Accelerated Policy Gradient: On the Convergence Rates of the Nesterov Momentum for Reinforcement Learning

Yen-Ju Chen [* 1]   Nai-Chieh Huang [* 1]   Ching-pei Lee [2]   Ping-Chun Hsieh [1]

Abstract

Various acceleration approaches for Policy Gradient (PG) have been analyzed within the realm of Reinforcement Learning (RL). However, the theoretical understanding of the widely used momentum-based acceleration method on PG remains largely open. In response to this gap, we adapt the celebrated Nesterov's accelerated gradient (NAG) method to policy optimization in RL, termed *Accelerated Policy Gradient* (APG). To demonstrate the potential of APG in achieving fast convergence, we formally prove that with the true gradient and under the softmax policy parametrization, APG converges to an optimal policy at rates: (i) $\tilde{O}(1/t^2)$ with nearly constant step sizes; (ii) $O(e^{-ct})$ with time-varying step sizes. To the best of our knowledge, this is the first characterization of the convergence rates of NAG in the context of RL. Notably, our analysis relies on one interesting finding: Regardless of the parameter initialization, APG ends up entering a locally nearly-concave regime, where APG can significantly benefit from the momentum, within finite iterations. Through numerical validation and experiments on the Atari 2600 benchmarks, we confirm that APG exhibits a $\tilde{O}(1/t^2)$ rate with nearly constant step sizes and a linear convergence rate with time-varying step sizes, significantly improving convergence over the standard PG.

1. Introduction

Policy gradient (PG) is a fundamental technique utilized in the field of reinforcement learning (RL) for policy optimization. It operates by directly optimizing the RL objectives to determine the optimal policy, employing first-order derivatives similar to the gradient descent algorithm in conventional optimization problems. Notably, PG has demonstrated empirical success (Mnih et al., 2016; Wang et al., 2016; Silver et al., 2014; Lillicrap et al., 2016; Schulman et al., 2017; Espeholt et al., 2018) and is supported by strong theoretical guarantees (Agarwal et al., 2021; Fazel et al., 2018; Liu et al., 2020; Bhandari & Russo, 2019; Mei et al., 2020; Wang et al., 2021; Mei et al., 2021a; 2022; Xiao, 2022; Chen & Maguluri, 2022; Lan, 2023). In a recent study, Mei et al. (2020) characterize the $O(1/t)$ convergence rate of PG in the non-regularized tabular softmax setting. This convergence behavior aligns with that of the gradient descent algorithm for convex minimization problems, despite that the RL objectives lack concave characteristics for maximization. Additionally, advancements in RL literature have introduced theories on improving the $O(1/t)$ convergence rate through various acceleration techniques. Khodadadian et al. (2021); Xiao (2022) offer sharp analyses of Natural Policy Gradient (NPG) with time-varying step sizes, demonstrating a notable linear convergence rate. On the other hand, Mei et al. (2021b) utilize the normalized gradient to accelerate the PG, also achieving a linear convergence rate.

In addition to the above-mentioned acceleration techniques, *momentum*, which is another popular acceleration method in the optimization literature, emerges as a widely adopted strategy for accelerating policy optimization in RL in practice due to its simplicity (Vieillard et al., 2020; Huang et al., 2020). Classic momentum methods, such as Nesterov momentum (Nesterov, 1983) and the heavy-ball method (Polyak, 1964), have been seamlessly integrated into optimization solvers and deep learning frameworks like PyTorch (Paszke et al., 2019). These approaches consistently outperform standard PG methods under A2C (Konda & Tsitsiklis, 1999) and PPO (Schulman et al., 2017), as evidenced by empirical studies (Henderson et al., 2018; Andrychowicz et al., 2021). Nesterov's Accelerated Gradient (NAG) method, introduced by Nesterov (1983), is a first-order approach

---

[*]Equal contribution [1]Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan [2]Department of Advanced Data Science, Institute of Statistical Mathematics, Tokyo, Japan. Correspondence to: Yen-Ju Chen <mru.11@nycu.edu.tw>, Nai-Chieh Huang <naich.cs09@nycu.edu.tw>, Ping-Chun Hsieh <pinghsieh@nycu.edu.tw>.

originally designed for convex minimization to improve the convergence rate to $O(1/t^2)$ from the $O(1/t)$ rate of gradient descent. Despite its wide applicability in various types of optimization problems over the past decades, to the best of our knowledge, NAG has never been theoretically analyzed in the context of RL for its convergence rate, mainly due to the non-concavity of RL objectives. Therefore, there exists one important and yet unexplored research question: *Could the Nesterov momentum achieve convergence rates better than or comparable to the existing PG and other acceleration methods in the context of RL?*

To answer this question, this paper introduces Accelerated Policy Gradient (APG), which utilizes Nesterov acceleration to address the policy optimization problem of RL. Despite the existing knowledge about the NAG methods from previous research (Beck & Teboulle, 2009a;b; Ghadimi & Lan, 2016; Li & Lin, 2015; Su et al., 2014; Carmon et al., 2018), there remain several fundamental challenges in establishing convergence rates within the realm of RL: (i) *NAG convergence results under nonconvex problems*: Although there is a plethora of theoretical works studying the convergence of NAG under general nonconvex problems, these results only establish convergence of first-order optimality measures. It is infeasible to characterize the convergence rate based solely on these results. (ii) *Inherent characteristics of the momentum term*: From an analytical perspective, the momentum term demonstrates intricate interactions with previous updates. As a result, accurately quantifying the specific impact of momentum during the execution of APG poses a considerable challenge. (iii) *The nature of the unbounded optimal parameter under softmax parameterization*: A crucial factor in characterizing the sub-optimality gap in the theory of optimization is the norm of the distance between the initial parameter and the optimal parameter (Beck & Teboulle, 2009a;b; Ghadimi & Lan, 2016). However, in the case of softmax parameterization in RL, the parameter of the optimal action of each state tends to approach infinity. As a result, the norm involved in the sub-optimality gap becomes infinite, thereby hindering characterization of convergence rates using existing results.

Our Contributions. Despite the above challenges, we present an affirmative answer to the research question described above and provide the first characterization of the convergence rate of NAG in the context of RL. Specifically, we present useful insights and novel techniques to tackle the above technical challenges: Regarding (i), we discover a fundamental property of the RL objective called *local near-concavity*, which indicates that the RL objective is nearly concave in the proximity of the optimal policy, despite its non-concave global landscape. For (ii), we show that the locally nearly-concave region is *absorbing* in the sense that even with the effect of the momentum term, the policy parameter could stay in the nearly-concave region in-

definitely once entered. This result is obtained by carefully quantifying the cumulative effect of each momentum term. Towards (iii), we introduce surrogate optimal parameters, which have bounded norms and induce nearly-optimal policies, and thereby characterize the convergence rate of APG. We summarize the contributions of this paper as follows:

- We introduce APG, which leverages the Nesterov momentum scheme to accelerate the convergence performance of PG for RL. To the best of our knowledge, this is the first formal attempt at understanding the convergence of the Nesterov momentum in the context of RL.

- To demonstrate the potential of APG in achieving fast convergence, we formally establish that APG enjoys a $\tilde{O}(1/t^2)$ convergence rate under softmax policy parameterization with nearly constant step sizes.[1] Furthermore, we offer that APG also achieves a linear convergence rate of $O(e^{-ct})$ if time-varying step sizes are allowed. These deliver better or comparable results in both constant step size and time-varying step size regimes. For ease of reference, a summarized comparison can be found in Table 1. To achieve this, we present several novel insights into RL and APG, including the local near-concavity property as well as the absorbing behavior of APG. Moreover, we show that these properties can also be applied to establish a $\tilde{O}(1/t)$ convergence rate of PG, which is of independent interest.

- Through numerical validation on MDP problems, we confirm that APG exhibits a $\tilde{O}(1/t^2)$ rate with nearly constant step sizes and a linear convergence rate with time-varying step sizes, thereby substantially improving the convergence behavior over the standard PG. Furthermore, we validate our APG algorithm on the benchmark Atari games and demonstrate the significant benefit provided by the Nesterov momentum.

2. Related Work

Policy Gradient. Policy gradient (Sutton et al., 1999) is a popular RL technique that directly optimizes the objective function by computing and using the gradient of the expected return with respect to the policy parameters. It has several popular variants, such as the REINFORCE algorithm (Williams, 1992), actor-critic methods (Konda & Tsitsiklis, 1999), trust region policy optimization (TRPO) (Schulman et al., 2015a), and proximal policy optimization (PPO) (Schulman et al., 2017). Recently, policy gradient methods have been shown to converge to the global optimal policy. Convergence properties of standard policy gradient methods under various settings have been proven by

---

[1]Note that this result does not contradict the $\Omega(1/t)$ lower bound of the sub-optimality gap of PG in (Mei et al., 2020). Please refer to Remark 10 for a detailed discussion.

Table 1. A summary of the convergence rates under tabular softmax policies under PG and various acceleration methods. Here NPG refers to Natural Policy Gradient (Kakade, 2001), and GNPG refers to Geometry-aware Normalized Policy Gradient (Mei et al., 2021b). For the sake of clarity, we use η_i's to denote the pre-constants in the step sizes of each algorithm.

Algorithm	Step Sizes	Gradient Update	Conv. Rate	Reference
APG	Nearly Constant[2]	$\omega^{(t)} + \eta_1 \cdot \frac{t}{t+1} \cdot \nabla_\theta V^{\pi_\omega^{(t)}}(\mu)$	$\tilde{O}(1/t^2)$	Theorem 2
PG	Constant	$\theta^{(t)} + \eta_2 \cdot \nabla_\theta V^{\pi_\theta}(\mu)$	$O(1/t)$	(Mei et al., 2020)
APG	Time-Varying	$\omega^{(t)} + \left[\eta_3 \cdot \beta^t \cdot \nabla_\theta V^{\pi_\omega^{(t)}}(\mu)\right]^{+K}$	$O(e^{-ct})$	Theorem 3
NPG	Time-Varying	$\theta^{(t)} + \eta_5 \cdot Q^{\pi_\theta^{(t)}}$	$O(e^{-ct})$	(Khodadadian et al., 2021; Mei et al., 2021a)
NPG	Time-Varying	$\theta^{(t)} + (\eta_6 + \log(1/\pi_\theta^{(t)})) \cdot Q^{\pi_\theta^{(t)}}$	$O(e^{-ct})$	(Khodadadian et al., 2021)
NPG [3]	Time-Varying	$\theta^{(t)} + \eta_7 \cdot \beta^t \cdot Q^{\pi_\theta^{(t)}}$	$O(e^{-ct})$	(Xiao, 2022)
GNPG	Time-Varying	$\theta^{(t)} + \eta_8 \cdot \left\|\nabla_\theta V^{\pi_\theta^{(t)}}(\mu)\right\|_2^{-1} \cdot \nabla_\theta V^{\pi_\theta^{(t)}}(\mu)$	$O(e^{-ct})$	(Mei et al., 2021b)

Agarwal et al. (2021). Furthermore, Mei et al. (2020) characterize a $O(1/t)$ convergence rate of policy gradient under non-regularized tabular softmax parameterization.

Acceleration Methods for RL. Numerous approaches aim to accelerate PG methods in RL. One notable approach is the Natural Policy Gradient (NPG), inspired by the natural gradient concept. NPG has been demonstrated to have: (i) a $O(1/t)$ rate with constant step sizes (Agarwal et al., 2021; Xiao, 2022); (ii) linear convergence with adaptively increasing step sizes (Khodadadian et al., 2021) or non-adaptive exponentially growing step sizes (Xiao, 2022). Gradient normalization is another effective method for accelerating PG. This approach involves normalizing the true policy gradient by its Euclidean norm. Mei et al. (2021b) propose Geometry-aware Normalized PG (GNPG) and demonstrate its linear convergence under softmax policies. For more details, please refer to Appendix J. To the best of our knowledge, the effects of momentum in the context of RL remain unclear. In this work, we utilize the Nesterov momentum to achieve substantial acceleration for PG. Please refer to Table 1 for a detailed comparison.

Nesterov's Accelerated Gradient. Accelerated gradient methods (Nesterov, 1983; Beck & Teboulle, 2009b; Su et al., 2014; Attouch & Peypouquet, 2016; Carmon et al., 2017; Jin et al., 2018) play a pivotal role in the optimization literature due to their ability to achieve convergence rates faster than that of the conventional gradient descent algorithm. Notably, in the convex minimization regime, accelerated gradient methods enjoy a convergence rate as fast as $O(1/t^2)$, surpassing the limited $O(1/t)$ rate offered by gradient descent. In order to further enhance the performance of accelerated gradient methods, several variants have been proposed. For instance, Beck & Teboulle (2009a) propose a variant of the accelerated gradient method with restart to achieve monotonicity in the objective value. Ghadimi & Lan (2016) present a unified analytical framework for a family of accelerated gradient methods that can be applied to solve

convex, non-convex, and stochastic optimization problems. Moreover, Li & Lin (2015) extend the restart approach of Beck & Teboulle (2009a) with monotonic objective values to further ensure sufficient objective improvement, in order to provide convergence guarantees on stationarity measures for non-convex problems. Other restart mechanisms have also been applied in multiple recent accelerated gradient methods (O'donoghue & Candes, 2015; Li & Lin, 2022). The above list of works is by no means exhaustive and is only meant to provide a brief overview of accelerated gradient methods. Our paper introduces APG, which combines Nesterov's accelerated gradient and PG methods for RL. This integration enables a substantial acceleration of the convergence rate compared to the standard PG method. It is essential to note that in RL problems, we are *maximizing* the objective function, which in general is neither convex nor concave. To connect with the literature about convex *minimization*, searching for properties akin to concavity becomes pivotal.

3. Preliminaries

Markov Decision Processes. For a finite set $\mathcal{X}$, we use $\Delta(\mathcal{X})$ to denote the probability simplex over $\mathcal{X}$. We consider that a finite Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma, \rho)$ is determined by: (i) a finite state space $\mathcal{S}$, (ii) a finite action space $\mathcal{A}^4$, (iii) a transition kernel $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, determining the transition probability $\mathcal{P}(s'|s, a)$ from each state-action pair (s, a) to the

---

[2]As $t/(t+1)$ is nearly constant for large t, we regard this design as constant step sizes, with a slight abuse of terminology.

[3]While (Xiao, 2022) examine the policy mirror descent (PMD), a generalized version of PG, under direct parameterization instead of softmax policies, it is known that by opting for KL-divergence as the Bregman divergence, the update rule of PMD can be reformulated in a manner akin to Natural Policy Gradient (NPG) expressed in the policy space.

[4]We consider the non-trivial case where $|\mathcal{A}| \geq 2$.

next state s', (iv) a reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, (v) a discount factor $\gamma \in [0, 1)$, and (vi) an initial state distribution $\rho \in \Delta(\mathcal{S})$. Without loss of generality, we presume the one-step reward to lie in the $[0, 1]$ interval. Given a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, the value of state s under π is defined as

$$V^\pi(s) := \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t r(s_t, a_t)\middle| \pi, s_0 = s\right]. \qquad (1)$$

We use $\boldsymbol{V}^\pi$ to denote the vector of $V^\pi(s)$ of all the states $s \in \mathcal{S}$. The goal of the learner (or agent) is to search for a policy that maximizes the following objective function as $V^\pi(\rho) := \mathbb{E}_{s \sim \rho}[V^\pi(s)]$. The Q-value (or action-value) and the advantage function of π at $(s, a) \in \mathcal{S} \times \mathcal{A}$ are defined as

$$\begin{aligned} Q^\pi(s, a) &:= r(s, a) + \gamma \sum_{s'} \mathcal{P}(s'|s, a) V^\pi(s'), \\ A^\pi(s, a) &:= Q^\pi(s, a) - V^\pi(s), \end{aligned} \qquad (2)$$

where the advantage function reflects the relative benefit of taking the action a at state s under policy π. The (discounted) state visitation distribution of π is defined as

$$d_{s_0}^\pi(s) := (1 - \gamma) \sum_{t=0}^\infty \gamma^t Pr(s_t = s|s_0, \pi, \mathcal{P}),$$

which reflects how frequently the learner would visit the state s under policy π, and we let $d_\rho^\pi(s) := \mathbb{E}_{s_0 \sim \rho}\left[d_{s_0}^\pi(s)\right]$ be the expected state visitation distribution under the initial state distribution ρ. Given ρ, there exists an optimal policy π^* such that[5]

$$V^{\pi^*}(\rho) = \max_{\pi : \mathcal{S} \to \Delta(\mathcal{A})} V^\pi(\rho). \qquad (3)$$

We denote $V^*(\rho) := V^{\pi^*}(\rho)$ and $Q^*(s, a) := Q^{\pi^*}(s, a)$. Additionally, we make the following assumption:

Assumption 1 (**Unique Optimal Action**). *There is a unique optimal action $a^*(s)$ for each state $s \in \mathcal{S}$.*

Surrogate Initial State Distribution. While obtaining the true initial state distribution ρ can be challenging in practice, it is fortunate that this issue can be addressed by considering other *surrogate* initial state distribution μ. Notably, we will show in the following theoretical proofs that even in the absence of knowledge about ρ, convergence guarantees of APG for $V^*(\rho)$ can still be obtained for almost every $\mu \in \Delta(\mathcal{S})$. This type of result bears high-level resemblance to those of gradient-based methods for general non-convex optimization (Lee et al., 2019; O'Neill & Wright, 2019). In this paper, we consider that μ is drawn uniformly at random from the $|\mathcal{S}|$-dimensional probability simplex.

---

[5]See, for example, Theorem 1.7 in (Agarwal et al., 2019) for the existence of such an optimal policy.

Softmax Parameterization. For unconstrained $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, the tabular softmax policy parameterization is defined as

$$\pi_\theta(\cdot|s) := \frac{\exp(\theta_{s,\cdot})}{\sum_{a' \in \mathcal{A}} \exp(\theta_{s,a'})}. \qquad (4)$$

The term "tabular" means that we store a parameter table $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ for each state-action pairs, and sometimes we would omit it and call (4) softmax parameterization.

Smoothness. As shown in prior works (Agarwal et al., 2021; Mei et al., 2020), the RL objective function $V^{\pi_\theta}(\rho)$ enjoys smoothness under softmax parameterization. Specifically, the objective function $\theta \to V^{\pi_\theta}(\rho)$ is $\frac{8}{(1-\gamma)^3}$-smooth.

Policy Gradient. In policy gradient methods, the parameters are updated by the gradient of $V^{\pi_\theta}(\mu)$ with respect to θ under a surrogate initial state distribution $\mu \in \Delta(\mathcal{S})$. Algorithm 1 presents the pseudo code of PG provided by (Mei et al., 2020).

Algorithm 1 Policy Gradient (PG) in (Mei et al., 2020)

Input: Step size $\eta = \frac{1}{L}$, where L is the Lipschitz constant of the gradient of the objective function $V^{\pi_\theta}(\mu)$.
Initialize: $\theta^{(1)}(s, a)$ for all (s, a).
for $t = 1$ to T **do**

$$\theta^{(t+1)} \leftarrow \theta^{(t)} + \eta \nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta = \theta^{(t)}} \qquad (5)$$

end for

Nesterov's Accelerated Gradient (NAG). Nesterov's Accelerated Gradient (NAG) (Nesterov, 1983) is an optimization algorithm that utilizes a variant of momentum known as Nesterov's momentum to expedite the convergence rate. Specifically, it computes an intermediate "lookahead" estimate of the gradient by evaluating the objective function at a point slightly ahead of the current estimate. We provide the pseudo code of NAG method as Algorithm 4 in Appendix A.

Remark 1 (**Step-Size Regimes**). Please note that the step-size regimes are categorized in Table 1 based on the gradient update presented in (5). Under the softmax parameterization, the update $\theta^{(t)} + \eta_5 \cdot Q^{\pi_\theta^{(t)}}$ is equivalent to $\theta^{(t)} + \eta_5 \cdot A^{\pi_\theta^{(t)}}$, as $A^{\pi_\theta^{(t)}}$ is derived by adding an action-independent value $V^{\pi_\theta^{(t)}}$ to $Q^{\pi_\theta^{(t)}}$. According to Lemma 11, it is evident that the step size is implicitly related to $1/\pi_\theta^{(t)}$ when being compared with (5). Therefore, we consider the NPG with the update $\theta^{(t)} + \eta_5 \cdot Q^{\pi_\theta^{(t)}}$ as an algorithm with time-varying step sizes.

Notations. Throughout the paper, we use $\|x\|$ to denote the L_2 norm of a real vector x.

4. Methodology

In this section, we present our proposed algorithm, Accelerated Policy Gradient (APG), which integrates Nesterov

acceleration with gradient-based reinforcement learning algorithms. In Section 4.1, we introduce our central algorithm, APG. Subsequently, in Section 4.2, we provide a motivating example in the bandit setting to illustrate the convergence behavior of APG.

Algorithm 2 Accelerated Policy Gradient (APG)

Input: Step size $\eta^{(t)} > 0$.
Initialize: $\theta^{(0)} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $\omega^{(0)} = \theta^{(0)}$.
for $t = 1$ to T **do**

$$\theta^{(t)} \leftarrow \omega^{(t-1)} + \eta^{(t)} \nabla_\theta V^{\pi_\theta}(\mu) \Big|_{\theta = \omega^{(t-1)}} \qquad (6)$$

$$\varphi^{(t)} \leftarrow \theta^{(t)} + \frac{t-1}{t+2}(\theta^{(t)} - \theta^{(t-1)}) \qquad (7)$$

$$\omega^{(t)} \leftarrow \begin{cases} \varphi^{(t)}, & \text{if } V^{\pi_\varphi^{(t)}}(\mu) \geq V^{\pi_\theta^{(t)}}(\mu), \\ \theta^{(t)}, & \text{otherwise.} \end{cases} \qquad (8)$$

end for

4.1. Accelerated Policy Gradient

We propose Accelerated Policy Gradient (APG) and present the pseudo code of our algorithm in Algorithm 2. Throughout this paper, we adopt the notation $\pi_\theta^{(t)} := \pi_{\theta^{(t)}}$, consistently applying this convention to denote various parameters including ω and φ. Our algorithm design draws inspiration from the monotone version of NAG, which was originally proposed in the seminal work (Beck & Teboulle, 2009a) with a restart mechanism and has been widely adopted by variants of NAG, e.g., (Li & Lin, 2015; O'donoghue & Candes, 2015; Li & Lin, 2022). We adapt these updates to the RL objective, specifically $V^{\pi_\theta}(\mu)$. It is important to note that we will specify the step sizes $\eta^{(t)}$ in Lemma 2, as presented in Section 5.

In Algorithm 2, the gradient update is performed in (6). Following this, (7) calculates the momentum for our parameters, which represents a fundamental technique employed in accelerated gradient methods. It is worth noting that in (6), the gradient is computed with respect to $\omega^{(t-1)}$, which is the parameter that the momentum brings us to, rather than $\theta^{(t)}$ itself. This distinction sets APG apart from the standard policy gradient updates (Algorithm 1). Moreover, (8) serves as the restart mechanism, playing the role of affirming our updates to improve the objective value.

Remark 2. One could also directly apply Nesterov's accelerated gradient (cf. Algorithm 4) to the PG setting. We provide the pseudo code as Algorithm 6 in Appendix H. However, in the absence of monotonicity, we must address the intertwined effects between the gradient and the momentum to achieve convergence results. Specifically, we empirically demonstrate that the non-monotonicity could occur during the updates of Algorithm 6. For more detailed

information, please refer to Appendix H.

4.2. A Motivating Example of APG

Prior to the exposition of convergence analysis, we aim to provide some insights into why APG has the potential to attain a fast convergence rate, especially under the intricate non-concave objectives in reinforcement learning.

Consider a simple two-action bandit with actions a^*, a_2 and the reward function $r(a^*) = 1, r(a_2) = 0$. Accordingly, the objective function we aim to optimize is $\mathbb{E}_{a \sim \pi_\theta}[r(a)] = \pi_\theta(a^*) = 1/(1+e^{-(\theta_{a^*} - \theta_{a_2})})$ under softmax parameterization. This can be expressed as a sigmoid function with the parameter $(\theta_{a^*} - \theta_{a_2})$, as illustrated in Figure 1. By deriving the Hessian matrix with respect to our policy parameters θ_{a^*} and θ_{a_2}, we could characterize the curvature of the objective function around the current policy parameters, which provides useful insights into its local concavity. Upon analyzing the Hessian matrix, we observe that it exhibits concavity when $\pi_\theta(a^*) > 0.5$. The detailed derivation is provided in Appendix F. The aforementioned observation implies that the objective function demonstrates *local concavity* when $\pi_\theta(a^*) > 0.5$. Since $\pi^*(a^*) = 1$, it follows that the objective function exhibits local concavity in the proximity of the optimal policy π^*. As a result, if one initializes the policy with a high probability assigned to the optimal action a^*, then the policy would directly fall in the locally concave part of the objective function. This allows us to apply the theoretical findings from the existing convergence rate of NAG in (Nesterov, 1983), which has demonstrated convergence rates of $O(1/t^2)$ for convex problems. Based on this insight, we establish the convergence rate of APG in the more general MDP setting in Section 5.
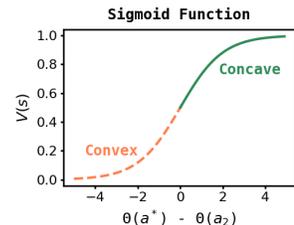
Figure 1. The value function $V(s)$ versus the policy parameter θ_{a^*} and θ_{a_2} under a 2-armed bandit problem.

5. Convergence Analysis

In this section, we take an important first step towards understanding the convergence behavior of APG and discuss the theoretical results of APG in the general MDP setting. Due to the space limit, we defer the proofs of the following theorems to Appendices C and D.

5.1. Asymptotic Convergence of APG

In this subsection, we will formally present the asymptotic convergence result of APG. This necessitates addressing several key challenges outlined in the introduction. We high-

light the challenges tackled in our analysis as follows: (C1) *The existing convergence results for non-convex problems under NAG are not directly applicable*: Note that the asymptotic convergence of standard PG is built on the standard convergence result of gradient descent for non-convex problems (i.e., convergence of stationarity measures), as shown in (Agarwal et al., 2021). While it appears natural to follow the same approach for APG, two fundamental challenges are that the existing results of NAG for non-convex problems typically show best-iterate convergence (e.g., (Ghadimi & Lan, 2016)) rather than last-iterate convergence, and moreover these results hold under the assumption of a bounded domain (e.g., see Theorem 2 of (Ghadimi & Lan, 2016)), which does not hold under the softmax parameterization in RL as the domain of the policy parameters and the optimal θ could be unbounded. These are yet additional salient differences between APG and PG. (C2) *Characterization of the cumulative effect of each momentum term*: Based on (C1), even if the limiting value functions exist, another crucial obstacle is to precisely quantify the memory effect of the momentum term on the policy's overall evolution. To address this challenge, we thoroughly examine the accumulation of the gradient and momentum terms, as well as the APG updates, to offer an accurate characterization of the momentum's memory effect on the policy.

Despite the above, we are still able to tackle these challenges and establish the asymptotic convergence to the global optimal policy of APG as follows. Recall the definition of the optimal objective in (3).

Theorem 1 (Asymptotic Convergence Under Softmax Parameterization). *Consider a tabular softmax parameterized policy π_θ. For APG with $\eta^{(t)} = \frac{t}{t+1} \frac{(1-\gamma)^3}{16}$ and μ initialized uniformly at random, the following holds almost surely:*

$$\lim_{t \to \infty} V^{\pi_\theta^{(t)}}(s) = V^*(s), \forall s \in \mathcal{S}.$$

The complete proof is provided in Appendix C.

Remark 3. Note that Theorem 1 suggests the use of step size $\eta^{(t)}$ that depends on $t/(t+1)$. This choice is related to one inherent issue of NAG: the choices of step sizes are typically different for the convex and the non-convex problems (e.g., (Ghadimi & Lan, 2016)). Recall from Section 4.2 that the RL objective function could be locally concave around the optimal policy despite its non-concavity of the global landscape. To enable the use of the same step size scheme throughout the whole training process, we find that incorporating the ratio $t/(t+1)$ could achieve the best of both concave and nonconcave cases.

Remark 4. In Theorem 1, the almost-sure statement comes from the random initialization of the surrogate initial state distribution μ. This condition resembles those of the convergence results of gradient-based methods for non-convex

problems (e.g., (Lee et al., 2019; O'Neill & Wright, 2019)).

5.2. Convergence Rate of APG

In this subsection, we leverage the asymptotic convergence of APG and proceed to characterize the convergence rate of APG under softmax parameterization. We assume that Assumption 1 holds for the subsequent results. For ease of notation, we denote the actions for each state $s \in \mathcal{S}$ as $a^*(s), a_2(s), \ldots, a_{|\mathcal{A}|}(s)$, ordered such that $Q^*(s, a^*(s)) > Q^*(s, a_2(s)) \geq \cdots \geq Q^*(s, a_{|\mathcal{A}|}(s))$. We will begin by introducing the two most fundamental concepts throughout this paper: the C-nearly concave property and the feasible update domain, as follows:

Definition 1 (C-Near Concavity). *For $C > 1$, a function f is said to be C-nearly concave at θ relative to a convex set $\mathcal{X} \ni \theta$ if for any $\theta' \in \mathcal{X}$, we have*

$$f(\theta') \leq f(\theta) + C \cdot \left\langle \nabla f(\theta), \theta' - \theta \right\rangle.$$

Definition 2. *We define the feasible update domain $\mathcal{U}$ as*

$$\mathcal{U} := \{ \boldsymbol{d} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} : d_{s,a^*(s)} > \max_{a \neq a^*(s)} d_{s,a}, \forall s \in \mathcal{S} \},$$

where $d_{s,a} \in \mathbb{R}$ is the update with respect to the parameter $\theta_{s,a}$.

In other words, a vector $\boldsymbol{d}$ in the feasible update domain $\mathcal{U}$ possesses the property that the update magnitude with respect to $\theta_{s,a^*(s)}$ is the greatest among the actions in each state $s \in \mathcal{S}$. We use $\mathcal{U}$ to help us characterize the locally nearly-concave regime as shown below, and subsequently show that the updates of APG would all fall into $\mathcal{U}$ after some finite time.

Lemma 1 (Locally C-Near Concavity; Informal). *Given any $C > 1$ and any $\boldsymbol{d} \in \mathcal{U}$. Let θ be a policy parameter satisfying the following conditions for all $s \in \mathcal{S}$: (i) $V^{\pi_\theta}(s) > Q^*(s, a_2(s))$; (ii) For some $M_{C,\boldsymbol{d}} > 0$ (depends on C and $\boldsymbol{d}$), $\theta_{s,a^*(s)} - \theta_{s,a} > M_{C,\boldsymbol{d}}$, for all $a \neq a^*(s)$. Then, the objective function $\theta \mapsto V^{\pi_\theta}(\mu)$ is C-nearly concave at θ along the direction $\boldsymbol{d}$.*

The formal information regarding the constant $M_{C,\boldsymbol{d}}$ mentioned in Lemma 1 is provided in Appendix D.

Remark 5. Notably, the optimal parameter under the softmax parameterization is unbounded. Consequently, it is challenging to determine whether a "neighborhood" exists near the optimal parameter such that the structure of the objective function exhibits desirable properties, such as concavity or near concavity. Lemma 1 delineates the sufficient conditions under which the objective function $V^{\pi_\theta}(\mu)$ achieves C-near concavity.

Remark 6. Note that the notion of *weak-quasi-convexity*, as defined in (Guminov et al., 2017; Hardt et al., 2018; Bu

& Mesbahi, 2020), shares similarities to our own definition in Definition 1. While their definition encompasses the direction from any to the optimal parameter, ours in Lemma 1 is more relaxed. Despite that the optimal parameter θ^* in our problem does not exist due to its unbounded nature, according to the definition of the feasible update domain $\mathcal{U}$, the direction toward the optimal policy π^* in the parameter space is always contained in $\mathcal{U}$.

With Lemma 1, our goal is to investigate whether APG updates can move along the local C-nearly concavity direction after some finite number of time steps. To address this, we establish Lemma 2, which guarantees the existence of a finite time T such that our policy will indeed achieve the condition stated in Lemma 1 through the APG updates and remain in this region.

Lemma 2. *Consider a tabular softmax parameterized policy. Under Assumption 1 and the setting of Theorem 1, the following holds almost surely: Given any $M > 0$, there exists a finite time T such that for all $t \geq T$, $s \in \mathcal{S}$, and $a \neq a^*(s)$, we have (i) $\theta^{(t)}_{s,a^*(s)} - \theta^{(t)}_{s,a} > M$, (ii) $V^{\pi^{(t)}_\theta}(s) > Q^*(s, a_2(s))$, (iii) $\left.\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a^*(s)}}\right|_{\theta=\omega^{(t)}} > 0 > \left.\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\right|_{\theta=\omega^{(t)}}$, (iv) $\omega^{(t)}_{s,a^*(s)} - \theta^{(t)}_{s,a^*(s)} \geq \omega^{(t)}_{s,a} - \theta^{(t)}_{s,a}$.*

Remark 7. Conditions (i) and (ii) are formulated to establish the local C-nearly concave conditions within a finite number of time steps. On the other hand, conditions (iii) and (iv) describe two essential properties for verifying that, after a finite time, all the update directions of APG fall within the feasible update domain $\mathcal{U}$. Consequently, the updates executed by APG are aligned with the nearly-concave structure. For a detailed proof regarding the examination of the feasible directions, please refer to Appendix D.

With the results of Lemmas 1 and 2, we are able to establish the main result for APG under softmax parameterization, which is a $\tilde{O}(1/t^2)$ convergence rate.

Theorem 2 (Convergence Rate of APG; Informal). *Consider a tabular softmax parameterized policy π_θ. Under APG with $\eta^{(t)} = \frac{t}{t+1} \cdot \frac{(1-\gamma)^3}{16}$ and μ is initialized uniformly at random, the following holds almost surely: There exists a finite time T such that for all $t \geq T$, we have*

$$
V^*(\rho) - V^{\pi^{(t)}_\theta}(\rho)
$$
$$
\leq \frac{1}{(1-\gamma)^3} \left\| \frac{d^{\pi^*}_\rho}{\mu} \right\|_\infty \left(\frac{2|\mathcal{S}|(|\mathcal{A}|-1)}{t^2 + |\mathcal{A}| - 1} + \right.
$$
$$
\left. \frac{512|\mathcal{S}|\ln^2(t) + 32\left\|2\theta^{(T)} - (2+T)\left(\omega^{(T)} - \theta^{(T)}\right)\right\|^2}{(1-\gamma)(t+1)t} \right).
$$

Remark 8. Notably, the intriguing local C-near concavity property is not specific to APG. Specifically, we can also establish that PG enters the local C-nearly concave regime

within finite time steps. Moreover, its updates align with the directions in $\mathcal{U}$, allowing us to view it as an optimization problem under the C-nearly concave objective. Accordingly, we can demonstrate that PG under softmax parameterization also enjoys C-near concavity and thereby achieves a convergence rate of $\tilde{O}(1/t)$. For more details, please refer to Appendix G.

Remark 9. It is important to note that the logarithmic factor in the sub-optimality gap is a consequence of the unbounded nature of the optimal parameter in softmax parameterization.

Remark 10. Regarding the fundamental capability of PG, Mei et al. (2020) has presented a lower bound on the sub-optimality gap for PG in their Theorem 10. This theorem asserts that the $O(1/t)$ convergence rate achievable by PG cannot be further improved. Despite the lower bound presented by Mei et al. (2020), our results for APG do not contradict theirs. Specifically, while both APG and PG are first-order methods relying solely on first-order derivatives for updates, it is crucial to highlight that APG encompasses a broader class of policy updates with the help of the momentum term in Nesterov acceleration. This allows APG to utilize the gradients with respect to parameters that PG cannot attain, thereby improving the convergence rate and overall performance of APG.

In addition to Theorem 2, we also demonstrate that APG exhibits a linear convergence when the time-varying step sizes are allowed. Specifically, we establish results similar to Theorem 2 and Lemma 2 with exponentially-growing step sizes under element-wise clipping updates, leading to the following theorem. For additional details, please refer to Appendix E.

Definition 3. *Given a vector $\boldsymbol{x} = [x_1, x_2, \cdots]$, we define the element-wise clipping operator by:*

$$
[\boldsymbol{x}]^{+y} = [clip(x_1, -y, y), clip(x_2, -y, y), \cdots],
$$

where

$$
clip(x, a, b) := \begin{cases} a, & \text{if } x < a, \\ x, & \text{if } a \leq x \leq b, \\ b, & \text{if } x > b. \end{cases} \quad (9)
$$

Theorem 3 (Convergence Rate of APG with Time-Varying Step Sizes). *Consider a tabular softmax parameterized policy π_θ. Under APG with (6) modified as*

$$
\theta^{(t)} \leftarrow \omega^{(t-1)} + \left[\eta^{(t)}\nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\omega^{(t-1)}}\right]^{+K},
$$

where $\eta^{(t)} = \beta^t \cdot \frac{(1-\gamma)^3}{8}$ for any $\exp\left(\frac{1}{8\sqrt{|\mathcal{A}||\mathcal{S}|}} \cdot \frac{1-\gamma}{4C_\infty-(1-\gamma)}\right) > \beta > 1$, μ initialized uniformly at random, and $K = \frac{1}{4\sqrt{|\mathcal{A}||\mathcal{S}|}} \cdot \frac{1-\gamma}{4C_\infty-(1-\gamma)} + 2\ln\beta$ where

7
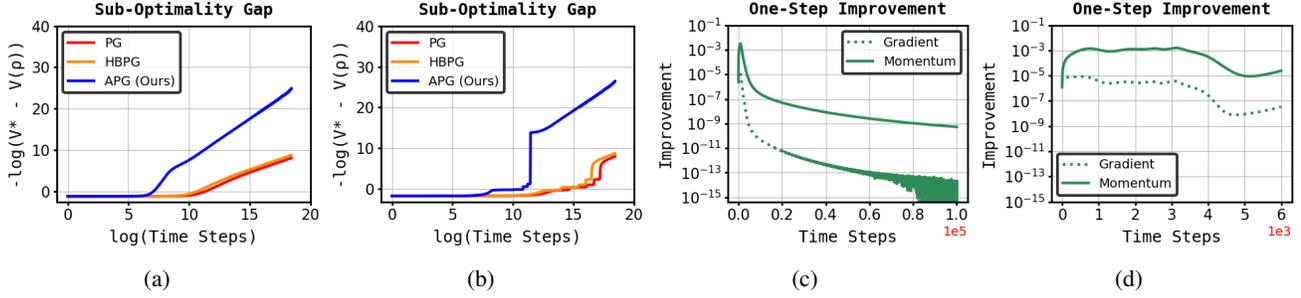
Figure 2. A comparison between the performance of APG, PG, and HBPG under an MDP with 5 states, 5 actions, with the uniform and hard policy initialization: (a)-(b) show the sub-optimality gaps under the uniform and the hard initialization, respectively; (c)-(d) show the one-step improvements of APG from the momentum (i.e., $V^{\pi_\omega^{(t)}}(\rho) - V^{\pi_\theta^{(t)}}(\rho)$) and the gradient (i.e., $V^{\pi_\theta^{(t+1)}}(\rho) - V^{\pi_\omega^{(t)}}(\rho)$), under the uniform and the hard initialization, respectively.
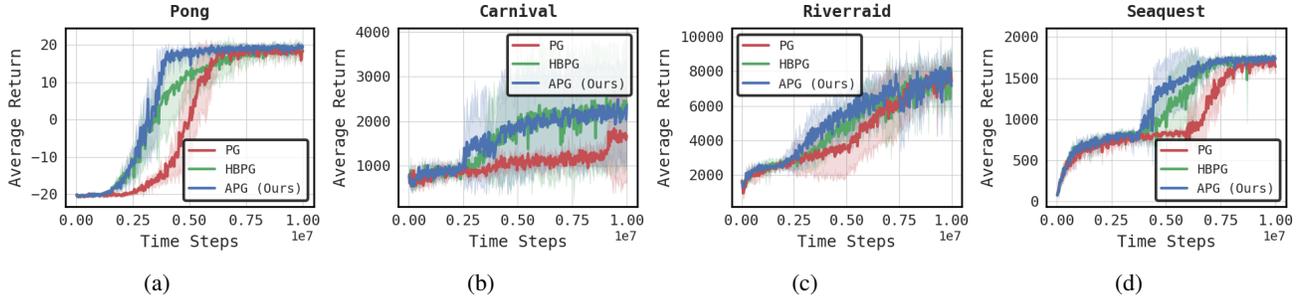
Figure 3. A comparison of the performance of APG and the benchmark algorithms in four Atari 2600 games. All the results are averaged over 5 random seeds (with the shaded area showing the range of mean $\pm$ std).

$C_\infty := \max_\pi \|\frac{d_\mu^\pi}{\mu}\|_\infty$, *the following holds almost surely: There exists a finite time T such that for all $t \geq T$, we have*

$$V^*(\rho) - V^{\pi_\theta^{(t)}}(\rho) = O(e^{-ct}),$$

for some constant $c > 0$.

Remark 11. It is worth noticing that the finite constant time T from Theorem 2 and Theorem 3 is problem-dependent. Specifically, this constant T shall depend on multiple factors, such as the initial state distribution ρ, the initial policy parameter θ_0, the cardinalities of the state space $|\mathcal{S}|$ and the action space $|\mathcal{A}|$, and the discount factor γ, among other factors. Despite this problem-dependent constant, the rates of convergence in Theorem 2 and Theorem 3 are $\tilde{O}(1/t^2)$ and $O(e^{-ct})$, under any problem instance.

