Accelerated Policy Gradient: On the Nesterov Momentum for Reinforcement Learning

Yen-Ju Chen [*1] **Nai-Chieh Huang** [*1] **Ping-Chun Hsieh** [1]

Abstract

Policy gradient methods have recently been shown to enjoy global convergence at a $\Theta(1/t)$ rate in the non-regularized tabular softmax setting. Accordingly, one important research question is whether this convergence rate can be further improved, with only first-order updates. In this paper, we answer the above question from the perspective of momentum by adapting the celebrated Nesterov's accelerated gradient (NAG) method to reinforcement learning (RL), termed *Accelerated Policy Gradient* (APG). To demonstrate the potential of APG in achieving faster global convergence, we start from the bandit setting and formally show that with the true gradient, APG with softmax policy parametrization converges to an optimal policy at a $\tilde{O}(1/t^2)$ rate. To the best of our knowledge, this is the first characterization of the global convergence rate of NAG in the context of RL. Notably, our analysis relies on one interesting finding: Regardless of the initialization, APG could end up reaching a locally-concave regime, where APG could benefit significantly from the momentum, within finite iterations. By means of numerical validation, we confirm that APG exhibits $\tilde{O}(1/t^2)$ rate in the bandit setting and still preserves the $\tilde{O}(1/t^2)$ rate in various Markov decision process instances, showing that APG could significantly improve the convergence behavior over the standard policy gradient.

[*]Equal contribution [1]Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan. 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>.

Workshop on New Frontiers in Learning, Control, and Dynamical Systems at the International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA, 2023.

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 the 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., 2021; 2022; Xiao, 2022). In a recent study by (Mei et al., 2020), they characterized the convergence rate of $\Theta(1/t)$ in the non-regularized tabular softmax setting. This convergence behavior aligns with that of the gradient descent algorithm for optimizing convex functions, despite that the RL objectives lack convex characteristics. Consequently, one critical open question arises as to whether this $\Theta(1/t)$ convergence rate can be further improved solely with first-order updates. In the realm of optimization, Nesterov's Accelerated Gradient (NAG) method, introduced by (Nesterov, 1983), is a first-order method originally designed for convex functions in order to improve the convergence rate to $O(1/t^2)$. Over the past decades since its introduction, to the best of our knowledge, NAG has never been formally analyzed or evaluated in the context of RL for its global convergence, mainly due to the non-concavity of the RL objective. Therefore, it is natural to ask the following research question: *Could Nesterov acceleration further improve the global convergence rate beyond the $\Theta(1/t)$ rate achieved by PG in 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; Krichene et al., 2015; Li & Lin, 2015; Su et al., 2014; Muehlebach & Jordan, 2019; Carmon et al., 2018), there remain several fundamental challenges in establishing the global convergence in the context 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 to a stationary point. Under these conditions, we cannot determine global convergence in RL. Furthermore, it is not possible to assess whether the convergence rate improves beyond $\Theta(1/t)$ based on these results. (ii) *The absence of monotonic improvement due to Nesterov acceleration*: Nesterov acceleration utilizes the momentum to enhance the convergence rate. Nevertheless, because of the presence of the momentum term, APG does not guarantee monotonic improvement in every iteration. This is a notable distinction from the standard PG, which exhibits monotonic improvement and ensures the existence of a limiting value function. Without monotonicity, the behavior of value functions in the limit remains uncertain. (iii) *Inherent characteristics of the momentum term*: From an analytical perspective, the momentum term demonstrates intricate interactions with the previous updates. As a result, accurately quantifying the specific impact of momentum during the execution of APG poses a considerable challenge. Moreover, despite the valuable insights provided by the non-uniform Polyak-Łojasiewicz (PL) condition in the field of RL proposed by (Mei et al., 2020), the complex influences of the momentum term present a significant obstacle in determining the convergence rate of APG. (iv) *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; Jaggi, 2013; Ghadimi & Lan, 2016). However, in the case of softmax parameterization, the parameter of the optimal action tends to approach infinity. As a result, the norm involved in the sub-optimality gap becomes infinite, thereby hindering the characterization of the desired convergence rate.

