\|_2\mathbf{1}_{\theta_t\in\mathcal{G}^0}|\mathcal{F}_t]\\
=&\tilde{\Delta}_t\mathbf{1}_{\tau>t}+\eta(1+\eta L)^{T-1-t}\mathbb{E}[\left\|e_t\right\|_2\mathbf{1}_{\theta_t\in\mathcal{G}^0}|\mathcal{F}_t]\\
=&\tilde{\Delta}_{\{t\wedge\tau\}}\mathbf{1}_{\tau>t}+\eta(1+\eta L)^{T-1-t}\mathbb{E}[\left\|e_t\right\|_2\mathbf{1}_{\theta_t\in\mathcal{G}^0}|\mathcal{F}_t]. \qquad (27)
\end{aligned}$$

Substituting the bounds (26) and (27) into (25), we find that

$$\begin{aligned}
\mathbb{E}[Y_{t+1}|\mathcal{F}_t] =&\tilde{\Delta}_{\{t\wedge\tau\}}+\eta(1+\eta L)^{T-1-t}\mathbb{E}[\left\|e_t\right\|_2\mathbf{1}_{\theta_t\in\mathcal{G}^0}|\mathcal{F}_t]+\eta\sum_{i=t+1}^{T-1}(1+\eta L)^{T-1-i}\mathbb{E}[\left\|\tilde{e}_i\right\|_2\mathbf{1}_{\theta_i\in\mathcal{G}^0}|\mathcal{F}_t]\\
=&\tilde{\Delta}_{\{t\wedge\tau\}}+\eta\sum_{i=t}^{T-1}(1+\eta L)^{T-1-i}\mathbb{E}[\left\|\tilde{e}_i\right\|_2\mathbf{1}_{\theta_i\in\mathcal{G}^0}|\mathcal{F}_t]\\
=&Y_t.
\end{aligned}$$

We have thus verified the martingale property. Hence, Y_t is also super-martingale. Finally, applying Doob's maximal inequality for super-martingales yields that

$$\begin{aligned}
\mathbb{P}(\tau\le T) =&\mathbb{P}(\max_{t\in\{1,\dots,T\}}\Delta_t\ge 20\bar{\Delta})\\
\le&\frac{\mathbb{E}[Y_1]}{20\bar{\Delta}}\\
\le&\frac{\eta\sum_{i=1}^{T-1}(1+\eta L)^{T-1-i}\mathbb{E}[\left\|e_i\right\|_2\mathbf{1}_{\theta_i\in\mathcal{G}^0}]+\bar{\Delta}}{20\bar{\Delta}}\\
\le&\frac{(1+\eta L)^{T-1}\eta\sum_{i=1}^{T-1}\mathbb{E}[\left\|e_i\right\|_2\mathbf{1}_{\theta_i\in\mathcal{G}^0}]+\bar{\Delta}}{20\bar{\Delta}}.
\end{aligned}$$

This completes the proof. □

D.2.1 Proof of Lemma 4.5

Proof. By combining Lemma D.3 and D.4, we obtain that

$$\begin{aligned}
\mathbb{P}(\delta_T\ge\epsilon)\le&\mathbb{P}(\mathbf{1}_{\tau>T}\delta_T\ge\epsilon)+\mathbb{P}(\mathbf{1}_{\tau\le T})\\
\le&\frac{\mathbb{E}[\mathbf{1}_{\tau>T}\delta_T]}{\epsilon}+\mathbb{P}(\tau\le T)\\
\le&\left(1-\frac{\eta C^0}{8}\right)^{T-1}\frac{\delta_1}{\epsilon}+\sum_{i=1}^{T}\frac{5\eta}{8\epsilon}\mathbb{E}\left[\left\|e_i\right\|_2^2\mathbf{1}_{\theta_i\in\mathcal{G}^0}\right]+\frac{(1+\eta L)^{T-1}\mathbb{E}[\sum_{i=1}^{T}\eta\left\|e_i\right\|_2\mathbf{1}_{\theta_i\in\mathcal{G}^0}]+\bar{\Delta}}{20\bar{\Delta}}\\
\le&\left(1-\frac{\eta C^0}{8}\right)^{\frac{8}{\eta C^0}\frac{\eta C^0 T}{8}}\frac{\delta_1}{\epsilon}+\sum_{i=1}^{T}\frac{5\eta}{8\epsilon}\mathbb{E}\left[\left\|e_i\right\|_2^2\mathbf{1}_{\theta_i\in\mathcal{G}^0}\right]+\frac{(1+\eta L)^{T-1}\mathbb{E}[\sum_{i=1}^{T}\eta\left\|e_i\right\|_2\mathbf{1}_{\theta_i\in\mathcal{G}^0}]+\bar{\Delta}}{20\bar{\Delta}}\\
\le&\frac{1}{2}^{\frac{\eta C^0 T}{8}}\frac{\delta_1}{\epsilon}+\sum_{i=1}^{T}\frac{5\eta}{8\epsilon}\mathbb{E}\left[\left\|e_i\right\|_2^2\mathbf{1}_{\theta_i\in\mathcal{G}^0}\right]+\frac{(1+\eta L)^{T-1}\mathbb{E}[\sum_{i=1}^{T}\eta\left\|e_i\right\|_2\mathbf{1}_{\theta_i\in\mathcal{G}^0}]+\bar{\Delta}}{20\bar{\Delta}},
\end{aligned}$$

where the second inequality holds due to Markov inequality, the last inequality holds because of $(1-\frac{1}{m})^m\le\frac{1}{2}$ for all $m\ge 1$ and $\frac{8}{\eta C^0}\ge 1$. □

D.3 Proof of Lemma 4.6

We first define the following notions:

$$\nabla J_1^H(\theta) := \mathbb{E}_{\tau\sim p(\tau|\theta,\mu)}\left[\left(\sum_{t=0}^{H-1}\nabla_\theta \log \pi_\theta(a_t|s_t)\right)\left(\sum_{t=0}^{H-1}\gamma^t\left(r(s_t,a_t)-\log\pi_\theta(a_t|s_t)\right)\right)\right],$$

$$\nabla J_1(\theta) := \mathbb{E}_{\tau\sim p(\tau|\theta,\mu)}\left[\left(\sum_{t=0}^{\infty}\nabla_\theta \log \pi_\theta(a_t|s_t)\right)\left(\sum_{t=0}^{\infty}\gamma^t\left(r(s_t,a_t)-\log\pi_\theta(a_t|s_t)\right)\right)\right],$$

$$\nabla J_2^H(\theta) := \mathbb{E}_{\tau\sim p(\tau|\theta,\mu)}\left[\sum_{h=0}^{\infty}-\gamma^t\nabla\log\pi_\theta(a_h^i,s_h^i)\right],$$

$$\nabla J_2^H(\theta) := \mathbb{E}_{\tau\sim p(\tau|\theta,\mu)}\left[\sum_{h=0}^{H-1}-\gamma^t\nabla\log\pi_\theta(a_h^i,s_h^i)\right],$$

$$\nabla \tilde{J}_\mu^H(\theta) := \nabla J_1^H(\theta) + \nabla J_2^H(\theta)$$

Next, we prove the following properties of the trajectory-based PG estimator for the entropy regularized objective.

Lemma D.5 *Given $\theta \in \mathcal{G}^0$ and denote $R := \max_{\theta\in\mathcal{G}^0} -\log\pi_\theta(a,s)$. For the truncated policy gradient estimator given in (11), the following properties hold:*

1. *$g(\tau_i^H|\theta,\mu)$ is Lipschitz continuous with the Lipschitz constant $M_h(1+R)/(1-\gamma)^2+M_h/(1-\gamma)$.*
2. *$\left\|g(\tau_i^H|\theta,\mu)\right\|_2 \le M_g(1+R)/(1-\gamma)^2 + M_g/(1-\gamma)$.*
3. *If the infinite-sum is well-defined, then*

$$\sum_{h=0}^{\infty}\sum_{j=h}^{\infty}\nabla\log\pi_\theta(a_h^i,s_h^i)\left(\gamma^j(r_j(s_j^i,a_j^i)-\log\pi_\theta(a_j^i|s_j^i))\right)+\sum_{h=0}^{\infty}-\gamma^h\nabla\log\pi_\theta(a_h^i|s_h^i) \quad (28)$$

   is an unbiased estimate of $\nabla\tilde{J}_\mu(\theta)$. Similarly, the truncated PGT estimate $g(\tau_i^H|\theta,\mu)$ given in (11) is an unbiased estimate of $\nabla\tilde{J}_\mu^H(\theta)$.
4. *$\left\|\nabla\tilde{J}_\mu^H(\theta)-\nabla\tilde{J}_\mu(\theta)\right\|_2 \le M_g(1+R)\left(\frac{H+1}{1-\gamma}+\frac{\gamma}{(1-\gamma)^2}\right)\gamma^H+\frac{M_g\gamma^H}{1-\gamma}$.*
5. *$\max\{\left\|\nabla\tilde{J}_\mu(\theta)\right\|_2, \left\|\nabla\tilde{J}_\mu^H(\theta)\right\|_2\} \le \frac{M_g(1+R)}{(1-\gamma)^2}+\frac{M_g}{1-\gamma}$.*

Proof. Regarding $g_1(\tau_i^H|\theta,\mu)$, it results from Lemma E.1 that

- $g_1(\tau_i^H|\theta,\mu)$ is Lipschitz continuous with Lipschitz constant $M_h(1+R)/(1-\gamma)^2$.
- $\left\|g_1(\tau_i^H|\theta,\mu)\right\|_2 \le M_g(1+R)/(1-\gamma)^2$.
- If the infinite-sum is well-defined, then

$$g_1(\tau_i^H|\theta,\mu) = \sum_{h=0}^{\infty}\sum_{j=h}^{\infty}\nabla\log\pi_\theta(a_h^i,s_h^i)\left(\gamma^j(r_j(s_j^i,a_j^i)-\log\pi_\theta(a_j^i|s_j^i))\right).$$

  is an unbiased estimate of $\nabla J_2(\theta)$. Similarly, the truncated estimator $g_1(\tau_i^H|\theta,\mu)$ given by

$$g_1(\tau_i^H|\theta,\mu) = \sum_{h=0}^{H-1}\sum_{j=h}^{H-1}\nabla\log\pi_\theta(a_h^i,s_h^i)\left(\gamma^j(r_j(s_j^i,a_j^i)-\log\pi_\theta(a_j^i|s_j^i))\right).$$

  is an unbiased estimate of $\nabla J_1^H(\theta)$.
- $\left\|\nabla J_1^H(\theta)-\nabla J_1(\theta)\right\|_2 \le M_g(1+R)\left(\frac{H+1}{1-\gamma}+\frac{\gamma}{(1-\gamma)^2}\right)\gamma^H$.
- $\max\{\left\|\nabla J_1(\theta)\right\|_2, \left\|\nabla J_1^H(\theta)\right\|_2\} \le \frac{M_g(1+R)}{(1-\gamma)^2}$.

On the other hand, the following properties are satisfied for $g_2(\tau_i^H|\theta,\mu)$,

- $g_2(\tau_i^H|\theta,\mu)$ is Lipschitz continuous with Lipschitz constant $M_h/(1-\gamma)$.
- $\left\|g_2(\tau_i^H|\theta,\mu)\right\|_2 \le M_g/(1-\gamma)$.
- If the infinite-sum is well-defined, then

$$g_2(\tau_i|\theta,\mu) = \sum_{h=0}^{\infty} -\gamma^t \nabla \log \pi_\theta(a_h^i, s_h^i)$$

  is an unbiased estimate of $\nabla J_2(\theta)$. Similarly, the truncated PGT estimate $g_2(\tau_i^H|\theta,\mu)$ given by (10) is an unbiased estimate of $\nabla J_2^H(\theta)$.
- $\left\|\nabla J_2^H(\theta) - \nabla J_2(\theta)\right\|_2 \le \frac{M_g \gamma^H}{1-\gamma}$.
- $\max\{\|\nabla J_2(\theta)\|_2, \left\|\nabla J_2^H(\theta)\right\|_2\} \le \frac{M_g}{1-\gamma}$.