6. Experiments

6.1. Numerical Validation of the Convergence Rates

In this subsection, we empirically validate the convergence rate of APG by conducting experiments on an MDP with 5 states and 5 actions. We validate the convergence behavior of APG and two popular methods, namely standard PG and

the algorithm that directly applies heavy-ball momentum (Polyak, 1964) to PG, which we refer to as Heavy-Ball PG (HBPG). Detailed configurations and the pseudo code of HBPG are provided in Appendix F. In the subsequent numerical validations, we focus on APG with nearly constant step sizes. The numerical validation for APG with time-varying step sizes is provided in Appendix E due to space limitations. Our code for the experiments is available at https://github.com/NYCU-RL-Bandits-Lab/APG.

We validate the convergence rate of APG, HBPG, and PG on an MDP with 5 states and 5 actions in two settings: (i) *Uniform initialization* (i.e., all actions have equal initial probability): The training curves of sub-optimality gap for PG, HBPG, and APG are depicted in Figure 2. Upon plotting the sub-optimality gaps of PG, HBPG, and APG under uniform initialization on a log-log graph in Figure 2(a), we observe that both PG and HBPG exhibit a slope of approximately 1, while APG demonstrates a slope of 2. These slopes match the respective convergence rates of $O(1/t)$ for PG and HBPG and $\tilde{O}(1/t^2)$ convergence rate for APG, as shown in Theorem 2. Figure 2(c) further confirms that the momentum term in APG contributes substantially in terms of policy improvement in the MDP case. (ii) *Hard*

initialization (i.e., the optimal action has the smallest initial probability): Figure 2(b) and Figure 2(d) show that APG could escape from sub-optimality much faster than PG and HBPG in the MDP case. This further showcases APG's superiority over PG and HBPG.

6.2. APG on Atari Games

In this subsection, we empirically evaluate the performance of APG on four Atari 2600 games from the Arcade Learning Environment (ALE) (Bellemare et al., 2013), including *Pong, Carnival, Riverraid, and Seaquest*. Each environment is associated with a $210 \times 160 \times 3$ binary state representation, which corresponds to the 210×160 grid and 3 channels. The detailed configuration is provided in Appendix F. In Figure 3, we observe that APG exhibits faster or comparable convergence compared to both HBPG and PG across all four environments. This further demonstrates the potential of APG in practice.

7. Concluding Remarks

Nesterov's Accelerated Gradient method, proposed almost four decades ago, provides a powerful first-order scheme for fast convergence under a broad class of optimization problems. However, since its introduction, NAG has never been formally analyzed or evaluated within the realm of RL, mainly due to the non-concavity of the RL objective. In this paper, we propose APG and take an important first step towards understanding NAG in RL. We rigorously present that APG can converge to a globally optimal policy at a $\tilde{O}(1/t^2)$ rate in the general MDP setting under softmax policies with nearly constant step sizes. Moreover, we showcase the linear convergence rate of $O(e^{-ct})$ with time-varying step sizes. This demonstrates the potential of APG in attaining fast convergence in RL.

On the other hand, our work also leaves open interesting research questions about NAG in RL. For example, as our paper mainly focuses on the exact gradient setting, another promising research direction is to extend our results of APG to the stochastic gradient setting, where the advantage function as well as the gradient are estimated from sampled transitions.

Impact Statement

This paper introduces research aimed at pushing the boundaries of the Machine Learning field. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Achiam, J. Spinning Up in Deep Reinforcement Learning. 2018.

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 32, 2019.

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(1):4431–4506, 2021.

Amari, S.-I. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

Andrychowicz, M., Raichuk, A., Stańczyk, P., Orsini, M., Girgin, S., Marinier, R., Hussenot, L., Geist, M., Pietquin, O., Michalski, M., et al. What matters for on-policy deep actor-critic methods? A large-scale study. In *International Conference on Learning Representations*, 2021.

Attouch, H. and Peypouquet, J. The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than 1/k^2. *SIAM Journal on Optimization*, 26(3):1824–1834, 2016.

Beck, A. and Teboulle, M. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Transactions on Image Processing*, 18(11):2419–2434, 2009a.

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009b.

Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

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

Bu, J. and Mesbahi, M. A note on nesterov's accelerated method in nonconvex optimization: a weak estimate sequence approach. *arXiv preprint arXiv:2006.08548*, 2020.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. "Convex Until Proven Guilty": Dimension-Free Acceleration of Gradient Descent on Non-Convex Functions. In *International Conference on Machine Learning*, pp. 654–663, 2017.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018.

Chen, Z. and Maguluri, S. T. Sample complexity of policy-based methods under off-policy sampling and linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pp. 11195–11214. PMLR, 2022.

Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International Conference on Machine Learning*, pp. 1407–1416, 2018.

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

Ghadimi, S. and Lan, G. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1-2):59–99, 2016.

Guminov, S., Gasnikov, A., and Kuruzov, I. Accelerated methods for α-weakly-quasi-convex problems. *arXiv preprint arXiv:1710.00797*, 2017.

Hardt, M., Ma, T., and Recht, B. Gradient descent learns linear dynamical systems. *The Journal of Machine Learning Research*, 19(1):1025–1068, 2018.

Henderson, P., Romoff, J., and Pineau, J. Where did my optimum go?: An empirical analysis of gradient descent optimization in policy gradient methods. *arXiv preprint arXiv:1810.02525*, 2018.

Huang, F., Gao, S., Pei, J., and Huang, H. Momentum-based policy gradient methods. In *International conference on machine learning*, pp. 4422–4433. PMLR, 2020.

Jin, C., Netrapalli, P., and Jordan, M. I. Accelerated gradient descent escapes saddle points faster than gradient descent. In *Conference On Learning Theory*, pp. 1042–1085, 2018.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, 2002.

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

Khodadadian, S., Jhunjhunwala, P. R., Varma, S. M., and Maguluri, S. T. On the linear convergence of natural policy gradient algorithm. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pp. 3794–3799. IEEE, 2021.

Konda, V. and Tsitsiklis, J. Actor-critic algorithms. *Advances in Neural Information Processing Systems*, 12, 1999.

Lan, G. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, 198(1):1059–1106, 2023.

Lee, J. D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M. I., and Recht, B. First-order methods almost always avoid strict saddle points. *Mathematical Programming*, 176:311–337, 2019.

Li, H. and Lin, Z. Accelerated proximal gradient methods for nonconvex programming. *Advances in Neural Information Processing Systems*, 28, 2015.

Li, H. and Lin, Z. Restarted Nonconvex Accelerated Gradient Descent: No More Polylogarithmic Factor in the $O(\varepsilon^{-7/4})$ Complexity. In *International Conference on Machine Learning*, pp. 12901–12916, 2022.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. In *International Conference on Machine Learning*, 2016.

Liu, Y., Zhang, K., Basar, T., and Yin, W. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33:7624–7636, 2020.

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

Mei, J., Dai, B., Xiao, C., Szepesvari, C., and Schuurmans, D. Understanding the effect of stochasticity in policy

optimization. *Advances in Neural Information Processing Systems*, 34:19339–19351, 2021a.

Mei, J., Gao, Y., Dai, B., Szepesvari, C., and Schuurmans, D. Leveraging non-uniformity in first-order non-convex optimization. In *International Conference on Machine Learning*, pp. 7555–7564. PMLR, 2021b.

Mei, J., Chung, W., Thomas, V., Dai, B., Szepesvari, C., and Schuurmans, D. The role of baselines in policy gradient optimization. *Advances in Neural Information Processing Systems*, 35:17818–17830, 2022.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, pp. 1928–1937, 2016.

Nesterov, Y. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. In *Soviet Mathematics Doklady*, volume 269, pp. 543–547, 1983.

O'donoghue, B. and Candes, E. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15:715–732, 2015.

O'Neill, M. and Wright, S. J. Behavior of accelerated gradient methods near critical points of nonconvex functions. *Mathematical Programming*, 176:403–427, 2019.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library, 2019.

Polyak, B. T. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

Raffin, A. Rl baselines3 zoo. `https://github.com/DLR-RM/rl-baselines3-zoo`, 2020.

Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., and Dormann, N. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL `http://jmlr.org/papers/v22/20-1364.html`.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International Conference on Machine Learning*, pp. 1889–1897, 2015a.

Schulman, J., Moritz, P., Levine, S., Jordan, M., and Abbeel, P. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015b.

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

Shani, L., Efroni, Y., and Mannor, S. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5668–5675, 2020.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. Deterministic policy gradient algorithms. In *International Conference on Machine Learning*, pp. 387–395, 2014.

Su, W., Boyd, S., and Candes, E. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Advances in Neural Information Processing Systems*, 27, 2014.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems*, 12, 1999.

Szepesvári, C. *Algorithms for reinforcement learning*. Springer nature, 2022.

Vieillard, N., Scherrer, B., Pietquin, O., and Geist, M. Momentum in reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 2529–2538. PMLR, 2020.

Wang, W., Han, J., Yang, Z., and Wang, Z. Global convergence of policy gradient for linear-quadratic mean-field control/game in continuous time. In *International Conference on Machine Learning*, pp. 10772–10782, 2021.

Wang, Z., Bapst, V., Heess, N., Mnih, V., Munos, R., Kavukcuoglu, K., and de Freitas, N. Sample efficient actor-critic with experience replay. In *International Conference on Learning Representations*, 2016.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Reinforcement Learning*, pp. 5–32, 1992.

Xiao, L. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282): 1–36, 2022.

Yuan, R., Du, S. S., Gower, R. M., Lazaric, A., and Xiao, L. Linear convergence of natural policy gradient methods with log-linear policies. *arXiv preprint arXiv:2210.01400*, 2022.

Appendix

Contents

A. Supporting Algorithms

For ease of exposition, we state the accelerated gradient algorithm of Ghadimi & Lan (2016) in Algorithm 3. We have made the following notational changes from the version in (Ghadimi & Lan, 2016) so that one could easily compare it with Algorithm 2: (i) The positions of the superscript and subscript are exchanged. (ii) The original gradient symbol is replaced with the gradient of our objective (i.e., $\nabla_\theta V^{\pi_\theta}(\mu)$). (iii) The iteration counter is changed from k to t. (iv) The original descent algorithm is now made an ascent algorithm (i.e., the sign in (11) and (12) is plus instead of minus). (v) We introduced a momentum variable $\bar{m}_{\text{init}}$ in (10) in order to consider a non-zero initial momentum when the algorithm enters the concave regime, which will be clarified in the subsequent discussion.

Algorithm 3 The Accelerated Policy Gradient (APG) Algorithm Revised From (Ghadimi & Lan, 2016)

Input: $\hat{\theta}^{(0)}, \bar{m}_{\text{init}} \in \mathbb{R}^n$, $\{\alpha^{(t)}\}$ with $\alpha^{(t)} \in (0,1]$ for all $t \geq 1$, $\{\beta^{(t)} > 0\}$, and $\{\lambda^{(t)} > 0\}$.

0. Set the initial points $\theta_{ag}^{(0)} = \hat{\theta}^{(0)} + \bar{m}_{\text{init}}$ and $t = 1$.

1. Set

$$\theta_{md}^{(t)} = (1 - \alpha^{(t)})\theta_{ag}^{(t-1)} + \alpha^{(t)}\hat{\theta}^{(t-1)} \tag{10}$$

2. Compute $\nabla V^{\pi_\theta}(\mu)$ and set

$$\hat{\theta}^{(t)} = \hat{\theta}^{(t-1)} + \lambda^{(t)}\nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\theta_{md}^{(t)}}, \tag{11}$$

$$\theta_{ag}^{(t)} = \theta_{md}^{(t)} + \beta^{(t)}\nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\theta_{md}^{(t)}}. \tag{12}$$

3. Set $t \leftarrow t + 1$ and go to step 1.

Lemma 3 (Equivalence between Algorithm 2 and Algorithm 3). *Suppose that from the T_{shift}-th iteration on, the restart mechanism of Algorithm 2 remains inactive, i.e., $\omega^{(t)} = \varphi^{(t)}$ for all $t \geq T_{shift}$. Then, starting at the T_{shift}-th iteration, Algorithm 2 is equivalent to initializing Algorithm 3 at $\theta^{(T_{shift})}$, with $\alpha^{(t)} = \frac{2}{t+T_{shift}+1}$, $\beta^{(t)} = \eta^{(t+T_{shift})}$, $\lambda^{(t)} = \beta^{(t)}/\alpha^{(t)}$, $\bar{m}_{init} = -\frac{\omega^{(T_{shift})}-\theta^{(T_{shift})}}{\alpha^{(1)}}$, and $\hat{\theta}^{(0)} = \theta^{(T_{shift})} - \bar{m}_{init}$, in the sense that $\omega^{(t)} = \theta_{md}^{(t-T_{shift}+1)}, \theta^{(t)} = \theta_{ag}^{(t-T_{shift})}$.*

Remark 12. Lemma 3 shows that our Algorithm 2 is equivalent to Algorithm 3 under some proper conditions, so we could leverage the results of Ghadimi & Lan (2016).

Proof of Lemma 3. Since $\alpha^{(t)}\lambda^{(t)} = \beta^{(t)}$, by subtracting (12) from (11) times $\alpha^{(t)}$, for all $t \geq 1$, we have

$$\alpha^{(t)}\hat{\theta}^{(t)} - \theta_{ag}^{(t)} = \alpha^{(t)}\hat{\theta}^{(t-1)} - \theta_{md}^{(t)}. \tag{13}$$

Then, substituting $\theta_{md}^{(t)}$ in (13) by (10) leads to

$$\hat{\theta}^{(t)} = \frac{\theta_{ag}^{(t)} - (1 - \alpha^{(t)})\theta_{ag}^{(t-1)}}{\alpha^{(t)}}. \tag{14}$$

Plugging (14) back into (10), for $t \geq 2$, we get

$$\begin{aligned}
\theta_{md}^{(t)} &= (1 - \alpha^{(t)})\theta_{ag}^{(t-1)} + \alpha^{(t)}\hat{\theta}^{(t-1)} \\
&= (1 - \alpha^{(t)})\theta_{ag}^{(t-1)} + \alpha^{(t)}\frac{\theta_{ag}^{(t-1)} - (1 - \alpha^{(t-1)})\theta_{ag}^{(t-2)}}{\alpha^{(t-1)}} \\
&= \theta_{ag}^{(t-1)} + \frac{\alpha^{(t)}(1 - \alpha^{(t-1)})}{\alpha^{(t-1)}}(\theta_{ag}^{(t-1)} - \theta_{ag}^{(t-2)}).
\end{aligned} \tag{15}$$

For $t = 1$, since $\theta_{ag}^{(0)} = \hat{\theta}^{(0)} + \bar{m}_{\text{init}}$, we get

$$\theta_{md}^{(1)} = (1 - \alpha^{(1)})\theta_{ag}^{(0)} + \alpha^{(1)}\hat{\theta}^{(0)}$$

14

$$= \theta_{ag}^{(0)} - \alpha^{(1)} \bar{m}_{\text{init}}. \tag{16}$$

By (12), (15) and (16), we could simplify Algorithm 3 into a two-variable update form:

$$\theta_{md}^{(t)} = \begin{cases} \theta_{ag}^{(0)} - \alpha^{(1)} \bar{m}_{\text{init}} & \text{if } t = 1, \\ \theta_{ag}^{(t-1)} + \frac{\alpha^{(t)}(1-\alpha^{(t-1)})}{\alpha^{(t-1)}}(\theta_{ag}^{(t-1)} - \theta_{ag}^{(t-2)}) & \text{if } t > 1. \end{cases}$$

$$\theta_{ag}^{(t)} = \theta_{md}^{(t)} + \beta^{(t)} \nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\theta_{md}^{(t)}} \qquad \text{for all } t \geq 1.$$

Finally, given $T_{\text{shift}} \in \mathbb{N}$, by applying $\alpha^{(t)} = \frac{2}{t+1+T_{\text{shift}}}$, $\beta^{(t)} = \eta^{(t)}$, and $\bar{m}_{\text{init}} = -\frac{\omega^{(T_{\text{shift}})} - \theta^{(T_{\text{shift}})}}{\alpha^{(1)}}$ in the update form above, we reach our desired result:

1. First, initialize $\theta_{md}^{(1)} = \theta_{ag}^{(0)} - \alpha^{(1)} \bar{m}_{\text{init}} = \theta_{ag}^{(0)} + (\omega^{(T_{\text{shift}})} - \theta^{(T_{\text{shift}})})$.

2. Then, run the following updates recursively:

$$\theta_{ag}^{(t)} = \theta_{md}^{(t)} + \eta^{(t)} \nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\theta_{md}^{(t)}} \tag{17}$$

$$\theta_{md}^{(t+1)} = \theta_{ag}^{(t)} + \frac{t-1+T_{\text{shift}}}{t+2+T_{\text{shift}}}(\theta_{ag}^{(t)} - \theta_{ag}^{(t-1)}). \tag{18}$$

Note that $\omega^{(T_{\text{shift}})} - \theta^{(T_{\text{shift}})}$ is the momentum at time $T_{\text{shift}} - 1$, so we select $\bar{m}_{\text{init}}$ as $-\frac{\omega^{(T_{\text{shift}})} - \theta^{(T_{\text{shift}})}}{\alpha^{(1)}}$. Furthermore, if $T_{\text{shift}} = 0$, it is exactly equivalent to the form shown in Algorithm 2. In summary, $\theta_{md}^{(t+1)}, \theta_{ag}^{(t)}$ in (17) and (18) corresponds to $\omega^{(t)}, \theta^{(t)}$ in Algorithm 2, respectively. $\square$

Below, we present the pseudo code of Nesterov's Accelerated Gradient (NAG) algorithm discussed in Section 3.

Algorithm 4 Nesterov's Accelerated Gradient (NAG) algorithm in (Su et al., 2014)

Input: Step size $\eta = \frac{1}{L}$, where L is the Lipschitz constant of the gradient of the objective function f.
Initialize: $\theta^{(0)}$ and $\omega^{(0)} = \theta^{(0)}$.
for $t = 1$ to T **do**

$$\theta^{(t)} = \omega^{(t-1)} - \eta \nabla f(\omega^{(t-1)})$$
$$\omega^{(t)} = \theta^{(t)} + \frac{t-1}{t+2}(\theta^{(t)} - \theta^{(t-1)})$$

end for

B. Supporting Lemmas

In this section, we provide useful properties of policy optimization, softmax parameterization, and Accelerated Policy Gradient (APG). Additionally, we also demonstrate the convergence rate of accelerated gradient (not necessarily restricted to reinforcement learning) for nearly concave objective functions.

B.1. Useful Properties of Policy Optimization

We first present properties related to general policy optimization that are not restricted to softmax parameterization.

Lemma 4 (**Performance Difference Lemma in (Kakade & Langford, 2002; Agarwal et al., 2021)**). *For any initial state distribution $\mu \in \Delta(\mathcal{S})$, the difference in the value under μ between two policies π and π' can be characterized as*

$$V^{\pi}(\mu) - V^{\pi'}(\mu) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{\mu}^{\pi}} \mathbb{E}_{a \sim \pi(\cdot|s)} \left[A^{\pi'}(s, a) \right].$$

In particular, by specifying μ as a one-hot distribution (i.e., $\mu(s_0) = 1$ for some $s_0 \in \mathcal{S}$ and $\mu(s') = 0$ for all $s' \neq s_0$), we have

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{s_0}^{\pi}} \mathbb{E}_{a \sim \pi(\cdot|s)} \left[A^{\pi'}(s, a) \right].$$

Lemma 5. $d_{\mu}^{\pi}(s) \geq (1 - \gamma) \cdot \mu(s), \ \forall \pi, \forall s \in \mathcal{S}$, *where $\mu(s)$ is any starting state distribution of the MDP.*

Proof of Lemma 5. Since $\gamma \in [0, 1)$, we have

$$
\begin{aligned}
d_{\mu}^{\pi}(s) &= \mathbb{E}_{s_0 \sim \mu} \left[d_{\mu}^{\pi}(s) \right] \\
&= \mathbb{E}_{s_0 \sim \mu} \left[(1 - \gamma) \cdot \sum_{t=0}^{\infty} \gamma^t \cdot \mathbb{P}(s_t = s \mid s_0, \pi) \right] \\
&\geq \mathbb{E}_{s_0 \sim \mu} \left[(1 - \gamma) \cdot \mathbb{P}(s_0 = s \mid s_0, \pi) \right] \\
&= (1 - \gamma) \cdot \mu(s). \qquad \square
\end{aligned}
$$

Lemma 6. $V^*(\rho) - V^{\pi}(\rho) \leq \frac{1}{1-\gamma} \cdot \left\| \frac{d_{\rho}^{\pi^*}}{\mu} \right\|_{\infty} \cdot (V^*(\mu) - V^{\pi}(\mu))$ *for any probability distributions ρ, μ and any policy π.*

Proof of Lemma 6.

$$V^*(\rho) - V^{\pi}(\rho) = \frac{1}{1-\gamma} \cdot \sum_{s \in \mathcal{S}} d_{\rho}^{\pi^*}(s) \sum_{a \in \mathcal{A}} \pi^*(a|s) \cdot A^{\pi}(s, a) \tag{19}$$

$$= \frac{1}{1-\gamma} \cdot \sum_{s \in \mathcal{S}} d_{\mu}^{\pi^*}(s) \cdot \frac{d_{\rho}^{\pi^*}(s)}{d_{\mu}^{\pi^*}(s)} \sum_{a' \in \mathcal{A}} \pi^*(a|s) \cdot A^{\pi}(s, a)$$

$$\leq \frac{1}{1-\gamma} \cdot \left\| \frac{d_{\rho}^{\pi^*}}{d_{\mu}^{\pi^*}} \right\|_{\infty} \cdot \sum_{s \in \mathcal{S}} d_{\mu}^{\pi^*}(s) \sum_{a' \in \mathcal{A}} \pi^*(a|s) \cdot A^{\pi}(s, a)$$

$$\leq \frac{1}{(1-\gamma)^2} \cdot \left\| \frac{d_{\rho}^{\pi^*}}{\mu} \right\|_{\infty} \cdot \sum_{s \in \mathcal{S}} d_{\mu}^{\pi^*}(s) \sum_{a' \in \mathcal{A}} \pi^*(a|s) \cdot A^{\pi}(s, a) \tag{20}$$

$$= \frac{1}{1-\gamma} \cdot \left\| \frac{d_{\rho}^{\pi^*}}{\mu} \right\|_{\infty} \cdot (V^*(\mu) - V^{\pi}(\mu)), \tag{21}$$

where (19) and (21) hold by Lemma 4 and (20) holds due to Lemma 5. $\qquad \square$

Lemma 7. *For any policy π, we have $\sum_{a \in \mathcal{A}} \pi(a|s) A^{\pi}(s, a) = 0, \ \forall s \in \mathcal{S}$.*

Proof of Lemma 7.

$$
\begin{aligned}
\sum_{a \in \mathcal{A}} \pi(a|s) A^\pi(s, a) &= \sum_{a \in \mathcal{A}} \pi(a|s) \Big(Q^\pi(s, a) - V^\pi(s) \Big) \\
&= \sum_{a \in \mathcal{A}} \pi(a|s) Q^\pi(s, a) - \sum_{a \in \mathcal{A}} \pi(a|s) V^\pi(s) \\
&= V^\pi(s) - V^\pi(s), \\
&= 0,
\end{aligned}
\tag{22}
$$

where (22) leverages the Bellman consistency equation in Lemma 1.4 of (Agarwal et al., 2019). □

Lemma 8. *Assume assumption 1 holds. Given any $s \in \mathcal{S}$, let $a^*(s)$ be the optimal action at the state s, and $a_i(s)$, where $2 \le i \le |\mathcal{A}|$, be such that $Q^*(s, a^*(s)) \ge Q^*(s, a_2(s)) \ge \cdots \ge Q^*(s, a_{|\mathcal{A}|}(s))$. Then, for any policy π and any $a_i(s)$ with $2 \le i \le |\mathcal{A}|$, we have $Q^*(s, a_2(s)) \ge Q^\pi(s, a_i(s))$. Furthermore, we also have $V^*(s) = Q^*(s, a^*(s))$ for all $s \in \mathcal{S}$.*

Proof of Lemma 8. Theorem 1.7 of (Agarwal et al., 2019) gives that

$$
Q^*(s, a) = \sup_{\bar{\pi}} Q^{\bar{\pi}}(s, a), \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}.
\tag{23}
$$

Therefore, for any policy π and $a_i(s)$ for $2 \le i \le |\mathcal{A}|$, we have

$$
\begin{aligned}
Q^\pi(s, a_i(s)) &\le \sup_{\pi'} Q^{\pi'}(s, a_i(s)) \\
&= Q^*(s, a_i(s)) \\
&\le Q^*(s, a_2(s)),
\end{aligned}
\tag{24}
\tag{25}
$$

where (24) holds by (23), and (25) is from the definition of $a_2(s)$. Furthermore, by page 14 of (Szepesvári, 2022) and the definition of $a^*(s)$, we have $V^*(s) = \arg\max_{a \in \mathcal{A}} Q^*(s, a) = Q^*(s, a^*(s))$ for all $s \in \mathcal{S}$. □

Lemma 9. *For any fixed state $s \in \mathcal{S}$, any fixed action $a \in \mathcal{A}$, and any policies π_1, π_2, we have*

$$
\left\| \pi_1(\cdot|s) - \pi_2(\cdot|s) \right\|_\infty \le 1 - \min\{\pi_1(a|s), \pi_2(a|s)\}.
$$

Proof of Lemma 9. For any $a' \in \mathcal{A}$, we have

$$
\left| \pi_1(a'|s) - \pi_2(a'|s) \right| \le \max\{\pi_1(a'|s), \pi_2(a'|s)\},
\tag{26}
$$

where (26) holds because both terms are nonnegative. In addition,

$$
\begin{aligned}
\left| \pi_1(a'|s) - \pi_2(a'|s) \right| &\le \max\{\pi_1(a'|s), \pi_2(a'|s)\} - \min\{\pi_1(a'|s), \pi_2(a'|s)\} \\
&\le 1 - \min\{\pi_1(a'|s), \pi_2(a'|s)\}.
\end{aligned}
\tag{27}
$$

Now, we fix $a \in \mathcal{A}$ and show that for any $a' \in \mathcal{A}$ we have

$$
\left| \pi_1(a'|s) - \pi_2(a'|s) \right| \le 1 - \min\{\pi_1(a|s), \pi_2(a|s)\}.
\tag{28}
$$

When $a' = a$, (28) holds by (27). When $a' \ne a$, we note that the maximum of $\pi_1(a'|s)$ and $\pi_2(a'|s)$ can be at most $1 - \pi_1(a|s)$ and $1 - \pi_2(a|s)$, respectively. Thus, by (26), we have

$$
\begin{aligned}
\left| \pi_1(a'|s) - \pi_2(a'|s) \right| &\le \max\{\pi_1(a'|s), \pi_2(a'|s)\} \\
&\le \max\{1 - \pi_1(a|s), 1 - \pi_2(a|s)\} \\
&= 1 - \min\{\pi_1(a|s), \pi_2(a|s)\}.
\end{aligned}
$$

The desired result is then obtained by taking the maximum over all $a' \in \mathcal{A}$. □

Lemma 10. *Given an infinite horizon discount MDP $\mathcal{M}$ as defined in Section 3. Suppose that for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, the reward function $r(s,a)$ of the MDP $\mathcal{M}$ lies in $[0, R_{max}]$ for some $R_{max} \geq 0$. Then, we have $V^\pi(s) \in [0, R_{max}/(1-\gamma)]$ for any $s \in \mathcal{S}$ and any policy π. Furthermore, $Q^\pi(s,a) \in [0, R_{max}/(1-\gamma)]$ and $|A^\pi(s,a)| \in [0, R_{max}/(1-\gamma)]$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and any policy π.*

Proof of Lemma 10. By the definition in (1) and that $r(s,a) \in [0, R_{\max}]$, for any policy π and any state $s \in \mathcal{S}$, we have

$$V^\pi(s) = \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t r(s_t, a_t) \Big| \pi, s_0 = s \right] \leq \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t R_{\max} \Big| \pi, s_0 = s \right] = \sum_{t=0}^\infty \gamma^t R_{\max} = \frac{R_{\max}}{1-\gamma}.$$

For the Q-value, by (2), given any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and any policy π,

$$Q^\pi(s,a) = r(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) V^\pi(s') \leq R_{\max} + \gamma \cdot \frac{R_{\max}}{1-\gamma} = \frac{R_{\max}}{1-\gamma}.$$

For the advantage function, by (2), we combine the above results of $V^\pi(s)$ and $Q^\pi(s,a)$ and obtain that $|A^\pi(s,a)| \in [0, R_{\max}/(1-\gamma)]$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and any policy π. $\qquad\square$

B.2. Useful Properties of Softmax Parameterization

This section provides results specific for softmax parameterization.

Lemma 11 (Lemma 1. in (Mei et al., 2020)). *The softmax policy gradient with respect to θ is*

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot A^{\pi_\theta}(s,a), \text{ for each } (s,a) \in \mathcal{S} \times \mathcal{A}.$$

Lemma 12 (Equation in (Agarwal et al., 2021), page 18). *The softmax policy gradient with respect to π_θ is*

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \pi_\theta(a|s)} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot A^{\pi_\theta}(s,a), \text{ for each } (s,a) \in \mathcal{S} \times \mathcal{A}.$$

Lemma 13 (Lemma 2. in (Mei et al., 2020)). *Under softmax policy parameterization, $\theta \to \pi_\theta^\top r$ is $5/2$-smooth for any $r \in [0,1]^{|\mathcal{A}|}$.*

Lemma 14 (Lemma 7. in (Mei et al., 2020)). *Under softmax policy parameterization, $\theta \to V^{\pi_\theta}(\rho)$ is $\frac{8}{(1-\gamma)^3}$-smooth for any $\gamma \in [0,1)$.*

Lemma 15 (Lemma 8. in (Mei et al., 2020)). *For all $\theta \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$, we have,*

$$\left\| \nabla_\theta V^{\pi_\theta} |_{\theta = \theta^{(t)}} \right\|_2 \geq \frac{\min_s \pi_\theta(a^*(s)|S))}{\sqrt{|\mathcal{S}|} \cdot \|d_\rho^{\pi^*}/d_\mu^{\pi^*}\|_\infty} \cdot (V^*(\rho) - V^{\pi_\theta}(\rho)).$$

Definition 4 (Definition 3. in (Mei et al., 2021b)). *The function $f : \Theta \to \mathbb{R}$ satisfies $\beta(\omega)$ non-uniform smoothness if f is differentiable and for all $\omega, \theta \in \Theta$,*

$$\left| f(\theta) - f(\omega) - \langle \frac{df(\omega)}{d\omega}, \theta - \omega \rangle \right| \leq \frac{\beta(\omega)}{2} \cdot \|\theta - \omega\|_2^2, \tag{29}$$

where β is a positive valued function: $\beta : \Theta \to (0, \infty)$.

Lemma 16. *If a function satisfies (29) in Definition 4, then f satisfies*

$$\left\| \frac{df(\theta)}{d\theta} - \frac{df(\omega)}{d\omega} \right\|_2 \leq \beta(\omega) \cdot \|\theta - \omega\|_2.$$

Proof of Lemma 16. Let $\omega, \theta \in \Theta$ and let $p \in \Theta$ be chosen later. By (29), we have the upper bound,

$$D := f(\theta + p) - f(\omega) + f(\omega - p) - f(\theta)$$

$$\leq \big\langle \frac{df(\omega)}{d\omega}, \theta + p - \omega \big\rangle + \frac{\beta(\omega)}{2} \|\theta + p - \omega\|_2^2 + \big\langle \frac{df(\theta)}{d\theta}, \omega - p - \theta \big\rangle + \frac{\beta(\omega)}{2} \|\omega - p - \theta\|_2^2$$

$$= -\big\langle \frac{df(\omega)}{d\omega} - \frac{df(\theta)}{d\theta}, \omega - \theta - p \big\rangle + \beta(\omega) \|\omega - \theta - p\|_2^2$$

and the lower bound

$$D = f(\theta + p) - f(\theta) + f(\omega - p) - f(\omega)$$

$$\geq \big\langle \frac{df(\theta)}{d\theta}, p \big\rangle - \frac{\beta(\omega)}{2} \|p\|_2^2 + \big\langle \frac{df(\omega)}{d\omega}, -p \big\rangle - \frac{\beta(\omega)}{2} \|p\|_2^2$$

$$= -\big\langle \frac{df(\omega)}{d\omega} - \frac{df(\theta)}{d\theta}, p \big\rangle - \beta(\omega) \|p\|_2^2.$$

By combining the above upper and lower bounds, we obtain

$$-\big\langle \frac{df(\omega)}{d\omega} - \frac{df(\theta)}{d\theta}, \omega - \theta - 2p \big\rangle \leq \beta(\omega) \|\omega - \theta - p\|_2^2 + \beta(\omega) \|p\|_2^2.$$

Taking $p = \frac{1}{2}[\omega - \theta - \frac{1}{\beta(\omega)}(\frac{df(\omega)}{d\omega} - \frac{df(\theta)}{d\theta})]$ implies that

$$\frac{1}{\beta(\omega)} \left\| \frac{df(\omega)}{d\omega} - \frac{df(\theta)}{d\theta} \right\|_2^2 \leq \frac{\beta(\omega)}{4} \left\| \omega - \theta + \frac{1}{\beta(\omega)}(\frac{df(\omega)}{d\omega} - \frac{df(\theta)}{d\theta}) \right\|_2^2 + \frac{\beta(\omega)}{4} \left\| \omega - \theta - \frac{1}{\beta(\omega)}(\frac{df(\omega)}{d\omega} - \frac{df(\theta)}{d\theta}) \right\|_2^2$$

$$= \frac{\beta(\omega)}{2} \|\omega - \theta\|_2^2 + \frac{1}{2\beta(\omega)} \left\| \frac{df(\omega)}{d\omega} - \frac{df(\theta)}{d\theta} \right\|_2^2.$$

By rearraging the terms, we have

$$\left\| \frac{df(\theta)}{d\theta} - \frac{df(\omega)}{d\omega} \right\|_2^2 \leq \beta(\omega) \cdot \|\theta - \omega\|_2^2.$$

By taking square root on each side, we obtain the desired result. $\qquad\square$

Lemma 17 (Lemma 6. in (Mei et al., 2021b)). *Denote $\omega_\zeta := \omega + \zeta \cdot (\theta - \omega)$ with some $\zeta \in [0,1]$. $\theta \to V^{\pi_\omega}(\mu)$ satisfies $\beta(\theta_\zeta)$ non-uniform smoothness with*

$$\beta(\theta_\zeta) = \left[3 + \frac{4 \cdot (C_\infty - (1 - \gamma))}{1 - \gamma}\right] \cdot \sqrt{|\mathcal{S}|} \cdot \|\nabla_\theta V^{\pi_\theta}|_{\theta = \omega_\zeta}\|_2,$$

where $C_\infty := \max_\pi \|\frac{d_\mu^\pi}{\mu}\|_\infty \leq \frac{1}{\min_s \mu(s)} < \infty$.

Lemma 18 (Lemma 7. in (Mei et al., 2021b)). *Let $\eta = \frac{1-\gamma}{6 \cdot (1-\gamma) + 8 \cdot (C_\infty - (1-\gamma))} \cdot \frac{1}{\sqrt{|\mathcal{S}|}}$ and*

$$\theta \leftarrow \omega + \eta \cdot \nabla_\theta V^{\pi_\theta}|_{\theta = \omega} / \|\nabla_\theta V^{\pi_\theta}|_{\theta = \omega}\|_2.$$

Denote $\omega_\zeta := \omega + \zeta \cdot (\theta - \omega)$ with some $\zeta \in [0,1]$. We have,

$$\|\nabla_\theta V^{\pi_\theta}|_{\theta = \omega_\zeta}\|_2 \leq 2 \|\nabla_\theta V^{\pi_\theta}|_{\theta = \omega}\|_2.$$

Lemma 19. *Consider softmax parameterization. Given $C > 1$ and a probability distribution μ over the state space $\mathcal{S}$, we let*

$$M_C := \ln \left[\frac{|\mathcal{S}||\mathcal{A}|^2}{(C-1)(1-\gamma)^2 \min_{s \in \mathcal{S}} \mu(s)} \right].$$

For any two parameters $\theta, \theta' \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, if

$$\min\{\theta'_{s,a^*(s)} - \theta'_{s,a}, \theta_{s,a^*(s)} - \theta_{s,a}\} > M_C, \quad \forall s \in \mathcal{S}, \forall a \neq a^*(s), \tag{30}$$

then we have

$$\left\| \frac{d_\mu^{\pi_{\theta'}}}{d_\mu^{\pi_\theta}} \right\|_\infty < C.$$

Remark 13. Lemma 19 is a general property of RL under softmax policies, regardless of the algorithm applied. In our subsequent proof of Theorem 2, we will let θ and θ' respectively represent the parameters before and after an APG update. Moreover, we also establish in Lemma 1 that the condition (30) is indeed achievevable for APG.

Proof of Lemma 19. Using Lemma 9 and (30), we have

$$
\left\| \pi_{\theta'} - \pi_\theta \right\|_\infty \leq \max_{s \in \mathcal{S}} \left\{ 1 - \min\{\pi_\theta(a^*(s)|s), \pi_{\theta'}(a^*(s)|s)\} \right\}
$$

$$
\leq \max_{s \in \mathcal{S}} \left\{ 1 - \min\{ \frac{e^{\theta_{s,a^*(s)}}}{e^{\theta_{s,a^*(s)}} + (|\mathcal{A}| - 1)e^{\theta_{s,a^*(s)} - M_C}}, \frac{e^{\theta'_{s,a^*(s)}}}{e^{\theta'_{s,a^*(s)}} + (|\mathcal{A}| - 1)e^{\theta'_{s,a^*(s)} - M_C}} \} \right\}
$$

$$
< \frac{|\mathcal{A}|e^{-M_C}}{1 + |\mathcal{A}|e^{-M_C}}
$$

$$
< \frac{(C - 1)(1 - \gamma)^2 \min_{s \in \mathcal{S}} \mu(s)}{|\mathcal{S}||\mathcal{A}|}. \tag{31}
$$

Next, we characterize the ratio $d_\mu^{\pi_{\theta'}}(s)/d_\mu^{\pi_\theta}(s)$ for a fixed s. We can regard $d_\mu^\pi(s)$ as the value function of a specific MDP. Specifically, we construct this MDP, denoted as $\mathcal{M}(s)$, as follows. We let $\mathcal{M}(s)$ share identical parameters with $\mathcal{M}$ defined in Section 3. The only distinction lies in its reward function $r_{\mathcal{M}(s)}$, which we define as $r_{\mathcal{M}(s)}(s', a') = (1 - \gamma)\mathbb{I}\{s' = s\}$, where $\mathbb{I}\{s' = s\}$ is the indicator function whose value is 1 if $s' = s$ and 0 otherwise. We can observe that

$$
\mathbb{E}\Big[\sum_t^\infty \gamma^t (1 - \gamma)\mathbb{I}\{s_t = s\} \Big| \pi, s_0 \sim \mu \Big] = (1 - \gamma) \sum_t^\infty \gamma^t Pr(s_t = s|\pi, s_0 \sim \mu). \tag{32}
$$

We denote $V_{\mathcal{M}(s)}^\pi(\mu)$ as the value function of π calculated under the MDP $\mathcal{M}(s)$. Moreover, we define $A_{\mathcal{M}(s)}^\pi$ as the advantage function and $d_{\mathcal{M}(s),\mu}^\pi$ as the state visitation distribution of π in $\mathcal{M}(s)$, respectively. By the definition of the state visitation distribution and the value function, together with (32), we obtain $d_\mu^\pi(s) = V_{\mathcal{M}(s)}^\pi(\mu)$ for any π. Consequently, we have

$$
\left| d_\mu^{\pi_{\theta'}}(s) - d_\mu^{\pi_\theta}(s) \right| = \left| V_{\mathcal{M}(s)}^{\pi_{\theta'}}(\mu) - V_{\mathcal{M}(s)}^{\pi_\theta}(\mu) \right|
$$

$$
= \Big| \sum_{s',a'} d_{\mathcal{M}(s),\mu}^{\pi_{\theta'}}(s')\pi_{\theta'}(a'|s')A_{\mathcal{M}(s)}^{\pi_\theta}(s', a') \Big| \tag{33}
$$

$$
= \Big| \sum_{s',a'} d_{\mathcal{M}(s),\mu}^{\pi_{\theta'}}(s')(\pi_{\theta'}(a'|s') - \pi_\theta(a'|s'))A_{\mathcal{M}(s)}^{\pi_\theta}(s', a') \Big| \tag{34}
$$

$$
\leq \sum_{s',a'} \Big| \underbrace{d_{\mathcal{M}(s),\mu}^{\pi_{\theta'}}(s')}_{\leq 1}(\pi_{\theta'}(a'|s') - \pi_\theta(a'|s')) \underbrace{A_{\mathcal{M}(s)}^{\pi_\theta}(s', a')}_{\leq 1 \text{ by Lemma 10}} \Big|
$$

$$
\leq \frac{1}{1 - \gamma} \sum_{s',a'} \Big| \pi_{\theta'}(a'|s') - \pi_\theta(a'|s') \Big|
$$

$$
\leq \frac{|\mathcal{S}||\mathcal{A}|}{1 - \gamma} \cdot \left\| \pi_{\theta'} - \pi_\theta \right\|_\infty
$$

$$
< (C - 1)(1 - \gamma) \min_{s' \in \mathcal{S}} \mu(s') \tag{35}
$$

$$
\leq (C - 1)(1 - \gamma)\mu(s), \tag{36}
$$

where (33) holds by leveraging Lemma 4 in the MDP $\mathcal{M}(s)$, (34) is from Lemma 7, and we used (31) to obtain (35). Finally, (36) leads to

$$
\frac{d_\mu^{\pi_{\theta'}}(s)}{d_\mu^{\pi_\theta}(s)} < \frac{d_\mu^{\pi_\theta}(s) + (C - 1)(1 - \gamma)\mu(s)}{d_\mu^{\pi_\theta}(s)}
$$

$$
= 1 + \frac{(C - 1)(1 - \gamma)\mu(s)}{d_\mu^{\pi_\theta}(s)}
$$

$$\leq 1 + \frac{(C-1)(1-\gamma)\mu(s)}{(1-\gamma)\mu(s)} \tag{37}$$

$$= C,$$

where (37) used the fact that $d_\mu^{\pi_\theta}(s) \geq (1-\gamma)\mu(s)$ for all $s \in \mathcal{S}$. Since s is arbitrary, the desired result is obtained by taking the maximum over all $s \in \mathcal{S}$. $\qquad\square$

According to Definition 2, $\mathcal{U}$ is a set consisting of vectors in the high-dimensional space of $\mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$. Fortunately, by focusing at one state s at a time, we can greatly simplify the analysis. We first define the following notations and then prove a critical lemma for our analysis.