Our Contributions. Despite the above challenges, we present an affirmative answer to the research question described above and provide the first characterization of the global convergence rate of NAG in the context of RL. As an important and highly non-trivial first step, we start from the bandit setting (i.e., single-state MDPs), which serves as a stylish setting subject to all the above technical challenges (i)-(iv), and establish that APG could achieve global convergence at a rate of $\tilde{O}(1/t^2)$. Specifically, we present useful insights and novel techniques to tackle the technical challenges: Regarding (i), we show that the RL objective enjoys local concavity in the proximity of the optimal policy, despite its non-concave global landscape. To better illustrate this, we start by presenting a motivating two-action bandit example, which demonstrates the local concavity directly via the corresponding sigmoid-type characteristic. Subsequently, we show that this intuitive argument could be extended to the general multi-action case. Regarding (ii) and (iii), we show that the locally-concave region is *absorbing* in the sense that even with the effect of the momentum term, the policy parameter could stay in the locally-concave region indefinitely once it enters this region. This result is obtained by carefully quantifying the cumulative effect of each momentum term. Regarding (iv), we introduce the concept of *effective domain*, which essentially captures the growth rate of the norm of the policy parameters, and thereby characterize the effective domain of APG.

We summarize the contributions of this paper as follows:

- We propose APG, which leverages the Nesterov's momentum scheme to accelerate the convergence performance of PG for RL.
- To demonstrate the potential of achieving fast global convergence, we start from the bandit setting and formally establish that APG enjoys a $\tilde{O}(1/t^2)$ convergence rate under softmax policy parameterization[1]. To achieve this, we present several novel insights into RL and APG, including the local concavity property as well as the absorbing behavior and the effective domain of APG. Moreover, we further show that the derived rate for APG is tight (up to a logarithmic factor) by providing a $\Omega(1/t^2)$ lower bound of the sub-optimality gap.
- Through numerical validation on both bandit and MDP problems, we confirm that APG exhibits $\tilde{O}(1/t^2)$ rate and hence substantially improves the convergence behavior over the standard PG.

2. Related Work

Policy Gradient. Policy gradient (Sutton et al., 1999) is a popular reinforcement learning 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., 2015), and proximal policy optimization (PPO) (Schulman et al., 2017). Recently, policy gradient methods have been shown to enjoy global convergence. The global convergence of standard policy gradient methods under various settings has been proven by (Agarwal et al., 2021). Furthermore, (Mei et al., 2020) characterizes a $O(1/t)$ convergence rate of policy gradient based on a Polyak-Lojasiewicz condition under the non-regularized tabular softmax parameterization. Moreover, (Fazel et al., 2018; Liu et al., 2020; Wang et al., 2021; Xiao, 2022) conduct theoretical analyses of several variants of policy gradient methods under various policy parameterizations and establish the global convergence guarantees

[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 Section 6.2 for a detailed discussion.

for these methods. In our work, we rigorously establish the accelerated $\tilde{O}(1/t^2)$ convergence rate for the proposed APG method under softmax parameterization.

Accelerated Gradient. Accelerated gradient methods (Nesterov, 1983; 2005; Beck & Teboulle, 2009b) play a pivotal role in the optimization literature due to their ability to achieve faster convergence rates when compared to the conventional gradient descent algorithm. Notably, in the convex regimes, the accelerated gradient methods enjoy a convergence rate as fast as $O(1/t^2)$, surpassing the limited convergence rate $O(1/t)$ offered by the gradient descent algorithm. The superior convergence behavior could also be characterized from the perspective of ordinary differential equations (Su et al., 2014; Krichene et al., 2015; Muehlebach & Jordan, 2019). Additionally, in order to enhance the performance of accelerated gradient methods, several variants have been proposed. For instance, (Beck & Teboulle, 2009a) proposes a variant of the proximal accelerated gradient method which incorporates monotonicity to further improve its efficiency. (Ghadimi & Lan, 2016) presents 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) proposes a monotone accelerated gradient approach with sufficient descent, providing convergence guarantees to stationary points for non-convex problems. The above list of works is by no means exhaustive and is only meant to provide a brief overview of the accelerated gradient methods. Our paper introduces APG, a novel approach that combines accelerated gradient methods and policy gradient methods for RL. This integration enables a substantial acceleration of the convergence rate compared to the standard policy gradient method.

3. Preliminaries

Markov Decision Processes. For a finite set $\mathcal{X}$, we use $\Delta(\mathcal{X})$ to denote a 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 $\mathcal{A}$, (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 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 $\rho \in \Delta(\mathcal{S})$. Given a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, the value of state s under π is defined as

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

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

$$Q^{\pi}(s,a) \coloneqq r(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) V^{\pi}(s') \ , \tag{2}$$

$$A^{\pi}(s,a) \coloneqq Q^{\pi}(s,a) - V^{\pi}(s), \tag{3}$$

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

$$d^{\pi}_{s_0}(s) \coloneqq (1-\gamma) \sum_{t=0}^{\infty} \gamma^t Pr(s_t = s | s_0, \pi, \mathcal{P}), \tag{4}$$

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

$$V^{\pi^*}(\rho) = \max_{\pi : \mathcal{S} \to \Delta(\mathcal{A})} V^{\pi}(\rho). \tag{5}$$

For ease of exposition, we denote $V^*(\rho) \coloneqq V^{\pi^*}(\rho)$.