Combining the above properties leads to the desired results. □

Next, we study the variance of the PG estimator (11).

Lemma D.6 *Under Assumption 4.1, the inequality* $\mathrm{Var}(g(\tau^H|\theta,\mu)) \le \sigma_e^2$ *holds for all* $\theta \in \mathcal{G}^0$, *where* $\sigma_e^2 = 2(\sigma_R^2 + \sigma_\pi^2)$.

Proof. One can write

$$\begin{aligned}\mathrm{Var}(g(\tau^H|\theta,\mu)) =& \mathbb{E}[(g_1(\tau^H|\theta,\mu) + g_2(\tau^H|\theta,\mu)) - \mathbb{E}[g_1(\tau^H|\theta,\mu)] - \mathbb{E}[g_2(\tau^H|\theta,\mu)])^2] \\ \le& 2\mathbb{E}[(g_1(\tau^H|\theta,\mu) - \mathbb{E}[g_1(\tau^H|\theta,\mu)])^2] + 2\mathbb{E}[(g_2(\tau^H|\theta,\mu)) - \mathbb{E}[g_2(\tau^H|\theta,\mu))])^2] \\ \le& 2\sigma_R^2 + 2\sigma_\pi^2.\end{aligned}$$

This completes the proof. □

We then show that the accumulated estimation error of the trajectory-based PG estimator (11) with the momentum can be well controlled.

D.3.1 Proof of Lemma 4.6

Proof. From Lemma C.2, using the time-invariant time-step η, the momentum coefficient β and $\mathrm{Var}(g(\tau^H|\theta,\mu)) \le \sigma_e^2$, we have

$$\mathbb{E}\left[\|e_t^H\|_2^2 \mathbf{1}_{\theta_i \in \mathcal{G}^0}\right] \le \mathbb{E}\left[\left((1-\beta)^2\|e_{t-1}^H\|_2^2 + \frac{2\beta^2\sigma_e^2}{B} + \frac{4b^2(1-\beta)^2}{B}\|\theta_t - \theta_{t-1}\|_2^2\right)\mathbf{1}_{\theta_i \in \mathcal{G}^0}\right].$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[\left(\|e_{t+1}^H\|_2^2 - \|e_t^H\|_2^2\right)\mathbf{1}_{\theta_i \in \mathcal{G}^0}\right] \le& \mathbb{E}\left[\left((1-\beta)^2 - 1)\|e_t^H\|_2^2 + \frac{2\beta^2\sigma_e^2}{B} + \frac{4b^2}{B}\|\theta_{t+1} - \theta_t\|_2^2\right)\mathbf{1}_{\theta_i \in \mathcal{G}^0}\right] \\ \le& \mathbb{E}\left[\left(-\beta\|e_t^H\|_2^2 + \frac{2\beta^2\sigma_e^2}{B} + \frac{4b^2}{B}\|\theta_{t+1} - \theta_t\|_2^2\right)\mathbf{1}_{\theta_i \in \mathcal{G}^0}\right].\end{aligned}$$

By taking $\beta = 48b^2\eta^2$ and dividing both sides by η^2, we obtain that

$$\mathbb{E}\left[\eta^{-2}(\|e_{t+1}^H\|_2^2 - \|e_t^H\|_2^2)\mathbf{1}_{\theta_i \in \mathcal{G}^0}\right] \le \mathbb{E}\left[\left(-48b^2\|e_t^H\|_2^2 + \frac{2c^2\eta^2\sigma_e^2}{B} + \frac{4b^2\eta^{-2}}{B}\|\theta_{t+1} - \theta_t\|_2^2\right)\mathbf{1}_{\theta_i \in \mathcal{G}^0}\right]. \tag{29}$$

By summing up the above inequality and dividing both sides by $32b^2$, we conclude that

$$\frac{1}{32\eta^2 b^2}\left(\mathbb{E}\left[\left(\left\|e_{t+1}^H\right\|_2^2 - \left\|e_1^H\right\|_2^2\right)\mathbf{1}_{\theta_i \in \mathcal{G}^0}\right]\right) \le \sum_{t=1}^{T}\mathbb{E}\left[\left(\frac{c^2\eta^2\sigma_e^2}{16b^2B} - \frac{3}{2}\left\|e_t^H\right\|_2^2 + \frac{1}{8B\eta^2}\left\|\theta_{t+1} - \theta_t\right\|_2^2\right)\mathbf{1}_{\theta_i \in \mathcal{G}^0}\right].$$

Then, by applying Lemma E.4 to

$$\tilde{J}_\mu^H(\theta) \coloneqq J_\mu^H(\theta) + \mathbb{E}_{s_0\sim\rho, a_t\sim\pi(\cdot|s_t), s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)}\left[\sum_{t=0}^{H-1} -\gamma^t \log\pi_\theta(a_t|s_t)\right],$$

it holds that

$$\begin{aligned}
&\frac{1}{32\eta^2 b^2}\left(\mathbb{E}\left[\left(\left\|e_{t+1}^H\right\|_2^2 - \left\|e_1^H\right\|_2^2\right)\mathbf{1}_{\theta_i\in\mathcal{G}^0}\right]\right)\\
&\le \sum_{\tau=1}^{T}\mathbb{E}\left[\left(\frac{c^2\eta^2\sigma_e^2}{16b^2B} - \frac{(6B-3)}{4B}\left\|e_t^H\right\|_2^2 + \frac{1}{B}\left(\tilde{J}_\mu^H(\theta_{t+1}) - \tilde{J}_\mu^H(\theta_t)\right)\mathbf{1}_{\theta_i\in\mathcal{G}^0}\right)\right]\\
&\le \sum_{t=1}^{T}\left(\frac{c^2\eta^2\sigma_e^2}{16b^2B} - \mathbb{E}\left[\frac{3}{4}\left\|e_t^H\right\|_2^2\mathbf{1}_{\theta_i\in\mathcal{G}^0}\right]\right) + \frac{1}{B}\left(\tilde{J}_\mu^H(\theta^*) - \tilde{J}_\mu^H(\theta_1)\right).
\end{aligned}$$

Rearranging the above inequality gives rise to

$$\sum_{t=1}^{T}\mathbb{E}\left[\frac{3}{4}\left\|e_t^H\right\|_2^2\mathbf{1}_{\theta_i\in\mathcal{G}^0}\right] \le \sum_{t=1}^{T}\frac{c^2\sigma_e^2\eta^2}{16b^2B} + \frac{1}{32\eta^2b^2}\mathbb{E}\left[\left\|e_1^H\right\|_2^2\mathbf{1}_{\theta_i\in\mathcal{G}^0}\right] + \frac{1}{B}\left(\tilde{J}_\mu^H(\theta^*) - \tilde{J}_\mu^H(\theta_1)\right).$$

Multiplying both sides by $\frac{4}{3}$ yields that

$$\sum_{t=1}^{T}\mathbb{E}\left[\left\|e_t^H\right\|_2^2\mathbf{1}_{\theta_i\in\mathcal{G}^0}\right] \le \frac{Tc^2\sigma_e^2\eta^2}{12b^2B} + \frac{1}{24b^2\eta^2}\mathbb{E}\left[\left\|e_1^H\right\|_2^2\mathbf{1}_{\theta_i\in\mathcal{G}^0}\right] + \frac{4}{3B}\left(\tilde{J}_\mu^H(\theta^*) - \tilde{J}_\mu^H(\theta_1)\right). \tag{30}$$

Since $\mathbb{E}\left[\left\|e_1^H\right\|_2^2\mathbf{1}_{\theta_i\in\mathcal{G}^0}\right] \le \frac{\sigma_e^2}{B}$, it can be concluded that

$$\sum_{t=1}^{T}\mathbb{E}\left[\left\|e_t^H\right\|_2^2\mathbf{1}_{\theta_i\in\mathcal{G}^0}\right] \le \frac{Tc^2\sigma_e^2\eta^2}{12b^2B} + \frac{\sigma_e^2}{24b^2\eta^2B} + \frac{4}{3B}\left(\tilde{J}_\mu^H(\theta^*) - \tilde{J}_\mu^H(\theta_1)\right). \tag{31}$$

Finally, from (23), we have $\tilde{J}_\rho(\theta) \le \frac{1+\lambda\log(|\mathcal{A}|)}{1-\gamma}$ for all $\theta\in\mathbb{R}^d$. In addition, we know $r(s,a) - \lambda\log\pi_\theta(a|s) \ge 0$ for all $s\in\mathcal{S}$ and $a\in\mathcal{A}$. Then, it holds that

$$0 \le \tilde{J}_\rho^H(\theta) \le \tilde{J}_\rho(\theta) \le \frac{1+\lambda\log(|\mathcal{A}|)}{1-\gamma}, \quad \forall\theta\in\mathbb{R}^{(|\mathcal{A}|-1)|\mathcal{S}|}.$$

Thus, $\left(\tilde{J}_\mu^H(\theta^*) - \tilde{J}_\mu^H(\theta_1)\right) \le \frac{1+\lambda\log(|\mathcal{A}|)}{1-\gamma}$. This completes the proof. □

D.3.2 Proof of Lemma 4.7

Proof. We first observe that $\eta \le \min\{\frac{L_J}{12}, \frac{8}{C^0}\}$ for $\epsilon \le 60\delta_0 \times 2^{-\left(\frac{L_J C^0}{96}\right)^4}$ and $T = \left(\frac{96}{L_J C^0}\log_{\frac{1}{2}}(\frac{\epsilon}{60\delta_0})\right)^{\frac{4}{3}}$. Then, lemma 4.5 can be leveraged to write

$$\mathbb{P}(\delta_T \ge \epsilon) \le \frac{1}{2}^{\frac{\eta C^0 T}{8}}\frac{\delta_0}{\epsilon} + \sum_{i=1}^{T}\frac{5\eta}{8\epsilon}\mathbb{E}[\left\|e_i\right\|_2^2\mathbf{1}_{\theta_i\in\mathcal{G}^0}] + \frac{(1+\eta L_J)^{T-1}\mathbb{E}[\sum_{i=1}^{T}\eta\left\|e_i\right\|_2\mathbf{1}_{\theta_i\in\mathcal{G}^0}] + \bar{\Delta}}{20\bar{\Delta}}.$$