Definition 5. *Consider any $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and $\boldsymbol{d} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. For all $s \in \mathcal{S}$, we define $\theta_{s_i,\cdot} \in \mathbb{R}^{|\mathcal{A}|}$ and $\boldsymbol{d}_{\boldsymbol{s}_i,\cdot} \in \mathbb{R}^{|\mathcal{A}|}$ to be respectively the subvectors of θ and $\boldsymbol{d}$ along the state s_i as follows:*

$$\theta_{s_i,\cdot} := [\theta_{s_i,a^*(s_i)}, \theta_{s_i,a_2(s_i)}, \cdots, \theta_{s_i,a_{|\mathcal{A}|}(s_i)}], \quad \boldsymbol{d}_{\boldsymbol{s}_i,\cdot} := [d_{s_i,a^*(s_i)}, d_{s_i,a_2(s_i)}, \cdots, d_{s_i,a_{|\mathcal{A}|}(s_i)}].$$

Lemma 20. *Assume that $|\mathcal{A}| \geq 2$. Consider any vector $\boldsymbol{d}$ in the feasible update domain $\mathcal{U}$ such that $\boldsymbol{d}_{\boldsymbol{s},\cdot}$ is a unit vector for all $s \in \mathcal{S}$. Let*

$$M_d := 2 \max_{s \in \mathcal{S}, i > 1} \left\{ \ln\left[2(|\mathcal{A}| - 1)\right] - \ln[d_{s,a^*(s)} - d_{s,a_i(s)}] \right\}. \tag{38}$$

Let $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ be a vector such that $\theta_{s,a^*(s)} - \theta_{s,a} > M_d$ for all $s \in \mathcal{S}$ and $a \neq a^*(s)$. Then, under softmax parameterization, we have the following

- *The function $\theta \mapsto \pi_\theta(a|s)$ is convex along the direction $\boldsymbol{d}$ for all $s \in \mathcal{S}$ and $a \neq a^*(s)$.*

- *The function $\theta \mapsto \pi_\theta(a^*(s)|s)$ is concave along the direction $\boldsymbol{d}$ for all $s \in \mathcal{S}$.*

Proof of Lemma 20. Given that the function $\theta \to \pi_\theta(a|s)$ is independent of any $s' \neq s$, for the first item it suffices to establish the convexity of the function $\theta_{s,\cdot} \to \pi_\theta(a|s)$ alone $\boldsymbol{d}_{\boldsymbol{s},\cdot}$ for all $a \neq a^*(s)$ for any given $s \in \mathcal{S}$. Following that, since $\pi_\theta(a^*(s)|s) = 1 - \sum_{a \neq a^*} \pi_\theta(a|s)$ can be viewed as a finite summation of concave functions along $\boldsymbol{d}$ with respect to θ, $\theta \to \pi_\theta(a^*(s)|s)$ is a concave along $\boldsymbol{d}_{\boldsymbol{s},\cdot}$, and the second item ensues.

We first note that $\ln[2(|\mathcal{A}|-1)] > 0$ because $|\mathcal{A}| \geq 2$, and $\ln[d_{s,a^*(s)} - d_{s,a_i(s)}] \leq 0$ because $d_{s,a^*(s)} - d_{s,a_i(s)} \in (0,1]$ for all i by the definition of $\mathcal{U}$ and the fact that $\|\boldsymbol{d}_{\boldsymbol{s},\cdot}\| \leq 1$. Therefore, we see that $M_d > 0$. Given an action $a_i(s) \neq a^*(s)$, since the convexity of a function is determined by its behavior along arbitrary lines in its domain (the domain $\mathcal{U}$ is clearly convex), it suffices to show that the following twice-differentiable function is concave (*i.e.*, the second derivative is non-positive) along the unit vector $\boldsymbol{d}_{\boldsymbol{s},\cdot}$ when $k \to 0$:

$$f(k) := \frac{e^{\theta_{s,a_i(s)} + k \cdot d_{s,a_i(s)}}}{e^{\theta_{s,a^*(s)} + k \cdot d_{s,a^*(s)}} + e^{\theta_{s,a_2(s)} + k \cdot d_{s,a_2(s)}} + \cdots + e^{\theta_{s,a_{|\mathcal{A}|}(s)} + k \cdot d_{s,a_{|\mathcal{A}|}(s)}}}$$

$$= \frac{1}{e^{(\theta_{s,a^*(s)} - \theta_{s,a_i(s)}) + k \cdot (d_{s,a^*(s)} - d_{s,a_i(s)})} + \sum_{j=2}^{|\mathcal{A}|} e^{(\theta_{s,a_j(s)} - \theta_{s,a_i(s)}) + k \cdot (d_{s,a_j(s)} - d_{s,a_i(s)})}}.$$

By defining $g(k) := 1/f(k)$ and taking the second derivative of $f(k)$, we have

$$f''(k) = \frac{2(g'(k))^2}{g(k)^3} - \frac{g''(k)}{g(k)^2}.$$

Let $k \to 0$, we then get

$$f''(0) = \frac{1}{g(0)^2} \left(\frac{2(g'(0))^2}{g(0)} - g''(0) \right).$$

Since $g(k) \geq 0$ for any $k \in \mathbb{R}$, $f''(0) \geq 0$ if and only if

$$2(g'(0))^2 - g''(0) \cdot g(0) \geq 0. \tag{39}$$

By direct calculation, we have $m'(0) = \sum_{j \neq i}(d_j - d_i)\exp(\theta_{s,a_j(s)} - \theta_{s,a_i(s)})$ and $m''(0) = \sum_{j \neq i}(d_j - d_i)^2 \exp(\theta_{s,a_j(s)} - \theta_{s,a_i(s)})$. Therefore, (39) leads to

$$
2(g'(0))^2 - g''(0) \cdot g(0)
$$

$$
= 2\sum_{j \neq i}(d_{s,a_j(s)} - d_{s,a_i(s)})^2 \exp(2\theta_{s,a_j(s)} - 2\theta_{s,a_i(s)}) \tag{40}
$$

$$
+ 2 \sum_{u,v \neq i, u < v} 2(d_{s,a_u(s)} - d_{s,a_i(s)})(d_{s,a_v(s)} - d_{s,a_i(s)})\exp(\theta_{s,a_u(s)} + \theta_{s,a_v(s)} - 2\theta_{s,a_i(s)})
$$

$$
- \sum_{j \neq i}(d_{s,a_j(s)} - d_{s,a_i(s)})^2 \exp(2\theta_{s,a_j(s)} - 2\theta_{s,a_i(s)}) - \sum_{j \neq i}(d_{s,a_j(s)} - d_{s,a_i(s)})^2 \exp(\theta_{s,a_j(s)} - \theta_{s,a_i(s)})
$$

$$
- \sum_{u,v \neq i, u < v}\left((d_{s,a_u(s)} - d_{s,a_i(s)})^2 + (d_{s,a_v(s)} - d_{s,a_i(s)})^2\right)\exp(\theta_{s,a_u(s)} + \theta_{s,a_v(s)} - 2\theta_{s.a_i(s)})
$$

$$
= \sum_{j \neq i}(d_{s,a_j(s)} - d_{s,a_i(s)})^2 \left(\exp(2\theta_{s,a_j(s)} - 2\theta_{s,a_i(s)}) - \exp(\theta_{s,a_j(s)} - \theta_{s,a_i(s)})\right)
$$

$$
+ \sum_{u,v \neq i, u < v} 2(d_{s,a_u(s)} - d_{s,a_i(s)})(d_{s,a_v(s)} - d_{s,a_i(s)})\exp(\theta_{s,a_u(s)} + \theta_{s,a_v(s)} - 2\theta_{s,a_i(s)})
$$

$$
- \sum_{u,v \neq i, u < v}(d_{s,a_u(s)} - d_{s,a_v(s)})^2 \exp(\theta_{s,a_u(s)} + \theta_{s,a_v(s)} - 2\theta_{s,a_i(s)})
$$

$$
\geq (d_{s,a^*(s)} - d_{s,a_i(s)})^2 \left(\exp(2\theta_{s,a^*(s)} - 2\theta_{s,a_i(s)}) - \exp(\theta_{s,a^*(s)} - \theta_{s,a_i(s)})\right)
$$

$$
- \sum_{j \neq i, a_j \neq a^*}(d_{s,a_j(s)} - d_{s,a_i(s)})^2 \exp(\theta_{s,a_j(s)} - \theta_{s,a_i(s)})
$$

$$
+ \sum_{u,v \neq i, u < v} 2(d_{s,a_u(s)} - d_{s,a_i(s)})(d_{s,a_v(s)} - d_{s,a_i(s)})\exp(\theta_{s,a_u(s)} + \theta_{s,a_v(s)} - 2\theta_{s,a_i(s)})
$$

$$
- \sum_{u,v \neq i, u < v}(d_{s,a_u(s)} - d_{s,a_v(s)})^2 \exp(\theta_{s,a_u(s)} + \theta_{s,a_v(s)} - 2\theta_{s,a_i(s)})
$$

$$
\geq (d_{s,a^*(s)} - d_{s,a_i(s)})^2 \left(\exp(2\theta_{s,a^*(s)} - 2\theta_{s,a_i(s)}) - \exp(\theta_{s,a^*(s)} - \theta_{s,a_i(s)})\right)
$$

$$
- \sum_{j \neq i, a_j \neq a^*} 2\exp(\theta_{s,a_j(s)} - \theta_{s,a_i(s)})
$$

$$
- \sum_{u,v \neq i, u < v} 2\exp(\theta_{s,a_u(s)} + \theta_{s,a_v(s)} - 2\theta_{s,a_i(s)}) - \sum_{u,v \neq i, u < v} 2\exp(\theta_{s,a_u(s)} + \theta_{s,a_v(s)} - 2\theta_{s,a_i(s)}) \tag{41}
$$

$$
\geq (d_{s,a^*(s)} - d_{s,a_i(s)})^2 \exp(\theta_{s,a^*(s)} - \theta_{s,a_i(s)})\left(\exp(\theta_{s,a^*(s)} - \theta_{s,a_i(s)}) - 1\right)
$$

$$
- 2(|\mathcal{A}| - 1)\max_{j \neq i, a_j \neq a^*}\exp(\theta_{s,a_j(s)} - \theta_{s,a_i(s)}) - 4\binom{|\mathcal{A}| - 1}{2}\max_{u,v \neq i, u < v}\exp(\theta_{s,a_u(s)} + \theta_{s,a_v(s)} - 2\theta_{s,a_i(s)})
$$

$$
\geq (d_{s,a^*(s)} - d_{s,a_i(s)})^2 \exp(\theta_{s,a^*(s)} - \theta_{s,a_i(s)})\left(\exp(\theta_{s,a^*(s)} - \theta_{s,a_i(s)}) - 1\right)
$$

$$
- 2(|\mathcal{A}| - 1)\exp(\theta_{s,a^*(s)} - \theta_{s,a_i(s)} - M_d) - 4\binom{|\mathcal{A}| - 1}{2}\exp(2\theta_{s,a^*(s)} - 2\theta_{s,a_i(s)} - M_d), \tag{42}
$$

$$
\geq \exp(\theta_{s,a^*(s)} - \theta_{s,a_i(s)})\Big[(d_{s,a^*(s)} - d_{s,a_i(s)})^2(\exp(\theta_{s,a^*(s)} - \theta_{s,a_i(s)}) - 1)
$$

$$
- 2(|\mathcal{A}| - 1)^2 \exp(\theta_{s,a^*(s)} - \theta_{s,a_i(s)} - M_d)\Big], \tag{43}
$$

where (41) holds by using the bounds $(d_{s,a_u(s)} - d_{s,a_i(s)})(d_{s,a_v(s)} - d_{s,a_i(s)}) \geq -1$ and $(d_{s,a_u(s)} - d_{s,a_v(s)})^2 \leq 2$ (both from the Cauchy-Schwarz inequality and that $\|d_{s,\cdot}\| = 1$), and (42) and (43) is from our hypothesis of $\theta_{s,a^*(s)} - \theta_{s,a} > M_d > 0$ for all $a \neq a^*(s)$.

Since $\exp(\theta_{s,a^*(s)} - \theta_{s,a_i(s)}) > 0$, to make (43) non-negative, we simply need to verify that

$$
(d_{s,a^*(s)} - d_{s,a_i(s)})^2(\exp(\theta_{s,a^*(s)} - \theta_{s,a_i(s)}) - 1) \geq 2(|\mathcal{A}| - 1)^2 \exp(\theta_{s,a^*(s)} - \theta_{s,a_i(s)} - M_d),
$$

which is equivalent to

$$\exp(-M_d) \leq \frac{(d_{s,a^*(s)} - d_{s,a_i(s)})^2 (\exp(\theta_{s,a^*(s)} - \theta_{s,a_i(s)}) - 1)}{2(|\mathcal{A}| - 1)^2 \exp(\theta_{s,a^*(s)} - \theta_{s,a_i(s)})}$$

$$= \frac{(d_{s,a^*(s)} - d_{s,a_i(s)})^2}{2(|\mathcal{A}| - 1)^2} \cdot \left(1 - \frac{1}{\exp(\theta_{s,a^*(s)} - \theta_{s,a_i(s)})}\right).$$

From (38), since $2(|\mathcal{A}| - 1) > \sqrt{2}$ and $-\ln[d_{s,a^*(s)} - d_{s,a_i(s)}] \geq 0$ for all $i > 1$ (as argued before), we get $M_d \geq \ln(2)$. Therefore, if $\theta_{s,a^*(s)} - \theta_{s,a_i(s)} > M_d$ for all $s \in \mathcal{S}$ and all $i > 1$, we indeed obtain

$$\exp(-M_d) \leq \frac{(d_{s,a^*(s)} - d_{s,a_i(s)})^2}{4(|\mathcal{A}| - 1)^2}$$

$$= \frac{(d_{s,a^*(s)} - d_{s,a_i(s)})^2}{2(|\mathcal{A}| - 1)^2}\left(1 - \frac{1}{2}\right)$$

$$\leq \frac{(d_{s,a^*(s)} - d_{s,a_i(s)})^2}{2(|\mathcal{A}| - 1)^2}\left(1 - \frac{1}{\exp(M_d)}\right) \tag{44}$$

$$\leq \frac{(d_{s,a^*(s)} - d_{s,a_i(s)})^2}{2(|\mathcal{A}| - 1)^2}\left(1 - \frac{1}{\exp(\theta_{s,a^*(s)} - \theta_{s,a_i(s)})}\right), \tag{45}$$

where (44) and (45) hold by $\theta_{s,a^*(s)} - \theta_{s,a_i(s)} > M_d \geq \ln(2)$ for all $i > 1$. $\qquad\square$

Lemma 21. *Under softmax parameterization, let $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ be a parameter such that $V^{\pi_\theta}(s) > Q^*(s, a_2(s))$ for all $s \in \mathcal{S}$. Then, for any $s \in \mathcal{S}$ and any $\theta' := \theta + d$ with $d \in \mathcal{U}$, we have*

$$\sum_{a \in \mathcal{A}} \left(\pi_{\theta'}(a|s) - \pi_\theta(a|s)\right) A^{\pi_\theta}(s, a) \geq 0.$$

Proof of Lemma 21. By Lemma 8 and our assumption on $V^{\pi_\theta}(s)$, we see that $V^{\pi_\theta}(s) > Q^*(s, a_2(s)) \geq Q^{\pi_\theta}(s, a_i(s))$ for all $i \geq 2$. From (2), this means

$$A^{\pi_\theta}(s, a) \leq 0, \quad \forall a \neq a^*(s). \tag{46}$$

Combining (46) with Lemma 7 then gives

$$A^{\pi_\theta}(s, a^*(s)) \geq 0. \tag{47}$$

Since $d \in \mathcal{U}$, we have $d_{s,a^*(s)} > d_{s,a}$ for all $s \in \mathcal{S}$ and all $a \neq a^*(s)$. Therefore,

$$\sum_{a \in \mathcal{A}} (\pi_{\theta'}(a|s) - \pi_\theta(a|s)) A^{\pi_\theta}(s, a)$$

$$= \sum_{a \in \mathcal{A}} \pi_{\theta'}(a|s) A^{\pi_\theta}(s, a) \tag{48}$$

$$= \sum_{a \in \mathcal{A}} \frac{\exp(\theta'_{s,a})}{\sum_{a'} \exp(\theta'_{s,a'})} A^{\pi_\theta}(s, a)$$

$$= \frac{\sum_{a'} \exp(\theta_{s,a'})}{\sum_{a'} \exp(\theta'_{s,a'})} \sum_{a \in \mathcal{A}} \frac{\exp(\theta_{s,a} + d_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})} A^{\pi_\theta}(s, a)$$

$$= \frac{\sum_{a'} \exp(\theta_{s,a'})}{\sum_{a'} \exp(\theta'_{s,a'} - \max_{a \neq a^*(s)} d_{s,a})} \sum_{a \in \mathcal{A}} \frac{\exp(\theta_{s,a} + (d_{s,a} - \max_{a \neq a^*(s)} d_{s,a}))}{\sum_{a'} \exp(\theta_{s,a'})} A^{\pi_\theta}(s, a)$$

$$= \frac{\sum_{a'} \exp(\theta_{s,a'})}{\sum_{a'} \exp(\theta'_{s,a'} - \max_{a \neq a^*(s)} d_{s,a})} \left(\frac{\exp(\theta_{s,a^*(s)} + (d_{s,a^*(s)} - \max_{a \neq a^*(s)} d_{s,a}))}{\sum_{a'} \exp(\theta_{s,a'})} A^{\pi_\theta}(s, a^*(s))\right.$$

$$\left. + \sum_{a \neq a^*(s)} \frac{\exp(\theta_{s,a} + (d_{s,a} - \max_{a \neq a^*(s)} d_{s,a}))}{\sum_{a'} \exp(\theta_{s,a'})} A^{\pi_\theta}(s, a)\right)$$

$$\geq \frac{\sum_{a'} \exp\left(\theta_{s,a'}\right)}{\sum_{a'} \exp\left(\theta'_{s,a'} - \max_{a \neq a^*(s)} d_{s,a}\right)} \left(\frac{\exp\left(\theta_{s,a^*(s)}\right)}{\sum_{a'} \exp\left(\theta_{s,a'}\right)} A^{\pi_\theta}(s, a^*(s)) + \sum_{a \neq a^*(s)} \frac{\exp\left(\theta_{s,a}\right)}{\sum_{a'} \exp\left(\theta_{s,a'}\right)} A^{\pi_\theta}(s, a)\right) \quad (49)$$

$$= \frac{\sum_{a'} \exp\left(\theta_{s,a'}\right)}{\sum_{a'} \exp\left(\theta'_{s,a'} - \max_{a \neq a^*(s)} d_{s,a}\right)} \sum_{a \in \mathcal{A}} \pi_\theta(a|s) A^{\pi_\theta}(s, a) = 0, \quad (50)$$

where (48) and (50) hold due to Lemma 7, and (49) holds by the definition of $\mathcal{U}$ that $d_{s,a^*(s)} > \max_{a \neq a^*(s)} d_{s,a} \geq d_{s,a(s)}$ for all $s \in \mathcal{S}$ and all $a \neq a^*$ and (46) and (47). $\qquad\square$

Lemma 22. *Under softmax parameterization, for APG, we have $V^{\pi_\theta^{(t+1)}}(s) \geq V^{\pi_\omega^{(t)}}(s)$ for any $s \in \mathcal{S}$ and any $t \geq 0$ as long as $\eta^{(t} > 0$.*

Proof of Lemma 22. By Lemma 4, it suffices to show that $\sum_{a \in \mathcal{A}} \pi_\theta^{(t+1)}(a|s) A^{\pi_\omega^{(t)}}(s, a) \geq 0$ for all $s \in \mathcal{S}$. Indeed, by defining

$$Z_\theta^{(t)}(s) := \sum_{a \in \mathcal{A}} \exp(\theta_{s,a}^{(t)}), \quad Z_\omega^{(t)}(s) := \sum_{a \in \mathcal{A}} \exp(\omega_{s,a}^{(t)}), \quad \partial_{s,a}^{(t)} := \left. \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} \right|_{\theta = \omega^{(t)}},$$

we have that for any $s \in \mathcal{S}$,

$$\sum_{a \in \mathcal{A}} \pi_\theta^{(t+1)}(a|s) A^{\pi_\omega^{(t)}}(s, a) = \sum_{a \in \mathcal{A}} \frac{\exp(\theta_{s,a}^{(t+1)})}{Z_\theta^{(t+1)}(s)} A^{\pi_\omega^{(t)}}(s, a)$$

$$= \frac{Z_\omega^{(t)}(s)}{Z_\theta^{(t+1)}(s)} \sum_{a \in \mathcal{A}} \frac{\exp(\theta_{s,a}^{(t+1)})}{Z_\omega^{(t)}(s)} A^{\pi_\omega^{(t)}}(s, a)$$

$$= \frac{Z_\omega^{(t)}(s)}{Z_\theta^{(t+1)}(s)} \sum_{a \in \mathcal{A}} \frac{\exp(\omega_{s,a}^{(t)} + \eta^{(t+1)} \partial_{s,a}^{(t)})}{Z_\omega^{(t)}(s)} A^{\pi_\omega^{(t)}}(s, a)$$

$$\geq \frac{Z_\omega^{(t)}(s)}{Z_\theta^{(t+1)}(s)} \sum_{a \in \mathcal{A}} \frac{\exp(\omega_{s,a}^{(t)})}{Z_\omega^{(t)}(s)} A^{\pi_\omega^{(t)}}(s, a) \quad (51)$$

$$= \frac{Z_\omega^{(t)}(s)}{Z_\theta^{(t+1)}(s)} \sum_{a \in \mathcal{A}} \pi_\omega^{(t)}(a|s) A^{\pi_\omega^{(t)}}(s, a) = 0, \quad (52)$$

where (51) holds by the fact that the sign of $\eta^{(t+1)} \partial_{s,a}^{(t)}$ is the same as $A^{\pi_\omega^{(t)}}(s, a)$ by Lemma 11, and (52) is true due to Lemma 7. $\qquad\square$

B.3. Useful Properties of Accelerated Policy Gradient

Throughout the appendices, We use $\nabla_{s,a}^{(t)}$ as the shorthand for $\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\big|_{\theta=\omega^{(t)}}$. By L-smooth, we mean a function is differentiable with its gradient L-Lipschitz continuous. For any pair of positive integers (j, t), we define

$$
G(j,t) := \begin{cases}
1 & \text{, if } t = j \text{ or } j = 0, \\
1 + \mathbb{I}\left\{V^{\pi_\varphi^{(j+h)}}(\mu) \geq V^{\pi_\theta^{(j+h)}}(\mu) \text{ for } h = 1\right\}\frac{j}{j+3} & \text{, if } t = j+1, \\[2mm]
\begin{aligned}
1 &+ \mathbb{I}\left\{V^{\pi_\varphi^{(j+h)}}(\mu) \geq V^{\pi_\theta^{(j+h)}}(\mu) \text{ for } h = 1\right\}\frac{j}{j+3} \\
&+ \mathbb{I}\left\{V^{\pi_\varphi^{(j+h)}}(\mu) \geq V^{\pi_\theta^{(j+h)}}(\mu) \text{ for all } h = 1,2\right\}\frac{(j+1)j}{(j+4)(j+3)}
\end{aligned} & \text{, if } t = j+2, \\[4mm]
\begin{aligned}
1 &+ \mathbb{I}\left\{V^{\pi_\varphi^{(j+h)}}(\mu) \geq V^{\pi_\theta^{(j+h)}}(\mu) \text{ for } h = 1\right\}\frac{j}{j+3} \\
&+ \mathbb{I}\left\{V^{\pi_\varphi^{(j+h)}}(\mu) \geq V^{\pi_\theta^{(j+h)}}(\mu) \text{ for all } h = 1,2\right\}\frac{(j+1)j}{(j+4)(j+3)} \\
&+ \mathbb{I}\left\{V^{\pi_\varphi^{(j+h)}}(\mu) \geq V^{\pi_\theta^{(j+h)}}(\mu) \text{ for all } h = 1,2,3\right\}\frac{(j+2)(j+1)j}{(j+5)(j+4)(j+3)}
\end{aligned} & \text{, if } t = j+3, \\[6mm]
\begin{aligned}
1 &+ \mathbb{I}\left\{V^{\pi_\varphi^{(j+h)}}(\mu) \geq V^{\pi_\theta^{(j+h)}}(\mu) \text{ for } h = 1\right\}\frac{j}{j+3} \\
&+ \mathbb{I}\left\{V^{\pi_\varphi^{(j+h)}}(\mu) \geq V^{\pi_\theta^{(j+h)}}(\mu) \text{ for all } h = 1,2\right\}\frac{(j+1)j}{(j+4)(j+3)} \\
&+ \mathbb{I}\left\{V^{\pi_\varphi^{(j+h)}}(\mu) \geq V^{\pi_\theta^{(j+h)}}(\mu) \text{ for all } h = 1,2,3\right\}\frac{(j+2)(j+1)j}{(j+5)(j+4)(j+3)} \\
&+ \sum_{k=4}^{t-j} \mathbb{I}\left\{V^{\pi_\varphi^{(j+h)}}(\mu) \geq V^{\pi_\theta^{(j+h)}}(\mu) \text{ for all } h = 1,2,\ldots,k\right\}\frac{(j+2)(j+1)j}{(j+k+2)(j+k+1)(j+k)}
\end{aligned} & \text{, if } t \geq j+4 \\[6mm]
0 & \text{, otherwise.}
\end{cases}
\tag{53}
$$

Lemma 23. *Under APG, we could express the policy parameter as follows:*

a) For $t \in \{1,2,3,4\}$, we have

$$
\theta_{s,a}^{(1)} = \eta^{(1)}\nabla_{s,a}^{(0)} + \theta_{s,a}^{(0)}, \tag{54}
$$

$$
\theta_{s,a}^{(2)} = \eta^{(2)}\nabla_{s,a}^{(1)} + \eta^{(1)}\nabla_{s,a}^{(0)} + \theta_{s,a}^{(0)} \tag{55}
$$

$$
\theta_{s,a}^{(3)} = \eta^{(3)}\nabla_{s,a}^{(2)} + \eta^{(2)}\Big(1 + \mathbb{I}\Big\{V^{\pi_\varphi^{(2)}}(\mu) \geq V^{\pi_\theta^{(2)}}(\mu)\Big\} \cdot \frac{1}{4}\Big)\nabla_{s,a}^{(1)} + \eta^{(1)}\nabla_{s,a}^{(0)} + \theta_{s,a}^{(0)} \tag{56}
$$

$$
\begin{aligned}
\theta_{s,a}^{(4)} =\ & \eta^{(4)}\nabla_{s,a}^{(3)} \\
&+ \eta^{(3)}\Big(1 + \mathbb{I}\Big\{V^{\pi_\varphi^{(3)}}(\mu) \geq V^{\pi_\theta^{(3)}}(\mu)\Big\} \cdot \frac{2}{5}\Big)\nabla_{s,a}^{(2)} \\
&+ \eta^{(2)}\Big(1 + \mathbb{I}\Big\{V^{\pi_\varphi^{(2)}}(\mu) \geq V^{\pi_\theta^{(2)}}(\mu)\Big\}\Big(\frac{1}{4} + \mathbb{I}\Big\{V^{\pi_\varphi^{(3)}}(\mu) \geq V^{\pi_\theta^{(3)}}(\mu)\Big\} \cdot \frac{2\cdot1}{5\cdot4}\Big)\Big)\nabla_{s,a}^{(1)} \\
&+ \eta^{(1)}\nabla_{s,a}^{(0)} \\
&+ \theta_{s,a}^{(0)}
\end{aligned}
\tag{57}
$$

b) For $t \geq 4$, we have

$$
\begin{aligned}
\theta_{s,a}^{(t+1)} =\ & \eta^{(t+1)}\nabla_{s,a}^{(t)} \\
&+ \eta^{(t)}\Big(1 + \mathbb{I}\Big\{V^{\pi_\varphi^{(t)}}(\mu) \geq V^{\pi_\theta^{(t)}}(\mu)\Big\} \cdot \frac{t-1}{t+2}\Big)\nabla_{s,a}^{(t-1)} \\
&+ \eta^{(t-1)}\Big(1 + \mathbb{I}\Big\{V^{\pi_\varphi^{(t-1)}}(\mu) \geq V^{\pi_\theta^{(t-1)}}(\mu)\Big\} \cdot \\
&\qquad \Big(\frac{t-2}{t+1} + \mathbb{I}\Big\{V^{\pi_\varphi^{(t)}}(\mu) \geq V^{\pi_\theta^{(t)}}(\mu)\Big\} \cdot \frac{(t-1)(t-2)}{(t+2)(t+1)}\Big)\Big)\nabla_{s,a}^{(t-2)} \\
&+ \sum_{j=1}^{t-3} \eta^{(j+1)}\Big(1 + \mathbb{I}\Big\{V^{\pi_\varphi^{(j+1)}}(\mu) \geq V^{\pi_\theta^{(j+1)}}(\mu)\Big\} \cdot \\
&\qquad \Big(\frac{j}{j+3} + \mathbb{I}\Big\{V^{\pi_\varphi^{(j+2)}}(\mu) \geq V^{\pi_\theta^{(j+2)}}(\mu)\Big\} \cdot
\end{aligned}
$$

$$\left(\frac{(j+1)j}{(j+4)(j+3)} + \mathbb{I}\left\{ V^{\pi_\varphi^{(j+3)}}(\mu) \geq V^{\pi_\theta^{(j+3)}}(\mu) \right\} \cdot$$

$$\left(\frac{(j+2)(j+1)j}{(j+5)(j+4)(j+3)} + \sum_{k=4}^{t-j} \mathbb{I}\left\{ V^{\pi_\varphi^{(j+k)}}(\mu) \geq V^{\pi_\theta^{(j+k)}}(\mu) \right\} \cdot$$

$$\left. \frac{(j+2)(j+1)j}{(j+k+2)(j+k+1)(j+k)} \right) \right) \right) \right) \nabla_{s,a}^{(j)}$$

$$+\eta^{(1)} \nabla_{s,a}^{(0)}$$

$$+\theta_{s,a}^{(0)}. \tag{58}$$

Therefore, in summary, we have

$$\theta_{s,a}^{(t+1)} = \sum_{j=0}^{t} G(j,t) \cdot \eta^{(j+1)} \nabla_{s,a}^{(j)} + \theta_{s,a}^{(0)}, \quad \forall t \geq 0. \tag{59}$$

Proof of Lemma 23. Regarding a), one could verify (54)-(57) by directly using the APG update in Algorithm 2.

For b), we prove this by induction. Specifically, suppose (58) holds for all iterations up to t. By the APG update, we know

$$\theta_{s,a}^{(t+1)} = \theta_{s,a}^{(t)} + \eta^{(t+1)} \nabla_{s,a}^{(t)} + \mathbb{I}\left\{ V^{\pi_\varphi^{(t)}}(\mu) \geq V^{\pi_\theta^{(t)}}(\mu) \right\} \frac{t-1}{t+2} (\theta_{s,a}^{(t)} - \theta_{s,a}^{(t-1)}). \tag{60}$$

By plugging into (60) the expressions of $\theta_{s,a}^{(t)}$ and $\theta_{s,a}^{(t-1)}$ from (58), we could see that (58) continues to hold at the $(t+1)$-th iteration. $\qquad\square$

Lemma 24. *Under APG for softmax parameterization, for any iteration k and any state-action pair (s,a), we have*

$$\sum_{a \in \mathcal{A}} \theta_{s,a}^{(k)} = \sum_{a \in \mathcal{A}} \theta_{s,a}^{(0)}.$$

Proof of Lemma 24. We prove this by induction through the following two claims.

Claim (a). $\sum_{a \in \mathcal{A}} \theta_{s,a}^{(1)} = \sum_{a \in \mathcal{A}} \theta_{s,a}^{(0)}$ and $\sum_{a \in \mathcal{A}} \theta_{s,a}^{(2)} = \sum_{a \in \mathcal{A}} \theta_{s,a}^{(0)}$.

Note that under APG, we have

$$\theta_{s,a}^{(1)} = \omega_{s,a}^{(0)} + \eta^{(1)} \left. \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} \right|_{\theta=\omega^{(0)}} = \theta_{s,a}^{(0)} + \eta^{(1)} \cdot \frac{1}{1-\gamma} d^{\pi_\theta^{(0)}}(s) \pi_\theta^{(0)}(a|s) A^{\pi_\theta^{(0)}}(s,a), \tag{61}$$

where the second equality holds by the initial condition of APG (i.e., $\omega^{(0)} = \theta^{(0)}$) as well as the softmax policy gradient in Lemma 11. By summing (61) over all the actions, we have $\sum_{a \in \mathcal{A}} \theta_{s,a}^{(1)} = \sum_{a \in \mathcal{A}} \theta_{s,a}^{(0)}$ from Lemma 7. Similarly, we have

$$\theta_{s,a}^{(2)} = \omega_{s,a}^{(1)} + \eta^{(2)} \cdot \left. \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} \right|_{\theta=\omega^{(1)}}$$

$$= \theta_{s,a}^{(1)} + \mathbb{I}\left\{ V^{\pi_\varphi^{(1)}}(\mu) \geq V^{\pi_\theta^{(1)}}(\mu) \right\} \cdot \frac{0}{3} \cdot (\theta_{s,a}^{(1)} - \theta_{s,a}^{(0)}) + \eta^{(2)} \cdot \frac{1}{1-\gamma} d^{\pi_\omega^{(1)}}(s) \pi_\omega^{(1)}(a|s) A^{\pi_\omega^{(1)}}(s,a). \tag{62}$$

By taking the sum of (62) over all the actions, we have $\sum_{a \in \mathcal{A}} \theta_{s,a}^{(2)} = \sum_{a \in \mathcal{A}} \theta_{s,a}^{(0)}$ by $\sum_{a \in \mathcal{A}} \theta_{s,a}^{(1)} = \sum_{a \in \mathcal{A}} \theta_{s,a}^{(0)}$ and Lemma 7.

Claim (b). If $\sum_{a \in \mathcal{A}} \theta_{s,a}^{(k)} = \sum_{a \in \mathcal{A}} \theta_{s,a}^{(0)}$ for all $k \in \{1, \cdots, M\}$, then $\sum_{a \in \mathcal{A}} \theta_{s,a}^{(M+1)} = \sum_{a \in \mathcal{A}} \theta_{s,a}^{(0)}$.

We use an argument similar to (62) as follows.

$$\theta_{s,a}^{(M+1)} = \omega_{s,a}^{(M)} + \eta^{(M+1)} \cdot \left. \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} \right|_{\theta=\omega^{(M)}}$$

$$
\begin{aligned}
= \theta_{s,a}^{(M)} &+ \mathbb{I}\left\{ V^{\pi_\varphi^{(M)}}(\mu) \geq V^{\pi_\theta^{(M)}}(\mu) \right\} \cdot \frac{M-1}{M+2} \cdot (\theta_{s,a}^{(M)} - \theta_{s,a}^{(M-1)}) \\
&+ \eta^{(M+1)} \cdot \frac{1}{1-\gamma} d^{\pi_\omega^{(M)}}(s) \pi_\omega^{(M)}(a|s) A^{\pi_\omega^{(M)}}(s,a).
\end{aligned}
\tag{63}
$$

By taking the sum of (63) over $a \in \mathcal{A}$, we see that $\sum_{a \in \mathcal{A}} \theta_{s,a}^{(M+1)} = \sum_{a \in \mathcal{A}} \theta_{s,a}^{(0)}$. $\qquad\square$

B.4. $O(1/t^2)$ Convergence Rate Under Nearly Concave Objectives

We present several theoretical results adapted from (Ghadimi & Lan, 2016) to our nearly concave regime in this subsection.

Theorem 4 (**Theorem 1(b) in (Ghadimi & Lan, 2016) with a slight modification**). *Let $\left\{ \theta_{md}^{(t)}, \theta_{ag}^{(t)} \right\}_{t \geq 1}$ be computed by Algorithm 3 for softmax parameterization and Γ_t be defined as:*

$$
\Gamma^{(t)} := \begin{cases} 1, & \text{if } t = 1, \\ (1 - \alpha^{(t)})\Gamma^{(t-1)}, & \text{if } t \geq 2. \end{cases}
\tag{64}
$$

Assume that the objective function $V^{\pi_\theta}(\mu)$ is C-nearly concave at $\theta_{md}^{(t)}$ along the direction $\theta_{ag}^{(t)} - \theta_{md}^{(t)}$ with a constant $C \in (1, \frac{3}{2}]$ for every $t \geq 1$. Suppose $\theta_{ag}^{(t)} - \theta_{md}^{(t)} \in \mathcal{U}$ for every $t \geq 1$. If $\alpha^{(t)}, \beta^{(t)}, \lambda^{(t)}$ are chosen such that

$$
0 < \alpha^{(t)}\lambda^{(t)} \leq \beta^{(t)} < \frac{2-C}{L}, \quad \frac{\alpha^{(1)}}{\lambda^{(1)}\Gamma^{(1)}} \geq \frac{\alpha^{(2)}}{\lambda^{(2)}\Gamma^{(2)}} \geq \cdots,
\tag{65}
$$

where L is the Lipschitz constant of the gradient of the objective. Then given any $t \geq 1$, and for any $\theta^{**} = \theta_{md}^{(t)} + \boldsymbol{d}$ with some $\boldsymbol{d} \in \mathcal{U}$, we have

$$
V^{\pi_{\theta^{**}}}(\mu) - V^{\pi_{\theta_{ag}^{(t)}}}(\mu) \leq \Gamma^{(t)} \frac{\alpha^{(1)} \left\| \theta^{**} - \theta_{ag}^{(0)} + \bar{m}_{init} \right\|^2}{\lambda^{(1)}}
\tag{66}
$$

where $\bar{m}_{init}, \theta_{ag}^{(0)} \in \mathbb{R}^n$ are input variables of Algorithm 3.