Although obtaining the true initial state distribution ρ in practical problems is challenging, it is fortunate that this challenge can be eased by considering other *surrogate* initial state distribution μ that are strictly positive for every state $s \in \mathcal{S}$. Notably, it can be demonstrated in the following theoretical proof that even in the absence of knowledge about ρ, convergence guarantees for $V^*(\rho)$ can still be obtained under the condition of strictly positive μ. Hence, we make the following assumption, which has also been adopted by (Agarwal et al., 2021) and (Mei et al., 2020).

Assumption 1. ***(Strict positivity of surrogate initial state distribution).*** *The surrogate initial state distribution satisfies* $\min_s \mu(s) > 0$.

Since $\mathcal{S} \times \mathcal{A}$ is finite, without loss of generality, we assume that the one-step reward is bounded in the $[0,1]$ interval:

Assumption 2. ***(Bounded reward).*** $r(s,a) \in [0,1], \forall s \in \mathcal{S}, a \in \mathcal{A}$.

For simplicity, we assume that the optimal action is unique. This assumption can be relaxed by considering the sum of probabilities of all optimal actions in the theoretical results.

Assumption 3. ***(Unique optimal action).*** *There is a unique optimal action a^* for each state $s \in \mathcal{S}$.*

Softmax Parameterization. For unconstrained $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, the softmax parameterization of θ is defined as $\pi_{\theta}(\cdot|s) \coloneqq \text{softmax}(\theta_{s,\cdot})$, where for all $a \in \mathcal{A}$. We use the shorthand for denoting the optimal policy $\pi^* \coloneqq \pi_{\theta^*}$, where θ^* is the optimal policy parameter.

$$\pi_{\theta}(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta_{s,a'})}. \tag{6}$$

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

Input: Learning rate $\eta = \frac{1}{L}$, where L is the Lipschitz constant of the objective function $V^{\pi_\theta}(\mu)$.
Initialize: $\theta_1(s,a)$ for all (s,a).
for $t = 1$ to T **do**

$$\theta_{t+1} \leftarrow \theta_t + \eta \nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\theta^{(t)}} \tag{7}$$

end for

Policy Gradient. Policy gradient (Sutton et al., 1999) is a policy search technique that involves defining a set of policies parametrized by a finite-dimensional vector θ and searching for an optimal policy π^* by exploring the space of parameters. This approach reduces the search for an optimal policy to a search in the parameters space. In policy gradient methods, the parameters are updated by the gradient of the function $f : \theta \to V^{\pi_\theta}(\mu)$ that maps policy parameters to the expected cumulative reward under an initial state distribution $\mu \in \Delta(\mathcal{S})$. The following Algorithm 1 presents the pseudo code of PG provided by (Mei et al., 2020).

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.

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. Additionally, in Section 4.3, we underline the main technical challenges involved in our analysis, particularly the absence of monotonic improvement that is typically observed in standard policy gradient methods.

4.1. Accelerated Policy Gradient

We propose Accelerated Policy Gradient (APG) and present the pseudo code of our algorithm in Algorithm 2. Our algorithm design draws inspiration from the renowned and elegant Nesterov's accelerated gradient updates as introduced in (Su et al., 2014). For the sake of comparison, we include the pseudo code of the approach in (Su et al., 2014) as Algorithm 4 in Appendix A. We adapt these updates to the reinforcement learning objective, specifically $V^{\pi_\theta}(\mu)$. It is important to note that we will specify the learning rate $\eta^{(t)}$ in Lemma 2, as presented in Section 5.

In Algorithm 2, the gradient update is performed in (8). Following this, (9) calculates the momentum for our parameters, which represents a fundamental technique employed in accelerated gradient methods. It is worth noting that in (8), 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 apart (8) from the standard policy gradient updates (Algorithm 1).

Algorithm 2 Accelerated Policy Gradient (APG)

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

$$\theta^{(t)} \leftarrow \omega^{(t-1)} + \eta^{(t)} \nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\omega^{(t-1)}} \tag{8}$$

$$\omega^{(t)} \leftarrow \theta^{(t)} + \frac{t-1}{t+2}(\theta^{(t)} - \theta^{(t-1)}) \tag{9}$$

end for

4.2. A Motivating Example of APG

Prior to the exposition of convergence analysis, we aim to provide further insights into why APG has the potential to attain a convergence rate of $\tilde{O}(1/t^2)$, especially under the intricate non-concave objectives in reinforcement learning.