By noticing that $T = \frac{8}{\eta C^0}\log_{\frac{1}{2}}(\frac{\epsilon}{60\delta_0})$, the above inequality can be simplified as

$$\begin{aligned}
\mathbb{P}(\delta_T \ge \epsilon) &\le \frac{1}{60} + \sum_{i=1}^{T}\frac{5\eta}{8\epsilon}\mathbb{E}[\left\|e_i\right\|_2^2\mathbf{1}_{\theta_i\in\mathcal{G}^0}] + \frac{(1+\eta L_J)^{\frac{\ln 2}{\eta L_J}\frac{8}{C^0}\log_{\frac{1}{2}}(\frac{\epsilon}{60\delta_0})\frac{L_J}{\ln 2}}\mathbb{E}[\sum_{i=1}^{T}\eta\left\|e_i\right\|_2\mathbf{1}_{\theta_i\in\mathcal{G}^0}] + \bar{\Delta}}{20\bar{\Delta}}\\
&\le \frac{1}{60} + \sum_{i=1}^{T}\frac{5\eta}{8\epsilon}\mathbb{E}[\left\|e_i\right\|_2^2\mathbf{1}_{\theta_i\in\mathcal{G}^0}] + \frac{(\frac{60\delta_0}{\epsilon})^{\frac{8}{C^0}\frac{L_J}{\ln 2}}\mathbb{E}[\sum_{i=1}^{T}\eta\left\|e_i\right\|_2\mathbf{1}_{\theta_i\in\mathcal{G}^0}] + \bar{\Delta}}{20\bar{\Delta}},
\end{aligned}$$

where the last inequality is due to $(1+x)^{\frac{1}{x}} \leq e$, $e^{\ln(2)} = 2$ and $2^{\log_{\frac{1}{2}} x} = \frac{1}{x}$ for all $x > 0$. The above inequality can be rewritten as

$$\begin{aligned}\mathbb{P}(\delta_T \geq \epsilon) \leq & \frac{1}{60} + \sum_{i=1}^{T} \frac{5\eta}{8\epsilon} \mathbb{E}[\left\|e_i^H\right\|_2^2 \mathbf{1}_{\theta_i \in \mathcal{G}^0}] + \frac{(\frac{60\delta_0}{\epsilon})^{\frac{8}{C^0}\frac{L_J}{\ln 2}} \mathbb{E}[\sum_{i=1}^{T} \eta \left\|e_i^H\right\|_2 \mathbf{1}_{\theta_i \in \mathcal{G}^0}] + \bar{\Delta}}{20\bar{\Delta}}, \\ & + \sum_{i=1}^{T} \frac{5\eta}{8\epsilon} \mathbb{E}[\left\|\nabla \tilde{J}_\mu^H(\theta_i) - \nabla \tilde{J}_\mu(\theta_i)\right\|_2^2 \mathbf{1}_{\theta_i \in \mathcal{G}^0}] \\ & + \frac{(1+\eta L)^{T-1} \sum_{i=1}^{T} \eta \mathbb{E}[\left\|\nabla \tilde{J}_\mu^H(\theta_i) - \nabla \tilde{J}_\mu(\theta_i)\right\|_2 \mathbf{1}_{\theta_i \in \mathcal{G}^0}] + \bar{\Delta}}{20\bar{\Delta}}. \end{aligned} \tag{32}$$

For the last two terms in (32), Lemma D.5 can be invoked to obtain, we have

$$\begin{aligned}& \sum_{i=0}^{T} \frac{5\eta}{8\epsilon} \mathbb{E}[\left\|\nabla \tilde{J}_\mu^H(\theta_i) - \nabla \tilde{J}_\mu(\theta_i)\right\|_2^2 \mathbf{1}_{\theta_i \in \mathcal{G}^0}] + \frac{(\frac{60\delta_0}{\epsilon})^{\frac{8}{C^0}\frac{L_J}{\ln 2}} \sum_{i=0}^{T} \eta \mathbb{E}[\left\|\nabla \tilde{J}_\mu^H(\theta_i) - \nabla \tilde{J}_\mu(\theta_i)\right\|_2 \mathbf{1}_{\theta_i \in \mathcal{G}^0}]}{20\bar{\Delta}} \\ & \leq \frac{5\log_{\frac{1}{2}}(\frac{\epsilon}{60\delta_0})}{C^0\epsilon} \left(M_g(1+R) \left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2} \right) \gamma^H + \frac{M_g \gamma^H}{1-\gamma} \right)^2 \\ & + \frac{2(\frac{60\delta_0}{\epsilon})^{\frac{8}{C^0}\frac{L_J}{\ln 2}} \log_{\frac{1}{2}}(\frac{\epsilon}{60\delta_0}) \left(M_g(1+R) \left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2} \right) \gamma^H + \frac{M_g \gamma^H}{1-\gamma} \right)}{5C^0\bar{\Delta}}. \end{aligned}$$

To guarantee that

$$\sum_{i=0}^{T} \frac{5\eta}{8\epsilon} \mathbb{E}[\left\|\nabla \tilde{J}_\mu^H(\theta_i) - \nabla \tilde{J}_\mu(\theta_i)\right\|_2^2 \mathbf{1}_{\theta_i \in \mathcal{G}^0}] + \frac{(\frac{60\delta_0}{\epsilon})^{\frac{8}{C^0}\frac{L_J}{\ln 2}} \sum_{i=0}^{T} \eta \mathbb{E}[\left\|\nabla \tilde{J}_\mu^H(\theta_i) - \nabla \tilde{J}_\mu(\theta_i)\right\|_2 \mathbf{1}_{\theta_i \in \mathcal{G}^0}]}{20\bar{\Delta}} \leq \frac{1}{60}, \tag{33}$$

it suffices to have

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

By the Cauchy-Schwarz inequality and the concavity of $\sqrt{x}$, we have

$$\sum_{t=1}^{T} \mathbb{E}\left[\left\|e_t^H\right\|_2 \mathbf{1}_{\theta_i \in \mathcal{G}^0}\right] \leq \sqrt{T}\mathbb{E}\left[\sqrt{\sum_{t=1}^{T} \left\|e_t^H\right\|_2^2 \mathbf{1}_{\theta_i \in \mathcal{G}^0}}\right] \leq \sqrt{T}\sqrt{\sum_{t=1}^{T} \mathbb{E}\left[\left\|e_t^H\right\|_2^2 \mathbf{1}_{\theta_i \in \mathcal{G}^0}\right]}.$$

Then, it follows from Lemma 4.6 that

$$\begin{aligned}& \sum_{i=1}^{T} \frac{5\eta}{8\epsilon} \mathbb{E}[\left\|e_i^H\right\|_2^2 \mathbf{1}_{\theta_i \in \mathcal{G}^0}] + \frac{(\frac{60\delta_0}{\epsilon})^{\frac{8}{C^0}\frac{L_J}{\ln 2}} \eta \sqrt{T \sum_{t=1}^{T} \mathbb{E}\left[\left\|e_t^H\right\|_2^2 \mathbf{1}_{\theta_i \in \mathcal{G}^0}\right]}}{20\bar{\Delta}} \\ & \leq \frac{5\eta}{8B\epsilon} \left(\frac{Tc^2\sigma_e^2\eta^2}{12b^2} + \frac{\sigma_e^2}{24b^2\eta^2} + \frac{4}{3}\frac{1+\lambda\log(|\mathcal{A}|)}{1-\gamma} \right) \\ & + \frac{(\frac{60\delta_0}{\epsilon})^{\frac{8}{C^0}\frac{L_J}{\ln 2}} \eta \sqrt{T\left(\frac{Tc^2\sigma_e^2\eta^2}{12b^2} + \frac{\sigma_e^2}{24b^2\eta^2} + \frac{4}{3}\frac{1+\lambda\log(|\mathcal{A}|)}{1-\gamma}\right)}}{20\bar{\Delta}\sqrt{B}}. \end{aligned}$$

By taking $\eta = \frac{L_J}{12T^{\frac{1}{4}}}$, we have $T = \left(\frac{96}{L_J C^0}\log_{\frac{1}{2}}(\frac{\epsilon}{60\delta_0})\right)^{\frac{4}{3}}$ and

$$\sum_{i=1}^{T}\frac{5\eta}{8\epsilon}\mathbb{E}[\left\|e_i^H\right\|_2^2\mathbf{1}_{\theta_i\epsilon\mathcal{G}^0}] + \frac{(\frac{60\delta_0}{\epsilon})^{\frac{8}{C^0}\frac{L_J}{\ln 2}}\eta\sqrt{T\sum_{t=1}^{T}\mathbb{E}\left[\left\|e_t^H\right\|_2^2\mathbf{1}_{\theta_i\epsilon\mathcal{G}^0}\right]}}{20\bar{\Delta}}$$
$$\leq \frac{5}{8B\epsilon}\left(\frac{L_J^3T^{\frac{1}{4}}c^2\sigma_e^2}{12^4b^2} + \frac{\sigma_e^2T^{\frac{1}{4}}}{2b^2L_J} + \frac{L_JT^{-\frac{1}{4}}}{9}\frac{1+\lambda\log(|\mathcal{A}|)}{1-\gamma}\right)$$
$$+ \frac{(\frac{60\delta_0}{\epsilon})^{\frac{8}{C^0}\frac{L_J}{\ln 2}}\sqrt{\frac{TL_J^4c^2\sigma_e^2}{12^5b^2} + \frac{\sigma_e^2T}{24b^2} + \frac{L_J^2T^{1/2}}{108}\frac{1+\lambda\log(|\mathcal{A}|)}{1-\gamma}}}{20\bar{\Delta}\sqrt{B}}.$$

To guarantee that

$$\sum_{i=1}^{T}\frac{5\eta}{8\epsilon}\mathbb{E}[\left\|e_i^H\right\|_2^2\mathbf{1}_{\theta_i\epsilon\mathcal{G}^0}] + \frac{(\frac{60\delta_0}{\epsilon})^{\frac{8}{C^0}\frac{L_J}{\ln 2}}\eta\sqrt{T\sum_{t=1}^{T}\mathbb{E}\left[\left\|e_t^H\right\|_2^2\mathbf{1}_{\theta_i\epsilon\mathcal{G}^0}\right]}}{20\bar{\Delta}} \leq \frac{1}{60}, \tag{34}$$

it suffices to have

$$B = \mathcal{O}\left(\max\left\{\frac{\left(\log_{\frac{1}{2}}(\epsilon)\right)^{\frac{1}{3}}}{\epsilon}, \frac{\left(\log_{\frac{1}{2}}(\epsilon)\right)^{\frac{4}{3}}}{\epsilon^{\frac{16L_J}{C^0\ln 2}}}\right\}\right).$$

By combining (32), (33) and (34), we obtain that $\mathbb{P}(\delta_T \geq \epsilon) \leq \frac{1}{10}$. Hence, the total sample complexity is

$$H\cdot T\cdot B = \mathcal{O}\left(\log_\gamma\left(\frac{\epsilon(1-\gamma)}{\log_{\frac{1}{2}}(\epsilon)}\right)\cdot\max\left\{\frac{\left(\log_{\frac{1}{2}}(\epsilon)\right)^{\frac{5}{3}}}{\epsilon}, \frac{\left(\log_{\frac{1}{2}}(\epsilon)\right)^{\frac{8}{3}}}{\epsilon^{\frac{16L_J}{C^0\ln 2}}}\right\}\right).$$

□

D.3.3 Maximum entropy RL with vanilla PG

Algorithm 5 PG for softmax parameterization with entropy (PG-SE)

1: **Inputs**: Iteration T, horizon H, batch size B, initial input θ_1, parameters η, β and initial distribution μ;
2: **Outputs**: θ_T;
3: **for** $t = 1, 2, \ldots, T-1$ **do**
4: Sample B trajectories $\{\tau_i^H\}_{i=1}^B$ from $p(\cdot|\theta_t, \mu)$;
5: Compute $u_t = \frac{1}{B}\sum_{i=1}^B g(\tau_i^H|\theta_t, \mu)$ where $g(\tau_i^H|\theta_t, \mu)$ is given by (11);
6: Update $\theta_{t+1} = \theta_t + \eta u_t$;
7: **end for**

Theorem D.7 *Under the conditions in Lemma 4.5, let θ_T be generated by Algorithm 5. Let $\eta \leq \min\{\frac{L_J}{12}, \frac{8}{C^0}\}$, $T = \frac{8}{\eta C^0}\log_{\frac{1}{2}}(\frac{\epsilon}{60\delta_0})$ where $L_J = \frac{(12+8\log|\mathcal{A}|)}{(1-\gamma)^3}$. Let $B = \mathcal{O}\left(\max\left\{\frac{\log_{\frac{1}{2}}(\frac{\epsilon}{60\delta_0})}{\epsilon}, \frac{(\log_{\frac{1}{2}}(\frac{\epsilon}{60\delta_0}))^2}{\epsilon^{\frac{16L_J}{C^0\ln 2}}}\right\}\right)$ and $H = \mathcal{O}\left(\log_\gamma\left(\frac{\epsilon(1-\gamma)}{\log_{\frac{1}{2}}(\epsilon)}\right)\right)$. Then, we have $\mathbb{P}(\delta_T \leq \epsilon) \geq \frac{9}{10}$. In total, it requires $\mathcal{O}\left(\log_\gamma\left(\frac{\epsilon(1-\gamma)}{\log_{\frac{1}{2}}(\epsilon)}\right)\cdot\max\left\{\frac{\left(\log_{\frac{1}{2}}(\epsilon)\right)^2}{\epsilon}, \frac{\left(\log_{\frac{1}{2}}(\epsilon)\right)^3}{\epsilon^{\frac{16L_J}{C^0\ln 2}}}\right\}\right)$ samples to achieve an ϵ-optimal policy.*