Proof of Theorem 4. First, by Lemma 14 and (12), we have

$$
\begin{aligned}
-V^{\pi_{\theta_{ag}^{(t)}}}(\mu) &\leq -V^{\pi_{\theta_{md}^{(t)}}}(\mu) - \left\langle \nabla_\theta V^{\pi_\theta}(\mu) \Big|_{\theta = \theta_{md}^{(t)}}, \theta_{ag}^{(t)} - \theta_{md}^{(t)} \right\rangle + \frac{L}{2} \left\| \theta_{ag}^{(t)} - \theta_{md}^{(t)} \right\|^2 \\
&= -V^{\pi_{\theta_{md}^{(t)}}}(\mu) - \beta^{(t)} \left\| \nabla_\theta V^{\pi_\theta}(\mu) \Big|_{\theta = \theta_{md}^{(t)}} \right\|^2 + \frac{L(\beta^{(t)})^2}{2} \left\| \nabla_\theta V^{\pi_\theta}(\mu) \Big|_{\theta = \theta_{md}^{(t)}} \right\|^2.
\end{aligned}
\tag{67}
$$

Therefore, by the near-concavity of the objective (from our assumption) and (10), we have

$$
\begin{aligned}
&-V^{\pi_{\theta_{md}^{(t)}}}(\mu) + \left[(1-\alpha^{(t)})V^{\pi_{\theta_{ag}^{(t-1)}}}(\mu) + \alpha^{(t)}V^{\pi_{\theta^{**}}}(\mu) \right] \\
&= \alpha^{(t)} \left[V^{\pi_{\theta^{**}}}(\mu) - V^{\pi_{\theta_{md}^{(t)}}}(\mu) \right] + (1-\alpha^{(t)}) \left[V^{\pi_{\theta_{ag}^{(t-1)}}}(\mu) - V^{\pi_{\theta_{md}^{(t)}}}(\mu) \right] \\
&\leq C \cdot \alpha^{(t)} \left\langle \nabla_\theta V^{\pi_\theta}(\mu) \Big|_{\theta = \theta_{md}^{(t)}}, \theta^{**} - \theta_{md}^{(t)} \right\rangle + C \cdot (1-\alpha^{(t)}) \left\langle \nabla_\theta V^{\pi_\theta}(\mu) \Big|_{\theta = \theta_{md}^{(t)}}, \theta_{ag}^{(t-1)} - \theta_{md}^{(t)} \right\rangle \\
&= C \cdot \left\langle \nabla_\theta V^{\pi_\theta}(\mu) \Big|_{\theta = \theta_{md}^{(t)}}, \alpha^{(t)}(\theta^{**} - \theta_{md}^{(t)}) + (1-\alpha^{(t)})(\theta_{ag}^{(t-1)} - \theta_{md}^{(t)}) \right\rangle \\
&= C \cdot \alpha^{(t)} \left\langle \nabla_\theta V^{\pi_\theta}(\mu) \Big|_{\theta = \theta_{md}^{(t)}}, \theta^{**} - \hat{\theta}^{(t-1)} \right\rangle
\end{aligned}
\tag{68}
$$

By (11), we have

$$
\left\| \theta^{**} - \hat{\theta}^{(t-1)} \right\|^2 - 2\lambda^{(t)} \left\langle \nabla_\theta V^{\pi_\theta}(\mu) \Big|_{\theta = \theta_{md}^{(t)}}, \theta^{**} - \hat{\theta}^{(t-1)} \right\rangle + \left(\lambda^{(t)}\right)^2 \left\| \nabla_\theta V^{\pi_\theta}(\mu) \Big|_{\theta = \theta_{md}^{(t)}} \right\|^2
$$

$$= \left\| \theta^{**} - \hat{\theta}^{(t-1)} - \lambda^{(t)} \nabla_\theta V^{\pi_\theta}(\mu) \Big|_{\theta=\theta_{md}^{(t)}} \right\|^2$$

$$= \left\| \theta^{**} - \hat{\theta}^{(t)} \right\|^2 ,$$

which directly leads to

$$\alpha^{(t)} \left\langle \nabla_\theta V^{\pi_\theta}(\mu) \Big|_{\theta=\theta_{md}^{(t)}}, \theta^{**} - \hat{\theta}^{(t-1)} \right\rangle$$

$$= \frac{\alpha^{(t)}}{2\lambda^{(t)}} \left[\left\| \theta^{**} - \hat{\theta}^{(t-1)} \right\|^2 - \left\| \theta^{**} - \hat{\theta}^{(t)} \right\|^2 \right] + \frac{\alpha^{(t)}\lambda^{(t)}}{2} \left\| \nabla_\theta V^{\pi_\theta}(\mu) \Big|_{\theta=\theta_{md}^{(t)}} \right\|^2 . \tag{69}$$

Combining (67)-(69), we have

$$-V^{\pi_{\theta_{ag}}^{(t)}}(\mu) \le - \left[(1-\alpha^{(t)}) V^{\pi_{\theta_{ag}}^{(t-1)}}(\mu) + \alpha^{(t)} V^{\pi_{\theta^{**}}}(\mu) \right]$$

$$+ C \cdot \frac{\alpha^{(t)}}{2\lambda^{(t)}} \left[\left\| \theta^{**} - \hat{\theta}^{(t-1)} \right\|^2 - \left\| \theta^{**} - \hat{\theta}^{(t)} \right\|^2 \right]$$

$$+ C \cdot \frac{\alpha^{(t)}\lambda^{(t)}}{2} \left\| \nabla_\theta V^{\pi_\theta}(\mu) \Big|_{\theta=\theta_{md}^{(t)}} \right\|^2$$

$$- \beta^{(t)} \left\| \nabla_\theta V^{\pi_\theta}(\mu) \Big|_{\theta=\theta_{md}^{(t)}} \right\|^2$$

$$+ \frac{L(\beta^{(t)})^2}{2} \left\| \nabla_\theta V^{\pi_\theta}(\mu) \Big|_{\theta=\theta_{md}^{(t)}} \right\|^2$$

$$\le - \left[(1-\alpha^{(t)}) V^{\pi_{\theta_{ag}}^{(t-1)}}(\mu) + \alpha^{(t)} V^{\pi_{\theta^{**}}}(\mu) \right]$$

$$+ C \cdot \frac{\alpha^{(t)}}{2\lambda^{(t)}} \left[\left\| \theta^{**} - \hat{\theta}^{(t-1)} \right\|^2 - \left\| \theta^{**} - \hat{\theta}^{(t)} \right\|^2 \right]$$

$$- \frac{\beta^{(t)}}{2} \left(2 - C - L\beta^{(t)} \right) \left\| \nabla_\theta V^{\pi_\theta}(\mu) \Big|_{\theta=\theta_{md}^{(t)}} \right\|^2$$

Adding $V^{\pi_{\theta^{**}}}(\mu)$ to both sides of the above inequality and using (64), we have

$$\frac{V^{\pi_{\theta^{**}}}(\mu) - V^{\pi_{\theta_{ag}}^{(t)}}(\mu)}{\Gamma^{(t)}} \le C \cdot \sum_{k=1}^{t} \frac{\alpha^{(k)}}{2\lambda^{(k)}\Gamma^{(k)}} \left[\left\| \theta^{**} - \hat{\theta}^{(k-1)} \right\|^2 - \left\| \theta^{**} - \hat{\theta}^{(k)} \right\|^2 \right]$$

$$- \sum_{k=1}^{t} \frac{\beta^{(k)}}{2\Gamma^{(k)}} \underbrace{\left(2 - C - L\beta^{(k)} \right)}_{>0, \text{ by (65)}} \left\| \nabla_\theta V^{\pi_\theta}(\mu) \Big|_{\theta=\theta_{md}^{(k)}} \right\|^2$$

$$\le \frac{\alpha^{(1)} C \cdot \left\| \theta^{**} - \hat{\theta}^{(0)} \right\|^2}{2\lambda^{(1)}} \tag{70}$$

$$\le \frac{\alpha^{(1)} \left\| \theta^{**} - \hat{\theta}^{(0)} \right\|^2}{\lambda^{(1)}} \tag{71}$$

$$= \frac{\alpha^{(1)} \left\| \theta^{**} - \theta_{ag}^{(0)} + \bar{m}_{\text{init}} \right\|^2}{\lambda^{(1)}}$$

where (70) holds by the fact that

$$\sum_{k=1}^{t} \frac{\alpha^{(k)}}{\lambda^{(k)}\Gamma^{(k)}} \left[\left\| \theta^{**} - \hat{\theta}^{(k-1)} \right\|^2 - \left\| \theta^{**} - \hat{\theta}^{(k)} \right\|^2 \right] \leq \frac{\alpha^{(1)} \left\| \theta^{**} - \hat{\theta}^{(0)} \right\|^2}{\lambda^{(1)}\Gamma^{(1)}} = \frac{\alpha^{(1)} \left\| \theta^{**} - \hat{\theta}^{(0)} \right\|^2}{\lambda^{(1)}},$$

and (71) is from our assumption of $C \leq \frac{3}{2} < 2$. Finally, we obtain the desired result by rearranging (71). $\qquad\square$

Corollary 1 (**Corollary 1 in (Ghadimi & Lan, 2016) with a slight modification**). *Given $T_{shift} \in \mathbb{N}$, suppose that $\left\{ \alpha^{(t)} \right\}$, $\left\{ \beta^{(t)} \right\}$ and $\left\{ \lambda^{(t)} \right\}$ in Algorithm 3 for softmax parameterization are set to*

$$\alpha^{(t)} = \frac{2}{(t+1)+T_{shift}}, \quad \beta^{(t)} = \frac{t+T_{shift}}{(t+1)+T_{shift}} \cdot \frac{1}{2L}, \quad \lambda^{(t)} = \frac{(t+1)+T_{shift}}{2} \cdot \beta^{(t)}, \quad where \ T_{shift} > 0, \qquad (72)$$

where L is the Lipschitz constant of the gradient of the objective. Assume that $V^{\pi_\theta}(\mu)$ is $\frac{3}{2}$-nearly concave at $\theta_{md}^{(t)}$ along the direction $\theta_{ag}^{(t)} - \theta_{md}^{(t)}$ for every $t \geq 1$. Suppose $\theta_{ag}^{(t)} - \theta_{md}^{(t)} \in \mathcal{U}$ for every $t \geq 1$. Then given any $t \geq 1$, and for any $\theta^{**} = \theta_{md}^{(t)} + \mathbf{d}$ with some $\mathbf{d} \in \mathcal{U}$, we have

$$V^{\pi_{\theta^{**}}}(\mu) - V^{\pi_\theta^{(t)}}(\mu) \leq \frac{8L \left\| \theta^{**} - \theta_{ag}^{(0)} + \bar{m}_{init} \right\|^2}{(t+T_{shift}+1)(t+T_{shift})} = O\left(\frac{1}{t^2} \right),$$

where $\bar{m}_{init}, \theta_{ag}^{(0)} \in \mathbb{R}^n$ are input variables of Algorithm 3.

Remark 14. In Theorem 4 and Corollary 1, addition to extending from concavity to near concavity, we have also made the following modifications: (i) We have introduced a iteration counter T_{shift}, since our objective is not nearly-concave initially and the theoretical result needs to be revised to account for the shifted initial step size when the nearly-concave region is entered. (ii) We have adjusted λ from $\frac{t}{2}$ to $\frac{t+1}{2}$ and β from $\frac{1}{2L}$ to $\frac{t+T_{\text{shift}}}{(t+1)+T_{\text{shift}}} \cdot \frac{1}{2L}$ to ensure applicability of both Lemma 3 and Theorem 4.

Remark 15. Theorem 4 and Corollary 1 are built upon *local near concavity* of the objective function. In Appendix D, we will show that such local near concavity indeed holds under APG in the MDP setting.

Proof of Corollary 1. We leverage Theorem 4 to reach our desired result. It suffices to show that the choice of $\left\{ \alpha^{(t)}, \lambda^{(t)}, \beta^{(t)} \right\}$ in (72) satisfy (65). Because $\alpha^{(t)} \cdot \lambda^{(t)} = \beta^{(t)} < \frac{1}{2L}$, the first part of (65) holds. By the definition of $\Gamma^{(t)}$ in (64), we have:

$$\Gamma^{(t)} = \frac{(2+T_{\text{shift}})(1+T_{\text{shift}})}{((t+1)+T_{\text{shift}})(t+T_{\text{shift}})}. \qquad (73)$$

Therefore,

$$\frac{\alpha^{(t)}}{\lambda^{(t)}\Gamma^{(t)}} = \frac{\frac{2}{(t+1)+T_{\text{shift}}}}{\frac{(t+1)+T_{\text{shift}}}{2} \cdot \frac{t+T_{\text{shift}}}{(t+1)+T_{\text{shift}}} \cdot \frac{1}{2L} \cdot \frac{(2+T_{\text{shift}})(1+T_{\text{shift}})}{((t+1)+T_{\text{shift}})(t+T_{\text{shift}})}} = \frac{8L}{(2+T_{\text{shift}})(1+T_{\text{shift}})},$$

which satisfies the second condition in (65). We hence reach the desired result by plugging (73) and the choice of $\lambda^{(1)}$ in (72) into (66). $\qquad\square$

C. Asymptotic Convergence

C.1. Supporting Lemmas for Asymptotic Convergence of APG

Lemma 25. *Under APG for softmax parameterzation with uniformly randomly initialized surrogate initial state distribution μ, and suppose there are $\eta' \geq \eta > 0$ such that $\eta^{(t)} \in [\eta, \eta']$ for all t. Then, almost surely, the limits of both $V^{\pi_\omega^{(t)}}(\mu)$ and $V^{\pi_\theta^{(t)}}(\mu)$ when t approaches infinity exist, $\lim_{t \to \infty} V^{\pi_\omega^{(t)}}(\mu) = \lim_{t \to \infty} V^{\pi_\theta^{(t)}}(\mu)$, and*

$$\lim_{t \to \infty} \left\| \nabla_\theta V^{\pi_\theta}(s)|_{\theta=\omega^{(t)}} \right\| = 0, \quad \lim_{t \to \infty} \left\| \nabla_\theta V^{\pi_\theta}(s)|_{\theta=\theta^{(t)}} \right\| = 0. \tag{74}$$

Proof of Lemma 25. For any $t \in \mathbb{N}$, we know that $V^{\pi_\theta^{(t+1)}}(s) \geq V^{\pi_\omega^{(t)}}(s)$ for all $s \in \mathcal{S}$ by Lemma 22. Then, since $V^\pi(\mu) = \sum_s \mu(s) V^\pi(s)$, we have $V^{\pi_\theta^{(t+1)}}(\mu) \geq V^{\pi_\omega^{(t)}}(\mu)$. This together with (8) implies that for any $t \in \mathbb{N}$, we have

$$V^{\pi_\omega^{(t+1)}}(\mu) \geq V^{\pi_\theta^{(t+1)}}(\mu) \geq V^{\pi_\omega^{(t)}}(\mu) \geq V^{\pi_\theta^{(t)}}(\mu). \tag{75}$$

By (75) and that $V^{\pi_\theta}(\mu) \leq \frac{1}{1-\gamma}$ for all θ from Lemma 10, we know from the monotone convergence theorem that the limits of both $V^{\pi_\omega^{(t)}}(\mu)$ and $V^{\pi_\theta^{(t)}}(\mu)$ exist and $\lim_{t \to \infty} V^{\pi_\omega^{(t)}}(\mu) = \lim_{t \to \infty} V^{\pi_\theta^{(t)}}(\mu)$. Therefore, by the update rule (6) of APG, the Lipschitz continuity of $\nabla_\theta V^{\pi_\theta}(\mu)$ from Lemma 14, and our assumption that $\eta^{(t)}$ is bounded away from zero, we further see

$$\lim_{t \to \infty} \left\| \nabla_\theta V^{\pi_\theta}(\mu)|_{\theta=\omega^{(t)}} \right\| = 0. \tag{76}$$

By (76), application of (6) again, the Lipschitz continuity of $\nabla_\theta V^{\pi_\theta}(\mu)$, and the boundedness of $\{\eta^{(t)}\}$, we also have

$$\lim_{t \to \infty} \left\| \nabla_\theta V^{\pi_\theta}(\mu)|_{\theta=\theta^{(t)}} \right\| = 0. \tag{77}$$

By the uniformly random initialization of μ, we know that $\min_{s \in \mathcal{S}} \mu(s) > 0$ almost surely. Thus, the expression of softmax policy gradient in Lemma 11, (76), (77), and Lemma 5 imply that with probability one, for all (s, a),

$$\pi_\omega^{(t)}(a|s) A^{\pi_\omega^{(t)}}(s, a) \to 0, \quad \text{as } t \to \infty, \tag{78}$$

$$\pi_\theta^{(t)}(a|s) A^{\pi_\theta^{(t)}}(s, a) \to 0, \quad \text{as } t \to \infty, \tag{79}$$

proving (74). $\qquad \square$

In the proofs of the subsequent lemmas (from Lemma 26 to Lemma 35), the upper bound for the step size is needed only for leveraging the result in Lemma 25. Other than that, these proofs only require the step sizes $\{\eta^{(t)}\}$ to be lower-bounded away from zero, but do not require an upper bound as in Lemma 25. We will demonstrate in Appendix E that (74) still holds under a time-varying step size setting, and therefore relaxing the requirement for an upper bound for the step sizes under such settings.

Lemma 26. *Under the setting of Lemma 25, the following statements hold almost surely:*

- *The limits $\lim_{t \to \infty} V^{\pi_\omega^{(t)}}(s)$, $\lim_{t \to \infty} Q^{\pi_\omega^{(t)}}(s, a)$, and $\lim_{t \to \infty} A^{\pi_\omega^{(t)}}(s, a)$ all exist for all $s \in \mathcal{S}$ and all $a \in \mathcal{A}$.*

- *The limits $\lim_{t \to \infty} V^{\pi_\theta^{(t)}}(s)$, $\lim_{t \to \infty} Q^{\pi_\theta^{(t)}}(s, a)$, and $\lim_{t \to \infty} A^{\pi_\theta^{(t)}}(s, a)$ all exist for all $s \in \mathcal{S}$ and all $a \in \mathcal{A}$.*

Furthermore, we have

$$\lim_{t \to \infty} V^{\pi_\omega^{(t)}}(s) = \lim_{t \to \infty} V^{\pi_\theta^{(t)}}(s); \quad \lim_{t \to \infty} Q^{\pi_\omega^{(t)}}(s, a) = \lim_{t \to \infty} Q^{\pi_\theta^{(t)}}(s, a); \quad \lim_{t \to \infty} A^{\pi_\omega^{(t)}}(s, a) = \lim_{t \to \infty} A^{\pi_\theta^{(t)}}(s, a)$$

for all $s \in \mathcal{S}$ and all $a \in \mathcal{A}$ when these limits exist when these limits exist.

Proof of Lemma 26. Recall that we use $\boldsymbol{V}^\pi$ to denote the $|\mathcal{S}|$-dimensional vector whose s-th component is $V^\pi(s)$ for all $s \in \mathcal{S}$. Our first goal is to show that the limit of the value vector $\boldsymbol{V}^{\pi_\omega^{(t)}}$ exists, *i.e.*, $\lim_{t \to \infty} V^{\pi_\omega^{(t)}}(s)$ exist for all $s \in \mathcal{S}$. We define

$$\mathcal{V} := \left\{ \boldsymbol{V}^\pi : \left| \pi(a|s) A^\pi(s, a) \right| = 0, \forall (s, a) \in \mathcal{S} \times \mathcal{A} \right\}. \tag{80}$$

Before starting our proof, we make the following claim.

Claim 1. *The cardinality of $\mathcal{V}$ is at most $|\mathcal{A}|^{|\mathcal{S}|}$.*

To prove Claim 1, we first observe that there are $|\mathcal{A}|^{|\mathcal{S}|}$ deterministic policies. In addition, by Lemma 7, we know that the value vector of any deterministic policy belongs to $\mathcal{V}$. For any stochastic policy π such that $\boldsymbol{V}^\pi \in \mathcal{V}$, by (80), it follows that for any (s, a) satisfying the condition $\pi(a|s) > 0$, we have $A^\pi(s, a) = 0$. Consequently, selecting the deterministic policy π' associated with π, wherein for all states $s \in \mathcal{S}$ we set $\pi'(a|s) = 1$ for some action a satisfying $\pi(a|s) > 0$, allows us to confirm that $\boldsymbol{V}^\pi = \boldsymbol{V}^{\pi'}$ using Lemma 4. Hence, for every value vector $\boldsymbol{V}^\pi$ in $\mathcal{V}$, there exists at least one corresponding deterministic policy. Since there are only $|\mathcal{A}|^{|\mathcal{S}|}$ deterministic policies, the number of distinct value vectors in $\mathcal{V}$ is at most $|\mathcal{A}|^{|\mathcal{S}|}$, establishing the claim.

With Claim 1 proven, we are now ready to prove Lemma 26. For any point x and any set Y, we define

$$d(x, Y) := \inf_{y \in Y} \|y - x\|_1$$

as the distance between x and Y measured by the L_1-norm. We first show that

$$\lim_{t \to \infty} d(\boldsymbol{V}^{\pi_\omega^{(t)}}, \mathcal{V}) = 0. \tag{81}$$

Given any $\varepsilon > 0$, let $\varepsilon' = \frac{(1-\gamma)\varepsilon}{|\mathcal{S}|^2}$. By (78), we know that there exists $T \geq 0$ such that

$$|\pi_\omega^{(t)}(a|s) A^{\pi_\omega^{(t)}}(s, a)| < \frac{\varepsilon'}{|\mathcal{A}|}, \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}, \quad \forall t \geq T. \tag{82}$$

Since $\sum_{a \in \mathcal{A}} \pi(a|s) = 1$ for any $s \in \mathcal{S}$ and any policy π, there must be at least one action at each state satisfying $\pi_\omega^{(t)}(a|s) \geq 1/|\mathcal{A}|$. We can thus observe from (82) that for any $t \geq T$ and any $s \in \mathcal{S}$, there must be an action $a_s^{(t)}$ such that $|A^{\pi_\omega^{(t)}}(s, a_s^{(t)})| < \varepsilon'$. Therefore, for every $t \geq T$, we pick $\boldsymbol{V}^{(t)} \in \mathcal{V}$ as the value vector corresponding to the deterministic policy $\pi^{(t)}$ with $\pi^{(t)}(a_s^{(t)}|s) = 1$ and $\pi^{(t)}(a|s) = 0$ for all $a \neq a_s^{(t)}$ for all $s \in \mathcal{S}$. (Note that the value vector of any deterministic policy belongs to $\mathcal{V}$.) Then, we can see that for every $t \geq T$,

$$
\begin{aligned}
d(\boldsymbol{V}^{\pi_\omega^{(t)}}, \mathcal{V}) &\leq \|\boldsymbol{V}^{(t)} - \boldsymbol{V}^{\pi_\omega^{(t)}}\|_1 \\
&= \sum_s \left| \frac{1}{1-\gamma} \sum_{s'} d_s^{\pi^{(t)}}(s') \sum_a \pi^{(t)}(a|s') A^{\pi_\omega^{(t)}}(s', a) \right| && (83) \\
&= \sum_s \left| \frac{1}{1-\gamma} \sum_{s'} \underbrace{d_s^{\pi^{(t)}}(s')}_{\in [0,1]} A^{\pi_\omega^{(t)}}(s', a_{s'}^{(t)}) \right| && (84) \\
&\leq \sum_s \frac{1}{1-\gamma} \sum_{s'} \left| A^{\pi_\omega^{(t)}}(s', a_{s'}^{(t)}) \right| \\
&= \frac{|\mathcal{S}|}{1-\gamma} \sum_{s'} \left| A^{\pi_\omega^{(t)}}(s', a_{s'}^{(t)}) \right| \\
&< \frac{|\mathcal{S}|^2}{1-\gamma} \varepsilon' = \varepsilon,
\end{aligned}
$$

where (83) holds by Lemma 4 and (84) is due to the fact that $\pi^{(t)}(a_s^{(t)}|s) = 1$. Since ε is arbitrary, we have proven (81).

Accordingly, since $\mathcal{V}$ is a finite and bounded (and thus compact) set according to Claim 1, there are two possible scenarios: (i) There exists $\boldsymbol{V} \in \mathcal{V}$ such that $\lim_{t \to \infty} \boldsymbol{V}^{\pi_\omega^{(t)}} = \boldsymbol{V}$; (ii) $\{\boldsymbol{V}^{\pi_\omega^{(t)}}\}_{t=1}^{\infty}$ has at least two limit points in $\mathcal{V}$.

For (i), we obtain the desired results. Regarding (ii), we prove that with probability one, this scenario will not occur. For any two distinct value vectors $\boldsymbol{V}^{\pi_1} \neq \boldsymbol{V}^{\pi_2}$ in $\mathcal{V}$, we know that the set of μ such that $V^{\pi_1}(\mu) = V^{\pi_2}(\mu)$ is a subset in $\Delta(\mathcal{S})$ with dimension strictly less than the dimension of $\Delta(\mathcal{S})$ due to the given hypothesis $\boldsymbol{V}^{\pi_1} \neq \boldsymbol{V}^{\pi_2}$. Thus, the set of μ such that $V^{\pi_1}(\mu) = V^{\pi_2}(\mu)$ is of measure zero in $\Delta(\mathcal{S})$. Additionally, since there are only finitely many distinct value vectors in $\mathcal{V}$ by Claim 1, the union of those sets of μ is still of measure zero in $\Delta(\mathcal{S})$. Hence, by the uniformly random initialization of

the surrogate initial state distribution μ, we have that with probability one, any two distinct value vectors $\boldsymbol{V}^{\pi_1} \neq \boldsymbol{V}^{\pi_2}$ in $\mathcal{V}$ satisfy $V^{\pi_1}(\mu) \neq V^{\pi_2}(\mu)$.

Let $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ be two distinct limit points in in scenario (ii), and $\{i_t\}$ and $\{j_t\}$ be two infinite index sequences such that

$$\lim_{t\to\infty} \boldsymbol{V}^{\pi_\omega^{(i_t)}} = \boldsymbol{V}_1, \quad \lim_{t\to\infty} \boldsymbol{V}^{\pi_\omega^{(j_t)}} = \boldsymbol{V}_2,$$

which implies

$$\lim_{t\to\infty} V^{\pi_\omega^{(i_t)}}(\mu) = V_1(\mu), \quad \lim_{t\to\infty} V^{\pi_\omega^{(j_t)}}(\mu) = V_2(\mu). \tag{85}$$

We have argued that with probability one, $V_1(\mu) \neq V_2(\mu)$, and let us assume without loss of generality that $V_1(\mu) > V_2(\mu)$ and denote $\epsilon := V_1(\mu) - V_2(\mu)$. From (85), almost surely there is $T_1 \geq 0$ such that

$$V^{\pi_\omega^{(i_t)}}(\mu) > V_1(\mu) - \frac{\epsilon}{2}, \quad \forall t \geq T_1. \tag{86}$$

By (75), we see that $V^{\pi_\omega^{(t)}}(\mu)$ is monotonically increasing, which together with (86) implies that

$$V^{\pi_\omega^{(t)}}(\mu) \geq V_2(\mu) + \frac{\epsilon}{2}, \quad \forall t \geq i_{T_1},$$

contradicting the second equality of (85) and that $\{j_t\}$ should be an infinite sequence. Therefore, $\{\boldsymbol{V}^{\pi_\omega^{(t)}}\}$ converges to a point. Since $\lim_{t\to\infty} V^{\pi_\omega^{(t)}}(s)$ exists for all $s \in \mathcal{S}$, we see that $\lim_{t\to\infty} Q^{\pi_\omega^{(t)}}(s,a)$ and therefore $\lim_{t\to\infty} A^{\pi_\omega^{(t)}}(s,a)$ also exist for all $s \in \mathcal{S}$ and all $a \in \mathcal{A}$ according to (1). Existence of the limit of $\{\boldsymbol{V}^{\pi_\theta^{(t)}}\}$, $\{Q^{\pi_\theta^{(t)}}(s,a)\}$, and $\{A^{\pi_\theta^{(t)}}(s,a)\}$ follows the same argument above with (78) replaced by (79).

Finally, we demonstrate that the limits of the value function, the Q-value function, and the advantage function with respect to $\omega^{(t)}$ and $\theta^{(t)}$ are equal. It suffices to show that $\lim_{t\to\infty} V^{\pi_\omega^{(t)}}(s) = \lim_{t\to\infty} V^{\pi_\theta^{(t)}}(s)$ for all $s \in \mathcal{S}$, *i.e.*, $\lim_{t\to\infty} \boldsymbol{V}^{\pi_\omega^{(t)}} = \lim_{t\to\infty} \boldsymbol{V}^{\pi_\theta^{(t)}}$. We will prove it by contradiction. Let

$$\boldsymbol{V}^{\pi_\omega^{(\infty)}} := \lim_{t\to\infty} \boldsymbol{V}^{\pi_\omega^{(t)}}, \quad \boldsymbol{V}^{\pi_\theta^{(\infty)}} := \lim_{t\to\infty} \boldsymbol{V}^{\pi_\theta^{(t)}}$$

and suppose that $\boldsymbol{V}^{\pi_\omega^{(\infty)}} \neq \boldsymbol{V}^{\pi_\theta^{(\infty)}}$. Recall that from our uniformly random initialization of μ, if $\boldsymbol{V}^{\pi_1} \neq \boldsymbol{V}^{\pi_2}$ in $\mathcal{V}$, then $V^{\pi_1}(\mu) \neq V^{\pi_2}(\mu)$ almost surely. We know that both $\boldsymbol{V}^{\pi_\omega^{(\infty)}}$ and $\boldsymbol{V}^{\pi_\theta^{(\infty)}}$ belong to $\mathcal{V}$. By the hypothesis, with probability one we have $V^{\pi_\omega^{(\infty)}}(\mu) \neq V^{\pi_\theta^{(\infty)}}(\mu)$, leading to the contradiction due to the fact that $\lim_{t\to\infty} V^{\pi_\omega^{(t)}}(\mu) = \lim_{t\to\infty} V^{\pi_\theta^{(t)}}(\mu)$ from (75). Hence, we obtain that with probability one, $\lim_{t\to\infty} V^{\pi_\omega^{(t)}}(s) = \lim_{t\to\infty} V^{\pi_\theta^{(t)}}(s)$ for all $s \in \mathcal{S}$, and by (2) we also obtain the desired results for both the Q-value function and the advantage function. $\square$

In the sequel, we use the following notations.

$$\begin{cases} A^{(\infty)}(s,a) & := \lim_{t\to\infty} A^{\pi_\omega^{(t)}}(s,a) = \lim_{t\to\infty} A^{\pi_\theta^{(t)}}(s,a), \\ Q^{(\infty)}(s,a) & := \lim_{t\to\infty} Q^{\pi_\omega^{(t)}}(s,a) = \lim_{t\to\infty} Q^{\pi_\theta^{(t)}}(s,a), \\ V^{(\infty)}(s) & := \lim_{t\to\infty} V^{\pi_\omega^{(t)}}(s) = \lim_{t\to\infty} V^{\pi_\theta^{(t)}}(s). \end{cases} \tag{87}$$

We divide the action space into the following subsets based on the advantage function:

$$\begin{aligned} I_s^+ &:= \{a \in \mathcal{A} : A^{(\infty)}(s,a) > 0\}, \\ I_s^- &:= \{a \in \mathcal{A} : A^{(\infty)}(s,a) < 0\}, \\ I_s^0 &:= \{a \in \mathcal{A} : A^{(\infty)}(s,a) = 0\}. \end{aligned}$$

The above action sets are well-defined as the limiting value functions exist by Lemma 26.

For each state s, we define

$$\Delta_s := \min_{a \in I_s^+ \cup I_s^-} |A^{(\infty)}(s,a)|.$$

Accordingly, we know that for each state $s \in \mathcal{S}$, there must exist some $\bar{T}_s$ such that the following hold:

- (i) For all $a \in I_s^+$,

$$A^{\pi_\omega^{(t)}}(s,a) \geq +\frac{\Delta_s}{4}, \quad \text{for all } t \geq \bar{T}_s, \tag{88}$$

- (ii) For all $a \in I_s^-$,

$$A^{\pi_\omega^{(t)}}(s,a) \leq -\frac{\Delta_s}{4}, \quad \text{for all } t \geq \bar{T}_s. \tag{89}$$

- (iii) For all $a \in I_s^0$,

$$|A^{\pi_\omega^{(t)}}(s,a)| \leq \frac{\Delta_s}{4}, \quad \text{for all } t \geq \bar{T}_s. \tag{90}$$

Lemma 27. *Under the setting of Lemma 25, for any state $s \in \mathcal{S}$, we have*

$$\lim_{t \to \infty} \sum_{a \in I_s^+ \cup I_s^-} \pi_\omega^{(t)}(a|s) = 0, \quad \lim_{t \to \infty} \sum_{a \in I_s^+ \cup I_s^-} \pi_\theta^{(t)}(a|s) = 0.$$

As a result, we also have

$$\lim_{t \to \infty} \sum_{a \in I_s^0} \pi_\omega^{(t)}(a|s) = 1, \quad \lim_{t \to \infty} \sum_{a \in I_s^0} \pi_\theta^{(t)}(a|s) = 1.$$

Proof of Lemma 27. For any $a \in I_s^+$, we have $\lim_{t \to \infty} A^{\pi_\omega^{(t)}}(s,a) = \lim_{t \to \infty} A^{\pi_\theta^{(t)}}(s,a) > 0$. By (78) and (79), this implies that $\pi_\omega^{(t)}(a|s) \to 0$ and $\pi_\theta^{(t)}(a|s) \to 0$ as $t \to \infty$ for all $a \in I_s^+$. Similarly, for any $a \in I_s^-$, we also have $\lim_{t \to \infty} A^{\pi_\omega^{(t)}}(s,a) = \lim_{t \to \infty} A^{\pi_\theta^{(t)}}(s,a) < 0$. Again, by (78) and (79), this implies that $\pi_\omega^{(t)}(a|s) \to 0$ and $\pi_\theta^{(t)}(a|s) \to 0$ as $t \to \infty$ for all $a \in I_s^-$. Hence, we conclude that $\sum_{a \in I_s^+ \cup I_s^-} \pi_\omega^{(t)}(a|s) \to 0$ and $\sum_{a \in I_s^+ \cup I_s^-} \pi_\theta^{(t)}(a|s) \to 0$ as $t \to \infty$ due to the finiteness of the cardinality of $\mathcal{A}$. $\qquad\square$

Lemma 28. *Consider any state $s \in \mathcal{S}$. Let a be an action in I_s^+. Under the setting of Lemma 25, $\{\theta_{s,a}^{(t)}\}_{t=1}^\infty$ and $\{\omega_{s,a}^{(t)}\}_{t=1}^\infty$ are bounded from below.*

Proof of Lemma 28. Recall the part of our definition in (88) of $\bar{T}_s$. From Lemma 11, we see that

$$\eta^{(t)} \nabla_{\theta_{s,a}} V^{\pi_\theta}(\mu)\Big|_{\theta=\omega^{(t-1)}} \geq 0, \quad \forall t \geq \bar{T}_s + 1. \tag{91}$$

We let $\delta_{\bar{T}_s} := \theta_{s,a}^{(\bar{T}_s)} - \theta_{s,a}^{(\bar{T}_s-1)}$. Regarding the case $\delta_{\bar{T}_s} \geq 0$, by (91) and the update rules in (6) and (7), the sequences $\{\theta_{s,a}^{(t)}\}_{t \geq \bar{T}_s+1}, \{\omega_{s,a}^{(t)}\}_{t \geq \bar{T}_s+1}$ are monotonically increasing and therefore lower-bounded.

On the other hand, if $\delta_{\bar{T}_s} < 0$, for any $M \in \mathbb{N}$, we have

$$\begin{aligned}
\theta_{s,a}^{(\bar{T}_s+M)} &= \omega_{s,a}^{(\bar{T}_s+M-1)} + \eta^{(\bar{T}_s+M)} \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\Big|_{\theta=\omega^{(\bar{T}_s+M-1)}} \\
&\geq \theta_{s,a}^{(\bar{T}_s+M-1)} + \mathbb{I}\left\{V^{\pi_\varphi^{(\bar{T}_s+M-1)}}(\mu) \geq V^{\pi_\theta^{(\bar{T}_s+M-1)}}(\mu)\right\} \cdot \frac{\bar{T}_s+M-2}{\bar{T}_s+M+1}(\theta_{s,a}^{(\bar{T}_s+M-1)} - \theta_{s,a}^{(\bar{T}_s+M-2)}) \\
&\geq \theta_{s,a}^{(\bar{T}_s)} + \frac{\bar{T}_s-1}{\bar{T}_s+2}\delta_{\bar{T}_s} + \frac{\bar{T}_s(\bar{T}_s-1)}{(\bar{T}_s+3)(\bar{T}_s+2)}\delta_{\bar{T}_s} + \cdots + \frac{(\bar{T}_s+M-2)\cdots(\bar{T}_s-1)}{(\bar{T}_s+M+1)\cdots(\bar{T}_s+2)}\delta_{\bar{T}_s} \\
&= \theta_{s,a}^{(\bar{T}_s)} + \left[\frac{\bar{T}_s-1}{\bar{T}_s+2} + \frac{\bar{T}_s(\bar{T}_s-1)}{(\bar{T}_s+3)(\bar{T}_s+2)} + \sum_{\tau=2}^{M} \frac{(\bar{T}_s+1)\bar{T}_s(\bar{T}_s-1)}{(\bar{T}_s+\tau+2)(\bar{T}_s+\tau+1)(\bar{T}_s+\tau)}\right]\delta_{\bar{T}_s},
\end{aligned} \tag{92}$$

where (92) holds by disregarding the updates contributed by the gradients taken at all $t > \bar{T}_s$, which are always nonnegative according to (91).

We note that for any $M \in \mathbb{N}$,

$$\sum_{\tau=2}^{M} \frac{(\bar{T}_s + 1)\bar{T}_s(\bar{T}_s - 1)}{(\bar{T}_s + \tau + 2)(\bar{T}_s + \tau + 1)(\bar{T}_s + \tau)}$$

$$= (\bar{T}_s + 1)\bar{T}_s(\bar{T}_s - 1) \sum_{\tau=2}^{M} \frac{1}{2}\left(\frac{1}{(\bar{T}_s + \tau)(\bar{T}_s + \tau + 1)} - \frac{1}{(\bar{T}_s + \tau + 1)(\bar{T}_s + \tau + 2)}\right)$$

$$= (\bar{T}_s + 1)\bar{T}_s(\bar{T}_s - 1) \cdot \frac{1}{2}\left(\frac{1}{(\bar{T}_s + 2)(\bar{T}_s + 3)} - \frac{1}{(\bar{T}_s + M + 1)(\bar{T}_s + M + 2)}\right)$$

$$\leq \frac{\bar{T}_s}{2}. \tag{93}$$

Therefore, it follows that for any $M \in \mathbb{N}$,

$$\theta_{s,a}^{(\bar{T}_s + M)} \geq \theta_{s,a}^{(\bar{T}_s)} - \left(2 + \frac{\bar{T}_s}{2}\right)|\delta_{\bar{T}_s}|.$$

Hence, $\theta_{s,a}^{(t)} \geq \theta_{s,a}^{(\bar{T}_s)} - (2 + \frac{\bar{T}_s}{2})|\delta_{\bar{T}_s}|$ for all $t \geq \bar{T}_s$. By (91), if there is some $\bar{T}_s' \geq \bar{T}_s$ such that $\theta_{s,a}^{(t)} - \theta_{s,a}^{(t-1)} \geq 0$ for any $a \in I_s^+$, we obtain that $\{\omega_{s,a}^{(t)}\}_{t=1}^{\infty}$ is also bounded from below since $a \in I_s^+$. Otherwise, if there is no such $\bar{T}_s'$, then $\{\theta\}_{t=1}^{\infty}$ is monotonically decreasing after time $\bar{T}_s'$. Thus, we can bound $\{\omega_{s,a}^{(t)}\}_{t=1}^{\infty}$ from below due to the momentum update (7), where $\omega_{s,a}^{(t)} \geq 2(\theta_{s,a}^{(\bar{T}_s)} - (2 + \frac{\bar{T}_s}{2})|\delta_{\bar{T}_s}|) - \max_{t \leq \bar{T}_s} \theta_{s,a}^{(t)}$ for all $t \geq \bar{T}_s$. Hence, we obtain that $\{\omega_{s,a}^{(t)}\}_{t=1}^{\infty}$ is alsobounded from below for all $s \in \mathcal{S}$ and $a \in I_s^+$. $\qquad\square$

Lemma 29. *Consider any state $s \in \mathcal{S}$. Let a be an action in I_s^-. Under the setting of Lemma 25, $\{\theta_{s,a}^{(t)}\}_{t=1}^{\infty}$ and $\{\omega_{s,a}^{(t)}\}_{t=1}^{\infty}$ are bounded from above.*

Proof of Lemma 29. We follow the same procedure as that in Lemma 28. Again, let $\delta_{\bar{T}_s} := \theta_{s,a}^{(\bar{T}_s)} - \theta_{s,a}^{(\bar{T}_s - 1)}$. From (88) and Lemma 11, we see that

$$\eta^{(t)} \nabla_{\theta_{s,a}} V^{\pi_\theta}(\mu)\Big|_{\theta = \omega^{(t-1)}} \leq 0, \quad \forall t \geq \bar{T}_s + 1. \tag{94}$$

Therefore, by (6), (7) and (94), we have

$$\theta_{s,a}^{(\bar{T}_s + 1)} = \omega_{s,a}^{(\bar{T}_s)} + \eta^{(\bar{T}_s + 1)} \cdot \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\Big|_{\theta = \omega^{(\bar{T}_s)}} \leq \omega_{s,a}^{(\bar{T}_s)}.$$

Regarding the case of $\delta_{\bar{T}_s} \leq 0$, the result directly holds by (94) as argued in the proof of Lemma 28. Considering $\delta_{\bar{T}_s} > 0$, for any $M \in \mathbb{N}$, we have

$$\theta_{s,a}^{(\bar{T}_s + M)} = \omega_{s,a}^{(\bar{T}_s + M - 1)} + \eta^{(\bar{T}_s + M)} \cdot \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\Big|_{\theta = \omega^{(\bar{T}_s + M - 1)}}$$

$$\leq \theta_{s,a}^{(\bar{T}_s + M - 1)} + \mathbb{I}\left\{V^{\pi_\varphi^{(T_0 + M - 1)}}(\mu) \geq V^{\pi_\theta^{(T_0 + M - 1)}}(\mu)\right\} \cdot \frac{\bar{T}_s + M - 2}{\bar{T}_s + M + 1}(\theta_{s,a}^{(\bar{T}_s + M - 1)} - \theta_{s,a}^{(\bar{T}_s + M - 2)})$$

$$\leq \theta_{s,a}^{(\bar{T}_s)} + \frac{\bar{T}_s - 1}{\bar{T}_s + 2}\delta_{\bar{T}_s} + \frac{\bar{T}_s(\bar{T}_s - 1)}{(\bar{T}_s + 3)(\bar{T}_s + 2)}\delta_{\bar{T}_s} + \cdots + \frac{(\bar{T}_s + M - 2)\cdots(\bar{T}_s - 1)}{(\bar{T}_s + M + 1)\cdots(\bar{T}_s + 2)}\delta_{\bar{T}_s}$$

$$= \theta_{s,a}^{(\bar{T}_s)} + \left[\frac{\bar{T}_s - 1}{\bar{T}_s + 2} + \frac{\bar{T}_s(\bar{T}_s - 1)}{(\bar{T}_s + 3)(\bar{T}_s + 2)} + \sum_{\tau=2}^{M} \frac{(\bar{T}_s + 1)\bar{T}_s(\bar{T}_s - 1)}{(\bar{T}_s + \tau + 2)(\bar{T}_s + \tau + 1)(\bar{T}_s + \tau)}\right]\delta_{\bar{T}_s}.$$

By (93), we know $\sum_{\tau=2}^{M} \frac{(\bar{T}_s + 1)\bar{T}_s(\bar{T}_s - 1)}{(\bar{T}_s + \tau + 2)(\bar{T}_s + \tau + 1)(\bar{T}_s + \tau)} \leq \frac{\bar{T}_s}{2}$. As a result, for any $M \in \mathbb{N}$,

$$\theta_{s,a}^{(\bar{T}_s + M)} \leq \theta_{s,a}^{(\bar{T}_s)} + (2 + \frac{\bar{T}_s}{2})|\delta_{\bar{T}_s}|.$$

Hence, $\theta_{s,a}^{(t)} \leq \theta_{s,a}^{(\bar{T}_s)} + (2 + \frac{\bar{T}_s}{2})|\delta_{\bar{T}_s}|$ for all $t \geq \bar{T}_s$. By using the same argument as in Lemma 28, we have $\omega_{s,a}^{(t)} \leq 2(\theta_{s,a}^{(\bar{T}_s)} + (2 + \frac{\bar{T}_s}{2})|\delta_{\bar{T}_s}|) - \min_{t \leq \bar{T}_s}$ for all $s \in I_s^-$. Hence, we obtain that $\{\omega_{s,a}^{(t)}\}_{t=1}^{\infty}$ is bounded from above for all $s \in \mathcal{S}$ and $a \in I_s^+$. $\qquad\square$

Lemma 30. *Consider any state $s \in \mathcal{S}$. Under the setting of Lemma 25, if I_s^+ is non-empty, then*

$$\lim_{t \to \infty} \max_{a \in I_s^0} \omega_{s,a}^{(t)} = \infty, \quad \lim_{t \to \infty} \max_{a \in I_s^0} \theta_{s,a}^{(t)} = \infty.$$

Proof. By Lemma 27, we know $\sum_{a \in I_s^0} \pi_\omega^{(t)}(a|s) \to 1$ and $\sum_{a \in I_s^0} \pi_\theta^{(t)}(a|s) \to 1$ as $t \to \infty$. Moreover, by Lemma 28, we know $\{\omega_{s,a}^{(t)}\}$ and $\{\theta_{s,a}^{(t)}\}$ are lower-bounded for all $a \in I_s^+$. Therefore, under the softmax policy parameterization, we obtain the desired results. $\qquad\square$

Recall from (89) that for all $a \in I_s^-$, we have $A^{\pi_\omega^{(t)}}(s,a) \le -\frac{\Delta_s}{4}$ for all $t \ge \bar{T}_s$.