Consider a simple two-action bandit with actions a^*, a_2 and reward function $r(a^*) = 1, r(a_2) = 0$. Accordingly, the objective we aim to optimize is $\mathbb{E}_{a\sim\pi_\theta}[r(a)] = \pi_\theta(a^*)$. 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 $\pi_\theta(a^*) \geq 0.5$. The detailed derivation is provided in Appendix E. The aforementioned observation implies that the objective function demonstrates *local concavity* when $\pi_\theta(a^*) \geq 0.5$. Since $\pi^*(a^*) = 1$, it follows that the objective function exhibits local concavity for 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 global convergence rate of APG in the general multi-action bandit setting in Section 5.

4.3. Non-Monotonic Improvement Under APG

In this subsection, we illustrate the difficulties involved in analyzing the convergence of APG, compared to the standard policy gradient methods through a numerical experiment. In contrast to the standard policy gradient (PG) method, which exhibits monotonic improvement, Accelerated Policy Gradient (APG) 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. The detailed configuration is provided in Appendix E. As shown in Figure 1, the one-step improvement becomes negative around epoch 180 and provides nearly zero improvement after that point. Notably, the asymptotic global 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 APG poses a fundamental challenge in analyzing and achieving global convergence to an optimal policy.

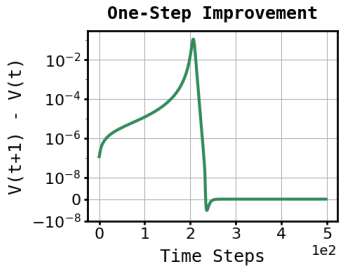


Figure 1. The one-step improvement of APG on a three-action bandit problem.

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 bandit setting under softmax parameterization. In the subsequent analysis, we assume that Assumption 1, 2, 3 are satisfied. Due to the space limit, we defer the proofs of the following theorems to Appendix C, 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 section. We highlight the features in our analysis as follows:

(C1) Lack of monotonic improvement under APG: Recall from Section 4.3 that APG 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. To establish the existence of the limiting value function, we demonstrate that the value will always converge to the reward of one of the arms, even in the scenario where monotonic improvement is lacking. Please see refer to Lemma 10 for further details.

(C2) The existing results of first-order stationary points 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 to a first-order stationary point), as shown in (Agarwal et al., 2021). While it appears natural to follow the same approach for APG, one fundamental challenge is that the existing results of NAG for non-convex problems 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. This is yet another salient difference between APG and PG. To address the issue of possibly unbounded domain , we need to characterize the *effective domain*, by considering the maximum growth rate of $\|\theta\|$ under APG. Please refer to Appendix B.2 for more comprehensive information.

(C3) 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 cumulation 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 all the three challenges and establish the asymptotic global convergence of APG as follows. Recall that optimal objective is defined by (5).

Theorem 1. ***(Global convergence under softmax parameterization)*** *Consider a tabular softmax parameterized policy π_θ. Under APG with $\eta^{(t)} = \frac{t}{t+1}\cdot\frac{1}{5}$, we have $V^{\pi_\theta^{(t)}}(s) \to V^*(s)$ as $t\to\infty$, for all $s\in\mathcal{S}$.*

The complete proof is provided in Appendix C. Specifically, we address the challenge (C1) in Appendix C.1, (C2) in

Appendix B.2, and (C3) in Appendix B.1 and C.2-C.3.

Remark 1. Note that Theorem 1 suggests the use of a time-varying learning rate $\eta^{(t)}$. This choice is related to one inherent issue of NAG: the choices of learning rate 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 could be locally concave around the optimal policy despite its non-concavity of the global landscape. To enable the use of the same learning rate scheme throughout the whole training process, we find that incorporating the ratio $t/(t+1)$ could achieve the best of both world.

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 in the bandit setting under softmax parameterization. In this case, (1) reduces to maximizing the expected reward $\mathbb{E}_{a\sim\pi_\theta}[r(a)] = \pi_\theta^\top r$.

Recall from Section 4.2, the objective function $\mathbb{E}_{a\sim\pi_\theta}[r(a)]$ can exhibit local concavity. In order to attain this regime of local concavity, we establish the sufficient condition in Lemma 1.

Lemma 1. ***(Local Concavity; Informal).*** *The function $\theta \to \pi_\theta^\top r$ is concave if $\theta_{a^*} - \theta_a > \delta$ for some $\delta > 0$, for all $a \neq a^*$.*

The information regarding the time-independent constant δ mentioned in Lemma 1 is provided in Appendix D.