Proof. By setting $T = \frac{8}{\eta C^0}\log_{\frac{1}{2}}(\frac{\epsilon}{60\delta_0})$, it results from (32) that

$$\begin{aligned}\mathbb{P}(\delta_T \geq \epsilon) \leq & \frac{1}{60} + \sum_{i=0}^{T}\frac{5\eta}{8\epsilon}\mathbb{E}[\left\|e_i^H\right\|_2^2 \mathbf{1}_{\theta_i\in\mathcal{G}^0}] + \frac{(\frac{60\delta_0}{\epsilon})^{\frac{8}{C^0}\frac{L_J}{\ln 2}}\mathbb{E}[\sum_{i=0}^{T}\eta\left\|e_i^H\right\|_2 \mathbf{1}_{\theta_i\in\mathcal{G}^0}] + \bar{\Delta}}{20\bar{\Delta}}, \\ & + \sum_{i=0}^{T}\frac{5\eta}{8\epsilon}\mathbb{E}[\left\|\nabla\tilde{J}_\mu^H(\theta_i) - \nabla\tilde{J}_\mu(\theta_i)\right\|_2^2 \mathbf{1}_{\theta_i\in\mathcal{G}^0}] \\ & + \frac{(1+\eta L)^{T-1}\sum_{i=0}^{T}\eta\mathbb{E}[\left\|\nabla\tilde{J}_\mu^H(\theta_i) - \nabla\tilde{J}_\mu(\theta_i)\right\|_2 \mathbf{1}_{\theta_i\in\mathcal{G}^0}] + \bar{\Delta}}{20\bar{\Delta}}. \end{aligned} \tag{35}$$

From (33), the inequality

$$\sum_{i=0}^{T}\frac{5\eta}{8\epsilon}\mathbb{E}[\left\|\nabla\tilde{J}_\mu^H(\theta_i) - \nabla\tilde{J}_\mu(\theta_i)\right\|_2^2 \mathbf{1}_{\theta_i\in\mathcal{G}^0}] + \frac{(\frac{60\delta_0}{\epsilon})^{\frac{8}{C^0}\frac{L_J}{\ln 2}}\sum_{i=0}^{T}\eta\mathbb{E}[\left\|\nabla\tilde{J}_\mu^H(\theta_i) - \nabla\tilde{J}_\mu(\theta_i)\right\|_2 \mathbf{1}_{\theta_i\in\mathcal{G}^0}]}{20\bar{\Delta}} \leq \frac{1}{60} \tag{36}$$

is guaranteed to hold if

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

For the second and the third terms in (32), since $\mathbb{E}[\left\|e_i^H\right\|_2^2 \mathbf{1}_{\theta_i\in\mathcal{G}^0}] \leq \frac{\sigma_e^2}{B}$, we have

$$\begin{aligned} & \sum_{i=0}^{T}\frac{5\eta}{8\epsilon}\mathbb{E}\left[\left\|e_i^H\right\|_2^2 \mathbf{1}_{\theta_i\in\mathcal{G}^0}\right] + \frac{(\frac{60\delta_0}{\epsilon})^{\frac{8}{C^0}\frac{L_J}{\ln 2}}\mathbb{E}\left[\sum_{i=0}^{T}\eta\left\|e_i^H\right\|_2 \mathbf{1}_{\theta_i\in\mathcal{G}^0}\right] + \bar{\Delta}}{20\bar{\Delta}} \\ \leq & \frac{5T\eta\sigma_e^2}{8B\epsilon} + \frac{(\frac{60\delta_0}{\epsilon})^{\frac{8}{C^0}\frac{L_J}{\ln 2}}\eta T\sigma_e}{20\bar{\Delta}\sqrt{B}} \\ = & \frac{5\sigma_e^2\log_{\frac{1}{2}}(\frac{\epsilon}{60\delta_0})}{C^0 B\epsilon} + \frac{2(\frac{60\delta_0}{\epsilon})^{\frac{8}{C^0}\frac{L_J}{\ln 2}}\sigma_e\log_{\frac{1}{2}}(\frac{\epsilon}{60\delta_0})}{5C^0\bar{\Delta}\sqrt{B}}. \end{aligned} \tag{37}$$

Similarly, to satisfy

$$\sum_{i=0}^{T}\frac{5\eta}{8\epsilon}\mathbb{E}[\left\|e_i^H\right\|_2^2 \mathbf{1}_{\theta_i\in\mathcal{G}^0}] + \frac{(\frac{60\delta_0}{\epsilon})^{\frac{8}{C^0}\frac{L_J}{\ln 2}}\mathbb{E}[\sum_{i=0}^{T}\eta\left\|e_i^H\right\|_2 \mathbf{1}_{\theta_i\in\mathcal{G}^0}] + \bar{\Delta}}{20\bar{\Delta}} \leq \frac{1}{60}, \tag{38}$$

it suffices to have

$$\begin{aligned} B = & \max\left\{\frac{600\sigma_e^2\log_{\frac{1}{2}}(\frac{\epsilon}{60\delta_0})}{C^0\epsilon}, \frac{1152\sigma_e^2(\frac{60\delta_0}{\epsilon})^{\frac{16}{C^0}\frac{L_J}{\ln 2}}(\log_{\frac{1}{2}}(\frac{\epsilon}{60\delta_0}))^2}{(C^0\bar{\Delta})^2}\right\} \\ = & \mathcal{O}\left(\max\left\{\frac{\log_{\frac{1}{2}}(\frac{\epsilon}{60\delta_0})}{\epsilon}, \frac{(\log_{\frac{1}{2}}(\frac{\epsilon}{60\delta_0}))^2}{\epsilon^{\frac{16L_J}{C^0\ln 2}}}\right\}\right). \end{aligned}$$

The total sample complexity is

$$H\cdot T\cdot B = \mathcal{O}\left(\log_\gamma\left(\frac{\epsilon(1-\gamma)}{\log_{\frac{1}{2}}(\epsilon)}\right)\cdot\max\left\{\frac{\left(\log_{\frac{1}{2}}(\epsilon)\right)^2}{\epsilon}, \frac{\left(\log_{\frac{1}{2}}(\epsilon)\right)^3}{\epsilon^{\frac{16L_J}{C^0\ln 2}}}\right\}\right).$$

□

E Additional supporting results

Lemma E.1 *For the truncated PGT policy gradient given in (2) and satisfying Assumptions 2.2 and 2.3, the following properties hold:*

1. $g(\tau_i^H|\theta)$ *is* $M_h/(1-\gamma)^2$*-Lipschitz continuous for all* $\theta \in \Theta$.
2. $\left\|g(\tau_i^H|\theta)\right\|_2 \leq M_g/(1-\gamma)^2$ *for all* $\theta \in \Theta$.
3. *If the infinite-sum is well-defined, then*

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

is an unbiased estimate of $\nabla J_\rho(\theta)$*. Similarly, the truncated PGT estimate* $g(\tau_i^H|\theta)$ *given by* (2) *is an unbiased estimate of* $\nabla J_\rho^H(\theta)$.

4. $\left\|\nabla J_\rho^H(\theta) - \nabla J_\rho(\theta)\right\|_2 \leq M_g\left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\right)\gamma^H$.
5. $\max\{\left\|\nabla J_\rho(\theta)\right\|_2, \left\|\nabla J_\rho^H(\theta)\right\|_2\} \leq \frac{M_g}{(1-\gamma)^2}$.

Proof. The first two properties are shown in Proposition 4.2 in [54]. The last three properties follow directly from Lemma B.1 in [27] as well as the equivalence between GPT [46] and GPOMDP [4]. □

Proposition E.2 (Lemma 1 in [9]) *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 E.3 (Lemma 6.1 in [54]) *Under Assumptions 2.2 and 3.1, let* $w(\tau|\theta_{t-1}, \theta_t) = p(\tau|\theta_{t-1})/p(\tau|\theta_t)$*, we have*

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

where $C_w = \sqrt{H(2HM_g^2 + M_h)(W+1)}$.

Lemma E.4 *Suppose that* $J_\rho^H(\theta)$ *is* L_J^H*-smooth. Given* $0 < \eta_t \leq \frac{1}{12L_J^H}$ *for all* $t \geq 1$*, let* $\{\theta_t\}_{t=1}^T$ *be generated by a general update of the form* $\theta_{t+1} = \mathrm{argmax}_{\theta\in\Theta}\{J_\rho^H(\theta_t) + \langle u_t, \theta - \theta_t\rangle - \frac{1}{2\eta_t}\left\|\theta - \theta_t\right\|_2^2\}$ *and let* $e_t^H = u_t - \nabla J_\rho^H(\theta_t)$*. We have*

$$J_\rho^H(\theta_{t+1}) \geq J_\rho^H(\theta_t) + \frac{1}{8\eta_t}\left\|\theta_{t+1} - \theta_t\right\|_2^2 - \frac{3\eta_t}{4}\left\|e_t^H\right\|_2^2.$$

Proof. Since $J_\rho^H(\theta)$ is L_J^H-smooth, one can write

$$\begin{aligned}
&J_\rho^H(\theta_{t+1}) - J_\rho^H(\theta_t) - \langle u_t, \theta_{t+1} - \theta_t\rangle \\
=&J_\rho^H(\theta_{t+1}) - J_\rho^H(\theta_t) - \langle \nabla J_\rho^H(\theta_t), \theta_{t+1} - \theta_t\rangle + \langle \sqrt{\eta_t}(\nabla J_\rho^H(\theta_t) - u_t), \frac{1}{\sqrt{\eta_t}}(\theta_{t+1} - \theta_t)\rangle \\
\geq& -\frac{L_J^H}{2}\left\|\theta_{t+1} - \theta_t\right\|^2 - \frac{b\eta_t}{2}\left\|\nabla J_\rho^H(\theta_t) - u_t\right\|_2^2 - \frac{1}{2b\eta_t}\left\|\theta_{t+1} - \theta_t\right\|_2^2 \\
=&(-\frac{1}{2b\eta_t} - \frac{L_J^H}{2})\left\|\theta_{t+1} - \theta_t\right\|_2^2 - \frac{b\eta_t}{2}\left\|e_t^H\right\|_2^2,
\end{aligned}$$

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

$$\begin{aligned}
J_\rho^H(\theta_{t+1}) \geq & J_\rho^H(\theta_t) + \langle u_t, \theta_{t+1} - \theta_t \rangle - (\frac{1}{2b\eta_t} + \frac{L_J^H}{2}) \|\theta_{t+1} - \theta_t\|_2^2 - \frac{b\eta_t}{2} \left\|e_t^H\right\|_2^2 \\
= & J_\rho^H(\theta_t) + \langle u_t, \theta_{t+1} - \theta_t \rangle - \frac{1}{2\eta_t} \|\theta_{t+1} - \theta_t\|_2^2 + (\frac{1-1/b}{2\eta_t} - \frac{L_J^H}{2}) \|\theta_{t+1} - \theta_t\|_2^2 - \frac{b\eta_t}{2} \left\|e_t^H\right\|_2^2 \\
= & \max_{\theta \in \Theta} \{ J_\rho^H(\theta_t) + \langle u_t, \theta - \theta_t \rangle - \frac{1}{2\eta_t} \|\theta - \theta_t\|_2^2 \} + (\frac{1-1/b}{2\eta_t} - \frac{L_J^H}{2}) \|\theta_{t+1} - \theta_t\|_2^2 - \frac{b\eta_t}{2} \left\|e_t^H\right\|^2 \\
\geq & J_\rho^H(\theta_t) + (\frac{1-1/b}{2\eta_t} - \frac{L_J^H}{2}) \|\theta_{t+1} - \theta_t\|_2^2 - \frac{b\eta_t}{2} \left\|e_t^H\right\|_2^2.
\end{aligned}$$

By choosing $b = \frac{3}{2}$ and using the fact that $0 < \eta_t \leq \frac{1}{12L_J^H}$, we have

$$\begin{aligned}
J_\rho^H(\theta_{t+1}) \geq & J_\rho^H(\theta_t) + (\frac{1}{6\eta_t} - \frac{L_J^H}{2}) \|\theta_{t+1} - \theta_t\|_2^2 - \frac{3\eta_t}{4} \left\|e_t^H\right\|_2^2 \\
\geq & J_\rho^H(\theta_t) + \frac{1}{8\eta_t} \|\theta_{t+1} - \theta_t\|_2^2 - \frac{3\eta_t}{4} \left\|e_t^H\right\|_2^2.
\end{aligned}$$

This completes the proof. □

F Smoothness proofs

Before proving the smoothness result, we first develop a lemma that is similar to Lemma D.2 in [2] but is generalized to a truncated value function.