Lemma 31. *Consider any state $s \in \mathcal{S}$. Under the setting of Lemma 25, if I_s^+ is non-empty, then for any $a \in I_s^-$, we have $\omega_{s,a}^{(t)} \to -\infty$ as $t \to \infty$.*

Proof of Lemma 31. We prove this by contradiction. Motivated by the proof of Lemma C11 in (Agarwal et al., 2021), our proof here extends their argument to the case with momentum by considering the cumulative effect of all the gradient terms on the policy parameter $\theta^{(t)}$.

We first show that $\lim_{t \to \infty} \theta_{s,a}^{(t)} = -\infty$. We prove

$$\liminf_{t \to \infty} \theta_{s,a}^{(t)} = -\infty \tag{95}$$

first by contradiction. Given an action $a \in I_s^-$, suppose that there exists ϑ such that $\theta_{s,a}^{(t)} > \vartheta$ for all $t \ge \bar{T}_s$. Then, by Lemmas 24 and 30, we know there must exist an action $a' \in \mathcal{A}$ such that $\liminf_{t \to \infty} \theta_{s,a'}^{(t)} = -\infty$. Let $\delta > 0$ be some positive scalar such that $\theta_{s,a}^{(\bar{T}_s)} \ge \vartheta - \delta$. For each $t \ge \bar{T}_s$, define

$$\nu(t) := \sup\{\tau : \theta_{s,a'}^{(\tau)} \ge \vartheta - \delta, \quad \bar{T}_s \le \tau \le t\}, \tag{96}$$

which is the latest iteration that $\theta_{s,a'}^{(\tau)}$ remains no smaller than $\vartheta - \delta$. We consider two cases: (i) $\{\nu(t)\}_{t=1}^\infty$ is bounded, and (ii) $\lim_{t \to \infty} \nu(t) = \infty$.

For case (i), we define the following index set

$$\mathcal{J}^{(t)} := \left\{\tau : \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a'}}\bigg|_{\theta = \omega^{(\tau)}} < 0, \quad \nu(t) < \tau < t\right\}. \tag{97}$$

and denote the cumulative effect (up to iteration t) of the gradient terms from those iterations in $\mathcal{J}^{(t)}$ as

$$Z^{(t)} := \sum_{t' \in \mathcal{J}^{(t)}} \eta^{(t'+1)} \cdot \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a'}}\bigg|_{\theta = \omega^{(t')}} \cdot G(t', t), \tag{98}$$

where $G(t', t)$ is defined in (53). If $\mathcal{J}^{(t)} = \varnothing$, we define $Z^{(t)} = 0$.

By setting $R_{\max} = 1$ in Lemma 10, we have the upper bounds $V^\pi(s) \le 1/(1-\gamma)$ for any $s \in \mathcal{S}$ and any π, $Q^\pi(s,a) \le 1/(1-\gamma)$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and any π, and also $|A(s,a)| \le 1/(1-\gamma)$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and any π. Leveraging these upper bounds and Lemma 11, we have

$$|\partial V^{\pi_\theta}(\mu)/\partial \theta_{s,a}| \le 1/(1-\gamma)^2, \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{99}$$

Accordingly, for any $t > \bar{T}_s$, we have

$$
Z^{(t)} \leq \sum_{t' \in \mathcal{J}^{(t)}} \eta^{(t'+1)} \cdot \left.\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a'}}\right|_{\theta=\omega^{(t')}} \cdot G(t',t) + \underbrace{\sum_{t':t' \notin \mathcal{J}^{(t)}, \nu(t) < t' < t} \eta^{(t'+1)} \cdot \left.\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a'}}\right|_{\theta=\omega^{(t')}} \cdot G(t',t)}_{\geq 0, \text{by the definition of } \mathcal{J}^{(t)}}
$$

$$
+ \sum_{t' \leq \nu(t)} \eta^{(t'+1)} \cdot \underbrace{\left(\left.\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a'}}\right|_{\theta=\omega^{(t')}} + \frac{1}{(1-\gamma)^2}\right) \cdot G(t',t)}_{\geq 0, \text{ by (99)}}
$$

$$
= \sum_{t' \leq t} \eta^{(t'+1)} \cdot \left.\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a'}}\right|_{\theta=\omega^{(t')}} \cdot G(t',t) + \sum_{t' \leq \nu(t)} \eta^{(t'+1)} \frac{1}{(1-\gamma)^2} G(t',t)
$$

$$
= (\theta_{s,a'}^{(t)} - \theta_{s,a'}^{(1)}) + \sum_{t' \leq \nu(t)} \eta^{(t'+1)} \frac{1}{(1-\gamma)^2} G(t',t), \tag{100}
$$

where (100) is from the update scheme of APG as in Algorithm 2. Since $\{\nu(t)\}_{t=1}^{\infty}$ is bounded, there is a finite constant M such that $\sum_{t' \leq \nu(t)} \eta^{(t'+1)} \frac{1}{(1-\gamma)^2} G(t',t) \leq M$ for all t. Therefore, by taking the limit infimum on (100), we know

$$
\liminf_{t \to \infty} Z^{(t)} = -\infty. \tag{101}
$$

Now we are ready to quantify $\theta_{s,a}^{(t)}$ for the action $a \in I_s^-$. For all $t' \in \mathcal{J}^{(t)}$, we have

$$
\frac{|(\partial V^{\pi_\theta}(\mu)/\partial \theta_{s,a})|_{\theta=\omega^{(t')}}|}{|(\partial V^{\pi_\theta}(\mu)/\partial \theta_{s,a'})|_{\theta=\omega^{(t')}}|} = \left|\frac{\pi_\omega^{(t')}(a|s)A^{\pi_\omega^{(t')}}(s,a)}{\pi_\omega^{(t')}(a'|s)A^{\pi_\omega^{(t')}}(s,a')}\right| \geq \exp(\vartheta - \theta_{s,a'}^{(t')}) \cdot \frac{(1-\gamma)\Delta_s}{4} \geq \exp(\delta) \cdot \frac{(1-\gamma)\Delta_s}{4}, \tag{102}
$$

where the first inequality follows from the softmax parameterization, the assumed upper bound of $\theta_{s,a}^{(t)}$, that $|A^{\pi_\omega^{(t')}}(s,a')| \leq 1/(1-\gamma)$, and that $A^{\pi_\omega^{(t')}}(s,a) \leq -\Delta_s/4$, and the second one holds by the definition of $\nu(t)$ in (96). For any $\mathcal{J}^{(t)} \neq \varnothing$, we have

$$
\theta_{s,a}^{(t)} - \theta_{s,a}^{(1)} = \sum_{t':1 \leq t' < \bar{T}_s} \eta^{(t'+1)} \cdot \left.\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\right|_{\theta=\omega^{(t')}} \cdot G(t',t) + \sum_{t':t' \geq \bar{T}_s} \eta^{(t'+1)} \cdot \left.\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\right|_{\theta=\omega^{(t')}} \cdot G(t',t)
$$

$$
\leq \sum_{t':1 \leq t' < \bar{T}_s} \eta^{(t'+1)} \cdot \left.\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\right|_{\theta=\omega^{(t')}} \cdot G(t',t) + \sum_{t':t' \in \mathcal{J}^{(t)}} \eta^{(t'+1)} \cdot \left.\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\right|_{\theta=\omega^{(t')}} \cdot G(t',t) \tag{103}
$$

$$
\leq \underbrace{\sum_{t':1 \leq t' < \bar{T}_s} \eta^{(t'+1)} \left.\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\right|_{\theta=\omega^{(t')}} G(t',t)}_{< \infty \text{ and does not depend on } t} + \exp(\delta)\frac{(1-\gamma)\Delta_s}{4} \underbrace{\sum_{t':t' \in \mathcal{J}^{(t)}} \eta^{(t'+1)} \left.\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a'}}\right|_{\theta=\omega^{(t')}} G(t',t)}_{\equiv Z^{(t)}},
$$

$$
\tag{104}
$$

where (103) holds by the fact that $A^{\pi_\omega^{(t')}}(s,a) < 0$ for all $t \geq \bar{T}_s$, and (104) is a direct result of (102). Therefore, by taking the limit infimum on (104), we have $\liminf_{t \to \infty} \theta_{s,a}^{(t)} = -\infty$, which leads to a contradiction.

For the case (ii), given that $\nu(t)$ becomes unbounded as $t \to \infty$, there are infinitely many $\tau \geq \bar{T}_s$ such that $\theta_{s,a'}^{(\tau)} \geq \vartheta - \delta$. In addition, recall that a' is an action that satisfies $\liminf_{t \to \infty} \theta_{s,a'}^{(t)} = -\infty$. Therefore, we can find a subsequence $\{t_i\}_{i=1}^{\infty}$ of time indices satisfying the following properties:

- $t_1 \geq \bar{T}_s$

- $\lim_{i \to \infty} \theta_{s,a'}^{(t_i)} = -\infty$.

- $\theta_{s,a'}^{(t_i)} < \vartheta - \delta$ for all $i \geq 1$.

- There is always a $\tau_i \geq \bar{T}_s$ satisfying that $\theta_{s,a'}^{(\tau_i)} \geq \vartheta - \delta$, $\tau_i \in (t_i, t_{i+1})$, and we specifically let $\tau_i := \inf\{\tau \in (t_i, t_{i+1}) : \theta_{s,a'}^{(\tau_i)} \geq \vartheta - \delta\}$ for each $i \geq 1$.

Since $\lim_{i \to \infty} \theta_{s,a'}^{(t_i)} = -\infty$, we can select $\{t_i\}$ in a way such that $\theta_{s,a'}^{(t_i)} - \theta_{s,a'}^{(t_i-1)} < 0$ for all i. Therefore,

$$
\begin{aligned}
\theta_{s,a'}^{(\tau_i)} - \theta_{s,a'}^{(t_i)} &= (\theta_{s,a'}^{(t_i)} - \theta_{s,a'}^{(t_i-1)}) \cdot (G(t_i - 1, \tau_i) - 1) + \sum_{t_i \leq t' < \tau_i} \eta^{(t'+1)} \cdot \left. \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a'}} \right|_{\theta=\omega^{(t')}} \cdot G(t', \tau_i) \\
&\leq \sum_{t_i \leq t' < \tau_i} \eta^{(t'+1)} \cdot \left. \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a'}} \right|_{\theta=\omega^{(t')}} \cdot G(t', \tau_i) \\
&\leq \sum_{t_i \leq t' < \tau_i} \left| \eta^{(t'+1)} \cdot \left. \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a'}} \right|_{\theta=\omega^{(t')}} \cdot G(t', \tau_i) \right| \\
&\leq \frac{4}{(1-\gamma)\Delta_s} \cdot \sum_{t_i \leq t' < \tau_i} \left| \eta^{(t'+1)} \cdot \left. \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} \right|_{\theta=\omega^{(t')}} \cdot G(t', \tau_i) \right|,
\end{aligned}
\tag{105}
$$

where (105) is from that

$$
\frac{|(\partial V^{\pi_\theta}(\mu)/\partial \theta_{s,a'})|_{\theta=\omega^{(t')}}|}{|(\partial V^{\pi_\theta}(\mu)/\partial \theta_{s,a})|_{\theta=\omega^{(t')}}|} = \left| \frac{\pi_\omega^{(t')}(a'|s) A^{\pi_\omega^{(t')}}(s,a')}{\pi_\omega^{(t')}(a|s) A^{\pi_\omega^{(t')}}(s,a)} \right| \leq \left| \frac{A^{\pi_\omega^{(t')}}(s,a')}{A^{\pi_\omega^{(t')}}(s,a)} \right| \leq \frac{4}{(1-\gamma)\Delta_s},
\tag{106}
$$

where the first inequality holds because $\theta_{s,a'}^{(t')} \leq \theta_{s,a}^{(t')}$, implying $\pi_\omega^{(t')}(a'|s) \leq \pi_\omega^{(t')}(a|s)$. Since $\theta_{s,a'}^{(\tau_i)} - \theta_{s,a'}^{(t_i)} \to \infty$ as $i \to \infty$, we have

$$
\lim_{i \to \infty} \sum_{t_i \leq t' < \tau_i} \left| \eta^{(t'+1)} \cdot \left. \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} \right|_{\theta=\omega^{(t')}} \cdot G(t', \tau_i) \right| = \lim_{i \to \infty} \left| \sum_{t_i \leq t' < \tau_i} \eta^{(t'+1)} \cdot \left. \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} \right|_{\theta=\omega^{(t')}} \cdot G(t', \tau_i) \right| = \infty,
\tag{107}
$$

where the first equality holds because $a \in I_s^-$ together with $t' \geq \bar{T}_s$ implies that $\left. \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} \right|_{\theta=\omega^{(t')}} < 0$ for all t' included in the summation. Now consider

$$
\theta_{s,a}^{(t)} - \theta_{s,a}^{(\bar{T}_s)} = (\theta_{s,a}^{(\bar{T}_s)} - \theta_{s,a}^{(\bar{T}_s-1)}) \cdot (G(\bar{T}_s - 1, t) - 1) + \sum_{t'=\bar{T}_s}^{t-1} \eta^{(t'+1)} \cdot \left. \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} \right|_{\theta=\omega^{(t')}} \cdot G(t', \tau_i).
$$

By an argument similar to (93), the first term above is at most $(\theta_{s,a}^{(\bar{T}_s)} - \theta_{s,a}^{(\bar{T}_s-1)}) \cdot O(\bar{T}_s)$ regardless of t. Combining this with (94) and (107), we obtain $\liminf_{t \to \infty} \theta_{s,a}^{(t)} = -\infty$, which also leads to a contradiction. Therefore, (95) is proven.

Then, we show that there exists a time $\bar{T}_s' \geq \bar{T}_s$ such that $\theta_{s,a}^{(\bar{T}_s'+1)} - \theta_{s,a}^{(\bar{T}_s')} \leq 0$. If such a $\bar{T}_s'$ does not exist, which means that $\{\theta_{s,a}^{(t)}\}$ is monotonically increasing, we see a contradiction with (95). Since $a \in I_s^-$, from (94), we know that if $\theta_{s,a}^{(\bar{T}_s'+1)} - \theta_{s,a}^{(\bar{T}_s')} \leq 0$, then $\theta_{s,a}^{(t+1)} - \theta_{s,a}^{(t)} \leq 0$ for all $t \geq \bar{T}_s'$, meaning that $\theta_{s,a}^{(t)}$ is monotonically decreasing after $\bar{T}_s'$. Therefore, combining this monotinicity after $\bar{T}_s'$ and (95), we obtain $\lim_{t \to \infty} \theta_{s,a}^{(t)} = -\infty$.

Finally, we show that $\lim_{t \to \infty} \omega_{s,a}^{(t)} = -\infty$. Since $\theta_{s,a}^{(t+1)} - \theta_{s,a}^{(t)} \leq 0$ for all $t \geq \bar{T}_s'$, by (7) and that $a \in I_s^-$, we know that $\omega_{s,a}^{(t)} \leq \theta_{s,a}^{(t)}$ and thus $\omega_{s,a}^{(t)}$ is decreasing for all $t \geq \bar{T}_s'$. Combining the above properties with $\lim_{t \to \infty} \theta_{s,a}^{(t)} = -\infty$, we arrive at $\lim_{t \to \infty} \omega_{s,a}^{(t)} = -\infty$. $\qquad\square$

In the following lemma, we use the notation $\Delta\theta_{s,a}^{(t)} := \theta_{s,a}^{(t+1)} - \theta_{s,a}^{(t)}$, for each state-action pair (s,a) and each $t \in \mathbb{N}$.

Lemma 32. *Consider any state s with non-empty I_s^+ and some $a_+ \in I_s^+$ and $a \in I_s^0$. Under the setting of Lemma 25, if $\omega_{s,a_+}^{(\tau)} > \omega_{s,a}^{(\tau)}$ and $\Delta\theta_{s,a_+}^{(\tau)} > \Delta\theta_{s,a}^{(\tau)}$ for some $\tau > \bar{T}_s$, then $\omega_{s,a_+}^{(t)} > \omega_{s,a}^{(t)}$ and $\Delta\theta_{s,a_+}^{(t)} > \Delta\theta_{s,a}^{(t)}$ for all $t > \tau$.*

Proof of Lemma 32. We prove this by induction. Suppose at some time $\tau > \bar{T}_s$, we have $\omega_{s,a_+}^{(\tau)} > \omega_{s,a}^{(\tau)}$ and $\Delta\theta_{s,a_+}^{(\tau)} > \Delta\theta_{s,a}^{(\tau)}$.

Recalling our definition in (88) and (90), we have from Lemma 11 that

$$
\left.\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a_+}}\right|_{\theta=\omega^{(\tau)}} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\omega^{(\tau)}}(s) \cdot \pi_\omega^{(\tau)}(a_+|s) \cdot A^{\pi_\omega^{(\tau)}}(s,a_+)
$$

$$
> \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\omega^{(\tau)}}(s) \cdot \pi_\omega^{(\tau)}(a|s) \cdot A^{\pi_\omega^{(\tau)}}(s,a) \tag{108}
$$

$$
= \left.\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\right|_{\theta=\omega^{(\tau)}}, \tag{109}
$$

where (108) holds by $\omega_{s,a_+}^{(\tau)} > \omega_{s,a}^{(\tau)}$ and the fact that $\tau > \bar{T}_s$ implies $A^{\pi_\omega^{(\tau)}}(s,a_+) \geq A^{\pi_\omega^{(\tau)}}(s,a)$ from (88). By the update rule (6), (109) and that $\omega_{s,a_+}^{(\tau)} > \omega_{s,a}^{(\tau)}$ imply

$$
\theta_{s,a_+}^{(\tau+1)} > \theta_{s,a}^{(\tau+1)}. \tag{110}
$$

Thus, from $\Delta\theta_{s,a_+}^{(\tau)} > \Delta\theta_{s,a}^{(\tau)}$, (7), (8) and (110), we have

$$
\omega_{s,a_+}^{(\tau+1)} = \theta_{s,a_+}^{(\tau+1)} + \mathbb{I}\left\{V^{\pi_\varphi^{(\tau+1)}}(\mu) \geq V^{\pi_\theta^{(\tau+1)}}(\mu)\right\} \cdot \frac{\tau}{\tau+3}(\theta_{s,a_+}^{(\tau+1)} - \theta_{s,a_+}^{(\tau)})
$$

$$
> \theta_{s,a}^{(\tau+1)} + \mathbb{I}\left\{V^{\pi_\varphi^{(\tau+1)}}(\mu) \geq V^{\pi_\theta^{(\tau+1)}}(\mu)\right\} \cdot \frac{\tau}{\tau+3}(\theta_{s,a}^{(\tau+1)} - \theta_{s,a}^{(\tau)}) = \omega_{s,a}^{(\tau+1)}. \tag{111}
$$

By (111), (4), and that $A^{\pi_\omega^{(\tau+1)}}(s,a_+) \geq A^{\pi_\omega^{(\tau+1)}}(s,a)$, we have

$$
\left.\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a_+}}\right|_{\theta=\omega^{(\tau+1)}} > \left.\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\right|_{\theta=\omega^{(\tau+1)}}. \tag{112}
$$

Furthermore, by (112) and that $\Delta\theta_{s,a_+}^{(\tau)} > \Delta\theta_{s,a}^{(\tau)}$, we have

$$
\Delta\theta_{s,a_+}^{(\tau+1)} = \mathbb{I}\left\{V^{\pi_\varphi^{(\tau+1)}}(\mu) \geq V^{\pi_\theta^{(\tau+1)}}(\mu)\right\} \cdot \frac{\tau}{\tau+3}\Delta\theta_{s,a_+}^{(\tau)} + \eta^{(\tau+2)}\left.\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a_+}}\right|_{\theta=\omega^{(\tau+1)}}
$$

$$
> \mathbb{I}\left\{V^{\pi_\varphi^{(\tau+1)}}(\mu) \geq V^{\pi_\theta^{(\tau+1)}}(\mu)\right\} \cdot \frac{\tau}{\tau+3}\Delta\theta_{s,a}^{(\tau)} + \eta^{(\tau+2)}\left.\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\right|_{\theta=\omega^{(\tau+1)}} = \Delta\theta_{s,a}^{(\tau+1)}.
$$

By repeating the above argument, we know $\omega_{s,a_+}^{(t)} > \omega_{s,a}^{(t)}$ and $\Delta\theta_{s,a_+}^{(t)} > \Delta\theta_{s,a}^{(t)}$ for all $t > \tau$. □

Next, we take a closer look at the actions in I_s^0. We further decompose I_s^0 into two subsets as follows: Consider any state s with a non-empty I_s^+, for any $a_+ \in I_s^+$, we define

$$
B_s^0(a_+) := \left\{a \in I_s^0 : \text{For any } t \geq \bar{T}_s, \text{ either } \omega_{s,a}^{(t)} \geq \omega_{s,a_+}^{(t)} \text{ or } \Delta\theta_{s,a}^{(t)} \geq \Delta\theta_{s,a_+}^{(t)}\right\}
$$

and use $\bar{B}_s^0(a_+)$ to denote its complement in I_s^0. We could thus also write $\bar{B}_s^0(a_+)$ as

$$
\bar{B}_s^0(a_+) := \left\{a \in I_s^0 : \omega_{s,a}^{(t)} < \omega_{s,a_+}^{(t)} \text{ and } \Delta\theta_{s,a}^{(t)} < \Delta\theta_{s,a_+}^{(t)} \text{ for some } t \geq \bar{T}_s\right\}.
$$

Lemma 33. *Consider any state s. Under the setting of Lemma 25, if I_s^+ is nonempty, then:*

(a) For all $a_+ \in I_s^+$, we have

$$
\lim_{t\to\infty} \sum_{a\in B_s^0(a_+)} \pi_\omega^{(t)}(a|s) = 1.
$$

(b) For all $a_+ \in I_s^+$, we have

$$
\lim_{t\to\infty} \max_{a\in B_s^0(a_+)} \omega_{s,a}^{(t)} = \infty.
$$

(c) For all $a_+ \in I_s^+$, we have

$$
\lim_{t\to\infty} \sum_{a\in B_s^0(a_+)} \omega_{s,a}^{(t)} = \infty.
$$

38

Proof of Lemma 33. Regarding **(a)**, by the definition of $\bar{B}_s^0(a_+)$, for each $a \in \bar{B}_s^0(a_+)$, there must exist some $T' \geq \bar{T}_s$ such that $\omega_{s,a_+}^{(T')} > \omega_{s,a}^{(T')}$ and $\Delta\theta_{s,a_+}^{(T')} > \Delta\theta_{s,a}^{(T')}$. Then, by Lemma 32, we know

$$\omega_{s,a_+}^{(t)} > \omega_{s,a}^{(t)} \quad \text{and} \quad \Delta\theta_{s,a_+}^{(t)} > \Delta\theta_{s,a}^{(t)}, \quad \text{for all } t \geq T'. \tag{113}$$

Moreover, by Lemma 27 and that $a_+ \in I_s^+$, we have $\pi_\omega^{(t)}(a_+|s) \to 0$ as $t \to \infty$. Based on (113), this further implies that $\pi_\omega^{(t)}(a|s) \to 0$ as $t \to \infty$, for all $a \in \bar{B}_s^0(a_+)$. Hence, together with lemma 27 we obtain the desired result.

Regarding **(b)**, by the result in **(a)** and the argument in the proof of Lemma 30, we directly see that the statement holds true.

Regarding **(c)**, let us consider any action $a \in B_s^0(a_+)$. By the definition of $B_s^0(a_+)$, at each iteration $t \geq \bar{T}_s$, either $\omega_{s,a}^{(t)} \geq \omega_{s,a_+}^{(t)}$ or $\Delta\theta_{s,a}^{(t)} \geq \Delta\theta_{s,a_+}^{(t)}$ holds. Given the result in **(b)**, it suffices to show that $\{\omega_{s,a}^{(t)}\}_{t=1}^\infty$ is bounded from below for all $a \in B_s^0(a_+)$. We prove it by contradiction. Suppose that there exists an action $\bar{a} \in B_s^0(a_+)$ such that $\liminf_{t\to\infty} \omega_{s,\bar{a}}^{(t)} = -\infty$. By Lemma 28, we know that $\{\omega_{s,a_+}^{(t)}\}_{t=1}^\infty$ is bounded from below. Therefore, once $\omega_{s,\bar{a}}^{(t)} < \omega_{s,a_+}^{(t)}$ for some t, we will have $\Delta\theta_{s,\bar{a}}^{(t)} \geq \Delta\theta_{s,a_+}^{(t)}$. Following an argument similar to that in the proof of Lemma 31 (for case (ii)) to choose a subsequence of $\omega_{s,\bar{a}}^{(t)}$ converging to negative infinity and since $\omega^{(t)}$ is $\theta^{(t)}$ add momentum with restart, we know that if momentum goes to negative infinity, by $\Delta\theta_{s,\bar{a}}^{(t)} \geq \Delta\theta_{s,a_+}^{(t)}$, we will have $\liminf_{t\to\infty} \omega_{s,a_+}^{(t)} = -\infty$. Otherwise, if $\theta^{(t)}$ goes to negative infinity, we can observe that the total momentum effect goes to negative infinity, which also leads to $\liminf_{t\to\infty} \omega_{s,a_+}^{(t)} = -\infty$ by $\Delta\theta_{s,\bar{a}}^{(t)} \geq \Delta\theta_{s,a_+}^{(t)}$. This leads to the desired contradiction and concludes our proof. □

Lemma 34. *Consider any state s. Under the setting of Lemma 25, for any $a_+ \in I_s^+$, the following hold.*

(a) There exists T_{a_+} such that for all $a \in \bar{B}_s^0(a_+)$,

$$\pi_\omega^{(t)}(a_+|s) > \pi_\omega^{(t)}(a|s), \quad \text{for all } t > T_{a_+}.$$

(b) There exists $T_{a_+}^\dagger$ such that for all $a \in \bar{B}_s^0(a_+)$,

$$|A^{\pi_\omega^{(t)}}(s,a)| < \frac{\pi_\omega^{(t)}(a_+|s)}{\pi_\omega^{(t)}(a|s)} \cdot \frac{\Delta_s}{16|\mathcal{A}|}, \quad \text{for all } t > T_{a_+}^\dagger. \tag{114}$$

This also implies that

$$\sum_{a \in \bar{B}_s^0(a_+)} \pi_\omega^{(t)}(a|s) A^{\pi_\omega^{(t)}}(s,a) > -\pi_\omega^{(t)}(a_+|s) \cdot \frac{\Delta_s}{16}, \quad \text{for all } t > T_{a_+}^\dagger. \tag{115}$$

Proof of Lemma 34. Regarding **(a)**, for each $a \in \bar{B}_s^0(a_+)$, we define

$$u_a(a_+) := \inf\{\tau \geq \bar{T}_s : \omega_{s,a_+}^{(\tau)} > \omega_{s,a}^{(\tau)} \text{ and } \Delta\theta_{s,a_+}^{(\tau)} > \Delta\theta_{s,a}^{(\tau)}\}.$$

By the definition of $\bar{B}_s^0(a_+)$, we know the following two facts: (i) $u_a(a_+)$ is finite for all $a \in \bar{B}_s^0(a_+)$. (ii) By Lemma 32, for all $t \geq u_a(a_+)$, we have $\omega_{s,a_+}^{(t)} > \omega_{s,a}^{(t)}$ and $\Delta\theta_{s,a_+}^{(t)} > \Delta\theta_{s,a}^{(t)}$. Therefore, by choosing $T_{a_+} := \max_{a \in \bar{B}_s^0(a_+)} u_a(a_+)$, we must have $\pi_\omega^{(t)}(a_+|s) > \pi_\omega^{(t)}(a|s)$ for all $t > T_{a_+}$.

Regarding **(b)**, we have from **(a)**

$$\frac{\pi_\omega^{(t)}(a_+|s)}{\pi_\omega^{(t)}(a|s)} > 1, \quad \forall t > T_{a_+}.$$

Thus, we know that for each $a \in \bar{B}_s^0(a_+) \subseteq I_s^0$, there exists some finite $t_a' > T_{a_+}$ such that

$$|A^{\pi_\omega^{(t)}}(s,a)| < \frac{\pi_\omega^{(t)}(a_+|s)}{\pi_\omega^{(t)}(a|s)} \cdot \frac{\Delta_s}{16|\mathcal{A}|}, \quad \text{for all } t \geq t_a'.$$

As a result, by choosing $T_{a_+}^\dagger := \max_{a \in \bar{B}_s^0(a_+)} t_a'$, we conclude that (114)-(115) indeed hold. □

39

Lemma 35. *Under the setting of Lemma 25, if I_s^+ is non-empty, then for any $a_+ \in I_s^+$, there exists some finite $\tilde{T}_{a_+}$ such that*

$$\sum_{a \in I_s^-} \pi_\omega^{(t)}(a|s) A^{\pi_\omega^{(t)}}(s,a) > -\pi_\omega^{(t)}(a_+|s)\frac{\Delta_s}{16}, \quad \forall t \geq \tilde{T}_{a_+}. \tag{116}$$

Proof. Consider any $a_- \in I_s^-$. By Lemmas 28 and 31, we know $\{\omega_{s,a_+}^{(t)}\}$ is bounded from below and $\omega_{s,a_-}^{(t)} \to -\infty$ as $t \to \infty$. This implies that $\pi_\omega^{(t)}(a_-|s)/\pi_\omega^{(t)}(a_+|s) \to 0$ as $t \to \infty$. Therefore, there exists some finite $t'_{a_-} \geq \bar{T}_s$ such that

$$\frac{\pi_\omega^{(t)}(a_-|s)}{\pi_\omega^{(t)}(a_+|s)} < \frac{\Delta_s(1-\gamma)}{16|\mathcal{A}|}, \quad \forall t \geq t'_{a_-}. \tag{117}$$

By the definition of $\bar{T}_s$ in (89) and Lemma 10, we know that

$$-\frac{1}{1-\gamma} \leq A^{\pi_\omega^{(t)}}(s,a_-) < 0, \quad \forall t \geq t'_{a_-}. \tag{118}$$

Combining (117) and (118) leads to

$$\pi_\omega^{(t)}(a_-|s) A^{\pi_\omega^{(t)}}(s,a_-) > -\pi_\omega^{(t)}(a_+|s)\frac{\Delta_s}{16|\mathcal{A}|}. \tag{119}$$

By choosing $\tilde{T}_{a_+} := \max_{a_- \in I_s^-} t'_{a_-}$ and summing (119) over $a_i \in I_s^-$, we obtain (116). $\qquad\square$

C.2. Putting Everything Together: Asymptotic Convergence of APG

Now we are ready to put everything together and prove Theorem 1, which we restate below.

Theorem 1 (**Asymptotic Convergence Under Softmax Parameterization**). *Consider a tabular softmax parameterized policy π_θ. For APG with $\eta^{(t)} = \frac{t}{t+1} \frac{(1-\gamma)^3}{16}$ and μ initialized uniformly at random, the following holds almost surely:*

$$\lim_{t \to \infty} V^{\pi_\theta^{(t)}}(s) = V^*(s), \forall s \in \mathcal{S}.$$

Proof of Theorem 1. By Lemma 4, it suffices to prove that $I_s^+ = \varnothing$ for all $s \in \mathcal{S}$. We prove this by contradiction. Suppose there exists at least one state $s \in \mathcal{S}$ with a non-empty I_s^+. Consider an action $a_+ \in I_s^+$. Recall the definitions of $\bar{T}_s$, $T_{a_+}, T_{a_+}^\dagger$, and $\tilde{T}_{a_+}$ from (88)-(90), Lemma 34, and Lemma 35. We define $T_{\max} := \max\{\bar{T}_s, T_{a_+}, T_{a_+}^\dagger, \tilde{T}_{a_+}\}$, then we have for all $t > T_{\max}$,

$$
0 = \sum_{a \in B_s^0(a_+)} \pi_\omega^{(t)}(a|s) A^{\pi_\omega^{(t)}}(s,a) + \underbrace{\sum_{a \in \bar{B}_s^0(a_+)} \pi_\omega^{(t)}(a|s) A^{\pi_\omega^{(t)}}(s,a)}_{> -\pi_\omega^{(t)}(a_+|s)\frac{\Delta_s}{16} \text{ by Lemma 34}}
$$

$$
+ \underbrace{\sum_{a \in I_s^+} \pi_\omega^{(t)}(a|s) A^{\pi_\omega^{(t)}}(s,a)}_{\geq \pi_\omega^{(t)}(a_+|s)\frac{\Delta_s}{4}} + \underbrace{\sum_{a \in I_s^-} \pi_\omega^{(t)}(a|s) A^{\pi_\omega^{(t)}}(s,a)}_{> -\pi_\omega^{(t)}(a_+|s)\frac{\Delta_s}{16} \text{ by Lemma 35}} \tag{120}
$$

$$
> \sum_{a \in B_s^0(a_+)} \pi_\omega^{(t)}(a|s) A^{\pi_\omega^{(t)}}(s,a) + \frac{1}{8} \cdot \pi_\omega^{(t)}(a_+|s)\Delta_s
$$

$$
> \sum_{a \in B_s^0(a_+)} \pi_\omega^{(t)}(a|s) A^{\pi_\omega^{(t)}}(s,a), \tag{121}
$$

where (120) uses Lemma 7. By Lemma 11, (121) further implies

$$
\sum_{a \in B_s^0(a_+)} \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} \bigg|_{\theta=\omega^{(t)}} < 0, \quad \forall t > T_{\max}. \tag{122}
$$

Moreover, for any $t > T_{\max} + 4$, we have

$$
\sum_{a \in B_s^0(a_+)} \theta_{s,a}^{(t)} - \theta_{s,a}^{(0)}
$$

$$
= \sum_{a \in B_s^0(a_+)} \sum_{t'=0}^{t-1} \left(\eta^{(t'+1)} \cdot \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} \bigg|_{\theta=\omega^{(t')}} \cdot G(t',t) \right) \tag{123}
$$

$$
= \sum_{t'=0}^{t-1} \eta^{(t'+1)} G(t',t) \cdot \left(\sum_{a \in B_s^0(a_+)} \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} \bigg|_{\theta=\omega^{(t')}} \right)
$$

$$
= \underbrace{\sum_{t'=0}^{T_{\max}} \eta^{(t'+1)} G(t',t) \cdot \left(\sum_{a \in B_s^0(a_+)} \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} \bigg|_{\theta=\omega^{(t')}} \right)}_{\text{Independent of } t \text{ and thus finite}} + \underbrace{\sum_{t'=T_{\max}+1}^{t-1} \eta^{(t'+1)} G(t',t) \cdot \left(\sum_{a \in B_s^0(a_+)} \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} \bigg|_{\theta=\omega^{(t')}} \right)}_{<0 \text{ by (122)}},
$$

$$\tag{124}$$

where (123) holds due to Lemma 23 and $t > T_{\max} + 4 > 4$. By taking the limit of t to infinity at both sides of (124), the left-hand side goes to positive infinity by Lemma 33 but the right-hand side is bounded from above. This leads to a contradiction and hence completes the proof. $\square$

Corollary 2. *Consider a tabular softmax parameterized policy and assume that Assumption 1 holds. Under the setting of Theorem 1, the following holds almost surely:*

$$
\lim_{t \to \infty} V^{\pi_\omega^{(t)}}(s) = V^*(s), \quad \lim_{t \to \infty} \pi_\theta(a^*(s)|s) = 1, \quad \lim_{t \to \infty} \pi_\omega(a^*(s)|s) = 1, \quad \forall s \in \mathcal{S}.
$$

Proof of Corollary 2. By Lemma 26 and Theorem 1, we know that almost surely

$$\lim_{t \to \infty} V^{\pi_\omega^{(t)}}(s) = \lim_{t \to \infty} V^{\pi_\theta^{(t)}}(s) = V^*(s), \quad \forall s \in \mathcal{S}. \tag{125}$$

From Assumption 1 and page 14 of (Szepesvári, 2022), (125) then implies that the policy weight of the optimal action $a^*(s)$ converges to 1, proving the desired results. $\qquad\square$

D. Convergence Rate of APG With Nearly Constant Step Sizes

D.1. Proofs of Lemmas 1 and 2

This subsection proves Lemmas 1 and 2 using the lemmas we have developed to prepare for proving Theorem 2. Before delving into the proof of Lemmas 1 and 2, we state a lemma useful for proving them.

Lemma 36. *Under the setting of Corollary 2, there exists a finite time T_V such that $V^{\pi_\theta^{(t)}}(s) > Q^*(s, a_2(s))$ for all $s \in \mathcal{S}$ and $t \geq T_V$ almost surely.*

Proof of Lemma 36. By Theorem 1, $V^{\pi_\theta^{(t)}}(s) \to V^*(s)$ for all $s \in \mathcal{S}$ almost surely. Hence, given $\varepsilon = V^*(s) - Q^*(s, a_2(s)) > 0$ by Lemma 8, almost surely there is a finite $T_V(s)$ such that $V^{\pi_\theta^{(t)}}(s) > Q^*(s, a_2(s))$ for any $t \geq T_V(s)$. Accordingly, by noting that $\mathcal{S}$ is a finite set and that any finite intersection of measure zero events is still measure zero, we let $T_V = \max_{s \in \mathcal{S}} T_V(s)$ and the proof is completed. $\qquad\square$

Lemma 1 (Locally C-Near Concavity; Formal). Given any $C > 1$ and a direction $\boldsymbol{d}$ satisfying the setting of Lemma 20. Let θ be a policy parameter satisfying the following conditions: (i) $V^{\pi_\theta}(s) > Q^*(s, a_2(s))$ for all $s \in \mathcal{S}$; (ii) $\theta_{s, a^*(s)} - \theta_{s,a} > M_{C,\boldsymbol{d}}$, for all $s \in \mathcal{S}$, and $a \neq a^*(s)$, where

$$M_{C,\boldsymbol{d}} := \max\left\{ 2 \max_{s \in \mathcal{S}, i > 1} \left\{ \ln\left[2(|\mathcal{A}| - 1)\right] - \ln[d_{s, a^*(s)} - d_{s, a_i(s)}] \right\}, \ln\left[\frac{|\mathcal{S}| |\mathcal{A}|^2}{(C-1)(1-\gamma)^2 \min_{s \in \mathcal{S}} \mu(s)}\right] \right\}. \quad (126)$$

Then, we have that the objective function $\theta \mapsto V^{\pi_\theta}(\mu)$ is C-nearly concave along the direction $\boldsymbol{d}$.