Remark 2. In simpler terms, Lemma 1 states that when the action probability of the optimal action a^* significantly outweighs the probability of the other actions, the objective function in the bandit setting enters a region of local concavity. It is crucial to emphasize that this lemma provides a sufficient condition for the objective function to demonstrate the local concavity. This condition is not specific to APG and can actually be applicable to other algorithms or scenarios as well.

After deriving Lemma 1, our goal is to investigate whether APG can reach the local concavity regime within a 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 local concavity through the APG updates and remain within this region without exiting.

Lemma 2. *Consider a tabular softmax parameterized policy π_θ. Under APG with $\eta^{(t)} = \frac{t}{t+1}\cdot\frac{1}{5}$, given any $\delta > 0$, there exists a finite time T such that for all $t > T$, we have $\theta_{a^*} - \theta_a > \delta$, for all $a \neq a^*$.*

In the proof of Lemma 2, it is crucial to require that the partial derivative is nonnegative for the optimal action and nonpositive for the sub-optimal action within finite time. This demands that our value function does not converge below any sub-optimal reward in the bandit setting. Therefore, we introduce Lemma 3 to assure that the optimal action parameter increases before our value function surpasses sub-optimal rewards.

Lemma 3. *Under APG, we have $\inf_{t\geq 0}\pi_\theta^{(t)}(a^*) > 0$.*

Remark 3. While (Mei et al., 2020) presents a similar argument to Lemma 3, it is crucial to highlight that the application and utilization of this argument vary between our proof and theirs. In their proof, they establish a lower bound on the one-step improvement using the PL condition and take the telescoping sum over multiple time steps, necessitating the characterization of the infimum of the probability of the optimal action over time. However, due to the intrinsic behavior of momentum in our approach, we are not allowed to characterize the improvement in the same manner as they do. Instead, we utilize this result to ensure that our policy can satisfy Lemma 2 and attain local concavity. This allows us to establish the convergence properties of our approach.

With the results of Lemma 1, 2 and 3, we are able to establish the main result in the bandit setting for APG under softmax parameterization, which is an $\tilde{O}(1/t^2)$ convergence rate.

Theorem 2. *Consider a tabular softmax parameterized policy π_θ. Under APG with $\eta^{(t)} = \frac{t}{t+1}\cdot\frac{1}{5}$, there exists a finite time T such that for all $t > T$, we have:*

$$\left(\pi^* - \pi_\theta^{(t)}\right)^\top r \leq \frac{|\mathcal{A}| - 1}{(t-T)^2 + |\mathcal{A}| - 1} \tag{10}$$

$$+ \frac{10(2+T)\left(\left\|\theta^{(T)}\right\| + 2\ln(t-T)\right)^2}{t(t+1)}. \tag{11}$$

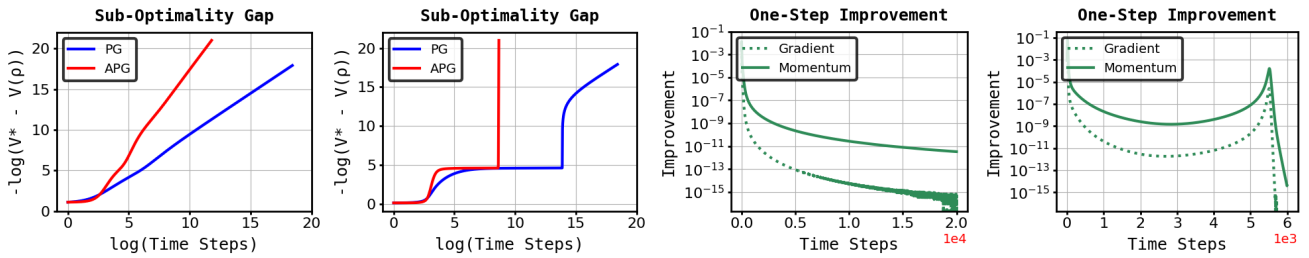
Remark 4. It is important to note that the logarithmic factor in the sub-optimality gap is a consequence of the unbounded nature of the optimal parameter in softmax parameterization. Furthermore, the finite time T mentioned in Theorem 2 guarantees that the policy enters the local concavity regime as established in Lemma 2. Therefore, for a concave function, we could utilize the $O(1/t^2)$ results of the original NAG.

5.3. Lower Bounds

In this subsection, we further present a lower bound of sub-optimality gap for APG as follows.