Lemma F.1 *Given $H > 0$, let $\pi_\alpha := \pi_{\theta+\alpha u}$ and $\tilde{V}^H(\alpha)$ be the corresponding truncated value function at a fixed state s_0, i.e.,*

$$\tilde{V}^H(\alpha) := V^{H,\pi_\alpha}(s_0) = \mathbb{E}_{a_h \sim \pi_\alpha(\cdot|s_h), s_{h+1} \sim \mathbb{P}(\cdot|s_h,a_h)} \left(\sum_{h=0}^{H-1} \gamma^h r(s_h, a_h) \middle| s_0 \right).$$

Assume that

$$\sum_{a \in \mathcal{A}} \left| \frac{d\pi_\alpha(a|s_0)}{d\alpha} \Big|_{\alpha=0} \right| \leq C_1, \quad \sum_{a \in \mathcal{A}} \left| \frac{d^2\pi_\alpha(a|s_0)}{(d\alpha)^2} \Big|_{\alpha=0} \right| \leq C_2.$$

Then,

$$\max_{\|u\|_2=1} \left| \frac{d^2\tilde{V}^H(\alpha)}{(d\alpha)^2} \right| \leq \frac{C_2}{(1-\gamma)^2} + \frac{2\gamma C_1^2}{(1-\gamma)^3}.$$

Proof. Let $\tilde{P}(\alpha)$ be the state-action transition matrix under π, i.e.,

$$[\tilde{P}(\alpha)]_{(s,a)\to(s',a')} = \pi_\alpha(a'|s')P(s'|s,a).$$

One can differentiate $\tilde{P}(\alpha)$ with respect to α to arrive at

$$\left[\frac{d\tilde{P}(\alpha)}{d\alpha} \Big|_{\alpha=0} \right]_{(s,a)\to(s',a')} = \frac{d\pi_\alpha(a'|s')}{d\alpha} \Big|_{\alpha=0} P(s'|s,a).$$

For an arbitrary vector x, it holds that

$$\left[\frac{d\tilde{P}(\alpha)}{d\alpha} \Big|_{\alpha=0} x \right]_{s,a} = \sum_{a',s'} \frac{d\pi_\alpha(a'|s')}{d\alpha} \Big|_{\alpha=0} P(s'|s,a)x_{a',s'}.$$

Therefore

$$\begin{aligned}\max_{\|u\|_2=1}\left|\left[\left.\frac{d\tilde{P}(\alpha)}{d\alpha}\right|_{\alpha=0}x\right]_{s,a}\right| &= \max_{\|u\|_2=1}\left|\sum_{a',s'}\left.\frac{d\pi_\alpha(a'|s')}{d\alpha}\right|_{\alpha=0}P(s'|s,a)x_{a',s'}\right| \\ &\le \sum_{a',s'}\left|\left.\frac{d\pi_\alpha(a'|s')}{d\alpha}\right|_{\alpha=0}\right|P(s'|s,a)\,|x_{a',s'}| \\ &\le \sum_{s'}P(s'|s,a)\,\|x\|_\infty\sum_{a'}\left|\left.\frac{d\pi_\alpha(a'|s')}{d\alpha}\right|_{\alpha=0}\right| \\ &\le \sum_{s'}P(s'|s,a)\,\|x\|_\infty C_1 \\ &\le C_1\,\|x\|_\infty .\end{aligned}$$

By definition of ℓ_∞ norm, we obtain that

$$\max_{\|u\|_2=1}\left\|\frac{d\tilde{P}(\alpha)}{d\alpha}x\right\| \le C_1\,\|x\|_\infty .$$

Similarly, differentiating $\tilde{P}(\alpha)$ twice with respect to α gives rise to

$$\left[\left.\frac{d^2\tilde{P}(\alpha)}{(d\alpha)^2}\right|_\alpha = 0\right]_{(s,a)\to(s',a')} = \left.\frac{d^2\pi_\alpha(a'|s')}{(d\alpha)^2}\right|_{\alpha=0}P(s'|s,a).$$

An identical argument leads to that, for arbitrary x,

$$\max_{\|u\|_2=1}\left\|\frac{d^2\tilde{P}(\alpha)}{(d\alpha)^2}x\right\|_2 \le C_2\,\|x\|_\infty .$$

Let $Q^{H,\alpha}(s_0,a_0)$ be the corresponding truncated Q-function for the policy π_α at the state s_0 under the action a_0. Observe that $Q^{H,\alpha}(s_0,a_0)$ can be written as

$$Q^{H,\alpha}(s_0,a_0) = e_{s_0,a_0}^\top M^H(\alpha)r,$$

where e_{s_0,a_0} is an indicator vector with 1 at the entry indexed by (s_0,a_0) and 0 otherwise, $M^H(\alpha) = \sum_{h=0}^{H-1}\gamma^h\tilde{P}(\alpha)^h$ and r is the reward vector. By differentiating $Q^{H,\alpha}(s_0,a_0)$ with respect to α, we obtain that

$$\frac{dQ^{H,\alpha}(s_0,a_0)}{d\alpha} = e_{s_0,a_0}^\top\sum_{h=0}^{H-1}\gamma^h h\tilde{P}(\alpha)^{h-1}\frac{d\tilde{P}(\alpha)}{d\alpha}r,$$

$$\frac{d^2Q^{H,\alpha}(s_0,a_0)}{(d\alpha)^2} = e_{s_0,a_0}^\top\sum_{h=1}^{H-1}\gamma^h h(h-1)\tilde{P}(\alpha)^{h-2}\frac{d\tilde{P}(\alpha)}{d\alpha}\frac{d\tilde{P}(\alpha)}{d\alpha}r + e_{s_0,a_0}^\top\sum_{h=0}^{H-1}\gamma^h h\tilde{P}(\alpha)^{h-1}\frac{d^2\tilde{P}(\alpha)}{(d\alpha)^2}r.$$

Since $\tilde{P}(\alpha)^h \ge 0$ (component-wise) and $\tilde{P}(\alpha)^h\mathbf{1} \le \mathbf{1}$ for all $h \ge 0$, i.e., each row of $\tilde{P}(\alpha)^h$ is positive and sums to 1, it holds that

$$\max_{\|u\|_2=1}\left\|M^H(\alpha)x\right\|_\infty \le \frac{1}{1-\gamma}\,\|x\|_\infty .$$

Therefore,

$$\begin{aligned}\max_{\|u\|_2=1}\left|\left.\frac{dQ^{H,\alpha}(s_0,a_0)}{d\alpha}\right|_{\alpha=0}\right| &= \max_{\|u\|_2=1}\left|e_{s_0,a_0}^\top\sum_{h=0}^{H-1}\gamma^h h\tilde{P}(\alpha)^{h-1}\frac{d\tilde{P}(\alpha)}{d\alpha}r\right| \\ &\le \left\|\sum_{h=0}^{H-1}\gamma^h h\tilde{P}(\alpha)^{h-1}\frac{d\tilde{P}(\alpha)}{d\alpha}r\right\|_\infty \\ &\le C_1\sum_{h=0}^{H-1}\gamma^h h \\ &\le \frac{\gamma C_1}{(1-\gamma)^2},\end{aligned}$$

and

$$\begin{aligned}
&\max_{\|u\|_2=1}\left|\frac{d^2Q^{H,\alpha}(s_0,a_0)}{(d\alpha)^2}\bigg|_{\alpha=0}\right| \\
&=\max_{\|u\|_2=1}\left|e_{s_0,a_0}^\top\sum_{h=1}^{H-1}\gamma^h h(h-1)\tilde{P}(\alpha)^{h-2}\frac{d\tilde{P}(\alpha)}{d\alpha}\frac{d\tilde{P}(\alpha)}{d\alpha}r+e_{s_0,a_0}^\top\sum_{h=1}^{H-1}\gamma^h h\tilde{P}(\alpha)^{h-1}\frac{d^2\tilde{P}(\alpha)}{(d\alpha)^2}r\right| \\
&\le\left\|\sum_{h=0}^{H-1}\gamma^h h(h-1)\tilde{P}(\alpha)^{h-2}\frac{d\tilde{P}(\alpha)}{d\alpha}\frac{d\tilde{P}(\alpha)}{d\alpha}r\right\|_\infty+\left\|\sum_{h=0}^{H-1}\gamma^h h\tilde{P}(\alpha)^{h-1}\frac{d^2\tilde{P}(\alpha)}{(d\alpha)^2}r\right\|_\infty \\
&\le C_1^2\sum_{h=1}^{H-1}\gamma^h h(h-1)+C_2\sum_{h=0}^{H-1}\gamma^h h \\
&\le\frac{2\gamma^2C_1^2}{(1-\gamma)^3}+\frac{\gamma C_2}{(1-\gamma)^2}.
\end{aligned}$$

Consider the identity

$$\tilde{V}^H(\alpha)=\sum_a\pi_\alpha(a|s_0)Q^{H,\alpha}(s_0,a).$$

Differentiating $\tilde{V}^H(\alpha)$ twice with respect to α yields that

$$\frac{d^2\tilde{V}^H(\alpha)}{(d\alpha)^2}=\sum_a\frac{d^2\pi_\alpha(a|s_0)}{(d\alpha)^2}Q^{H,\alpha}(s_0,a)+2\sum_a\frac{d\pi_\alpha(a|s_0)}{d\alpha}\frac{dQ^{H,\alpha}(s_0,a)}{d\alpha}+\sum_a\pi_\alpha(a|s_0)\frac{d^2Q^{H,\alpha}(s_0,a)}{(d\alpha)^2}$$

Hence,

$$\begin{aligned}
\max_{\|u\|_2=1}\left|\frac{d^2\tilde{V}^H(\alpha)}{(d\alpha)^2}\right|\le&\frac{C_2}{1-\gamma}+\frac{2\gamma C_1^2}{(1-\gamma)^2}+\frac{2\gamma^2C_1^2}{(1-\gamma)^3}+\frac{\gamma C_2}{(1-\gamma)^2} \\
=&\frac{C_2}{(1-\gamma)^2}+\frac{2\gamma C_1^2}{(1-\gamma)^3}.
\end{aligned}$$

This completes the proof. □

Using the above lemma, we establish smoothness for the value functions under the softmax parameterization with the log barrier regularization, which is similar to Lemma D.4 in [2]. We prove it here for the completeness.