Proof of Lemma 1. Given any $s \in \mathcal{S}$, according to Lemma 20, provided that $\theta_{s, a^*} - \theta_{s, a} > M_{\boldsymbol{d}}$ for all $a \neq a^*$, the function $\theta \to \pi_\theta(a^*(s)|s)$ is concave along the direction $\boldsymbol{d}$, and the functions $\theta \to \pi_\theta(a|s)$ for all $a \neq a^*(s)$ are convex along the direction $\boldsymbol{d}$. Suppose θ is a parameter satisfies $\theta_{s, a^*(s)} - \theta_{s, a} > M_{C, \boldsymbol{d}}$ for all $s \in \mathcal{S}$ and all $a \neq a^*(s)$, and consider θ' as a parameter obtained from updating θ along the direction $\boldsymbol{d}$. To show that $\theta \to V^{\pi_\theta}(\mu)$ is C-nearly concave along the desired direction, we have to show that

$$V^{\pi_{\theta'}}(\mu) - V^{\pi_\theta}(\mu) \leq C \cdot \left\langle \nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\theta}, \theta' - \theta \right\rangle.$$

By Lemmas 4 and 7,

$$V^{\pi_{\theta'}}(\mu) - V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma} \sum_{s,a} d_\mu^{\pi_{\theta'}}(s) \pi_{\theta'}(a|s) A^{\pi_\theta}(s, a)$$

$$= \frac{1}{1-\gamma} \sum_{s,a} d_\mu^{\pi_{\theta'}}(s) (\pi_{\theta'}(a|s) - \pi_\theta(a|s)) A^{\pi_\theta}(s, a)$$

$$= \sum_{s,a} \frac{d_\mu^{\pi_{\theta'}}(s)}{d_\mu^{\pi_\theta}(s)} (\pi_{\theta'}(a|s) - \pi_\theta(a|s)) \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s) A^{\pi_\theta}(s, a)$$

$$\leq \left\| \frac{d_\mu^{\pi_{\theta'}}}{d_\mu^{\pi_\theta}} \right\|_\infty \left\langle \nabla_{\pi_\theta} V^{\pi_\theta}(\mu)|_{\pi=\pi_\theta}, \pi_{\theta'} - \pi_\theta \right\rangle, \quad (127)$$

where (127) holds by Lemmas 12 and 21.

Since θ' is obtained by updating θ with the direction $\boldsymbol{d}$, we have

$$\theta'_{s, a^*(s)} - \theta'_{s, a} > \theta_{s, a^*(s)} - \theta_{s, a} > M_{C, \boldsymbol{d}} \geq \left[\frac{|\mathcal{S}| |\mathcal{A}|^2}{(C-1)(1-\gamma)^2 \min_{s \in \mathcal{S}} \mu(s)}\right].$$

Thus, by Lemma 19, we have that $\|d_\mu^{\pi_{\theta'}}/d_\mu^{\pi_\theta}\|_\infty < C$. Moreover, by Lemma 12, we have

$$\left\langle \nabla_{\pi_\theta} V^{\pi_\theta}(\mu)|_{\pi=\pi_\theta}, \pi_{\theta'} - \pi_\theta \right\rangle = \left\langle \nabla_\theta V^{\pi_\theta}(\mu)|_{\theta=\theta}, \frac{\pi_{\theta'} - \pi_\theta}{\pi_\theta} \right\rangle.$$

43

Hence, it remains to show that

$$\left\langle \nabla_\theta V^{\pi_\theta}(\mu)|_{\theta=\theta}, \frac{\pi_{\theta'} - \pi_\theta}{\pi_\theta} \right\rangle \leq \left\langle \nabla_\theta V^{\pi_\theta}(\mu)|_{\theta=\theta}, \theta' - \theta \right\rangle. \tag{128}$$

According to the concavity of $\theta \to \pi_\theta(a^*(s)|s)$ along $\boldsymbol{d}$ for all $s \in \mathcal{S}$, we have

$$
\begin{aligned}
\frac{\pi_{\theta'}(a^*(s)|s) - \pi_\theta(a^*(s)|s)}{\pi_\theta(a^*(s)|s)} &\leq \frac{\left\langle \nabla_\theta \pi_\theta(a^*(s)|s)|_{\theta=\theta}, \theta' - \theta \right\rangle}{\pi_\theta(a^*(s)|s)} \\
&= \frac{\sum_a \frac{\partial \pi_\theta(a^*(s)|s)}{\partial \theta_{s,a}} \cdot (\theta'_{s,a} - \theta_{s,a})}{\pi_\theta(a^*(s)|s)} \\
&= (1 - \pi_\theta(a^*(s)|s)) \cdot (\theta'_{s,a^*(s)} - \theta_{s,a^*(s)}) - \sum_{a \neq a^*(s)} \pi_\theta(a|s) \cdot (\theta'_{s,a} - \theta_{s,a}) \\
&= (\theta'_{s,a^*(s)} - \theta_{s,a^*(s)}) - \sum_a \pi_\theta(a|s) \cdot (\theta'_{s,a} - \theta_{s,a}). \tag{129}
\end{aligned}
$$

On the other hand, since $\theta \to \pi_\theta(a_i|s)$ is convex along $\boldsymbol{d}$ for all $a_i \neq a^*(s)$, we have

$$
\begin{aligned}
\frac{\pi_{\theta'}(a_i|s) - \pi_\theta(a_i|s)}{\pi_\theta(a_i|s)} &\geq \frac{\left\langle \nabla_\theta \pi_\theta(a_i|s)\big|_{\theta=\theta}, \theta' - \theta \right\rangle}{\pi_\theta(a_i|s)} \\
&= \frac{\sum_a \frac{\partial \pi_\theta(a_i|s)}{\partial \theta_{s,a}} \cdot (\theta'_{s,a} - \theta_{s,a})}{\pi_\theta(a_i|s)} \\
&= (1 - \pi_\theta(a_i|s)) \cdot (\theta'_{s,a_i} - \theta_{s,a_i}) - \sum_{a \neq a_i} \pi_\theta(a|s) \cdot (\theta'_{s,a} - \theta_{s,a}) \\
&= (\theta'_{s,a_i} - \theta_{s,a_i}) - \sum_a \pi_\theta(a|s) \cdot (\theta'_{s,a} - \theta_{s,a}). \tag{130}
\end{aligned}
$$

Finally, we put everything together. By (46), (47), (129) and (130), we have

$$
\begin{aligned}
\left\langle \nabla_\theta V^{\pi_\theta}(\mu)\big|_{\theta=\theta}, \frac{\pi_{\theta'} - \pi_\theta}{\pi_\theta} \right\rangle &= \sum_{s,a} \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\bigg|_{\theta=\theta_{s,a}} \frac{\pi_{\theta'}(a|s) - \pi_\theta(a|s)}{\pi_\theta(a|s)} \\
&\leq \sum_{s,a} \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\bigg|_{\theta=\theta_{s,a}} \left((\theta'_{s,a} - \theta_{s,a}) - \sum_{a'} \pi_\theta(a'|s) \cdot (\theta'_{s,a'} - \theta_{s,a'}) \right) \tag{131} \\
&= \left\langle \nabla_\theta V^{\pi_\theta}(\mu)\big|_{\theta=\theta}, \theta' - \theta \right\rangle \\
&\quad - \sum_s \sum_{a'} \pi_\theta(a'|s)(\theta'_{s,a'} - \theta_{s,a'}) \cdot \sum_a \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\bigg|_{\theta=\theta_{s,a}} \\
&= \left\langle \nabla_\theta V^{\pi_\theta}(\mu)\big|_{\theta=\theta}, \theta' - \theta \right\rangle, \tag{132}
\end{aligned}
$$

where (131) is from (46), (47), (129) and (130), and (132) is because

$$
\begin{aligned}
\sum_a \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\bigg|_{\theta=\theta_{s,a}} &= \sum_a \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s) \pi_\theta(a|s) A^{\pi_\theta}(s,a) \\
&= \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s) \sum_a \pi_\theta(a|s) A^{\pi_\theta}(s,a) = 0
\end{aligned}
$$

from Lemmas 7 and 11. We have therefore obtained (128) and thus the proof is completed. $\square$

Lemma 2. *Consider a tabular softmax parameterized policy. Under Assumption 1 and the setting of Theorem 1, the following holds almost surely: Given any $M > 0$, there exists a finite time T such that for all $t \geq T$, $s \in \mathcal{S}$, and $a \neq a^*(s)$, we have (i) $\theta^{(t)}_{s,a^*(s)} - \theta^{(t)}_{s,a} > M$, (ii) $V^{\pi_\theta^{(t)}}(s) > Q^*(s, a_2(s))$, (iii) $\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a^*(s)}}\Big|_{\theta=\omega^{(t)}} > 0 > \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\Big|_{\theta=\omega^{(t)}}$, (iv) $\omega^{(t)}_{s,a^*(s)} - \theta^{(t)}_{s,a^*(s)} \geq \omega^{(t)}_{s,a} - \theta^{(t)}_{s,a}$.*

Proof of Lemma 2. We first show the $\theta^{(t)}_{s,a^*(s)} - \theta^{(t)}_{s,a} > M$ part. By Corollary 2, almost surely we have

$$\lim_{t\to\infty} \pi_\theta^{(t)}(a^*(s)|s) = 1, \quad \forall s \in \mathcal{S}. \tag{133}$$

Under this event, assume for contradiction that there is a $M > 0$ such that no such a finite T exists. Then there is an infinite sequence $\{t_i\}$ such that for each t_i, there is a state s_{t_i} and an action $a_{t_i} \neq a^*(s_{t_i})$ satisfying that $\theta^{(t_i)}_{s_{t_i},a^*(s_{t_i})} - \theta^{(t_i)}_{s_{t_i},a_{t_i}} \leq M$. Since the action space and the state space are finite, there must be a state s, an action $\tilde{a}$, and an infinite subsequence $\{t_{i_j}\}_{j=1}^\infty$ of $\{t_i\}$ such that $\theta^{(t_{i_j})}_{s,a^*(s)} - \theta^{(t_{i_j})}_{s,\tilde{a}} \leq M$ for all $j \in \mathbb{N}$. For each of such t_{i_j}, we have

$$\pi_\theta^{(t_{i_j})}(a^*(s)|s) = \frac{\exp(\theta^{(t_{i_j})}_{s,a^*(s)})}{\sum_a \exp(\theta^{(t_{i_j})}_{s,a})} \leq \frac{\exp(\theta^{(t_{i_j})}_{s,\tilde{a}} + M)}{\exp(\theta^{(t_{i_j})}_{s,\tilde{a}} + M) + \exp(\theta^{(t_{i_j})}_{s,\tilde{a}})} = \frac{\exp(M)}{\exp(M) + 1} < 1,$$

which contradicts with (133). Therefore, with probability one, for any $M > 0$, there is a corresponding $T_M < \infty$ such that

$$\theta^{(t)}_{s,a^*(s)} - \theta^{(t)}_{s,a} > M, \quad \forall s \in \mathcal{S}, a \neq a^*(s), \quad \forall t \geq T_M. \tag{134}$$

From Lemma 36, we directly see that with probability one, there is $T_V < \infty$ such that

$$V^{\pi_\theta^{(t)}}(s) > Q^*(s, a_2(s)), \quad \forall s \in \mathcal{S}, \quad \forall t \geq T_V. \tag{135}$$

Given any $s \in \mathcal{S}$, by Corollary 2 and (135), we know that almost surely there is a finite $T_1(s)$ such that $V^{\pi_\omega^{(t)}}(s) > Q^*(s, a_2(s))$ for all $t \geq T_1(s)$. By Lemma 8, it then implies

$$Q^{\pi_\omega^{(t)}}(s, a) \leq Q^*(s, a_2(s)) < V^{\pi_\omega^{(t)}}(s), \quad \forall a \neq a^*(s), \quad \forall t \geq T_1(s).$$

Hence, we have from (2) that $A^{\pi_\omega^{(t)}}(s, a) < 0$ for all $a \neq a^*(s)$ and all $t \geq T_1(s)$. This together with Lemma 7 and that $\pi_{\omega^{(t)}}(a|s) > 0$ for all a (from the nature of softmax parameterization) then implies $A^{\pi_\omega^{(t)}}(s, a^*(s)) > 0$. Therefore, by Lemma 11, we see that with probability one,

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a^*(s)}}\Big|_{\theta=\omega^{(t)}} > 0 > \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\Big|_{\theta=\omega^{(t)}}, \quad \forall a \neq a^*(s), \quad \forall t \geq T_1(s). \tag{136}$$

By taking $T_1 \coloneqq \max_{s \in \mathcal{S}} T_1(s)$, (136) implies that with probability one,

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a^*(s)}}\Big|_{\theta=\omega^{(t)}} > 0 > \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\Big|_{\theta=\omega^{(t)}}, \quad \forall a \neq a^*(s), \quad \forall s \in \mathcal{S}, \quad \forall t \geq T_1. \tag{137}$$

We now discuss the part regarding (iv). We claim that

Claim 2. *For any $t \geq T_1$ in (137), if there is $(s, a') \in \mathcal{S} \times \mathcal{A}$ such that*

$$\omega^{(t)}_{s,a^*(s)} - \theta^{(t)}_{s,a^*(s)} \geq \omega^{(t)}_{s,a'} - \theta^{(t)}_{s,a'}, \tag{138}$$

then

$$\omega^{(t+1)}_{s,a^*(s)} - \theta^{(t+1)}_{s,a^*(s)} \geq \omega^{(t+1)}_{s,a'} - \theta^{(t+1)}_{s,a'}.$$

If $a' = a^*(s)$, there is nothing to prove. When $a' \neq a^*(s)$, by considering the update (6), we have

$$\theta_{s,a}^{(t+1)} = \theta_{s,a}^{(t)} + (\omega_{s,a}^{(t)} - \theta_{s,a}^{(t)}) + \eta^{(t+1)} \left. \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} \right|_{\theta=\omega^{(t)}}, \qquad \forall a \in \mathcal{A}. \tag{139}$$

Let $\mathbb{I}^{(t+1)} := \mathbb{I}\left\{ V^{\pi_\varphi^{(t+1)}}(\mu) \geq V^{\pi_\theta^{(t+1)}}(\mu) \right\}$. Combining (139) with (7) and (8) then leads to

$$\begin{aligned}
\omega_{s,a^*(s)}^{(t+1)} - \theta_{s,a^*(s)}^{(t+1)} &= \mathbb{I}^{(t+1)} \frac{t}{t+3} (\theta_{s,a^*(s)}^{(t+1)} - \theta_{s,a^*(s)}^{(t)}) \\
&= \mathbb{I}^{(t+1)} \frac{t}{t+3} \left(\omega_{s,a^*(s)}^{(t)} - \theta_{s,a^*(s)}^{(t)} + \eta^{(t+1)} \left. \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a^*(s)}} \right|_{\theta=\omega^{(t)}} \right) \\
&\geq \mathbb{I}^{(t+1)} \frac{t}{t+3} \left(\omega_{s,a'}^{(t)} - \theta_{s,a'}^{(t)} + \eta^{(t+1)} \left. \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a'}} \right|_{\theta=\omega^{(t)}} \right) \\
&= \mathbb{I}^{(t+1)} \frac{t}{t+3} (\theta_{s,a'}^{(t+1)} - \theta_{s,a'}^{(t)}) \\
&= \omega_{s,a'}^{(t+1)} - \theta_{s,a'}^{(t+1)},
\end{aligned} \tag{140}$$

proving Claim 2, where (140) follows from (137) and (138).

With Claim 2, we are now ready to prove that with probability one, there exists a finite $T_\omega \geq T_1$ such that

$$\omega_{s,a^*(s)}^{(t)} - \theta_{s,a^*(s)}^{(t)} \geq \omega_{s,a}^{(t)} - \theta_{s,a}^{(t)}, \quad \forall s \in \mathcal{S}, \quad \forall a \in \mathcal{A}, \quad \forall t \geq T_\omega. \tag{141}$$

Suppose for contradiction that there is no such a T_ω. Then, as $\mathcal{S}$ and $\mathcal{A}$ are finite sets, there is an infinite sequence $\{t_i\}$ and a state-action pair $(s, \tilde{a})$ such that

$$\omega_{s,a^*(s)}^{(t_i)} - \theta_{s,a^*(s)}^{(t_i)} < \omega_{s,\tilde{a}}^{(t_i)} - \theta_{s,\tilde{a}}^{(t_i)}, \quad \forall i \geq 0. \tag{142}$$

If there is a $\hat{t} \geq T_1$ such that $\omega_{s,a^*(s)}^{(\hat{t})} - \theta_{s,a^*(s)}^{(\hat{t})} \geq \omega_{s,\tilde{a}}^{(\hat{t})} - \theta_{s,\tilde{a}}^{(\hat{t})}$, Claim 2 implies $\omega_{s,a^*(s)}^{(t)} - \theta_{s,a^*(s)}^{(t)} \geq \omega_{s,\tilde{a}}^{(t)} - \theta_{s,\tilde{a}}^{(t)}$ for every $t \geq \hat{t}$, which contradicts (142). Therefore,

$$\omega_{s,a^*(s)}^{(t)} - \theta_{s,a^*(s)}^{(t)} < \omega_{s,\tilde{a}}^{(t)} - \theta_{s,\tilde{a}}^{(t)}, \quad \forall t \geq T_1. \tag{143}$$

The strict inequality in (143) implies that

$$\mathbb{I}^{(t)} \neq 0, \quad \forall t \geq T_1, \tag{144}$$

for it otherwise leads to $\omega^{(t)} - \theta^{(t)} = 0$, and thus $\omega_{s,a^*(s)}^{(t)} - \theta_{s,a^*(s)}^{(t)} = \omega_{s,\tilde{a}}^{(t)} - \theta_{s,\tilde{a}}^{(t)}$ by (8).

For any $N > T_1$, we have

$$\begin{aligned}
\theta_{s,a^*(s)}^{(N)} &= \theta_{s,a^*(s)}^{(T_1(s))} + \sum_{t=T_1}^{N-1} (\theta_{s,a^*}^{(t+1)} - \theta_{s,a^*}^{(t)}) \\
&= \theta_{s,a^*}^{(T_1)} + \sum_{t=T_1}^{N-1} \frac{t+3}{t} (\omega_{s,a^*}^{(t+1)} - \theta_{s,a^*}^{(t+1)}) \\
&< \theta_{s,a^*}^{(T_1)} + \sum_{t=T_1}^{N-1} \frac{t+3}{t} (\omega_{s,\tilde{a}}^{(t+1)} - \theta_{s,\tilde{a}}^{(t+1)}) \\
&= \theta_{s,a^*}^{(T_1)} + \sum_{t=T_1}^{N-1} (\theta_{s,\tilde{a}}^{(t+1)} - \theta_{s,\tilde{a}}^{(t)}) \\
&= \theta_{s,a^*}^{(T_1)} - \theta_{s,\tilde{a}}^{(T_1)} + \theta_{s,\tilde{a}}^{(N)},
\end{aligned}$$

(145)
(146)
(147)
(148)

where (145) and (147) used (7), (8) and (144), and (146) is from (143). Letting $N \to \infty$ in (148), we see a direct contradiction with (133). Hence, there is a $T_\omega \geq T_1$ such that (141) holds.

Finally, the proof is completed by letting $T = \max\{T_M, T_V, T_\omega\}$. $\qquad \square$

D.2. Convergence Rate of APG

We need thew following lemma to faciliate the proof of Theorem 2.

Lemma 37. *Given any state* $s \in \mathcal{S}$ *and a corresponding vector* $\boldsymbol{d_{s,\cdot}} := [d_{s,a^*(s)}, d_{s,a_2(s)}, \cdots, d_{s,a_{|\mathcal{A}|}(s)}]$ *satisfying (i)* $d_{s,a^*(s)} + \sum_{i=2}^{|\mathcal{A}|} d_{s,a_i(s)} = 0$, *and (ii)* $\|\boldsymbol{d_{s,\cdot}}\|_2^2 = 1$, *we have*

- *If* $d_{s,a^*(s)} > 0 \geq \max_{i \geq 2} d_{s,a_i(s)}$, *then* $\min_{i \geq 2}\{d_{s,a^*(s)} - d_{s,a_i(s)}\} > |\mathcal{A}|^{-\frac{3}{2}}$.

- *If* $d_{s,a^*(s)} > (1 + \varepsilon_r) \max_{i \geq 2} d_{s,a_i(s)} > 0$, *then* $\min_{i \geq 2}\{d_{s,a^*(s)} - d_{s,a_i(s)}\} > \frac{\varepsilon_r}{1+\varepsilon_r}|\mathcal{A}|^{-\frac{3}{2}}$.

Proof of Lemma 37. We fix some state $s \in \mathcal{S}$. Since $d_{s,a^*(s)} + \sum_{i=2}^{|\mathcal{A}|} d_{s,a_i(s)} = 0$, we have

$$|d_{s,a}| \leq \sum_{a':d_{a'}>0} d_{s,a'}, \quad \forall a \in \mathcal{A}. \tag{149}$$

By taking square on both side of (149), we obtain from $\|\boldsymbol{d_{s,\cdot}}\|_2^2 = 1$ that

$$1 = \sum_{a \in \mathcal{A}} d_{s,a}^2 \leq \sum_{a \in \mathcal{A}} \Big(\sum_{a':d_{a'}>0} d_{s,a'}\Big)^2 = |\mathcal{A}|\Big(\sum_{a':d_{a'}>0} d_{s,a'}\Big)^2. \tag{150}$$

Using the assumption of $d_{s,a^*(s)} > \max_{i \geq 2} d_{s,a_i(s)}$ and taking the square root of (150), we get

$$d_{s,a^*(s)} > \frac{1}{|\mathcal{A}|}\Big(\sum_{a':d_{a'}>0} d_{s,a'}\Big) \geq |A|^{-\frac{3}{2}}. \tag{151}$$

Thus, if $d_{s,a^*(s)} > 0 \geq \max_{i \geq 2} d_{s,a_i(s)}$, by (151), we have

$$\min_{i \geq 2}\{d_{s,a^*(s)} - d_{s,a_i(s)}\} \geq d_{s,a^*(s)} > |\mathcal{A}|^{-\frac{3}{2}}.$$

On the other hand, if $d_{s,a^*(s)} > (1 + \varepsilon_r) \max_{i \geq 2} d_{s,a_i(s)} > 0$, again by (151), we have

$$\min_{i \geq 2}\{d_{s,a^*(s)} - d_{s,a_i(s)}\} > d_{s,a^*(s)} - \frac{d_{s,a^*(s)}}{(1 + \varepsilon_r)} \geq \frac{\varepsilon_r}{1 + \varepsilon_r}|\mathcal{A}|^{-\frac{3}{2}}.$$

$\square$

With all the tools prepared, now we restate and then prove Theorem 2.

Theorem 2 (Convergence Rate of APG; Formal). Consider reinforcement learning with tabular softmax parameterized policies π_θ. Under APG with $\eta^{(t)} = \frac{t}{t+1} \cdot \frac{(1-\gamma)^3}{16}$ and μ initialized uniformly at random, the following holds almost surely: There exists a finite time T such that for all $t \geq T$, we have

$$V^*(\rho) - V^{\pi_\theta^{(t)}}(\rho) \leq \frac{1}{(1-\gamma)^3}\left\|\frac{d_\rho^{\pi^*}}{\mu}\right\|_\infty \left(\frac{2|\mathcal{S}|(|\mathcal{A}|-1)}{t^2 + |\mathcal{A}| - 1} + \frac{512|\mathcal{S}|\ln^2(t) + 32\left\|2\theta^{(T)} - (2+T)(\omega^{(T)} - \theta^{(T)})\right\|^2}{(1-\gamma)(t+1)t}\right).$$

Proof of Theorem 2. We will first show that there exists a finite time T such that for all $t \geq T$, the restart mechanism is inactive and the objective function $V^{\pi_\theta}(\mu)$ will be $\frac{3}{2}$-nearly concave at each iterate along the APG update directions. We will then leverage Lemma 3 and Corollary 1 to characterize the convergence rate.

Let

$$M' := \ln\left[\frac{2|\mathcal{S}||\mathcal{A}|^2}{(1-\gamma)^2 \min_{s \in S} \mu(s)}\right].$$

By Lemma 2, there exists a finite time T' such that for all $t \geq T'$, $s \in \mathcal{S}$, and $a \neq a^*(s)$, we have (i) $\theta_{s,a^*(s)} - \theta_{s,a} > M'$, (ii) $V^{\pi_\theta^{(t)}}(s) > Q^*(s, a_2(s))$, (iii) $\left.\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a^*(s)}}\right|_{\theta=\omega^{(t)}} > 0 > \left.\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\right|_{\theta=\omega^{(t)}}$, (iv) $\omega_{s,a^*(s)}^{(t)} - \theta_{s,a^*(s)}^{(t)} \geq \omega_{s,a}^{(t)} - \theta_{s,a}^{(t)}$.

We first show that the restart mechanism is inactive for all $t > T'$. The momentum term at time t is

$$\frac{t-1}{t+2}(\theta^{(t)} - \theta^{(t-1)}) = \frac{t-1}{t+2}\left(\omega^{(t-1)} - \theta^{(t-1)} + \eta^{(t)}\nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\omega^{(t-1)}}\right). \tag{152}$$

We can observe that (152) lies in the feasible update domain $\mathcal{U}$ for any $t > T'$ by the properties (iii) and (iv) above. Hence, by combining Lemmas 4, 7 and 21, we know that $V^{\pi_{\varphi^{(t)}}}(\mu) \geq V^{\pi_{\theta^{(t)}}}(\mu)$, meaning that there is no restart at $t > T'$. Additionally, by (iii), we know that the gradient updates also lie in $\mathcal{U}$.

To show that the near concavity of the objective, we separately consider two cases for each state $s \in \mathcal{S}$: (I) $\max_{a \neq a^*(s)}\left\{\omega_{s,a}^{(T')} - \theta_{s,a}^{(T')} + \eta^{(T'+1)}\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\Big|_{\theta=\omega^{(T')}}\right\} \leq 0$; (II) $\max_{a \neq a^*(s)}\left\{\omega_{s,a}^{(T')} - \theta_{s,a}^{(T')} + \eta^{(T'+1)}\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\Big|_{\theta=\omega^{(T')}}\right\} > 0$. For case (I), we can see that for any $t \geq T'$, $\max_{a \neq a^*(s)}\left\{\omega_{s,a}^{(t)} - \theta_{s,a}^{(t)} + \eta^{(t+1)}\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\Big|_{\theta=\omega^{(t)}}\right\} \leq 0$ by the APG updates and (iii). Let

$$M_{(i)} := \max\left\{2\ln\left[2(|\mathcal{A}|-1)\right] - 2\ln\left[|\mathcal{A}|^{-\frac{3}{2}}\right], M'\right\}. \tag{153}$$

From Lemma 2, we know that there is a finite time $T_s \geq T'$ such that (i)-(iv) above with M' replaced by $M_{(i)}$ remain true for all $t > T_s$. Let $d_{s,\cdot}^{(t)} := \frac{\theta_{s,\cdot}^{(t)} - \theta_{s,\cdot}^{(t-1)}}{\|\theta_{s,\cdot}^{(t)} - \theta_{s,\cdot}^{(t-1)}\|_2}$ be the scaled APG update direction at time t such that $\|d_{s,\cdot}^{(t)}\|_2 = 1$. By Lemma 24, we have $\sum_{a \in \mathcal{A}} d_{s,a}^{(t)} = 0$ for any t and $s \in \mathcal{S}$. We also know that in this case, $\max_{a \neq a^*(s)} d_{s,a}^{(t)} \leq 0$ for all $t > T_s$. Lemma 37 then leads to $\min_{a \neq a^*(s)}\{d_{s,a^*(s)}^{(t)} - d_{s,a}^{(t)}\} > |\mathcal{A}|^{-\frac{3}{2}}$ for any $t > T_s$. By noting that M' is exactly the second term in the maximum in (126) with $C = 3/2$, we observe that the setting of $M_{(i)}$ in (153) satisfies all the conditions in Lemma 1, so the objective function $\theta \to V^{\pi_\theta}(\mu)$ is $\frac{3}{2}$-nearly concave along our APG update directions for all $t > T_s$ for the state s in the case (I).

Regarding case (II), let

$$\varepsilon_{T'} := \frac{\omega_{s,a^*(s)}^{(T')} - \theta_{s,a^*(s)}^{(T')} + \eta^{(t'+1)}\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a^*(s)}}\Big|_{\theta=\omega^{(T')}}}{\max_{a \neq a^*(s)}\left\{\omega_{s,a}^{(T')} - \theta_{s,a}^{(T')} + \eta^{(t'+1)}\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\Big|_{\theta=\omega^{(T')}}\right\}} - 1.$$

Note that $\varepsilon_{T'} > 0$ due to (iii) and (iv) and the hypothesis of this case. Let

$$M_{(ii)} := \max\left\{2\ln\left[2(|\mathcal{A}|-1)\right] - 2\ln\left[\frac{\varepsilon_{T'}}{1+\varepsilon_{T'}}|\mathcal{A}|^{-\frac{3}{2}}\right], M'\right\}. \tag{154}$$

Again by Lemma 2, we know that there is a finite time $T_s' \geq T'$ such that (i)-(iv) above with M' replaced by $M_{(ii)}$ remain true for all $t > T_s'$. Similarly, we demonstrate that the objective function $\theta \to V^{\pi_\theta}(\mu)$ is $\frac{3}{2}$-nearly concave along our APG update directions for all $t > T_s'$ for the state s in this case. Following the argument for (I), we only need to show that $\frac{\varepsilon_{T'}}{1+\varepsilon_{T'}}|\mathcal{A}|^{-\frac{3}{2}}$ is a lower bound for $\min_{a \neq a^*(s)}\{d_{s,a^*(s)}^{(t)} - d_{s,a}^{(t)}\}$ for every $t > T_s'$. We use the same definition of $d_{s,\cdot}^{(t)}$ as above. Then, for any $k \in \mathbb{N}_+$ and $s \in \mathcal{S}$ satisfying $\max_{a \neq a^*(s)} d_{s,a}^{(T'+k)} \leq 0$, from the argument for (I), $|\mathcal{A}|^{-\frac{3}{2}}$ is a lower bound for $\min_{a \neq a^*(s)}\{d_{s,a^*(s)}^{(t)} - d_{s,a}^{(t)}\}$. Moreover, since $\frac{\varepsilon_{T'}}{1+\varepsilon_{T'}} < 1$, $\frac{\varepsilon_{T'}}{1+\varepsilon_{T'}}|\mathcal{A}|^{-\frac{3}{2}}$ is also a valid lower bound for it. On the other hand, for any $k \in \mathbb{N}_+$ and $s \in \mathcal{S}$ satisfying $\max_{a \neq a^*(s)} d_{s,a}^{(T'+k)} > 0$, by the APG updates, (iii) and (iv) abvove, and that there is no restart after time T' (as proven above), we have

$$\frac{d_{s,a^*(s)}^{(T'+k)}}{\max_{a \neq a^*(s)}\{d_{s,a}^{(T'+k)}\}}$$
$$= \frac{\theta_{s,a^*(s)}^{(T'+k)} - \theta_{s,a^*(s)}^{(T'+k-1)}}{\max_{a \neq a^*(s)}\{\theta_{s,a}^{(T'+k)} - \theta_{s,a}^{(T'+k-1)}\}}$$

$$= \frac{\left(\prod_{j=1}^{k-1}\frac{T'+k-1-j}{T'+k+2-j}\right)(\omega_{s,a^*(s)}^{(T')}-\theta_{s,a^*(s)}^{(T')})+\sum_{t=1}^{k}\left(\left(\prod_{j=1}^{t-1}\frac{T'+k-1-j}{T'+k+2-j}\right)\eta^{(T'+k-t+1)}\overbrace{\frac{\partial V^{\pi_\theta}(\mu)}{\partial\theta_{s,a^*(s)}}\Big|_{\theta=\omega^{(T'+k-t)}}}^{>0}\right)}{\max_{a\neq a^*(s)}\left\{\left(\prod_{j=1}^{k-1}\frac{T'+k-1-j}{T'+k+2-j}\right)(\omega_{s,a}^{(T')}-\theta_{s,a}^{(T')})+\sum_{t=1}^{k}\left(\left(\prod_{j=1}^{t-1}\frac{T'+k-1-j}{T'+k+2-j}\right)\eta^{(T'+k-t+1)}\underbrace{\frac{\partial V^{\pi_\theta}(\mu)}{\partial\theta_{s,a}}\Big|_{\theta=\omega^{(T'+k-t)}}}_{<0}\right)\right\}}$$

$$> \frac{\left(\prod_{j=1}^{k-1}\frac{T'+k-1-j}{T'+k+2-j}\right)\left(\omega_{s,a^*(s)}^{(T')}-\theta_{s,a^*(s)}^{(T')}+\eta^{(T'+1)}\frac{\partial V^{\pi_\theta}(\mu)}{\partial\theta_{s,a^*(s)}}\Big|_{\theta=\omega^{(T')}}\right)}{\max_{a\neq a^*(s)}\left\{\left(\prod_{j=1}^{k-1}\frac{T'+k-1-j}{T'+k+2-j}\right)\left(\omega_{s,a}^{(T')}-\theta_{s,a}^{(T')}+\eta^{(T'+1)}\frac{\partial V^{\pi_\theta}(\mu)}{\partial\theta_{s,a}}\Big|_{\theta=\omega^{(T')}}\right)\right\}}$$

$$= \frac{\omega_{s,a^*(s)}^{(T')}-\theta_{s,a^*(s)}^{(T')}+\eta^{(T'+1)}\frac{\partial V^{\pi_\theta}(\mu)}{\partial\theta_{s,a^*(s)}}\Big|_{\theta=\omega^{(T')}}}{\max_{a\neq a^*(s)}\left\{\omega_{s,a}^{(T')}-\theta_{s,a}^{(T')}+\eta^{(T'+1)}\frac{\partial V^{\pi_\theta}(\mu)}{\partial\theta_{s,a}}\Big|_{\theta=\omega^{(T')}}\right\}}$$

$$= 1 + \varepsilon_{T'}. \tag{155}$$

Therefore, by Lemma 37 and (155), we obtain $\min_{a\neq a^*(s)}\{d_{s,a^*(s)}^{(t)}-d_{s,a}^{(t)}\}>\frac{\varepsilon_r}{1+\varepsilon_r}|\mathcal{A}|^{-\frac{3}{2}}$ for any $t>T_s'$. Finally, let $T:=\max_{s\in\mathcal{S}}\max\{T_s,T_s'\}+1$, we obtain a finite time T (since $|\mathcal{S}|$ is finite) such that the objective function $V^{\pi_\theta}(\mu)$ is indeed $\frac{3}{2}$-nearly concave along the APG update directions and the restart mechanism is inactive for all $t\geq T$.

As the objective function is $\frac{3}{2}$-nearly concave along the APG update directions and there will be no restart mechanism for all $t\geq T$, we now apply Lemma 3 to establish the asymptotic convergence rate. It is necessary to account for a shift in the initial step size due to the passage of time T. Specifically, if we divide the update process into two phases based on time T, the latter phase will commence with a modified step size. Specifically, the latter phase could be considered as Algorithm 2 with $\theta^{(0)}=\theta^{(T)}$, $\omega^{(0)}=\omega^{(T)}$ and $\eta^{(0)}=\eta^{(T+1)}$. We define the surrogate solution $\theta^{**(t)}$ at time t such that $\theta_{s,\cdot}^{**(t)}:=[\theta_{s,a^*(s)}^{**(t)},\theta_{s,a_2(s)}^{**(t)},\cdots,\theta_{s,a_{|\mathcal{A}|}(s)}^{**(t)}]$ satisfies $\theta_{s,\cdot}^{**(t)}=[2\ln(t),0,0,\cdots,0]$ for all $s\in\mathcal{S}$.

By Lemmas 3 and 14 and Corollary 1 with $T_{\text{shift}}=T$ and $\theta^{**}=\theta^{**(t)}$, we have that

$$V^{\pi_{\theta^{**}}^{(t+T)}}(\mu)-V^{\pi_\theta^{(t+T)}}(\mu) \leq \frac{16\left\|2\theta^{**(t+T)}-2\theta^{(T)}-(2+T)(\omega^{(T)}-\theta^{(T)})\right\|^2}{(1-\gamma)^3(t+T+1)(t+T)}$$

$$\leq \frac{128\left\|\theta^{**(t+T)}\right\|^2+32\left\|2\theta^{(T)}-(2+T)(\omega^{(T)}-\theta^{(T)})\right\|^2}{(1-\gamma)^3(t+T+1)(t+T)}$$

$$= \frac{512|\mathcal{S}|\ln^2(t+T)+32\left\|2\theta^{(T)}-(2+T)(\omega^{(T)}-\theta^{(T)})\right\|^2}{(1-\gamma)^3(t+T+1)(t+T)}. \tag{156}$$

Additionally, the sub-optimality gap between the optimal V^* and the surrogate optimal solution can be bounded by

$$V^*(\mu)-V^{\pi_{\theta^{**}}^{(t)}}(\mu) = \frac{1}{1-\gamma}\sum_s d_\mu^{\pi^*}(s)\sum_a(\pi^*(a|s)-\pi_{\theta^{**}}^{(t)}(a|s))\cdot A_{\theta^{**}}^{\pi^{(t)}}(s,a) \tag{157}$$

$$\leq \frac{1}{1-\gamma}\sum_s \underbrace{d_\mu^{\pi^*}(s)}_{\leq 1}\sum_a|\pi^*(a|s)-\pi_{\theta^{**}}^{(t)}(a|s)|\cdot \underbrace{\left|A_{\theta^{**}}^{\pi^{(t)}}(s,a)\right|}_{\leq\frac{1}{1-\gamma}\text{ by Lemma 10}}$$

$$\leq \frac{1}{(1-\gamma)^2}\sum_s\sum_a|\pi^*(a|s)-\pi_{\theta^{**}}^{(t)}(a|s)|$$

$$= \frac{1}{(1-\gamma)^2}\sum_s\left(|\pi^*(a^*(s)|s)-\pi_{\theta^{**}}^{(t)}(a^*(s)|s)|+\sum_{a\neq a^*(s)}|\pi^*(a|s)-\pi_{\theta^{**}}^{(t)}(a|s)|\right)$$

$$= \frac{1}{(1-\gamma)^2}\sum_s 2\left(1-\pi_{\theta^{**}}^{(t)}(a^*(s)|s)\right) \tag{158}$$

$$\leq \frac{2|\mathcal{S}|}{(1-\gamma)^2}\left(1-\frac{\exp(2\ln(t))}{\exp(2\ln(t))+\exp(0)\cdot(|\mathcal{A}|-1)}\right) \tag{159}$$

$$= \frac{2|\mathcal{S}|}{(1-\gamma)^2} \left(\frac{|\mathcal{A}|-1}{t^2+|\mathcal{A}|-1} \right), \tag{160}$$

where (157) holds by Lemma 4 and Lemma 7, (158) holds due to $\pi^*(a^*(s)|s) = 1$ and $\pi^*(a|s) = 0$ for all $a \neq a^*(s)$.