Theorem 3. *Consider a simple two-armed bandit with actions a^*, a_2, reward function $r(a^*) = 1, r(a_2) = 0$, and initial policy parameters $\theta_{a^*}^{(0)} = \theta_{a_2}^{(0)} = 0$. Under APG with $\eta^{(t)} = \frac{t}{t+1}\cdot\frac{1}{5}$, for all $t > 0$, we have:*

$$\left(\pi^* - \pi_\theta^{(t)}\right)^\top r = \Omega(\frac{1}{t^2}) \tag{12}$$



(a) (b) (c) (d)

Figure 2. A comparison between the performance of APG and PG under a 3-armed bandit with uniform initialization ($\theta^{(0)} = [0, 0, 0]$) and hard initialization ($\theta^{(0)} = [1, 3, 5]$ and hence the optimal action has the smallest initial probability): (a)-(b) show the sub-optimality gaps of APG and PG under uniform and hard initializations, respectively; (c)-(d) show the one-step improvements of APG from the momentum (i.e., ${\pi_\omega^{(t)}}^\top r - {\pi_\theta^{(t)}}^\top r$) and the gradient (i.e., ${\pi_\theta^{(t+1)}}^\top r - {\pi_\omega^{(t)}}^\top r$) under uniform and hard initializations, respectively.

The above lower bound indicates that the $\tilde{O}(1/t^2)$ convergence rate of APG in Theorem 2 is actually tight up to a logarithmic factor.

Remark 5. Notably, Theorem 3 suggests that the lower bound $\Omega(1/t^2)$ of APG holds for *some* MDPs (and not necessarily for all MDPs). This result is in essence different from the lower bound of PG in (Mei et al., 2020), which shows that the lower bound $\Omega(1/t)$ of PG holds for *any* MDP. On the other hand, as the lower bound in Theorem 3 focuses on APG, one interesting research question is to explore whether there exists any first-order method that could achieve a convergence rate beyond $\tilde{O}(1/t^2)$ under softmax policies.

6. Discussions

6.1. Numerical Validation of the Convergence Rates

In this subsection, we empirically validate the convergence rate of APG by conducting experiments on a 3-armed bandit as well as an MDP with 5 states and 5 actions. The detailed configuration is provided in Appendix E. Codes are available at `https://github.com/NYCU-RL-Bandits-Lab/APG`.

(Bandit) To validate the convergence rate of both APG and PG, we first conduct a 3-armed bandit experiment with both a uniform initialization ($\theta^{(0)} = [0, 0, 0]$) and a hard initialization ($\theta^{(0)} = [1, 3, 5]$ and hence the optimal action has the smallest initial probability). First, upon plotting the sub-optimality gaps of PG and APG under uniforma initialization on a log-log graph in Figure 2(a), we observe that they exhibit a slope of approximately 1 and 2, respectively, matching the convergence rate of $O(1/t)$ and $\tilde{O}(1/t^2)$ shown in Theorem 2. Under the hard initialization, Figure 2(b) shows that APG could escape from sub-optimality much faster than PG and thereby enjoys fast convergence. Moreover, Figure 2(c)-2(d) further show that the momentum term in APG does contribute substantially in terms of policy improvement, under both initializations.

(MDP) We proceed to validate the convergence rate on an MDP with 5 states and 5 actions: (i) *Uniform initialization*: The training curves of value functions and sub-optimality gap for both APG and PG are depicted in Figure 3. As shown by the log-log graph in Figure 3(a), the sub-optimality gap curve of APG exhibits a remarkable alignment with the $\tilde{O}(1/t^2)$ curve. Moreover, Figures 3(b) and 3(c) present the training curves of the value functions for APG and PG. Notably, the scale of the required training steps in Figure 3(b) is considerably smaller than that of Figure 3(c), highlighting the significantly faster convergence of APG compared to PG. 3(d) further confirms that the momentum term in APG still contributes substantially in terms of policy improvement in the MDP case. The above observations demonstrate the potential and efficacy of the Nesterov acceleration employed in Algorithm 2. (ii) *Hard initialization*: We also evaluate APG and PG under a hard policy initialization. Figure 4 shows that APG could still escape from sub-optimality much faster than PG in the MDP case. This further showcases APG's superiority over PG.