Lemma F.2 *For the softmax parameterization with*

$$\begin{aligned}
L_{\lambda,\rho}(\theta)&=J_\rho(\theta)+\frac{\lambda}{|\mathcal{S}||\mathcal{A}|}\sum_{s,a}\log\pi_\theta(a|s), \\
L_{\lambda,\rho}^H(\theta)&=J_\rho^H(\theta)+\frac{\lambda}{|\mathcal{S}||\mathcal{A}|}\sum_{s,a}\log\pi_\theta(a|s),
\end{aligned}$$

the functions $L_{\lambda,\rho}(\theta)$ and $L_{\lambda,\rho}^H(\theta)$ are $\left(\frac{8}{(1-\gamma)^3}+\frac{2\lambda}{|\mathcal{S}|}\right)$-smooth.

Proof. Let $\theta_s\in\mathbb{R}^{|\mathcal{A}|}$ denote the parameters associated with a given state s. We have that

$$\nabla_{\theta_s}\pi_\theta(a|s)=\pi_\theta(a|s)\left(e_a-\pi(\cdot|s)\right)$$

and

$$\nabla_{\theta_s}^2\pi_\theta(a|s)=\pi_\theta(a|s)\left(e_ae_a^\top-e_a\pi(\cdot|s)^\top-\pi(\cdot|s)e_a^\top+2\pi(\cdot|s)\pi(\cdot|s)^\top-\mathrm{diag}(\pi(\cdot|s))\right),$$

where e_a is a standard basis vector and $\pi(\cdot|s)$ is a vector of probabilities. Differentiating $\pi_\alpha(a|s)$ once with respect to α leads to

$$\begin{aligned}
\sum_{a\in\mathcal{A}}\left|\frac{d\pi_\alpha(a|s)}{d\alpha}\bigg|_{\alpha=0}\right|&\le\sum_{a\in\mathcal{A}}\left|u^\top\nabla_{\theta+\alpha u}\pi_\alpha(a|s)\bigg|_{\alpha=0}\right| \\
&\le\sum_{a\in\mathcal{A}}\pi_\theta(a|s)\left|u_s^\top e_a-u_s^\top\pi(\cdot|s)\right| \\
&\le\max_{a\in\mathcal{A}}\left(\left|u_s^\top e_a\right|+\left|u_s^\top\pi(\cdot|s)\right|\right)\le 2.
\end{aligned}$$

Similarly, differentiating once again with respect to α concludes that

$$\begin{aligned}\sum_{a\in\mathcal{A}}\left|\frac{d^2\pi_\alpha(a|s)}{(d\alpha)^2}\bigg|_{\alpha=0}\right| \le& \sum_{a\in\mathcal{A}}\left|u^\top\nabla^2_{\theta+\alpha u}\pi_\alpha(a|s)\Big|_{\alpha=0}u\right| \\ \le& \max_{a\in\mathcal{A}}\big(|u_s^\top e_a e_a^\top u_s| + |u_s^\top e_a\pi(\cdot|s)^\top u_s| + |u_s^\top\pi(\cdot|s)e_a^\top u_s| \\ &+2|u_s^\top\pi(\cdot|S)\pi(\cdot|s)^\top u_s| + |u_s^\top\mathrm{diag}(\pi(\cdot|s))u_s|\big) \\ \le& 6.\end{aligned}$$

Now it follows from Lemma F.1 with $C_1 = 2$ and $C_2 = 6$ that

$$\begin{aligned}\max_{\|u\|_2=1}\left|\frac{d^2\tilde{V}(\alpha)}{(d\alpha)^2}\bigg|_{\alpha=0}\right| &\le \frac{8}{(1-\gamma)^3}, \\ \max_{\|u\|_2=1}\left|\frac{d^2\tilde{V}^H(\alpha)}{(d\alpha)^2}\bigg|_{\alpha=0}\right| &\le \frac{8}{(1-\gamma)^3}.\end{aligned}$$

Thus, $V^{H,\pi_\theta}(s_0)$ and $V^{\pi_\theta}(s_0)$ are both $\frac{8}{(1-\gamma)^3}$-smooth for all states s_0, and hence $J_\rho(\theta)$ and $J_\rho^H(\theta)$ are also $\frac{8}{(1-\gamma)^3}$-smooth.

For the regularizer $\frac{\lambda R(\theta)}{|\mathcal{S}|}$ with $R(\theta) = \frac{1}{\mathcal{A}}\sum_{s,a}\log\pi_\theta(a|S)$, we have

$$\begin{aligned}\nabla^2_{\theta_s}R(\theta) &= -\mathrm{diag}(\pi_\theta(\cdot|s)) + \pi_\theta(\cdot|s)\pi_\theta(\cdot|s)^\top, \\ \nabla_{\theta_s}\nabla_{\theta_{s'}}R(\theta) &= 0, \quad \text{for } s\neq s'.\end{aligned}$$

Therefore

$$|u^\top\nabla^2_\theta R(\theta)u| = \left|\sum_s u_s^\top\nabla^2_{\theta_s}R(\theta)u_s\right| \le 2\|u\|_2^2, \quad \forall u\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}.$$

Thus, $\frac{\lambda R}{|\mathcal{S}|}$ is $\frac{2\lambda}{|\mathcal{S}|}$-smooth. This completes the proof. □

The following result is similar to Lemma 7 in [30] but applies to the restricted soft-max parameterization and the truncated horizon.

Lemma F.3 *Consider the restricted soft-max parameterization with the entropy regularization*

$$\begin{aligned}\tilde{J}_\rho(\theta) &= J_\rho(\theta) + \lambda\mathbb{H}(\rho,\theta), \\ \tilde{J}_\rho^H(\theta) &= J_\rho^H(\theta) + \lambda\mathbb{H}^H(\rho,\theta),\end{aligned}$$

where

$$\begin{aligned}\mathbb{H}(\rho,\theta) &\coloneqq \mathbb{E}_{s_0\sim\rho,a_t\sim\pi_\theta(\cdot|s_t),s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)}\left[\sum_{t=0}^{\infty}-\gamma^t\log\pi_\theta(a_t|s_t)\right], \\ \mathbb{H}^H(\rho,\theta) &\coloneqq \mathbb{E}_{s_0\sim\rho,a_t\sim\pi_\theta(\cdot|s_t),s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)}\left[\sum_{t=0}^{H-1}-\gamma^t\log\pi_\theta(a_t|s_t)\right].\end{aligned}$$

The functions $\tilde{J}_\rho(\theta)$ and $\tilde{J}_\rho^H(\theta)$ are $\left(\frac{(12+8\log|\mathcal{A}|)}{(1-\gamma)^3}\right)$-smooth.

Proof. In light of Lemma F.2, $J_\rho(\theta)$ and $J_\rho^H(\theta)$ are both $\frac{8}{(1-\gamma)^3}$-smooth. It remains to prove the smoothness of $\mathbb{H}^H(s,\theta)$.

Define $\mathbb{H}^H_{\pi_{\theta_\alpha}}(s) \coloneqq \mathbb{H}^H(s,\theta)$ and $\theta_\alpha \coloneqq \theta+\alpha u$, where $\alpha\in\mathbb{R}$ and $u\in\mathbb{R}^{(|\mathcal{A}|-1)|\mathcal{S}|}$. One can write

$$\begin{aligned}\mathbb{H}^H_{\pi_{\theta_\alpha}}(s) =& \mathbb{E}_{s_0\sim s,a_t\sim\pi_{\theta_\alpha}(\cdot|s_t),s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)}\left[\sum_{t=0}^{H-1}-\gamma^t\log\pi_{\theta_\alpha}(a_t|s_t)\right] \\ =& e_s^\top M^H(\alpha)h_{\theta_\alpha},\end{aligned}$$

where $M^H(\alpha) = \sum_{h=0}^{H-1} \gamma^h P(\alpha)^h$, $[P(\alpha)]_{(s\to s')} = \sum_a \pi_{\theta_\alpha}(a|s)\mathcal{P}(s'|s,a)$ and $h_{\theta_\alpha}(s) = -\sum_a \pi_{\theta_\alpha}(a|s)\log \pi_{\theta_\alpha}(a|s)$, for all $s \in \mathcal{S}$. It results from Lemma 7 in [30] that

$$\left\| \frac{\partial P(\alpha)}{\partial \alpha}\Big|_{\alpha=0} x \right\|_\infty \le 2\|u\|_2 \|x\|_\infty, \quad \left\| \frac{\partial^2 P(\alpha)}{\partial \alpha^2}\Big|_{\alpha=0} x \right\|_\infty \le 6\|u\|_2 \|x\|_\infty. \tag{39}$$

Since $h_{\theta_\alpha}(s) \in [0, \log|\mathcal{A}|]$, we have

$$\|h_{\theta_\alpha}\|_\infty = \max_s |h_{\theta_\alpha}(s)| \le |\mathcal{A}|. \tag{40}$$

For any state $s \in \mathcal{S}$, it holds that

$$\begin{aligned}
\left|\frac{\partial h_{\theta_\alpha}(s)}{\partial \alpha}\right| &= \left|\langle \frac{\partial h_{\theta_\alpha}(s)}{\partial \theta_\alpha}, \frac{\partial \theta_\alpha}{\partial \alpha}\rangle\right| \\
&= \left|\langle \frac{\partial h_{\theta_\alpha}(s)}{\partial \theta_\alpha(\cdot|s)}, u(s,\cdot)\rangle\right| \\
&= \left|(H^-(\pi_{\theta_\alpha}(\cdot|s))\log \pi_{\theta_\alpha}(\cdot|s))^\top u(s,\cdot)\right| \\
&\le \|(H^-(\pi_{\theta_\alpha}(\cdot|s))\log \pi_{\theta_\alpha}(\cdot|s))\|_1 \cdot \|u(s,\cdot)\|_\infty .
\end{aligned}$$

The ℓ_1 norm is upper bounded as

$$\begin{aligned}
\|(H^-(\pi_{\theta_\alpha}(\cdot|s))\log \pi_{\theta_\alpha}(\cdot|s))\|_1 &\le \sum_a \pi_{\theta_\alpha}(a|s)\cdot \left|\log \pi_{\theta_\alpha}(a|s) - \pi_{\theta_\alpha}(\cdot|s)^\top \log \pi_{\theta_\alpha}(\cdot|s)\right| \\
&\le \sum_a \pi_{\theta_\alpha}(a|s)\cdot \left(|\log \pi_{\theta_\alpha}(a|s)| + \left|\pi_{\theta_\alpha}(\cdot|s)^\top \log \pi_{\theta_\alpha}(\cdot|s)\right|\right) \\
&\le -2\sum_a \pi_{\theta_\alpha}(a|s)\cdot \log \pi_{\theta_\alpha}(a|s) \\
&\le 2\log|\mathcal{A}|.
\end{aligned} \tag{41}$$

Therefore,

$$\begin{aligned}
\left\|\frac{\partial h_{\theta_\alpha}(s)}{\partial \alpha}\right\|_\infty &= \max_s \left|\frac{\partial h_{\theta_\alpha}(s)}{\partial \alpha}\right| \\
&\le \max_s \|(H^-(\pi_{\theta_\alpha}(\cdot|s))\log \pi_{\theta_\alpha}(\cdot|s))\|_1 \cdot \|u(s,\cdot)\|_\infty \\
&\le 2\cdot \log|\mathcal{A}| \cdot \|u\|_2 .
\end{aligned} \tag{42}$$