Finally, by Lemma 6, (156) and (160), we can then change the convergence rate bound for $V^\pi(\mu)$ to for $V^\pi(\rho)$ and get the convergence rate:

$$
V^*(\rho) - V^{\pi_\theta^{(t)}}(\rho) \leq \frac{1}{1-\gamma} \cdot \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty \cdot \left(V^*(\mu) - V^{\pi_\theta^{(t)}}(\mu) \right)
$$

$$
= \frac{1}{1-\gamma} \cdot \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty \cdot \left(\left(V^{\pi_\theta^{**(t)}}(\mu) - V^{\pi_\theta^{(t)}}(\mu) \right) + \left(V^*(\mu) - V^{\pi_\theta^{**(t)}}(\mu) \right) \right)
$$

$$
\leq \frac{1}{1-\gamma} \cdot \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty
$$

$$
\cdot \left(\frac{2|\mathcal{S}|}{(1-\gamma)^2} \frac{|\mathcal{A}|-1}{t^2+|\mathcal{A}|-1} \right.
$$

$$
\left. + \frac{512|\mathcal{S}|\ln^2(t) + 32 \left\| 2\theta^{(T)} - (2+T)\left(\omega^{(T)} - \theta^{(T)} \right) \right\|^2}{(1-\gamma)^3(t+1)t} \right)
$$

$$
= \frac{1}{(1-\gamma)^3} \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty \left(\frac{2|\mathcal{S}|(|\mathcal{A}|-1)}{t^2+|\mathcal{A}|-1} + \frac{512|\mathcal{S}|\ln^2(t) + 32 \left\| 2\theta^{(T)} - (2+T)\left(\omega^{(T)} - \theta^{(T)} \right) \right\|^2}{(1-\gamma)(t+1)t} \right)
$$

for all $t \geq T$, and the proof is complete. $\qquad\square$

E. APG With Time-Varying Step Sizes

E.1. Asymptotic Convergence of APG With Time-Varying Step Sizes

Lemma 38. *Given any $t \in \mathbb{N}_+$, let $\eta^{(t)} = \beta^t \cdot \frac{(1-\gamma)^3}{8}$ for any $\exp(\frac{1}{8\sqrt{|\mathcal{A}||\mathcal{S}|}} \cdot \frac{1-\gamma}{4C_\infty-(1-\gamma)}) > \beta > 1$ and $K =$*
$\frac{1}{4\sqrt{|\mathcal{A}||\mathcal{S}|}} \cdot \frac{1-\gamma}{4C_\infty-(1-\gamma)} + 2\ln\beta$ *where $C_\infty := \max_\pi \|\frac{d_\mu^\pi}{\mu}\|_\infty$. Consider*

$$\theta \leftarrow \omega + \left[\eta^{(t)}\nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\omega}\right]^{+K}$$

and denote $\omega_\zeta := \omega + \zeta \cdot (\theta - \omega)$ with some $\zeta \in [0,1]$. Under softmax parameterization, we have,

$$\|\nabla_\theta V^{\pi_\theta}|_{\theta=\omega_\zeta}\|_2 \le 2\|\nabla_\theta V^{\pi_\theta}|_{\theta=\omega}\|_2.$$

Proof of Lemma 38. By the proofs of Lemmas 3 and 7 in (Mei et al., 2021b), we know that it suffices to replace the upper bound of $\|\theta' - \theta\|_2$ in (234) in (Mei et al., 2021b) with our updates. Additionally, the upper bound of the norm of our updates is $\sqrt{|\mathcal{S}||\mathcal{A}|}K$. Thus, we have the bound as follows:

$$\|\nabla_\theta V^{\pi_\theta}|_{\theta=\omega_\zeta}\|_2 \le \frac{1}{1 - \frac{4C_\infty-(1-\gamma)}{1-\gamma}\sqrt{|\mathcal{S}|}\sqrt{|\mathcal{S}||\mathcal{A}|}K}\|\nabla_\theta V^{\pi_\theta}|_{\theta=\omega}\|_2. \tag{161}$$

We elaborate on the bound of $\frac{1}{1-\frac{4C_\infty-(1-\gamma)}{1-\gamma}|\mathcal{S}|\sqrt{|\mathcal{A}|}K}$,

$$\frac{1}{1 - \frac{4C_\infty-(1-\gamma)}{1-\gamma}|\mathcal{S}|\sqrt{|\mathcal{A}|}K} = \frac{1}{1 - \frac{4C_\infty-(1-\gamma)}{1-\gamma}|\mathcal{S}|\sqrt{|\mathcal{A}|}(\frac{1}{4\sqrt{|\mathcal{A}||\mathcal{S}|}} \cdot \frac{1-\gamma}{4C_\infty-(1-\gamma)} + 2\ln\beta)}$$

$$= \frac{1}{1 - \frac{1}{4} - \frac{4C_\infty-(1-\gamma)}{1-\gamma}|\mathcal{S}|\sqrt{|\mathcal{A}|}2\ln\beta}$$

$$\le \frac{1}{1 - \frac{1}{4} - \frac{4C_\infty-(1-\gamma)}{1-\gamma}|\mathcal{S}|\sqrt{|\mathcal{A}|} \cdot 2 \cdot (\frac{1}{8\sqrt{|\mathcal{A}||\mathcal{S}|}} \cdot \frac{1-\gamma}{4C_\infty-(1-\gamma)})} \tag{162}$$

$$= \frac{1}{1 - \frac{1}{2}} = 2, \tag{163}$$

where (162) holds by plugging in the upper bound of β. Hence, we combine (161) and (163) and obtain the results. $\square$

Lemma 39. *Consider the setting of Theorem 3. Then, almost surely, the limits of both $V^{\pi_\omega^{(t)}}(\mu)$ and $V^{\pi_\theta^{(t)}}(\mu)$ when t approaches infinity exist, $\lim_{t\to\infty} V^{\pi_\omega^{(t)}}(\mu) = \lim_{t\to\infty} V^{\pi_\theta^{(t)}}(\mu)$, and*

$$\lim_{t\to\infty}\left\|\nabla_\theta V^{\pi_\theta}(s)|_{\theta=\omega^{(t)}}\right\| = 0, \quad \lim_{t\to\infty}\left\|\nabla_\theta V^{\pi_\theta}(s)|_{\theta=\theta^{(t)}}\right\| = 0. \tag{164}$$

Proof of Lemma 39. Following the argument in Lemma 25, we have $\lim_{t\to\infty} V^{\pi_\omega^{(t)}}(\mu) = \lim_{t\to\infty} V^{\pi_\theta^{(t)}}(\mu)$. We prove $\lim_{t\to\infty}\left\|\nabla_\theta V^{\pi_\theta}(s)|_{\theta=\omega^{(t)}}\right\| = 0$ by contradiction. Suppose that the limit of $\left\|\nabla_\theta V^{\pi_\theta}(s)|_{\theta=\omega^{(t)}}\right\|$ does not exist or is not 0. Then, we know that there exists a sequence $\{t_i\}_{i=1}^\infty$ and $\varepsilon_0 > 0$ such that for every t_i, there is a state-action pair (s_{t_i}, a_{t_i}) satisfying

$$\left.\frac{\partial V^{\pi_\theta}(\mu)}{\partial\theta_{s_{t_i},a_{t_i}}}\right|_{\theta=\omega^{(t_i)}} > \varepsilon_0. \tag{165}$$

Note that (165) is satisfied since the summation of the gradient over all actions at any given state s is zero. We define $\partial_{s,a}^{(t)} := \frac{\partial V^{\pi_\theta}(\mu)}{\partial\theta_{s,a}}|_{\theta=\omega^{(t)}}$ and $Z_\phi^{(t)}(s) := \sum_{a\in\mathcal{A}}\exp(\phi_{s,a}^{(t)})$ for any policy parameter $\phi^{(t)}$. Then, for any $t_i \in \{t_i\}_{i=1}^\infty$,

$$V^{\pi_\theta^{(t_i+1)}}(\mu) - V^{\pi_\omega^{(t_i)}}(\mu) = \frac{1}{1-\gamma}\sum_s d_\mu^{\pi_\theta^{(t_i+1)}}(s)\sum_{a\in\mathcal{A}}\pi_\theta^{(t_i+1)}(a|s)A^{\pi_\omega^{(t_i)}}(s,a) \tag{166}$$

51

$$\geq \frac{1}{1-\gamma} d_\mu^{\pi_\theta^{(t_i+1)}}(s_{t_i}) \sum_{a \in \mathcal{A}} \pi_\theta^{(t_i+1)}(a|s_{t_i}) A^{\pi_\omega^{(t_i)}}(s_{t_i}, a) \tag{167}$$

$$\geq \mu(s_{t_i}) \sum_{a \in \mathcal{A}} \pi_\theta^{(t_i+1)}(a|s_{t_i}) A^{\pi_\omega^{(t_i)}}(s_{t_i}, a) \tag{168}$$

$$= \mu(s_{t_i}) \frac{Z_\omega^{(t_i)}(s_{t_i})}{Z_\theta^{(t_i+1)}(s_{t_i})} \sum_{a \in \mathcal{A}} \frac{\exp(\omega_{s_{t_i},a}^{(t_i)}) + \mathrm{clip}\{\eta^{(t_i+1)} \partial_{s_{t_i},a}^{(t_i)}, -K, K\}}{Z_\omega^{(t_i)}(s_{t_i})} A^{\pi_\omega^{(t_i)}}(s_{t_i}, a)$$

$$\geq \mu(s_{t_i}) \frac{Z_\omega^{(t_i)}(s_{t_i})}{Z_\theta^{(t_i+1)}(s_{t_i})} \left[\left(\frac{\exp(\omega_{s_{t_i},a_{t_i}}^{(t_i)}) + \mathrm{clip}\{\eta^{(t_i+1)} \partial_{s_{t_i},a_{t_i}}^{(t_i)}, -K, K\}}{Z_\omega^{(t_i)}(s_{t_i})} \right) A^{\pi_\omega^{(t_i)}}(s_{t_i}, a_{t_i}) \right.$$

$$\left. + \sum_{a \neq a_{t_i}} \frac{\exp(\omega_{s_{t_i},a}^{(t_i)})}{Z_\omega^{(t_i)}(s_{t_i})} A^{\pi_\omega^{(t_i)}}(s_{t_i}, a) \right] \tag{169}$$

$$= \mu(s_{t_i}) \frac{Z_\omega^{(t_i)}(s_{t_i})}{Z_\theta^{(t_i+1)}(s_{t_i})} (\exp(\min\{\eta^{(t_i+1)} \partial_{s_{t_i},a_{t_i}}^{(t_i)}, K\}) - 1) \pi_\omega^{(t_i)}(a_{t_i}|s_{t_i}) A^{\pi_\omega^{(t_i)}}(s_{t_i}, a_{t_i}) \tag{170}$$

$$\geq \mu(s_{t_i}) \frac{\sum_{a \in \mathcal{A}} \exp(\omega_{s_{t_i},a}^{(t_i)})}{\sum_{a \in \mathcal{A}} \exp(\omega_{s_{t_i},a}^{(t_i)} + K)} (\exp(\min\{\eta^{(t_i+1)} \partial_{s_{t_i},a_{t_i}}^{(t_i)}, K\}) - 1) \frac{(1-\gamma)\varepsilon_0}{d_\mu^{\pi_\theta}(s_{t_i})} \tag{171}$$

$$\geq \mu(s_{t_i}) \exp(-K)(\exp(\min\{\eta^{(t_i+1)} \varepsilon_0, K\}) - 1) \frac{(1-\gamma)\varepsilon_0}{d_\mu^{\pi_\theta}(s_{t_i})} > 0, \tag{172}$$

where (166) holds by Lemma 4, (167) holds by Lemma 22, (168) holds by Lemma 5, (169) is due to the fact that the signs of the updates are the same as the advantage function values with respect to the corresponding actions (see Lemmas 4 and 11), (170) is true because of $\partial_{s_{t_i},a_{t_i}}^{(t_i)} > 0$ and Lemma 7, (171) holds due to that the maximum magnitude of the updates is K by clipping, (165), and Lemma 11. We can observe that (172) is a value bounded away from 0, which leads to a constant improvement at time t_i. This contradicts with $\lim_{t \to \infty} V^{\pi_\omega^{(t)}}(\mu) = \lim_{t \to \infty} V^{\pi_\theta^{(t)}}(\mu)$. Hence, we have

$$\lim_{t \to \infty} \left\| \nabla_\theta V^{\pi_\theta}(s)|_{\theta=\omega^{(t)}} \right\| = 0. \tag{173}$$

Finally, we show that $\lim_{t \to \infty} \left\| \nabla_\theta V^{\pi_\theta}(s)|_{\theta=\theta^{(t)}} \right\| = 0$. By the triangle inequality, we have

$$\left\| \nabla_\theta V^{\pi_\theta}(s)|_{\theta=\theta^{(t)}} \right\| \leq \left\| \nabla_\theta V^{\pi_\theta}(s)|_{\theta=\theta^{(t)}} - \nabla_\theta V^{\pi_\theta}(s)|_{\theta=\omega^{(t-1)}} \right\| + \left\| \nabla_\theta V^{\pi_\theta}(s)|_{\theta=\omega^{(t-1)}} \right\|. \tag{174}$$

By Definition 4 and lemmas 16, 17 and 38, we have

$$\left\| \nabla_\theta V^{\pi_\theta}(s)|_{\theta=\theta^{(t)}} - \nabla_\theta V^{\pi_\theta}(s)|_{\theta=\omega^{(t-1)}} \right\| \leq \left[3 + \frac{4 \cdot (C_\infty - (1-\gamma))}{1-\gamma} \sqrt{|\mathcal{S}|} \right] \cdot 2 \cdot \left\| \nabla_\theta V^{\pi_\theta}(s)|_{\theta=\omega^{(t-1)}} \right\| \cdot \left\| \theta^{(t)} - \omega^{(t-1)} \right\|. \tag{175}$$

By the clipped update, we know that $\left\| \theta^{(t)} - \omega^{(t-1)} \right\|$ is always upper bounded by $\sqrt{|\mathcal{S}||\mathcal{A}|} K$. Thus, (175) approaches 0 as $t \to \infty$ by (173), leading to that the RHS of (174) goes to 0 as $t \to \infty$. Hence, we have proven (164). $\square$

Theorem 5 (**Asymptotic Convergence Under Softmax Parameterization**). *Consider the setting of Theorem 3, the following holds almost surely:* $\lim_{t \to \infty} V^{\pi_\theta^{(t)}}(s) = V^*(s)$, *for all* $s \in \mathcal{S}$.

Proof of Theorem 5. We note that the proofs of Theorem 1 and Lemma 26-35 are all based on the results of Lemma 25. Now we have provided the counterpart of Lemma 25 as Lemma 39, the requirement for the step size upper bounded in these lemmas can be removed. By using arguments similar to those in the proof of Theorem 1 and Lemma 26-35, we have the results. $\square$

E.2. Convergence Rate of APG With Time-Varying Step Sizes

Lemma 40. *Consider a tabular softmax parameterized policy. Under Assumption 1 and the setting of Theorem 5, the following holds almost surely: Given any $M > 0$, there exists a finite time T such that for all $t \geq T$, $s \in \mathcal{S}$, and*

$a \neq a^*(s)$, we have (i) $\theta^{(t)}_{s,a^*(s)} - \theta^{(t)}_{s,a} > M$, (ii) $V^{\pi^{(t)}_\theta}(s) > Q^*(s, a_2(s))$, (iii) $\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a^*(s)}}\Big|_{\theta=\omega^{(t)}} > 0 > \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\Big|_{\theta=\omega^{(t)}}$, (iv) $\omega^{(t)}_{s,a^*(s)} - \theta^{(t)}_{s,a^*(s)} \geq \omega^{(t)}_{s,a} - \theta^{(t)}_{s,a}$.

Proof of Lemma 40. Applying Theorem 5 and the logic in its proof to the argument of the proof of Lemma 2, we can obtain the stated results. $\qquad\square$

Theorem 3 (**Convergence Rate of APG with Time-Varying Step Sizes**). *Consider a tabular softmax parameterized policy* π_θ. *Under APG with* (6) *modified as*

$$\theta^{(t)} \leftarrow \omega^{(t-1)} + \left[\eta^{(t)}\nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\omega^{(t-1)}}\right]^{+K},$$

where $\eta^{(t)} = \beta^t \cdot \frac{(1-\gamma)^3}{8}$ *for any* $\exp(\frac{1}{8\sqrt{|\mathcal{A}||\mathcal{S}|}} \cdot \frac{1-\gamma}{4C_\infty - (1-\gamma)}) > \beta > 1$, μ *initialized uniformly at random, and* $K = \frac{1}{4\sqrt{|\mathcal{A}||\mathcal{S}|}} \cdot \frac{1-\gamma}{4C_\infty-(1-\gamma)} + 2\ln\beta$ *where* $C_\infty := \max_\pi \|\frac{d^\pi_\mu}{\mu}\|_\infty$, *the following holds almost surely: There exists a finite time* T *such that for all* $t \geq T$, *we have*

$$V^*(\rho) - V^{\pi^{(t)}_\theta}(\rho) = O(e^{-ct}),$$

for some constant $c > 0$.

Proof of Theorem 3. By Lemma 4, we have

$$V^*(\mu) - V^{\pi^{(t)}_\theta}(\mu) = \frac{1}{1-\gamma}\sum_{s,a} d^{\pi^*}_\mu(s)\pi^*(a|s)A^{\pi^{(t)}_\theta}(s,a)$$

$$= \frac{1}{1-\gamma}\sum_{s,a} d^{\pi^*}_\mu(s)(\pi^*(a|s) - \pi^{(t)}_\theta(a|s))A^{\pi^{(t)}_\theta}(s,a)$$

$$\leq \frac{1}{1-\gamma}\sum_{s,a} \left| d^{\pi^*}_\mu(s)(\pi^*(a|s) - \pi^{(t)}_\theta(a|s))A^{\pi^{(t)}_\theta}(s,a)\right|$$

$$\leq \frac{2}{(1-\gamma)^2} \left\|d^{\pi^*}_\mu\right\|_\infty \sum_{s,a} \left|\pi^*(a|s) - \pi^{(t)}_\theta(a|s)\right|$$

$$= \frac{2}{(1-\gamma)^2} \left\|d^{\pi^*}_\mu\right\|_\infty \sum_s \left(\left|\pi^*(a^*|s) - \pi^{(t)}_\theta(a^*|s)\right| + \sum_{a\neq a^*}\left|\pi^*(a|s) - \pi^{(t)}_\theta(a|s)\right|\right)$$

$$= \frac{2}{(1-\gamma)^2} \left\|d^{\pi^*}_\mu\right\|_\infty \sum_s 2(1 - \pi^{(t)}_\theta(a^*|s)).$$

To obtain the desired outcome, it is sufficient to show that there exist $\bar{c}, \tilde{c} > 0$ such that

$$1 - \pi^{(t)}_\theta(a^*(s)|s) \leq \tilde{c}e^{-\bar{c}t} \tag{176}$$

for all $s \in \mathcal{S}$ and all $t \geq T$. Given

$$M := \ln \frac{|\mathcal{A}| - 1}{\exp\left(\frac{1}{4\sqrt{|\mathcal{A}||\mathcal{S}|}} \cdot \frac{1-\gamma}{4C_\infty-(1-\gamma)}\right) - 1} > 0,$$

where $C_\infty := \max_\pi \|\frac{d^\pi_\mu}{\mu}\|_\infty$. Selecting T_M from Lemma 40 such that $\theta^{(t)}_{s,a^*(s)} - \theta^{(t)}_{s,a} > M$ for all $t \geq T_M, s \in \mathcal{S}$, and $a \in \mathcal{A}$, for contradiction we assume that after time-step T_M, for any $\bar{c}, \tilde{c} > 0$ there exists some state s' and an infinite time sequence $\{t_i\}_{i=1}^\infty$ such that

$$1 - \pi^{(t)}_\theta(a^*(s')|s') > \tilde{c}e^{-\bar{c}t}, \quad \forall t \in \{t_i\}_{i=1}^\infty. \tag{177}$$

Define

$$C := \inf_{s' \in \mathcal{S}, t \geq T} \left\{ \frac{(1-\gamma)^2}{8} \cdot d_\mu^{\pi_\omega^{(t)}}(s') \cdot \pi_\omega^{(t)}(a^*(s')|s') \cdot (Q^{\pi_\theta^{(t)}}(s', a^*(s')) - \max_{a \neq a^*(s')} Q^{\pi_\theta^{(t)}}(s', a^*(s'))) \right\},$$

then we have $C > 0$ since $d_\mu^{\pi_\omega^{(t)}}(s') \geq (1-\gamma)\mu(s')$ by Lemma 5, $\pi_\omega^{(t)}(a^*(s')|s') > \frac{1}{|\mathcal{A}|}$ since $\theta_{s', a^*(s')} > \theta_{s', a}$ for all $a \neq a^*(s')$ by property (i) of Lemma 40, and $Q^{\pi_\theta^{(t)}}(s', a^*(s')) \geq V^{\pi_\theta^{(t)}}(s') > Q^*(s', a^*(s'))) > \max_{a \neq a^*(s')} Q^{\pi_\theta^{(t)}}(s', a^*(s'))$ by properties (ii) of Lemma 40 and Lemma 8.

By choosing $\tilde{c} = \frac{K}{C}$ and $\bar{c} = \beta$ and using the corresponding s' and $\{t_i\}$, we have:

$$\begin{aligned}
A^{\pi_\theta^{(t)}}(s', a^*(s')) &= Q^{\pi_\theta^{(t)}}(s', a^*(s')) - V^{\pi_\theta^{(t)}}(s') \\
&= Q^{\pi_\theta^{(t)}}(s', a^*(s')) - \sum_{a \in \mathcal{A}} \pi_\theta^{(t)}(a(s')|s') Q^{\pi_\theta^{(t)}}(s', a) \\
&= (1 - \pi_\theta^{(t)}(a^*(s')|s')) Q^{\pi_\theta^{(t)}}(s', a^*(s')) - \sum_{a \neq a^*(s')} \pi_\theta^{(t)}(a|s') Q^{\pi_\theta^{(t)}}(s', a) \\
&\geq (1 - \pi_\theta^{(t)}(a^*(s')|s')) Q^{\pi_\theta^{(t)}}(s', a^*(s')) - \sum_{a \neq a^*(s')} \pi_\theta^{(t)}(a|s') \cdot \max_{a \neq a^*} Q^{\pi_\theta^{(t)}}(s', a) \\
&= (1 - \pi_\theta^{(t)}(a^*(s')|s')) Q^{\pi_\theta^{(t)}}(s', a^*(s')) - (1 - \pi_\theta^{(t)}(a^*|s')) \cdot \max_{a \neq a^*(s')} Q^{\pi_\theta^{(t)}}(s', a) \\
&= (1 - \pi_\theta^{(t)}(a^*(s')|s')) \underbrace{(Q^{\pi_\theta^{(t)}}(s', a^*(s')) - \max_{a \neq a^*(s')} Q^{\pi_\theta^{(t)}}(s', a))}_{\text{bounded away from 0}} \\
&> \frac{K}{C\beta^t} \cdot (Q^{\pi_\theta^{(t)}}(s', a^*(s')) - \max_{a \neq a^*(s')} Q^{\pi_\theta^{(t)}}(s', a)). \tag{178}
\end{aligned}$$

By following (178), for all $t \in \{t_i\}_{i=1}^\infty$, we have

$$\begin{aligned}
&\theta_{s', a^*(s')}^{(t+1)} - \theta_{s', a^*(s')}^{(t)} \\
&= \underbrace{(\theta_{s', a^*(s')}^{(t+1)} - \omega_{s', a^*(s')}^{(t)})}_{\text{gradient}} + \underbrace{(\omega_{s', a^*(s')}^{(t)} - \theta_{s', a^*(s')}^{(t)})}_{\text{momentum}} \\
&= \min \left\{ K, \eta^{(t)} \nabla_\theta V^{\pi_\theta}(\mu) \Big|_{\theta = \omega^{(t)}} \right\} + (\omega_{s', a^*(s')}^{(t)} - \theta_{s', a^*(s')}^{(t)}) \\
&= \min \left\{ K, \eta^{(t)} \cdot \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\omega^{(t)}}(s') \cdot \pi_\omega^{(t)}(a^*|s') \cdot A^{\pi_\omega^{(t)}}(s', a^*) \right\} + (\omega_{s', a^*(s')}^{(t)} - \theta_{s', a^*(s')}^{(t)}) \\
&= K + (\omega_{s', a^*(s')}^{(t)} - \theta_{s', a^*(s')}^{(t)}) \tag{179} \\
&\geq K + \max_{a \neq a^*(s')} \left\{ \omega_{s', a}^{(t)} - \theta_{s', a}^{(t)} \right\}, \tag{180}
\end{aligned}$$

where (179) holds by

$$\begin{aligned}
&\eta^{(t)} \cdot \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\omega^{(t)}}(s') \cdot \pi_\omega^{(t)}(a^*|s') \cdot A^{\pi_\omega^{(t)}}(s', a^*) \\
&> \beta^t \cdot \frac{(1-\gamma)^2}{8} \cdot d_\mu^{\pi_\omega^{(t)}}(s') \cdot \pi_\omega^{(t)}(a^*|s') \cdot \frac{K}{C\beta^t} \cdot (Q^{\pi_\theta^{(t)}}(s', a^*(s')) - \max_{a \neq a^*} Q^{\pi_\theta^{(t)}}(s', a^*(s'))) \\
&= \frac{\frac{(1-\gamma)^2}{8} \cdot d_\mu^{\pi_\omega^{(t)}}(s') \cdot \pi_\omega^{(t)}(a^*(s')|s') \cdot (Q^{\pi_\theta^{(t)}}(s', a^*(s')) - \max_{a \neq a^*(s')} Q^{\pi_\theta^{(t)}}(s', a^*(s')))}{C} \cdot K \\
&\geq K,
\end{aligned}$$

and (180) holds by Lemma 40.

Now we discuss two possible cases:

- The infinite time sequence $\{t_i\}_{i=1}^{\infty}$ contains all $t > T_M$. (i.e., $\{t_i\}_{i=1}^{\infty} = \{T_M + 1, T_M + 2, T_M + 3, \cdots\}$):
  In this case, by (180) and Lemma 40, we have that

$$
\begin{aligned}
\theta_{s',a^*(s')}^{(t+1)} - \theta_{s',a^*(s')}^{(t)} &\geq K + \max_{a \neq a^*(s')} \left\{ \omega_{s',a}^{(t)} - \theta_{s',a}^{(t)} \right\} \\
&> K + \max_{a \neq a^*(s')} \left\{ \underbrace{(\theta_{s',a}^{(t+1)} - \omega_{s',a}^{(t)})}_{\text{gradient}<0} \right\} + \max_{a \neq a^*(s')} \left\{ \underbrace{(\omega_{s',a}^{(t)} - \theta_{s',a}^{(t)})}_{\text{momentum}} \right\} \\
&\geq K + \max_{a \neq a^*(s')} \left\{ \theta_{s',a}^{(t+1)} - \theta_{s',a}^{(t)} \right\}
\end{aligned}
$$

This indicates a constant growth of the policy parameter $\theta_{s',a^*(s')}$. Consequently, by softmax parameterization, it follows that $1 - \pi_\theta^{(t)}(a^*|s') = O(e^{-\bar{c}t})$ for some $\bar{c} > 0$, thereby completing our contradiction.

- There exists at least one $(t_i - 1) > T_M$ such that $(t_i - 1) \notin \{t_i\}_{i=1}^{\infty}$ (i.e., $1 - \pi_\theta^{(t_i-1)}(a^*(s)|s') \leq \frac{K}{C\beta^{t_i-1}}$):
  In this case, we set $K = \frac{1}{4\sqrt{|\mathcal{A}||\mathcal{S}|}} \cdot \frac{1-\gamma}{4C_\infty - (1-\gamma)} + 2\ln\beta$ to ensure that the one step improvement $\pi_\theta^{(t_i+1)}(a^*(s)|s') - \pi_\theta^{(t_i)}(a^*(s)|s')$ is sufficiently large once we follow the update $\theta_{s',a^*(s')}^{(t_i+1)} = \omega_{s',a^*(s')}^{(t_i)} + K$ shown in (179). By Lemma 40, we have

$$
\begin{aligned}
&\pi_\theta^{(t+1)}(a^*(s)|s') \\
&= \frac{\exp(\theta_{s',a^*(s')}^{(t+1)})}{\sum_{a \neq a^*(s')} \exp(\theta_{s',a}^{(t+1)})} \\
&= \frac{\exp(\theta_{s',a^*(s')}^{(t)} + (\theta_{s',a^*(s')}^{(t+1)} - \omega_{s',a^*(s')}^{(t)}) + (\omega_{s',a^*(s')}^{(t)} - \theta_{s',a^*(s')}^{(t)}))}{\sum_{a \in \mathcal{A}} \exp(\theta_{s',a}^{(t)} + (\theta_{s',a}^{(t+1)} - \omega_{s',a}^{(t)}) + (\omega_{s',a}^{(t)} - \theta_{s',a}^{(t)}))} \\
&= \frac{\exp(\theta_{s',a^*(s')}^{(t)})}{\sum_{a \in \mathcal{A}} \exp(\theta_{s',a}^{(t)} + \underbrace{(\theta_{s',a}^{(t+1)} - \omega_{s',a}^{(t)}) + (\omega_{s',a}^{(t)} - \theta_{s',a}^{(t)}) - (\theta_{s',a^*(s')}^{(t+1)} - \omega_{s',a^*(s')}^{(t)}) - (\omega_{s',a^*(s')}^{(t)} - \theta_{s',a^*(s')}^{(t)}))}_{\leq 0}} \\
&\geq \frac{\exp(\theta_{s',a^*(s')}^{(t)})}{\sum_{a \neq a^*(s')} \exp(\theta_{s',a}^{(t)})} \\
&= \pi_\theta^{(t)}(a^*(s)|s').
\end{aligned} \tag{181}
$$

Following (181), since $\pi_\theta^{(t)}(a^*(s)|s')$ is increasing for all $t > T_M$, we have

$$
\sum_{a \neq a^*(s')} \pi_\theta^{(t_i)}(a|s') = 1 - \pi_\theta^{(t_i)}(a^*(s)|s') < 1 - \pi_\theta^{(t_i-1)}(a^*(s)|s') \leq \frac{K}{C\beta^{t_i-1}}. \tag{182}
$$

Moreover, by Lemma 40, we have

$$
\frac{\sum_{a \neq a^*(s')} \exp(\theta_{s',a}^{(t_i)})}{\exp(\frac{1}{4\sqrt{|\mathcal{A}||\mathcal{S}|}} \cdot \frac{1-\gamma}{4C_\infty - (1-\gamma)}) \exp(\theta_{s',a^*(s')}^{(t_i)})} \leq \sum_{a \neq a^*(s')} \pi_\theta^{(t_i)}(a|s') \leq \frac{K}{C\beta^{t_i-1}}, \tag{183}
$$

where the first inequality in (183) holds by the setting of M.

Combining (180), (182) and (183), we get

$$
\begin{aligned}
1 - \pi_\theta^{(t_i+1)}(a^*(s)|s') &= \sum_{a \neq a^*(s')} \pi_\theta^{(t_i+1)}(a|s') \\
&= \frac{\sum_{a \neq a^*(s')} \exp(\theta_{s',a}^{(t_i)} + (\theta_{s',a}^{(t_i+1)} - \theta_{s',a}^{(t_i)}))}{\sum_{a \in \mathcal{A}} \exp(\theta_{s',a}^{(t_i)} + (\theta_{s',a}^{(t_i+1)} - \theta_{s',a}^{(t_i)}))}
\end{aligned}
$$

55

$$
= \frac{\sum_{a \neq a^*(s')} \exp(\theta_{s',a}^{(t_i)} + \overbrace{(\theta_{s',a}^{(t_i+1)} - \omega_{s',a}^{(t_i)})}^{<0} + \overbrace{(\omega_{s',a}^{(t_i)} - \theta_{s',a}^{(t_i)}) - (\omega_{s',a^*(s')}^{(t_i)} - \theta_{s',a^*(s')}^{(t_i)}))}^{\leq 0}}{\sum_{a \in \mathcal{A}} \exp(\theta_{s',a}^{(t_i)} + (\theta_{s',a}^{(t_i+1)} - \omega_{s',a}^{(t_i)}) + (\omega_{s',a}^{(t_i)} - \theta_{s',a}^{(t_i)}) - (\omega_{s',a^*(s')}^{(t_i)} - \theta_{s',a^*(s')}^{(t_i)}))}
$$

$$
\leq \frac{\sum_{a \neq a^*(s')} \exp(\theta_{s',a}^{(t_i)})}{\exp(\theta_{s',a^*(s')}^{(t_i)} + (\theta_{s',a^*(s')}^{(t_i+1)} - \omega_{s',a}^{(t_i)}) + (\omega_{s',a^*(s')}^{(t_i)} - \theta_{a^*(s')}^{(t_i)}) - (\omega_{s',a^*(s')}^{(t_i)} - \theta_{s',a^*(s')}^{(t_i)}))}
$$

$$
\leq \frac{\sum_{a \neq a^*(s')} \exp(\theta_{s',a}^{(t_i)}))}{\exp(\theta_{s',a^*(s')}^{(t_i)} + (\theta_{s',a^*(s')}^{(t_i+1)} - \omega_{s',a^*(s')}^{(t_i)}))}
$$

$$
= \frac{\sum_{a \neq a^*(s')} \exp(\theta_{s',a}^{(t_i)})}{\exp(\theta_{s',a^*(s')}^{(t_i)})} \cdot \exp(-K)
$$

$$
\leq \frac{\exp(\frac{1}{2\sqrt{|\mathcal{A}||\mathcal{S}|}} \cdot \frac{1-\gamma}{4C_\infty - (1-\gamma)})K}{C\beta^{t_i-1}} \cdot \exp(-K)
$$

$$
= \frac{K}{C\beta^{t_i+1}}, \tag{184}
$$

where (184) holds by setting $K = \frac{1}{4\sqrt{|\mathcal{A}||\mathcal{S}|}} \cdot \frac{1-\gamma}{4C_\infty - (1-\gamma)} + 2\ln\beta$.

Hence we get that once $1 - \pi_\theta^{(t)}(a^*(s)|s') > \frac{K}{C\beta^t}$ occurs for any $t \in \{t_i\}_{i=1}^\infty$, we could ensure that $1 - \pi_\theta^{(t+1)}(a^*(s)|s') \leq \frac{K}{C\beta^{t+1}}$ (i.e., $(t+1) \notin \{t_i\}_{i=1}^\infty$ for all $t \in \{t_i\}_{i=1}^\infty$ with $t > t_i$), In other words, for all $t \in \{t_i\}_{i=1}^\infty$, we have $t-1 \notin \{t_i\}_{i=1}^\infty$ and hence $1 - \pi_\theta^{(t-1)}(a^*(s)|s') \leq \frac{K}{C\beta^{t-1}}$. Combining with the monotonicity of $1 - \pi_\theta^{(t-1)}(a^*(s)|s')$ in (181), we get $1 - \pi_\theta^{(t)}(a^*(s)|s') \leq 1 - \pi_\theta^{(t-1)}(a^*(s)|s') \leq \frac{K}{C\beta^{t-1}}$. We thus see that $1 - \pi_\theta^{(t)}(a^*(s)|s') \leq \frac{K}{C\beta^{t-1}}$ for all $t > T_M$. This means that for any such s' that has a companion infinite sequence $\{t_i\}$ that satisfies (177) with $\tilde{c} = K/C$ and $\bar{c} = \beta$, the condition (176) actually holds for s' with $\tilde{c} = \beta K/C$ and $\bar{c} = \beta$, so for this setting of $\tilde{c}$ and c we have obtained a contradiction and thus completed the proof. $\qquad\square$

E.3. Numerical Validation of the Linear Convergence Rate

In this subsection, we empirically validate the convergence rate of APG with an time-varying step sizes scheduling by conducting experiments on an MDP with five states and five actions. This MDP is identical to the one discussed in Section 6.1, and detailed configuration information is provided in Appendix F. The figure illustrating the logarithmic sub-optimality gap of APG with an time-varying step sizes scheduling under uniform initialization is presented in Figure 4. It is evident that after ten iterations, the curve tends to form a linear pattern. This behavior aligns with the linear convergence rate demonstrated in Theorem 3.
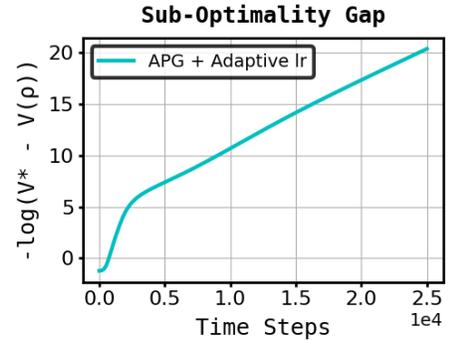
Figure 4. The sub-optimality gap of APG with time-varying step sizes under an MDP of five states and five actions with uniform initialization.

F. Detailed Explanation of the Motivating Example and the Experimental Configurations

F.1. Motivating Examples of APG (Section 4.2)

Consider a simple two-action bandit with actions a^* and a_2 and a reward function $r(a^*) = 1, r(a_2) = 0$. Accordingly, the objective we aim to optimize is $\mathbb{E}_{a \sim \pi_\theta}[r(a)] = \pi_\theta(a^*)$. Under softmax parameterization, the Hessian matrix with respect to the parameters θ_{a^*} and θ_{a_2} is:

$$
\mathbf{H} = \begin{bmatrix} \frac{\partial^2 \pi_\theta(a^*)}{\partial \theta_{a^*} \partial \theta_{a^*}} & \frac{\partial^2 \pi_\theta(a^*)}{\partial \theta_{a^*} \partial \theta_{a_2}} \\ \frac{\partial^2 \pi_\theta(a^*)}{\partial \theta_{a_2} \partial \theta_{a^*}} & \frac{\partial^2 \pi_\theta(a^*)}{\partial \theta_{a_2} \partial \theta_{a_2}} \end{bmatrix}
$$
$$
= \begin{bmatrix} \pi_\theta(a^*)(1 - \pi_\theta(a^*))(1 - 2\pi_\theta(a^*)) & \pi_\theta(a^*)(1 - \pi_\theta(a^*))(2\pi_\theta(a^*) - 1) \\ \pi_\theta(a^*)(1 - \pi_\theta(a^*))(2\pi_\theta(a^*) - 1) & \pi_\theta(a^*)(1 - \pi_\theta(a^*))(1 - 2\pi_\theta(a^*)) \end{bmatrix}.
$$

After straightforward calculation, we see that its eigenvalues are $\lambda_1 = 2\pi_\theta(a^*)(1 - \pi_\theta(a^*))(1 - 2\pi_\theta(a^*)), \lambda_2 = 0$. Therefore, when $\pi_\theta(a^*) \geq 0.5$, we see that $\lambda_1, \lambda_2 \leq 0$ and thus the Hessian is negative semi-definite, which is equivalent to that the function $\mathbb{E}_{a \sim \pi_\theta}[r(a)]$ is concave.

F.2. Setting of the Experiment for Numerical Validation of the Convergence Rates of APG (Section 6.1)

(Bandit) We conduct a three-action bandit experiment with actions $\mathcal{A} = [a^*, a_2, a_3]$ with the corresponding rewards being $r = [r(a^*), r(a_2), r(a_3)] = [1, 0.99, 0]$. We consider two initializations for the policy parameters. The first one is the uniform initialization ($\theta^{(0)} = [0, 0, 0], \pi^{(0)} = [1/3, 1/3, 1/3]$) and the second one is a hard initialization such that $\theta^{(0)} = [1, 3, 5], \pi^{(0)} = [0.01588, 0.11731, 0.86681]$ and hence the optimal action has the smallest initial probability.

(MDP) We conduct an experiment on an MDP with five states and five actions under the initial state distribution $\rho = [0.3, 0.2, 0.1, 0.15, 0.25]$. The reward function, the initial policy parameters, and the transition probability can be found in Tables 2-9.

$r(s, a)$	a_1	a_2	a_3	a_4	a_5
s_1	1.0	0.8	0.6	0.7	0.4
s_2	0.5	0.3	0.1	1.0	0.6
s_3	0.6	0.9	0.8	0.7	1.0
s_4	0.1	0.2	0.6	0.7	0.4
s_5	0.8	0.4	0.6	0.2	0.9

Table 2. Experimental settings: Reward function.

$\theta_{s,a}^{(0)}$	a_1	a_2	a_3	a_4	a_5
s_1	1	2	3	4	5
s_2	3	4	5	1	2
s_3	5	2	3	4	1
s_4	5	4	2	1	3
s_5	2	4	3	5	1

Table 3. Experimental settings: Hard initialization.

$\theta_{s,a}^{(0)}$	a_1	a_2	a_3	a_4	a_5
s_1	0	0	0	0	0
s_2	0	0	0	0	0
s_3	0	0	0	0	0
s_4	0	0	0	0	0
s_5	0	0	0	0	0

Table 4. Experimental settings: Uniform initialization.

$P(s\|s_1,a)$	a_1	a_2	a_3	a_4	a_5
s_1	0.1	0.6	0.5	0.4	0.2
s_2	0.5	0.1	0.1	0.3	0.1
s_3	0.1	0.1	0.1	0.1	0.1
s_4	0.2	0.1	0.2	0.1	0.1
s_5	0.1	0.1	0.1	0.1	0.5

Table 5. Experimental settings: Transition probability $P(\cdot|s_1,\cdot)$.

$P(s\|s_2,a)$	a_1	a_2	a_3	a_4	a_5
s_1	0.1	0.4	0.1	0.4	0.2
s_2	0.5	0.1	0.4	0.1	0.2
s_3	0.2	0.2	0.3	0.1	0.2
s_4	0.1	0.2	0.1	0.1	0.2
s_5	0.1	0.1	0.1	0.3	0.2

Table 6. Experimental settings: Transition probability $P(\cdot|s_2,\cdot)$.

$P(s\|s_3,a)$	a_1	a_2	a_3	a_4	a_5
s_1	0.6	0.2	0.3	0.1	0.2
s_2	0.1	0.4	0.3	0.4	0.1
s_3	0.1	0.1	0.2	0.3	0.1
s_4	0.1	0.2	0.1	0.1	0.1
s_5	0.1	0.1	0.1	0.1	0.5

Table 7. Experimental settings: Transition probability $P(\cdot|s_3,\cdot)$.

$P(s\|s_4,a)$	a_1	a_2	a_3	a_4	a_5
s_1	0.6	0.1	0.2	0.4	0.5
s_2	0.1	0.5	0.1	0.3	0.1
s_3	0.1	0.1	0.1	0.1	0.1
s_4	0.1	0.2	0.1	0.1	0.2
s_5	0.1	0.1	0.5	0.1	0.1

Table 8. Experimental settings: Transition probability $P(\cdot|s_4,\cdot)$.

$P(s\|s_5,a)$	a_1	a_2	a_3	a_4	a_5
s_1	0.2	0.4	0.4	0.1	0.2
s_2	0.2	0.1	0.1	0.4	0.5
s_3	0.2	0.2	0.1	0.2	0.1
s_4	0.2	0.2	0.3	0.1	0.1
s_5	0.2	0.1	0.1	0.2	0.1

Table 9. Experimental settings: Transition probability $P(\cdot|s_5,\cdot)$.

F.3. APG on Atari 2600 Games (Section 6.2)

We empirically evaluate the performance of APG on four Atari 2600 games from the Arcade Learning Environment (ALE) (Bellemare et al., 2013).

Our implementation for APG, HBPG, and PG are based on the code base of Stable Baselines3 and RL Baselines3 Zoo (Raffin, 2020; Raffin et al., 2021). The hyperparameters, exactly set to their default values, are detailed in Table 10. Pseudo code of HBPG is given in Algorithm 5.