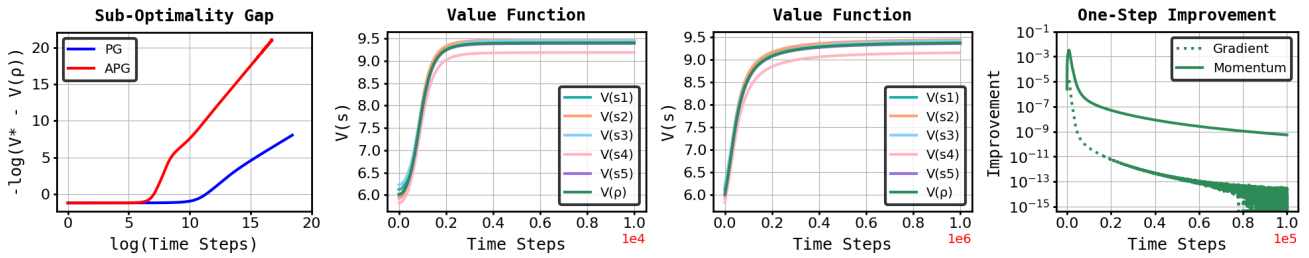
6.2. Lower Bounds of Policy Gradient

Regarding the fundamental capability of PG, (Mei et al., 2020) has presented a lower bound of sub-optimality gap for PG. For ease of exposition, we restate the theorem in (Mei et al., 2020) as follows.

Theorem 4. ***(Lower bound of sub-optimality gap for PG in Theorem 10 of (Mei et al., 2020)).*** *Take any MDP. For large enough $t \geq 1$, using Algorithm 1 with $\eta \in (0, 1]$,*

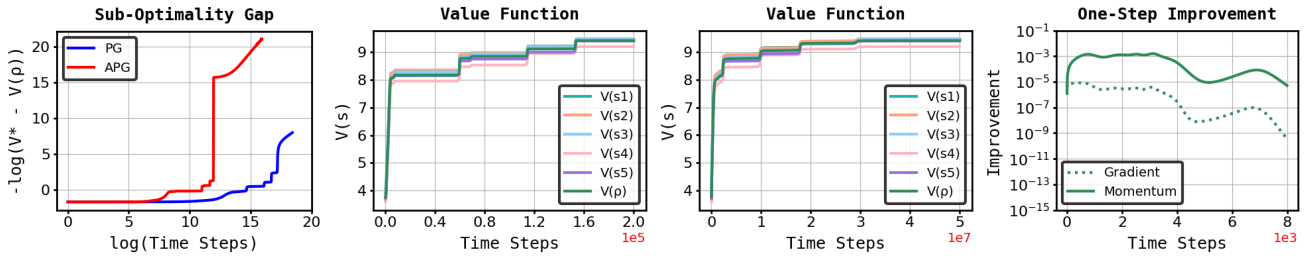
$$V^*(\mu) - V^{\pi_\theta^{(t)}}(\mu) \geq \frac{(1-\gamma)^5 \cdot (\Delta^*)^2}{12 \cdot t}, \qquad (13)$$

where $\Delta^* := \min_{s \in \mathcal{S}, a \neq a^*(s)} \{Q^*(s, a^*(s)) - Q^*(s, a)\} > 0$ is the optimal value gap of the MDP.



(a) (b) (c) (d)

Figure 3. A comparison between the performance of APG and PG under an MDP with 5 states, 5 actions, and *uniform policy initialization*: (a) shows the sub-optimality gap of APG and PG via a log-log plot; (b)-(c) show the per-state value functions of APG and PG (and the optimal objective value $V^*(\rho) \approx 9.41$); (d) presents the one-step improvement of APG from the momentum (i.e., $V^{\pi_\omega^{(t)}}(\rho) - V^{\pi_\theta^{(t)}}(\rho)$) and the gradient (i.e., $V^{\pi_\theta^{(t+1)}}(\rho) - V^{\pi_\omega^{(t)}}(\rho)$).



(a) (b) (c) (d)

Figure 4. A comparison between the performance of APG and PG under an MDP with 5 states, 5 actions, and a *hard policy initialization* (and the detailed captions of (a)-(d) exactly follow those in Figures 3(a)-3(d)).

Recall that PG has been shown to have a $O(1/t)$ convergence rate (Mei et al., 2020). Therefore, Theorem 4 indicates that the $O(1/t)$ convergence rate achievable by PG cannot be further improved.

On the other hand, despite the lower bound shown in Theorem 4 by (Mei et al., 2020), our results of APG do not contradict theirs. Specifically, while both APG and PG are first-order methods that rely 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 gradient with respect to parameters that PG cannot attain. As a result, APG exhibits improved convergence behavior compared to PG. Our findings extend beyond the scope of PG, demonstrating the advantages of APG in terms of convergence rate and overall performance.