The second derivative with respect to α is

$$\begin{aligned}
\left|\frac{\partial^2 h_{\theta_\alpha}(s)}{\partial \alpha^2}\right| &= \left|\left(\frac{\partial}{\partial \theta_\alpha}\left\{\frac{\partial h_{\theta_\alpha}(s)}{\partial \alpha}\right\}\right)^\top \frac{\partial \theta_\alpha}{\partial \alpha}\right| \\
&= \left|\left(\frac{\partial^2 h_{\theta_\alpha}(s)}{\partial \theta_\alpha^2}\frac{\partial \theta_\alpha}{\partial \alpha}\right)^\top \frac{\partial \theta_\alpha}{\partial \alpha}\right| \\
&= \left|u(s,\cdot)^\top \frac{\partial^2 h_{\theta_\alpha}(s)}{\partial \theta_\alpha^2(s,\cdot)} u(s,\cdot)\right|.
\end{aligned}$$

Denote the Hessian $T(s,\theta_\alpha) = \frac{\alpha^2 h_{\theta_\alpha}(s)}{\partial \theta_\alpha^2(s,\cdot)}$. Then,

$$\begin{aligned}
T(s,\theta_\alpha) &= \frac{\alpha^2 h_{\theta_\alpha}(s)}{\partial \theta_\alpha^2(s,\cdot)} \\
&= \frac{\partial}{\partial \theta_\alpha(s,\cdot)}\left\{\frac{\partial h_{\theta_\alpha}(s)}{\partial \theta_\alpha(s,\cdot)}\right\} \\
&= \frac{\partial}{\partial \theta_\alpha(s,\cdot)}\left\{\left(\frac{\partial \pi_{\theta_\alpha}(\cdot|s)}{\partial \theta_\alpha(s,\cdot)}\right)^\top \frac{\partial h_{\theta_\alpha}(s)}{\partial \pi_{\theta_\alpha}(s,\cdot)}\right\} \\
&= \frac{\partial}{\partial \theta_\alpha(s,\cdot)}\left\{H^-(\pi_{\theta_\alpha}(\cdot|s))(-\log \pi_{\theta_\alpha}(\cdot|s))\right\}.
\end{aligned}$$

Note that $T(s,\theta_\alpha) \in \mathbb{R}^{|\mathcal{A}-1||\mathcal{A}-1|}$. For all $i,j \in \mathcal{A}/a_1$, we have

$$\begin{aligned}
T_{i,j} =& \frac{d\{\pi_{\theta_\alpha}(i|s)\cdot(-\log\pi_{\theta_\alpha}(i|s)-h_{\theta_\alpha}(s))\}}{d\theta_\alpha(s,j)}\\
=& \frac{d\pi_{\theta_\alpha}(i|s)}{d\theta_\alpha(s,j)}(-\log\pi_{\theta_\alpha}(i|s)-h_{\theta_\alpha}(s)+\pi_{\theta_\alpha}(i|s))\cdot\frac{d\{(-\log\pi_{\theta_\alpha}(i|s)-h_{\theta_\alpha}(s))\}}{d\theta_\alpha(s,j)}\\
=&(\delta_{i,j}\pi_{\theta_\alpha}(j|s)-\pi_{\theta_\alpha}(i|s)\pi_{\theta_\alpha}(j|s))\cdot(-\log\pi_{\theta_\alpha}(i|s)-h_{\theta_\alpha}(s))\\
&+\pi_{\theta_\alpha}(i|s)\left(-\frac{1}{\pi_{\theta_\alpha}(i|s)}\cdot(\delta_{i,j}\pi_{\theta_\alpha}(j|s)-\pi_{\theta_\alpha}(i|s)\pi_{\theta_\alpha}(i|s)-\pi_{\theta_\alpha}(j|s)\cdot(-\log\pi_{\theta_\alpha}(j|s)-h_{\theta_\alpha}(s)))\right)\\
=&\delta_{i,j}\pi_{\theta_\alpha}(j|s)\cdot(-\log\pi_{\theta_\alpha}(i|s)-h_{\theta_\alpha}(s)-1)-\pi_{\theta_\alpha}(i|s)\pi_{\theta_\alpha}(j|s)\cdot(-\log\pi_{\theta_\alpha}(i|s)-h_{\theta_\alpha}(s)-1)\\
&-\pi_{\theta_\alpha}(i|s)\pi_{\theta_\alpha}(j|s)\cdot(-\log-\pi_{\theta_\alpha}(j|s)-h_{\theta_\alpha}(s)).
\end{aligned}$$

For every vector $y \in \mathbb{R}^{|\mathcal{A}|}$,

$$\begin{aligned}
\left|y^\top T(s,\theta_\alpha)\,y\right| =& \left|\sum_{i=1}^{|\mathcal{A}|-1}\sum_{j=1}^{|\mathcal{A}|-1} T_{i,j}y(i)y(j)\right|\\
\le& \left|\sum_i \pi_{\theta_\alpha}(i\mid s)\cdot(-\log\pi_{\theta_\alpha}(i\mid s)-h_{\theta_\alpha}(s)-1)\cdot y(i)^2\right|\\
&+2\cdot\left|\sum_i \pi_{\theta_\alpha}(i\mid s)\cdot y(i)\sum_j \pi_{\theta_\alpha}(j\mid s)\cdot(-\log\pi_{\theta_\alpha}(j\mid s)-h_{\theta_\alpha}(s))\cdot y(j)\right|+\left(\pi_{\theta_\alpha}(\cdot\mid s)^\top y\right)^2\\
=&\left(H^-\left(\pi_{\theta_\alpha}(\cdot\mid s)\right)\left(-\log\pi_{\theta_\alpha}(\cdot\mid s)\right)-\pi_{\theta_\alpha}(\cdot\mid s)\right)^\top(y\odot y)\mid\\
&+2\cdot\left|\left(\pi_{\theta_\alpha}(\cdot\mid s)^\top y\right)\cdot\left(H^-\left(\pi_{\theta_\alpha}(\cdot\mid s)\right)\left(-\log\pi_{\theta_\alpha}(\cdot\mid s)\right)\right)^\top y\right|+\left(\pi_{\theta_\alpha}(\cdot\mid s)^\top y\right)^2\\
\le&\|H^-\left(\pi_{\theta_\alpha}(\cdot\mid s)\right)\left(-\log\pi_{\theta_\alpha}(\cdot\mid s)\right)\|_\infty\cdot\|y\odot y\|_1+\|\pi_{\theta_\alpha}(\cdot\mid s)\|_\infty\cdot\|y\odot y\|_1\\
&+2\cdot\|\pi_{\theta_\alpha}(\cdot\mid s)\|_1\cdot\|y\|_\infty\cdot\|H^-\left(\pi_{\theta_\alpha}(\cdot\mid s)\right)\left(-\log\pi_{\theta_\alpha}(\cdot\mid s)\right)\|_1\cdot\|y\|_\infty+\|\pi_{\theta_\alpha}(\cdot\mid s)\|_2^2\cdot\|y\|_2^2,
\end{aligned}$$

where the last inequality is due to Hölder's inequality and $\odot$ is Hadamard (component-wise) product. Note that $\|y\odot y\|_1 = \|y\|_2^2, \|\pi_{\theta_\alpha}(\cdot\mid s)\|_\infty \le \|\pi_{\theta_\alpha}(\cdot\mid s)\|_1$ and $\|\pi_{\theta_\alpha}(\cdot\mid s)\|_2 \le \|\pi_{\theta_\alpha}(\cdot\mid s)\|_1 = 1$. The ℓ_∞ norm is upper bounded as

$$\begin{aligned}
\|H^-\left(\pi_{\theta_\alpha}(\cdot\mid s)\right)\left(-\log\pi_{\theta_\alpha}(\cdot\mid s)\right)\|_\infty \le& \max_a\left|\pi_{\theta_\alpha}(a\mid s)\cdot\left(-\log\pi_{\theta_\alpha}(a\mid s)+\pi_{\theta_\alpha}(\cdot\mid s)^\top\log\pi_{\theta_\alpha}(\cdot\mid s)\right)\right|\\
\le& \max_a -\pi_{\theta_\alpha}(a\mid s)\cdot\log\pi_{\theta_\alpha}(a\mid s)-\pi_{\theta_\alpha}(\cdot\mid s)^\top\log\pi_{\theta_\alpha}(\cdot\mid s)\\
\le& \frac{1}{e}+\log A, \qquad (43)
\end{aligned}$$

where the last inequality is due to $-x\cdot\log x \le \frac{1}{e}$ for all $x\in[0,1]$. Therefore,

$$\begin{aligned}
\left|y^\top T(s,\theta_\alpha)\,y\right| \le& \|H^-\left(\pi_{\theta_\alpha}(\cdot\mid s)\right)\left(-\log\pi_{\theta_\alpha}(\cdot\mid s)\right)\|_\infty\cdot\|y\|_2^2\\
&+\|y\|_2^2+2\cdot\|H^-\left(\pi_{\theta_\alpha}(\cdot\mid s)\right)\left(-\log\pi_{\theta_\alpha}(\cdot\mid s)\right)\|_1\cdot\|y\|_2^2+\|y\|_2^2\\
\le&\left(\frac{1}{e}+\log A+2\right)\cdot\|y\|_2^2+2\cdot\|H^-\left(\pi_{\theta_\alpha}(\cdot\mid s)\right)\left(-\log\pi_{\theta_\alpha}(\cdot\mid s)\right)\|_1\cdot\|y\|_2^2\\
\le&\left(\frac{1}{e}+\log A+2+2\cdot\log A\right)\cdot\|y\|_2^2\\
\le&3\cdot(1+\log A)\cdot\|y\|_2^2.
\end{aligned}$$

where the second inequality is due to (43) and the third inequality is due to (41). Thus,

$$\begin{aligned}\left\|\frac{\partial^2 h_{\theta_\alpha}}{\partial \alpha^2}\right\|_\infty &= \max_s \left|\frac{\partial^2 h_{\theta_\alpha}(s)}{\partial \alpha^2}\right| \\ &= \max_s \left|u(s,\cdot)^\top \frac{\partial^2 h_{\theta_\alpha}(s)}{\partial \theta_\alpha^2(s,\cdot)} u(s,\cdot)\right| \\ &= \max_s \left|u(s,\cdot)^\top T\left(s,\theta_\alpha\right) u(s,\cdot)\right| \\ &\le 3\cdot(1+\log A)\cdot \max_s \|u(s,\cdot)\|_2^2 \\ &\le 3\cdot(1+\log A)\cdot \|u\|_2^2. \end{aligned} \tag{44}$$

Taking the derivative of $\mathbb{H}^H_{\pi_{\theta_\alpha}}(s)$ with respect to α yields that

$$\frac{\partial \mathbb{H}^H_{\pi_{\theta_\alpha}}(s)}{\partial \alpha} = e_s^\top \sum_{h=0}^{H-1} \gamma^h h P(\alpha)^{h-1} \frac{\partial P(\alpha)}{\partial \alpha} h_{\theta_\alpha} + e_s^\top M^H(\alpha) \frac{\partial h_{\theta_\alpha}}{\partial \alpha}.$$

Taking the second derivative with respect to α gives rise to

$$\begin{aligned}\frac{\partial^2 \mathbb{H}^H_{\pi_{\theta+\alpha}}(s)}{\partial \alpha^2} =& e_s^\top \sum_{h=1}^{H-1} \gamma^h h(h-1) P(\alpha)^{h-2} \frac{\partial P(\alpha)}{\partial \alpha}\frac{\partial P(\alpha)}{\partial \alpha} h_{\theta_\alpha} + e_s^\top \sum_{h=0}^{H-1} \gamma^h h P(\alpha)^{h-1} \frac{\partial^2 P(\alpha)}{\partial \alpha^2} h_{\theta_\alpha} \\ &+ e_s^\top \sum_{h=0}^{H-1} \gamma^h h P(\alpha)^{h-1} \frac{\partial P(\alpha)}{\partial \alpha}\frac{\partial h_{\theta_\alpha}}{\partial \alpha} + e_s^\top M^H(\alpha) \frac{\partial^2 h_{\theta_\alpha}}{\partial \alpha^2}.\end{aligned}$$