In Atari, one major challenge of implementing the exact APG is how to get a good estimate of the true gradient $\nabla_\theta V^{\pi_\theta}(\mu)$ given the large state and action spaces of the Atari games. While it is theoretically possible to get a nearly true gradient by Monte-Carlo estimation, it is not feasible in practice due to the resulting prohibitively large computation time. To address this, we made a few design choices in the practical implementation:

- To ensure a sufficiently accurate gradient estimate with a reasonable amount of training time, we use a moderately large batch size of 80 for APG (also for PG and HBPG for a fair comparison) and leverage RMSProp, which is designed to better average the stochastic gradients over successive mini-batches than the vanilla stochastic gradient descent.

- To estimate the advantage function (required in the policy gradient), we use the generalized advantage estimator (GAE) (Schulman et al., 2015b), which is a popular technique for advantage estimation in deep RL.

Hyperparameters	APG (Ours)	HBPG	PG
Batch Size	80	80	80
Entropy coefficient	0.01	0.01	0.01
Gamma (discount factor)	0.99	0.99	0.99
Lambda for GAE	1.0	1.0	1.0
Learning rate	0.0007	0.0007	0.0007
Number of environments	16	16	16
Time steps	10^7	10^7	10^7
Value function coefficient	0.25	0.25	0.25

Table 10. Hyperparameters of APG, HBPG and PG in Carnival, Pong, Riverraid, and Seaquest.

Algorithm 5 Heavy-Ball Policy Gradient (HBPG)

Input: Step size $\eta = \frac{1}{L}$, where L is the Lipschitz constant of the gradient of the objective function f, momentum factor β.
Initialize: $\theta^{(0)}$ and let $\theta^{(-1)} = \theta^{(0)}$.
for $t = 1$ to T **do**
$\quad \theta^{(t)} = \theta^{(t-1)} + \eta \nabla_\theta V^{\pi_\theta}(\mu)\big|_{\theta=\theta^{(t-1)}} + \beta(\theta^{(t-1)} - \theta^{(t-2)})$
end for

G. Convergence Rate of Policy Gradient Under Softmax Parameterization

The Gradient Descent (GD) method within the realm of optimization exhibits a convergence rate of $O(1/t)$ when applied to convex objectives. Remarkably, (Mei et al., 2020) have demonstrated that the Policy Gradient (PG) method, even when dealing with nonconvex RL objectives, can achieve the same convergence rate, leveraging the non-uniform Polyak–Łojasiewicz (PL) condition under softmax parameterization. In this paper, we employ a primary proof technique involving the characterization of local C-nearly concavity and subsequently establish that the Accelerated Policy Gradient (APG) method attains this convergence regime. Our investigation also reveals that the PG method with softmax parameterization shares this intriguing property.

In this section, we will begin by providing a proof demonstrating the $\tilde{O}(1/t)$ convergence rate for PG when optimizing the value objectives that are C-nearly concave, while employing the softmax parameterized policy. Additionally, we will offer a proof establishing the capability of PG to reach the local C-nearly concavity regime within a finite number of time steps, considering the general MDP setting.

G.1. Convergence Rate Under Nearly Concave Objectives for Policy Gradient

Theorem 6. *Let $\left\{\theta^{(t)}\right\}_{t\geq 1}$ be computed by PG and that $V^{\pi_\theta}(\mu)$ is C-nearly concave along the gradient update at time t with a constant $C > 1$ for every $t \geq 1$. Then given any $T \geq 1$, and for any θ^{**} such that for all $t \leq T$ the direction $\theta^{**} - \theta^{(t)}$ satisfies the C-nearly concavity and $V^{\pi_{\theta^{**}}}(\mu) \geq V^{\pi_\theta^{(t)}}(\mu)$, we have*

$$V^{\pi_{\theta^{**}}}(\mu) - V^{\pi_\theta^{(T)}}(\mu) \leq \frac{2LC^2\left(\|\theta^{**}\| + \kappa \ln T\right)^2}{T-1},$$

where L is the Lipschitz constant of the objective and $\kappa \in \mathbb{R}$ is a finite constant.

Proof of Theorem 6. Initially, given any T, by (5) and Lemma 14, for any $t < T$, we have

$$-V^{\pi_\theta^{(t+1)}}(\mu) \leq -V^{\pi_\theta^{(t)}}(\mu) - \left\langle \nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\theta^{(t)}}, \theta^{(t+1)} - \theta^{(t)} \right\rangle + \frac{L}{2}\left\|\theta^{(t+1)} - \theta^{(t)}\right\|^2$$

$$= -V^{\pi_\theta^{(t)}}(\mu) - \eta\left\|\nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\theta^{(t)}}\right\|^2 + \frac{L\eta^2}{2}\left\|\nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\theta^{(t)}}\right\|^2$$

$$= -V^{\pi_\theta^{(t)}}(\mu) - \frac{1}{2L}\left\|\nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\theta^{(t)}}\right\|^2. \tag{185}$$

Additionally, by the hypothesis of the nearly concavity of objective along the direction to θ^{**}, we have

$$-V^{\pi_\theta^{(t)}}(\mu) \leq -V^{\pi_\theta^{**}}(\mu) + C \cdot \left\langle \nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\theta^{(t)}}, \theta^{**} - \theta^{(t)} \right\rangle. \tag{186}$$

Define

$$\delta^{(t)} := V^{\pi_\theta^{**}}(\mu) - V^{\pi_\theta^{(t)}}(\mu).$$

Then (185) leads to

$$\delta^{(t+1)} - \delta^{(t)} \leq -\frac{1}{2L}\left\|\nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\theta^{(t)}}\right\|^2. \tag{187}$$

Moreover, (186) leads to

$$\delta^{(t)} \leq C \cdot \left\langle \nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\theta^{(t)}}, \theta^{**} - \theta^{(t)} \right\rangle$$

$$\leq C\left\|\nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\theta^{(t)}}\right\|\left\|\theta^{**} - \theta^{(t)}\right\|$$

$$\leq C\left\|\nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\theta^{(t)}}\right\|\left(\|\theta^{**}\| + \left\|\theta^{(t)}\right\|\right)$$

$$\leq C\left\|\nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\theta^{(t)}}\right\|\left(\|\theta^{**}\| + \max_{t'\leq t}\left\|\theta^{(t')}\right\|\right). \tag{188}$$

By substituting and rearranging (187) and (188), we have

$$\delta^{(t+1)} \leq \delta^{(t)} - \frac{1}{2LC^2 \left(\|\theta^{**}\| + \max_{t' \leq t} \|\theta^{(t')}\| \right)^2} \cdot {\delta^{(t)}}^2. \tag{189}$$

Consequently, by rearranging and dividing $\delta^{(t)}\delta^{(t+1)}$ on each side of (189), we have

$$\frac{1}{\delta^{(t+1)}} \geq \frac{1}{\delta^{(t)}} + \frac{\delta^{(t)}}{\left(2LC^2 \left(\|\theta^{**}\| + \max_{t' \leq t} \|\theta^{(t')}\| \right)^2 \right) \delta^{(t+1)}}.$$

Furthermore, since PG enjoys monotonic improvement, we have

$$\frac{1}{\delta^{(t+1)}} - \frac{1}{\delta^{(t)}} \geq \frac{\delta^{(t)}}{\delta^{(t+1)}} \frac{1}{2LC^2 \left(\|\theta^{**}\| + \max_{t' \leq t} \|\theta^{(t')}\| \right)^2} \geq \frac{1}{2LC^2 \left(\|\theta^{**}\| + \max_{t' \leq T} \|\theta^{(t')}\| \right)^2}. \tag{190}$$

By implementing telescoping on (190) from $t = 1$ to $T - 1$, we have

$$\delta^{(T)} = V^{\pi_\theta^{**}}(\mu) - V^{\pi_\theta^{(T)}}(\mu) \leq \frac{2LC^2 \left(\|\theta^{**}\| + \max_{t' \leq T} \left\|\theta^{(t')}\right\| \right)^2}{T - 1}. \tag{191}$$

Finally, we claim that the norm of $\theta^{(t)}$ grows in a rate of $O(\ln t)$. Without loss of generality, we assume that $\sum_{a \in \mathcal{A}} \theta_{s,a}^{(t)} = 0$ for all $s \in \mathcal{S}$ and $t \geq 1$ (the gradient updates sum up to 0). Now we prove it by contradiction. Assume that the norm of $\theta^{(t)}$ does not grow in a rate of $O(\ln t)$, then for any $M \in \mathbb{R}$, there must exist a finite time T_1 such that $\|\theta^{(t)}\|/\ln t > M$ for all $t > T_1$. Moreover, by the asymptotic convergence results (Theorem 5.1) of (Agarwal et al., 2021) and Assumption 1, we have that there exist an upper bound $\bar{M}$ and a finite time T_2 such that

$$\theta_{s,a}^{(t)} < \bar{M}, \text{ for all } a \neq a^*(s), s \in \mathcal{S}, t \in \mathbb{N}, \tag{192}$$

$$\theta_{s,a^*(s)}^{(t)} > \theta_{s,a}^{(t)}, \text{ for all } a \neq a^*(s), s \in \mathcal{S}, t > T_2. \tag{193}$$

Hence, by choosing $M = 2|\mathcal{A}|$, then we have

$$\frac{\|\theta^{(t)}\|}{\ln t} > 2|\mathcal{A}|, \text{ for all } t > \max\{T_1, T_2\}. \tag{194}$$

Note that by (192) and (193) and the fact that $\|\theta^{(t)}\| > 2|\mathcal{A}| \cdot \ln t$, we have that $\theta_{s,a^*(s)} > 2\ln t$. Hence, for all $t > \max\{T_1, T_2\}, s \in \mathcal{S}$, we have that

$$1 - \pi_\theta^{(t)}(a^*(s)|s) \leq \frac{\sum_{a \neq a^*(s)} \exp(\theta_{s,a})}{\exp(\theta_{s,a^*(s)}) + \sum_{a \neq a^*(s)} \exp(\theta_{s,a})} \leq \frac{(|\mathcal{A}| - 1)\exp(\bar{M})}{\exp(2\ln(t))} = \frac{(|\mathcal{A}| - 1)\exp(\bar{M})}{t^2},$$

where the last inequality holds by the positivity of the exponential function $\exp$. Therefore, we can obtain the bound of the gradient with respect to $\theta_{s,a^*(s)}$

$$\left| \frac{\partial V^{\pi_\theta^{(t)}}(\mu)}{\partial \theta_{s,a^*(s)}} \right| = \left| \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta^{(t)}}(s) \cdot \pi_\theta^{(t)}(a^*(s)|s) \cdot A^{\pi_\theta^{(t)}}(s, a^*(s)) \right|$$

$$\leq \left| \frac{1}{1-\gamma} \cdot 1 \cdot 1 \cdot (Q^{\pi_\theta^{(t)}}(s, a^*(s)) - V^{\pi_\theta^{(t)}}(s)) \right|$$

$$= \frac{1}{1-\gamma} \cdot \left| Q^{\pi_\theta^{(t)}}(s, a^*(s)) - \pi_\theta^{(t)}(a^*(s)|s) Q^{\pi_\theta^{(t)}}(s, a^*(s)) \right.$$

$$\left. - \sum_{a \neq a^*(s)} \pi_\theta^{(t)}(a|s) Q^{\pi_\theta^{(t)}}(s, a)) \right|$$

61

$$\leq \frac{1 - \pi_\theta^{(t)}(a^*(s)|s)}{1 - \gamma} \left(\left| Q^{\pi_\theta^{(t)}}(s, a^*(s)) \right| + \max_a \left| Q^{\pi_\theta^{(t)}}(s, a) \right| \right)$$

$$\leq \frac{2(|\mathcal{A}| - 1)\exp(\bar{M})}{t^2(1 - \gamma)^2}, \tag{195}$$

$$\left| \frac{\partial V^{\pi_\theta^{(t)}}(\mu)}{\partial \theta_{s,a}} \right| = \left| \frac{1}{1 - \gamma} \cdot d_\mu^{\pi_\theta^{(t)}}(s) \cdot \pi_\theta^{(t)}(a|s) \cdot A^{\pi_\theta^{(t)}}(s, a) \right|$$

$$\leq \left| \frac{1}{1 - \gamma} \cdot 1 \cdot (1 - \pi_\theta^{(t)}(a^*(s)|s)) \cdot \frac{1}{1 - \gamma} \right|$$

$$= \frac{(|\mathcal{A}| - 1)\exp(\bar{M})}{t^2(1 - \gamma)^2}, \text{ for all } a \neq a^*(s). \tag{196}$$

Furthermore, by the update rule of PG, for all $t > \max\{T_1, T_2\}, s \in \mathcal{S}$, we have

$$\left| \theta_{s,a^*(s)}^{(t+k)} \right| = \left| \theta_{s,a^*(s)}^{(t)} + \sum_{i=t}^{t+k-1} \eta \frac{\partial V^{\pi_\theta^{(i)}}(\mu)}{\partial \theta_{s,a^*(s)}} \right| \leq \left| \theta_{s,a^*(s)}^{(t)} \right| + \underbrace{\sum_{i=t}^{\infty} \eta \left| \frac{\partial V^{\pi_\theta^{(i)}}(\mu)}{\partial \theta_{s,a^*(s)}} \right|}_{<\infty, \text{ by } (195)},$$

$$\left| \theta_{s,a}^{(t+k)} \right| = \left| \theta_{s,a}^{(t)} + \sum_{i=t}^{t+k-1} \eta \frac{\partial V^{\pi_\theta^{(i)}}(\mu)}{\partial \theta_{s,a}} \right| \leq \left| \theta_{s,a}^{(t)} \right| + \underbrace{\sum_{i=t}^{\infty} \eta \left| \frac{\partial V^{\pi_\theta^{(i)}}(\mu)}{\partial \theta_{s,a}} \right|}_{<\infty, \text{ by } (196)}.$$

which leads to the contradiction that $\|\theta^{(t)}\| \to \infty$ in (194).

Therefore, for any T, we have $\max_{t' \leq T} \|\theta^{(t')}\| \in O(\ln T)$. Hence, follow from (191),

$$V^{\pi_\theta^{**}}(\mu) - V^{\pi_\theta^{(T)}}(\mu) \leq \frac{2LC^2 \left(\|\theta^{**}\| + \max_{t' \leq T} \left\| \theta^{(t')} \right\| \right)^2}{T - 1}$$

$$\leq \frac{2LC^2 \left(\|\theta^{**}\| + \kappa \ln T \right)^2}{T - 1},$$

where $\kappa \in \mathbb{R}$ is a finite constant.

$\square$

Remark 16. It is crucial to emphasize that when using PG with softmax parameterization, the parameters cannot exhibit excessive growth. More specifically, we can demonstrate that the growth rate of the parameters remains bounded by $O(\ln t)$. This remarkable bounding property allows us to relax the objective structure to a C-nearly concave form while still achieving a convergence rate of $\tilde{O}(1/t)$. However, it is worth noting that due to the unbounded nature of softmax parameterization, it must also compensate for a logarithmic factor in its convergence rate, which may not be tightly aligned with results in the optimization regime.

G.2. Convergence Rate of Policy Gradient

Theorem 7 (**Asymptotic Convergence for Softmax Parameterization in (Agarwal et al., 2021)**). *Assume we follow the gradient ascent update rule as specified in* (5) *and that* $\mu(s) > 0$ *for all states* s*. Suppose* $\eta^{(t)} \leq \frac{(1-\gamma)^3}{8}$*, then we have that for all states* s*,* $V^{(t)}(s) \to V^*(s)$ *as* $t \to \infty$*.*

Lemma 41. *Consider a tabular softmax parameterized policy* π_θ*. Under PG with* $\eta^{(t)} = \frac{(1-\gamma)^3}{8}$ *and Assumption 1, given any* $M > 0$*, there exists a finite time* T *such that for all* $t \geq T, s \in \mathcal{S}$*, and* $a \neq a^*(s)$*, we have (i)* $\theta_{s,a^*(s)} - \theta_{s,a} > M$*; (ii)* $V^{\pi_\theta^{(t)}}(s) > Q^*(s, a_2(s))$*; (iii)* $\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a^*(s)}}\Big|_{\theta=\theta^{(t)}} > 0 > \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\Big|_{\theta=\theta^{(t)}}$*.*

Proof of Lemma 41. By Assumption 1, there is a unique optimal action, so Theorem 7 implies that the policy weight probability of the optimal action approaches 1. For $\theta_{s,a^*(s)} - \theta_{s,a} > M$, since we also have the asymptotic convergence property, we can see that the same strategy works here as the proof shown in Lemma 2.

Moreover, by Lemma 8, we can show that (ii) can imply the third point. Regarding the second point, it is the direct result of Theorem 7 because $V^{\pi_\theta^{(t)}}$ can always enter a small enough neighborhood such that the values always are greater than $Q^*(s, a_2(s))$. Hence, by Lemmas 8 and 11, we obtain the results. $\qquad\square$

Theorem 8 (**Convergence Rate of PG**). *Consider a tabular softmax parameterized policy π_θ. Under PG with $\eta^{(t)} = \frac{(1-\gamma)^3}{8}$, there exists a finite time T such that for all $t \geq T$, we have*

$$V^*(\rho) - V^{\pi_\theta^{(t)}}(\rho) \leq \frac{1}{1-\gamma} \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty \left(\frac{16C^2(\kappa+2)^2[\ln(t-T)]^2}{(1-\gamma)^3(t-T-1)} + \frac{2|\mathcal{S}|}{(1-\gamma)^2} \left(\frac{|\mathcal{A}|-1}{(t-T)^2+|\mathcal{A}|-1} \right) \right),$$

where $\kappa \in \mathbb{R}$ is a finite constant.

Proof of Theorem 8. Let

$$M' := \ln \left[\frac{2|\mathcal{S}||\mathcal{A}|^2}{(1-\gamma)^2 \min_{s \in S} \mu(s)} \right].$$

By Lemma 41, there exists a finite time T' such that for all $t \geq T'$, $s \in \mathcal{S}$, and $a \neq a^*(s)$, we have (i) $\theta_{s,a^*(s)} - \theta_{s,a} > M'$, (ii) $V^{\pi_\theta^{(t)}}(s) > Q^*(s, a_2(s))$, (iii) $\left. \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a^*(s)}} \right|_{\theta=\theta^{(t)}} > 0 > \left. \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} \right|_{\theta=\theta^{(t)}}$. By (iii), we know that the PG updates (5) at time $t \geq T'$ lie in the feasible update domain $\mathcal{U}$. Furthermore, by the first point in Lemma 37 and (iii), we obtain that all the PG updates after time T' under state rescaling, i.e., $\|\boldsymbol{d}_{\boldsymbol{s},\cdot}\|_2^2 = 1$, we have $\min_{i \geq 2}\{d_{s,a^*(s)} - d_{s,a_i(s)}\} > |\mathcal{A}|^{-\frac{3}{2}}$. Thus, we further define

$$M := \max \left\{ 2\ln \left[2(|\mathcal{A}|-1) \right] - 2\ln \left[|\mathcal{A}|^{-\frac{3}{2}} \right], M' \right\},$$

by Lemma 2 again, we know that there is a finite time $T \geq T'$ such that for all $t \geq T'$, $s \in \mathcal{S}$, and $a \neq a^*(s)$, we have (i) $\theta_{s,a^*(s)} - \theta_{s,a} > M'$, (ii) $V^{\pi_\theta^{(t)}}(s) > Q^*(s, a_2(s))$, (iii) $\left. \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a^*(s)}} \right|_{\theta=\theta^{(t)}} > 0 > \left. \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} \right|_{\theta=\theta^{(t)}}$, (iv) the state rescaled updates satisfy $\min_{i \geq 2}\{d_{s,a^*(s)} - d_{s,a_i(s)}\} > |\mathcal{A}|^{-\frac{3}{2}}$. Therefore, by Lemma 1, we know that the objective function $\theta \to V^{\pi_\theta}(\mu)$ is $\frac{3}{2}$-nearly concave along the PG update directions for all $t \geq T$.

By Theorem 6,

$$V^{\pi_{\theta^{**}}^{(t)}}(\mu) - V^{\pi_\theta^{(t)}}(\mu) \leq \frac{2LC^2 \left(\left\| \theta^{**(t)} \right\| + \kappa \ln(t-T) \right)^2}{(t-T)-1}$$

$$\leq \frac{16C^2 \left(2\ln(t-T) + \kappa \ln(t-T) \right)^2}{(1-\gamma)^3(t-T-1)}$$

$$= \frac{16C^2(\kappa+2)^2[\ln(t-T)]^2}{(1-\gamma)^3(t-T-1)},$$

where $\theta^{(t)**} := [2\ln(t-T), 0, 0, \cdots, 0]$ is a chosen surrogate optimal solution at time t. Additionally, we have the sub-optimality gap between the original optimal solution and the surrogate optimal solution can be bounded as

$$V^*(\mu) - V^{\pi_{\theta^{**}}^{(t)}}(\mu) = \frac{1}{1-\gamma} \sum_s d_\mu^{\pi^*}(s) \sum_a (\pi^*(a|s) - \pi_{\theta^{**}}^{(t)}(a|s)) \cdot A^{\pi_{\theta^{**}}^{(t)}}(s,a)$$

$$\leq \frac{2|\mathcal{S}|}{(1-\gamma)^2} \left(1 - \frac{\exp(2\ln(t-T))}{\exp(2\ln(t-T)) + \exp(0) \cdot (|\mathcal{A}|-1)} \right) \tag{197}$$

$$= \frac{2|\mathcal{S}|}{(1-\gamma)^2} \left(\frac{|\mathcal{A}|-1}{(t-T)^2+|\mathcal{A}|-1} \right),$$

where (197) consider the upper bound of $A^{\pi_{\theta^{**}}^{(t)}}(s,a) \leq \frac{1}{1-\gamma}$ and the sum of differences between sub-optimal actions can be at most the difference between the optimal actions, which causes a preconstant 2. By Lemma 6, we can change the convergence rate from $V^\pi(\mu)$ to $V^\pi(\rho)$,

$$V^*(\rho) - V^{\pi_\theta^{(t)}}(\rho) \leq \frac{1}{1-\gamma} \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty \left(\frac{16C^2(\kappa+2)^2[\ln(t-T)]^2}{(1-\gamma)^3(t-T-1)} + \frac{2|\mathcal{S}|}{(1-\gamma)^2} \left(\frac{|\mathcal{A}|-1}{(t-T)^2+|\mathcal{A}|-1} \right) \right),$$

and we complete the proof. $\qquad\square$

H. Challenges of APG Without Momentum Restart Mechanisms

In this section, we provide insights into the challenges encountered when applying APG without momentum restart mechanisms. Furthermore, we've established the convergence rate of $\tilde{O}(1/t^2)$ under APG without restart mechanisms in the bandit setting. More specifically, In Appendix H.1, we introduce the algorithm, NAPG, which is the original APG without restart mechanisms. Subsequently, in Appendix H.2, we provide a numerical validation in the bandit setting to illustrate the non-monotonic improvement of NAPG. Additionally, in Appendix H.3, we illustrate the challenges of establishing the limiting value functions in the MDPs setting.

H.1. APG Without Restart Mechanisms

For ease of exposition, this subsection presents the algorithm for Nesterov Accelerated Policy Gradient (NAPG) without restart mechanisms as in Algorithm 6. As depicted in Algorithm 6. It is worth noting that without these restart mechanisms, Algorithm 6 aligns precisely with the Nesterov acceleration method outlined in (Su et al., 2014).

Algorithm 6 Nesterov Accelerated Policy Gradient (NAPG) *Without Restart Mechanisms*

Input: Step size $\eta^{(t)} > 0$.
Initialize: $\theta^{(0)} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $\omega^{(0)} = \theta^{(0)}$.
for $t = 1$ to T **do**

$$\theta^{(t)} \leftarrow \omega^{(t-1)} + \eta^{(t)} \nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\omega^{(t-1)}}$$
$$\omega^{(t)} \leftarrow \theta^{(t)} + \frac{t-1}{t+2}(\theta^{(t)} - \theta^{(t-1)})$$

end for

H.2. Numerical Validation of the Non-Monotonic Improvement Under NAPG

In this subsection, we illustrate the difficulties involved in analyzing the convergence of NAPG, compared to the standard policy gradient methods through a numerical experiment. In contrast to the standard policy gradient (PG) method, which exhibits monotonic improvement, NAPG could experience non-monotonic progress as a result of the momentum term, which could lead to negative performance changes. To further demonstrate this phenomenon, we conduct a 3-action bandit experiment with a highly sub-optimal initialization, where the weight of the optimal action of the initial policy is extremely small. More specifically, we conduct a 3-action bandit experiment with actions $\mathcal{A} = [a^*, a_2, a_3]$, where the corresponding rewards are $r = [r(a^*), r(a_2), r(a_3)] = [1, 0.8, 0]$. We initialize the policy parameters as $\theta^{(0)} = [0, 3, 10]$, which represents a highly sub-optimal initialization. Remarkably, the weight of the optimal action in the initial policy $\pi^{(0)} \approx [0.00005, 0.00091, 0.99904]$ is exceedingly small. As shown in Figure 5, the one-step improvement becomes negative around epoch 180 and provides

Figure 5. The one-step improvement of APG on a three-action bandit problem.

nearly zero improvement after that point. Notably, the asymptotic convergence of the standard PG is largely built on the monotonic improvement property, as shown in (Agarwal et al., 2021). With that said, the absence of monotonic improvement in NAPG poses a fundamental challenge when analyzing the convergence to an optimal policy.

H.3. Challenges of Establishing the Limiting Value Functions in the MDPs Setting Under NAPG

In this subsection, we illustrate the challenges involved in establishing the limiting value functions under NAPG in the MDPs setting. Recall from Appendix H.2 that NAPG is not guaranteed to achieve monotonic improvement in each iteration

due to the momentum. This is one salient difference from the standard PG, which inherently enjoys strict improvement and hence the existence of the limiting value functions (i.e., $\lim_{t\to\infty} V^{\pi_{\theta(t)}}(s)$) by Monotone Convergence Theorem (Agarwal et al., 2021). Without monotonicity, it remains unknown if the limiting value functions even exist.

Furthermore, while the gradient step ensures a guaranteed monotonic improvement, which can help counteract the non-monotonicity resulting from the momentum step, it is still a challenging task to quantify this non-monotonic behavior due to the cumulative effect of the momentum. To be more specific, we demonstrate that the influence exerted by the momentum term can lead to a substantial and challenging-to-control effect.

Let $\delta^{(T_0)} := \theta_{s,a}^{(T_0)} - \theta_{s,a}^{(T_0-1)}$, by the update rule of NAPG, we have

$$\theta_{s,a}^{(T_0+1)} = \theta_{s,a}^{(T_0)} + \frac{T_0-1}{T_0+2}\delta^{(T_0)} + \eta^{(T_0+1)} \cdot \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\bigg|_{\theta=\omega^{(T_0)}} \tag{198}$$

$$\theta_{s,a}^{(T_0+2)} = \theta_{s,a}^{(T_0+1)} + \frac{T_0}{T_0+3}\delta^{(T_0+1)} + \eta^{(T_0+2)} \cdot \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\bigg|_{\theta=\omega^{(T_0+1)}} \tag{199}$$

$$= \theta_{s,a}^{(T_0)} + \frac{T_0-1}{T_0+2}\delta^{(T_0)} + \eta^{(T_0+1)} \cdot \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\bigg|_{\theta=\omega^{(T_0)}}$$
$$+ \frac{T_0}{T_0+3}\left(\frac{T_0-1}{T_0+2}\delta^{(T_0)} + \eta^{(T_0+1)} \cdot \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\bigg|_{\theta=\omega^{(T_0)}}\right)$$
$$+ \eta^{(T_0+2)} \cdot \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\bigg|_{\theta=\omega^{(T_0+1)}} \tag{200}$$

$$= \theta_{s,a}^{(T_0)} + \left(\frac{T_0-1}{T_0+2} + \frac{T_0}{T_0+3}\frac{T_0-1}{T_0+2}\right) \cdot \delta^{(T_0)}$$
$$+ \left(1 + \frac{T_0}{T_0+3}\right) \cdot \eta^{(T_0+1)} \cdot \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\bigg|_{\theta=\omega^{(T_0)}}$$
$$+ \eta^{(T_0+2)} \cdot \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\bigg|_{\theta=\omega^{(T_0+1)}} \tag{201}$$

$$\theta_{s,a}^{(T_0+M)} = \theta_{s,a}^{(T_0)} + \left(\frac{T_0-1}{T_0+2} + \frac{T_0}{T_0+3}\frac{T_0-1}{T_0+2} + \sum_{\tau=3}^{M}\frac{T_0+1}{T_0+\tau+1}\frac{T_0}{T_0+\tau}\frac{T_0-1}{T_0+\tau-1}\right) \cdot \delta^{(T_0)}$$
$$+ \cdots \tag{202}$$

where (202) holds by expanding (198)-(201) iteratively.

Note that for large enough $M \in \mathbb{N}$,

$$\sum_{\tau=3}^{M}\frac{(T_0+1)T_0(T_0-1)}{(T_0+\tau+2)(T_0+\tau+1)(T_0+\tau)}$$
$$= (T_0+1)T_0(T_0-1)\sum_{\tau=3}^{M}\frac{1}{2}\left(\frac{1}{(T_0+\tau)(T_0+\tau+1)} - \frac{1}{(T_0+\tau+1)(T_0+\tau+2)}\right)$$
$$= (T_0+1)T_0(T_0-1)\cdot\frac{1}{2}\left(\frac{1}{(T_0+3)(T_0+4)} - \frac{1}{(T_0+M+1)(T_0+M+2)}\right)$$
$$= \Theta(T_0). \tag{203}$$

Follow from (202) and (203), we could deduce that the momentum term $\delta^{(T_0)}$ in time T_0 results in an effect that is directly proportional to T_0. Therefore, it becomes crucial to establish the upper bound of the gradient norm to confine the adverse effects stemming from the momentum term. However, this endeavor involves addressing the intertwined challenge of estimating both the gradient norm and the momentum norm simultaneously.

I. Additional Experiments

I.1. Stochastic APG (SAPG)

In this subsection, we conduct an empirical evaluation of the performance of Stochastic Accelerated Policy Gradient (SAPG) on an MDP with 5 states and 5 actions, utilizing the true gradient. In the subsequent experimental results, we set the batch size to $B = 1$ and perform the experiments with 50 different seeds. It's important to highlight that the MDP used is identical to the one discussed in Section 6.1, and detailed configuration information is provided in Appendix F.

I.1.1. IMPLEMENTATION DETAIL

For ease of exposition, we provide the pseudo code for Stochastic Accelerated Policy Gradient (SAPG).

Algorithm 7 Stochastic Accelerated Policy Gradient (SAPG)

Input: Step size $\eta^{(t)} > 0$, batch size $B \in \mathbb{N}$.

Initialize: $\theta^{(0)} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $\omega^{(0)} = \theta^{(0)}$.

for $t = 1$ to T **do**

Sample a batch of $\mathcal{S}^{(t)} \times \mathcal{A}^{(t)} := \{(s_1^{(t)}, a_1^{(t)}), (s_2^{(t)}, a_2^{(t)}), \ldots, (s_B^{(t)}, a_B^{(t)})\}$ where state $s^{(t)}$ is sampled with probability $d_\mu^{\pi^{(t)}}(\cdot)$ and action $a^{(t)}$ is sampled with probability $\pi^{(t)}(\cdot|s^{(t)})$.

Calculate stochastic gradient $v^{(t)} = \sum_{s,a \in \mathcal{S}^{(t)} \times \mathcal{A}^{(t)}} \left. \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} \right|_{\theta = \omega^{(t-1)}}$.

$$\theta^{(t)} \leftarrow \omega^{(t-1)} + \eta^{(t)} v^{(t)}$$

$$\varphi^{(t)} \leftarrow \theta^{(t)} + \frac{t-1}{t+2}(\theta^{(t)} - \theta^{(t-1)})$$

$$\omega^{(t)} \leftarrow \begin{cases} \varphi^{(t)}, & \text{if } V^{\pi_\varphi^{(t)}}(\mu) \geq V^{\pi_\theta^{(t)}}(\mu), \\ \theta^{(t)}, & \text{otherwise.} \end{cases}$$

end for

(a)   (b)   (c)
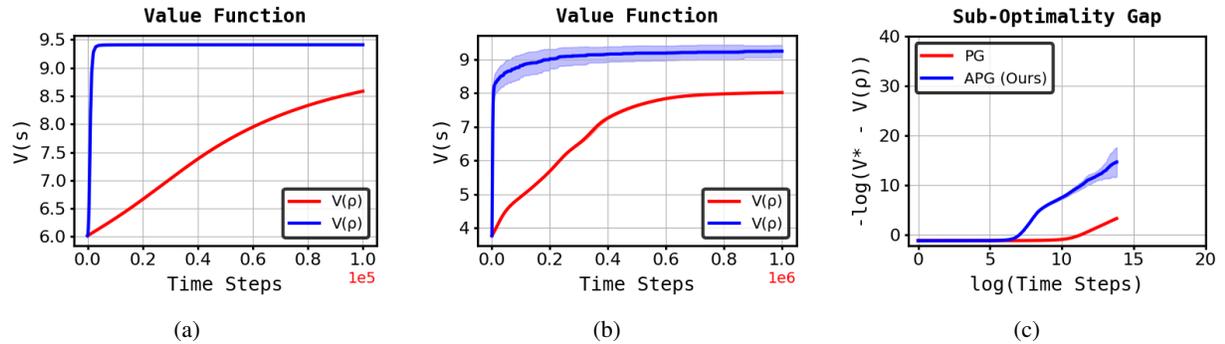
Figure 6. A comparison between the performance of SAPG and SPG under an MDP with 5 states, 5 actions, with the uniform and hard policy initialization: (a)-(b) show the value function under the uniform and the hard initialization, respectively. The optimal objective value $V^*(\rho) \approx 9.41$; (c) show the sub-optimality gaps under uniform initialization.

I.1.2. SAPG UNDER AN MDP WITH 5 STATES AND 5 ACTIONS

In Figure 6, it is evident that SAPG outperforms Stochastic PG (SPG) with a remarkable speed of convergence. Furthermore, the results presented in Figure 6(c) suggest that SAPG may exhibit a convergence rate of $O(\frac{1}{t^\alpha})$, where $2 > \alpha > 1$. This observation opens up the possibility of extending and exploring this potential in future research. Additionally, it's noteworthy

that under hard policy initialization (where the optimal action has the smallest initial probability), SPG tends to get stuck at a local optimum for an extended period, whereas SAPG does not exhibit this issue.

I.2. Empirical comparison between APG and NPG

To further highlight the empirical effectiveness of APG compared to NPG, we provide an additional experimental comparison in both BipedalWalker and the Atari games. It is worth noting that our implementation on Atari 2600 games for APG, NPG, and PG are based on the Stable Baselines3 and RL Baselines3 Zoo (Raffin, 2020; Raffin et al., 2021). The hyperparameters, exactly set to their default values, are detailed in Appendix F. Additionally, for BipedalWalker, we utilize the codebase of Spinning Up (Achiam, 2018).

In Figure 7, we observe that in BipedalWalker and Carnival, although NPG improves fast during the initial training phase, it tends to get stuck thereafter. Additionally, APG achieves acceleration compared to PG and reaches higher performance than both NPG and PG at the $1e7$ timesteps. Furthermore, in Riverraid, we observe that APG demonstrates notable acceleration compared to NPG.

(a) (b) (c)
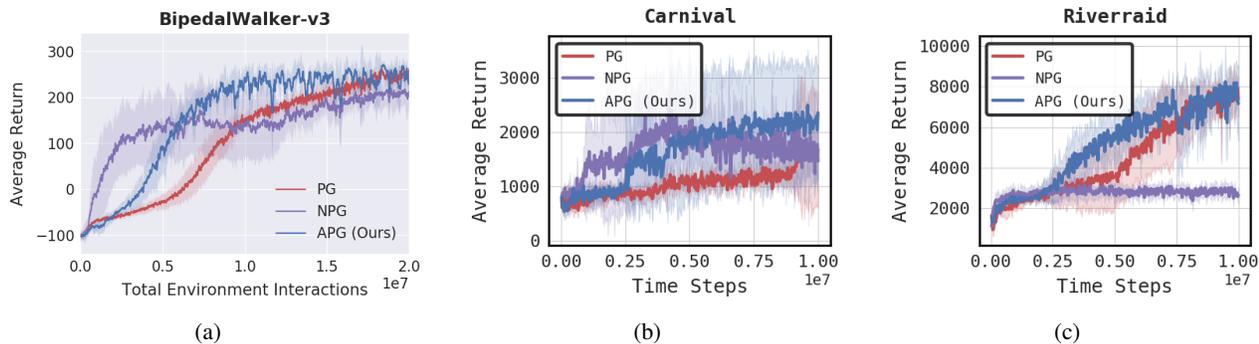
Figure 7. A comparison of the performance of APG and the benchmark algorithms in BipedalWalker and two Atari 2600 games. All the results are averaged over 5 random seeds (with the shaded area showing the range of mean $\pm$ std).

J. A Detailed Comparison of the Acceleration Methods for RL

In the optimization literature, there are several major categories of acceleration methods, including momentum, regularization, natural gradient, and normalization. Interestingly, these approaches all play an important role in the context of RL.

J.1. Momentum

Empirical Aspect: In practice, momentum is one of the most commonly used approaches for acceleration in policy optimization for RL, mainly due to its simplicity. Specifically, a variety of classic momentum methods (e.g., Nesterov momentum, heavy-ball method, and AdaGrad) have been already implemented in optimization solvers and deep learning frameworks (e.g., PyTorch) as a basic machinery and shown to improve the empirical convergence results than the standard policy gradient methods like A2C and PPO (e.g., see (Henderson et al., 2018)).

Theoretical Aspect: However, despite the popularity of these approaches in empirical RL, it has remained largely unknown whether momentum could indeed improve the theoretical convergence rates of RL algorithms. To address this fundamental question, we establish the very first convergence rate of Nesterov momentum in RL (termed APG in our paper) and show that APG could improve the rate of PG from $O(1/t)$ to $\tilde{O}(1/t^2)$ (in the constant step-size regime). Given the wide application of momentum, our analytical framework and the insights from our convergence result (e.g., local near-concavity) serve as an important first step towards better understanding the momentum approach in RL.

Theoretical Advantage: In addition, we would like to highlight that in the constant step-size regime (which is one of the most widely adopted setting in practice), our proposed APG indeed achieves a superior convergence rate of $\tilde{O}(1/t^2)$, which is better than the $O(1/t)$ rate of vanilla PG and the $O(1/t)$ rate of NPG with constant step sizes.

J.2. Regularization

Theoretical Aspect: In the convex optimization literature, it is known that adding a strongly convex regularizer achieves acceleration by changing the optimization landscape. In the context of RL, despite the non-concave objective, regularization has also been shown to achieve faster convergence, such as the log-barrier regularization with $O(1/\sqrt{t})$ rate (Agarwal et al., 2021) and the entropy regularization (i.e., adding a policy entropy bonus to the reward) with linear convergence rate $O(e^{-ct})$ (Mei et al., 2020), both in the exact gradient setting with constant step sizes. Additionally, Lan (2023) presents linear convergence results when employing strongly convex regularizers.

Despite the above convergence results, one common attack on regularization approaches is that they essentially change the objective function to that of regularized MDPs and hence do not directly address the original objective in unregularized RL, as pointed out by (Mei et al., 2021b; Xiao, 2022).

J.3. Natural Gradients (A Special Case of Policy Mirror Descent)

Natural policy gradient (NPG), another RL acceleration technique, borrows the idea from the natural gradient, which uses the inverse of Fisher information matrix as the preconditioner of the gradient and can be viewed as achieving approximate steepest descent in the distribution space (Amari, 1998; Kakade, 2001). Moreover, under direct policy parameterization, NPG is also known as a special case of policy mirror descent (PMD) with KL divergence as the proximal term (Shani et al., 2020; Xiao, 2022). Regarding the convergence rates, NPG has been shown to achieve: (i) $O(1/t)$ rate with constant step sizes (Agarwal et al., 2021; Xiao, 2022); (ii) linear convergence either with adaptively increasing step sizes (Khodadadian et al., 2021) or non-adaptive exponentially growing step sizes (Xiao, 2022). Notably, under direct policy parameterization and geometrically increasing step sizes, similar linear convergence results can also be established for the more general PMD method (Xiao, 2022).

Empirical Issues: Despite the above convergence rates, as also pointed out by (Mei et al., 2021b), it is known that NPG in general requires solving a costly optimization problem in each iteration (cf. (Kakade, 2001; Agarwal et al., 2021)) even with the help of compatible approximation theorem (Kakade, 2001) (unless the policy parametrization adopts some special forms, e.g., log-linear policies (Yuan et al., 2022)). Notably, the TRPO algorithm, a variant of NPG, is known to suffer from this additional computational overhead. Similarly, the PMD in general also requires solving a constrained optimization problem in each policy update (Xiao, 2022). This computational complexity could be a critical factor for RL in practice.

J.4. Normalization

Theoretical Aspect: In the exact gradient setting, the normalization approach, which exploits the non-uniform smoothness by normalizing the true policy gradient by its L2 gradient norm, has also been shown to achieve acceleration in RL. Specifically, the geometry-aware normalized PG (GNPG) has been shown to exhibit linear convergence rate under softmax policies (Mei et al., 2021b). This rate could be largely attributed to the effective increasing step size induced by the normalization technique.

Empirical Issues: Despite the linear convergence property, it has been already shown that GNPG could suffer from divergence in the standard on-policy stochastic setting, even in simple 1-state MDPs. This phenomenon is mainly due to the committal behavior induced by the increasing effective step size. This behavior could substantially hinder the use of GNPG in practice.