Since $\tilde{P}(\alpha)^h \ge 0$ (component-wise) and $\tilde{P}(\alpha)^h \mathbf{1} \le \mathbf{1}$ for all $h \ge 0$, i.e., each row of $\tilde{P}(\alpha)^h$ is positive and sums to 1, it holds that

$$\max_{\|u\|_2=1} \left\|M^H(\alpha)x\right\|_\infty \le \frac{1}{1-\gamma}\|x\|_\infty.$$

Therefore, for the last term,

$$\begin{aligned}\left|e_s^\top M^H(\alpha)\frac{\partial^2 h_{\theta_\alpha}}{\partial \alpha^2}\Big|_{\alpha=0}\right| &\le \|e_s\|_1 \cdot \left\| M^H(\alpha)\frac{\partial^2 h_{\theta_\alpha}}{\partial \alpha^2}\Big|_{\alpha=0}\right\|_\infty \\ &\le \frac{1}{1-\gamma}\cdot \left\|\frac{\partial^2 h_{\theta_\alpha}}{\partial \alpha^2}\Big|_{\alpha=0}\right\|_\infty \\ &\le \frac{3\cdot(1+\log A)}{1-\gamma}\cdot \|u\|_2^2.\end{aligned}$$

where the last inequality is due to (44). For the second last term,

$$\begin{aligned}\left|e_s^\top \sum_{h=0}^{H-1} \gamma^h h P(\alpha)^{h-1}\frac{\partial P(\alpha)}{\partial \alpha}\frac{\partial h_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0}\right| &\le \left\|\sum_{h=0}^{H-1} \gamma^h h P(\alpha)^{h-1}\frac{\partial P(\alpha)}{\partial \alpha}\frac{\partial h_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0}\right\|_\infty \\ &\le \Big(\sum_{h=0}^{H-1}\gamma^h h\Big)\left\|P(\alpha)^{h-1}\frac{\partial P(\alpha)}{\partial \alpha}\frac{\partial h_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0}\right\|_\infty \\ &\le \frac{\gamma}{(1-\gamma)^2}\cdot \left\|P(\alpha)^{h-1}\frac{\partial P(\alpha)}{\partial \alpha}\frac{\partial h_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0}\right\|_\infty \\ &\le \frac{\gamma}{(1-\gamma)^2}\cdot \left\|\frac{\partial P(\alpha)}{\partial \alpha}\frac{\partial h_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0}\right\|_\infty \\ &\le \frac{2\gamma\cdot\|u\|_2}{(1-\gamma)^2}\cdot \left\|\frac{\partial h_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0}\right\|_\infty \\ &\le \frac{2\gamma\cdot\|u\|_2}{(1-\gamma)^2}\cdot 2\cdot\log A\cdot\|u\|_2 = \frac{4\gamma\cdot\log A}{(1-\gamma)^2}\cdot\|u\|_2^2,\end{aligned}$$

where the fifth inequality is due to (39) and the last inequality is due to (42). For the second term,

$$\begin{aligned}
\left| e_s^\top \sum_{h=0}^{H-1} \gamma^h h P(\alpha)^{h-1} \frac{\partial^2 P(\alpha)}{\partial \alpha^2} h_{\theta_\alpha} \Big|_{\alpha=0} \right| &\le \left\| \sum_{h=0}^{H-1} \gamma^h h P(\alpha)^{h-1} \frac{\partial^2 P(\alpha)}{\partial \alpha^2} h_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty \\
&\le (\sum_{h=0}^{H-1} \gamma^h h) \cdot \left\| P(\alpha)^{h-1} \frac{\partial^2 P(\alpha)}{\partial \alpha^2} h_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty \\
&\le \frac{\gamma}{(1-\gamma)^2} \cdot \left\| P(\alpha)^{h-1} \frac{\partial^2 P(\alpha)}{\partial \alpha^2} h_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty \\
&\le \frac{\gamma}{(1-\gamma)^2} \cdot \left\| \frac{\partial^2 P(\alpha)}{\partial \alpha^2} h_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty \\
&\le \frac{6\gamma \cdot \|u\|_2^2}{(1-\gamma)^2} \cdot \left\| h_{\theta_\alpha}|_{\alpha=0} \right\|_\infty \\
&\le \frac{6\gamma \cdot \log A}{(1-\gamma)^2} \cdot \|u\|_2^2,
\end{aligned}$$

where the fifth inequality is due to (39) and the last inequality is due to (40). For the first term,

$$\begin{aligned}
\left| e_s^\top \sum_{h=1}^{H-1} \gamma^h h(h-1) P(\alpha)^{h-2} \frac{\partial P(\alpha)}{\partial \alpha} \frac{\partial P(\alpha)}{\partial \alpha} h_{\theta_\alpha} \Big|_{\alpha=0} \right| &\le \left\| \sum_{h=1}^{H-1} \gamma^h h(h-1) P(\alpha)^{h-2} \frac{\partial P(\alpha)}{\partial \alpha} \frac{\partial P(\alpha)}{\partial \alpha} h_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty \\
&\le \sum_{h=1}^{H-1} \gamma^h h(h-1) \left\| P(\alpha)^{h-2} \frac{\partial P(\alpha)}{\partial \alpha} \frac{\partial P(\alpha)}{\partial \alpha} h_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty \\
&\le \frac{2\gamma^2}{(1-\gamma)^3} \left\| P(\alpha)^{h-2} \frac{\partial P(\alpha)}{\partial \alpha} \frac{\partial P(\alpha)}{\partial \alpha} h_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty \\
&\le \frac{2\gamma^2}{(1-\gamma)^3} \left\| \frac{\partial P(\alpha)}{\partial \alpha} \frac{\partial P(\alpha)}{\partial \alpha} h_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty \\
&\le \frac{8 \log |\mathcal{A}| \gamma^2 \|u\|_2^2}{(1-\gamma)^3},
\end{aligned}$$

where the forth inequality is due to (39). Combining the above results, we obtain that

$$\begin{aligned}
\left| \frac{\partial^2 \mathbb{H}_{\pi_{\theta_\alpha}}^H(s)}{\partial \alpha^2}|_{\alpha=0} \right| &\le \left(2\gamma^2 \cdot \frac{4 \cdot \log A}{(1-\gamma)^3} + \gamma \cdot \frac{6 \cdot \log A}{(1-\gamma)^2} + \gamma \cdot \frac{4 \cdot \log A}{(1-\gamma)^2} + \frac{3 \cdot (1 + \log A)}{1-\gamma} \right) \cdot \|u\|_2^2 \\
&\le \left(\frac{8 \cdot \log A}{(1-\gamma)^3} + \frac{3}{1-\gamma} \right) \cdot \|u\|_2^2 \\
&\le \frac{4 + 8 \cdot \log A}{(1-\gamma)^3} \cdot \|u\|_2^2,
\end{aligned}$$

which implies the following relations for all $y \in \mathbb{R}^{|\mathcal{S}|\cdot(|\mathcal{A}|-1)}$ and θ:

$$\begin{aligned}
\left| y^\top \frac{\partial^2 \mathbb{H}^H_{\pi_{\theta_\alpha}}(s)}{\partial \theta^2} y \right| &= \left| \left(\frac{y}{\|y\|_2} \right)^\top \frac{\partial^2 \mathbb{H}^H_{\pi_{\theta_\alpha}}(s)}{\partial \theta^2} \left(\frac{y}{\|y\|_2} \right) \right| \cdot \|y\|_2^2 \\
&\leq \max_{\|u\|_2=1} \left| \left\langle \frac{\partial^2 \mathbb{H}^H_{\pi_{\theta_\alpha}}(s)}{\partial \theta^2} u, u \right\rangle \right| \cdot \|y\|_2^2 \\
&= \max_{\|u\|_2=1} \left| \left\langle \frac{\partial^2 \mathbb{H}^H_{\pi_{\theta_\alpha}}(s)}{\partial \theta_\alpha^2} \Big|_{\alpha=0} \frac{\partial \theta_\alpha}{\partial \alpha}, \frac{\partial \theta_\alpha}{\partial \alpha} \right\rangle \right| \cdot \|y\|_2^2 \\
&= \max_{\|u\|_2=1} \left| \left\langle \frac{\partial}{\partial \theta_\alpha} \left\{ \frac{\partial \mathbb{H}^H_{\pi_{\theta_\alpha}}(s)}{\partial \alpha} \right\} \Big|_{\alpha=0}, \frac{\partial \theta_\alpha}{\partial \alpha} \right\rangle \right| \cdot \|y\|_2^2 \\
&= \max_{\|u\|_2=1} \left| \frac{\partial^2 \mathbb{H}^H_{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2} \Big|_{\alpha=0} \right| \cdot \|y\|_2^2 \\
&\leq \frac{4 + 8 \cdot \log A}{(1-\gamma)^3} \cdot \|y\|_2^2.
\end{aligned}$$

This proves the smoothness of $\mathbb{H}^H(s,\theta)$. The smoothness of $\mathbb{H}(s,\theta)$ follows by observing that $\mathbb{H}^\infty(s,\theta) = \mathbb{H}(s,\theta)$. □

G Related work

Momentum-based policy gradient. Conventional approaches to reducing the high variance in PG methods include adding the baselines [46, 49] and using the actor-critic algorithms [23, 7, 33]. The idea of variance reduction, inspired by its successes in the stochastic nonconvex optimization [18, 3, 36, 12, 31], is also incorporated to improve the PG methods [57, 32, 54]. In addition, momentum techniques, which are demonstrated as a powerful and generic recipe for accelerating stochastic gradient methods for nonconvex optimization [35, 22, 37], have also been extended to improve PG methods both in theory and in practice [52, 59, 34, 17]. A new STORM-PG method is proposed in [59], which incorporates momentum in the update and matches the sample complexity of the SRVR-PG method proposed in [54]. In addition, [34] applies the idea of SARAH and considers a more general setting with regularization. Finally, [17] proposes a class of momentum-based policy gradient algorithms, with the adaptive time-step and the small batch size, that matches the sample complexity in [54]. However, the 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 bad performance, in contrast to the more desired global convergence guarantees studied in the current work.

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 [46, 24, 32]. It was not until very recently that a series of works emerged to establish the global convergence properties of these algorithms. [13] 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. [5] generalizes the results in [13] 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, [30, 29] show that the value function satisfies a non-uniform Łojasiewicz inequality and the fast global convergence rate can be achieved if the exact PG is available. In addition, [2] 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 [27] for both PG and natural PG methods. When overparameterized neural networks is used for function approximation, the global convergence are proved for the (natural) PG methods [47] and for the trust-region policy optimization [26]. Very recently, a series of non-asymptotic global convergence results [16, 55, 51, 56, 14] 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 [60, 61].

Policy gradient for maximum entropy reinforcement learning Maximum entropy reinforcement learning introduces a non-greedy entropy regularization term to improve the learned policies in terms of deep exploration, pretraining performance, robustness to disturbances, etc. [15, 38, 10, 11]. Recently, the global convergence and linear convergence rate of policy gradient methods for the maximum entropy RL with the soft-max parameterization are shown in [30]. In [8], it is shown that the natural policy gradient methods could also achieve a linear global convergence rate and even a quadratic convergence rate in the small-ϵ region with a clear dependence on the salient parameters, such as $|\mathcal{S}|, |\mathcal{A}|, \gamma, \lambda$. However, the analyses in those works rely heavily on the assumption of access to exact and deterministic policy gradients. It remains unclear whether the power of the entropy regularization is still maintained when a practical trajectory-based policy gradient estimator is used to estimate the policy gradient.