6.3. Challenges of Convergence Analysis of APG for the General MDPs

Notably, our analysis in Section 5 paves the way towards characterizing the convergence rate of APG for general MDPs. We expect that the key steps and challenges include: (i) *Showing that the limiting value functions exist*: This could be done via showing convergence to stationary points, under a good characterization of the effective domain of APG. (ii) *Establishing the asymptotic convergence for the MDP case*: With (i), this could be achieved by reusing the arguments in Appendices C.2-C.3. (iii) *Establishing the convergence rate for the MDP case*: With (ii), this could be achieved by extending the *local concavity* to the Q-functions. That said, the challenge lies in that the ordering of the Q-values could vary continuously during learning.

7. Concluding Remarks

The Nesterov's Accelerated Gradient method, proposed in the optimization literature almost four decades ago, provides a powerful first-order scheme for fast convergence under a broad class of optimization problems. Over the past decades since its introduction, NAG has never been formally analyzed or evaluated in the context of RL for its global convergence, 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 show that APG can converge to a globally optimal policy at a $\tilde{O}(1/t^2)$ rate in the multi-action bandit setting. This demonstrates the potential of APG in attaining fast convergence in RL.

On the other hand, our work also leaves open several interesting research questions: (i) Given the convergence rate in the bandit setting, one important future work would be to extend the result in Section 5 to the MDP case. (ii) Given that our convergence rate is tight up to a logarithmic factor, it remains open whether this limitation could be addressed by closing this logarithmic gap. (iii) As this 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.

Acknowledgements

This material is based upon work partially supported by the National Science and Technology Council (NSTC), Taiwan under Contract No. 110-2628-E-A49-014 and Contract No. 111-2628-E-A49-019, and based upon work partially supported by the Higher Education Sprout Project of the National Yang Ming Chiao Tung University and Ministry of Education (MOE), Taiwan.

References

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.

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.

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

Boyd, S., Boyd, S. P., and Vandenberghe, L. *Convex optimization*. Cambridge University Press, 2004.

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

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.

Jaggi, M. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, pp. 427–435, 2013.

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

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

Krichene, W., Bayen, A., and Bartlett, P. L. Accelerated mirror descent in continuous and discrete time. *Advances in Neural Information Processing Systems*, 28, 2015.

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

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, 2021.

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.

Muehlebach, M. and Jordan, M. A dynamical systems perspective on Nesterov acceleration. In *International Conference on Machine Learning*, pp. 4656–4662, 2019.

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.

Nesterov, Y. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103:127–152, 2005.

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

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

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.

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.

Appendix

A. Supporting Algorithm

For ease of exposition, we restate the accelerated gradient algorithm stated in (Ghadimi & Lan, 2016) as follows. Note that we've made several revisions so that one could easily compare Algorithm 2 and Algorithm 3: (i) We have exchanged the positions of the superscript and subscript. (ii) We've replaced the original gradient symbol with the gradient of our objective (i.e. $\nabla_\theta V^{\pi_\theta}(\mu)$). (iii) We've replaced the time variable k with t. (iv) We've changed the algorithm into ascent algorithm (i.e. the sign in (15) and (16) is plus instead of minus.)

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

Input: $\theta^{(0)} \in \mathbb{R}^n$, $\{\alpha^{(t)}\}$ s.t. $\alpha^{(1)} = 1$ and $\alpha^{(t)} \in (0,1)$ for any $t \geq 2$, $\{\beta^{(t)} > 0\}$, and $\{\lambda^{(t)} > 0\}$.

0. Set the initial points $\theta_{ag}^{(0)} = \theta^{(0)}$ and $t = 1$.

1. Set

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

2. Compute $\nabla\Psi(\theta_{md}^{(t)})$ and set

$$\theta^{(t)} = \theta^{(t-1)} + \lambda^{(t)}\nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\theta_{md}^{(t)}}, \tag{15}$$

$$\theta_{ag}^{(t)} = \theta_{md}^{(t)} + \beta^{(t)}\nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\theta_{md}^{(t)}}. \tag{16}$$

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

Lemma 4. *(Equivalence between Algorithm 2 and Algorithm 3) Using Algorithm 2 and setting $\alpha^{(t)}\lambda^{(t)} = \beta^{(t)}$ and $\alpha^{(t)} = \frac{2}{t+1}, \forall t \geq 1$ leads to Algorithm 3 where $\eta^{(t)} = \beta^{(t)}$.*

Remark 6. Lemma 4 shows that our Algorithm 2 is equivalent to Algorithm 3 so that one can leverage the theoretical result stated in (Ghadimi & Lan, 2016) and adopt the general accelerated algorithm simultaneously.

Proof of Lemma 4. Since $\alpha^{(t)}\lambda^{(t)} = \beta^{(t)}$, by subtracting (16) from $\alpha^{(t)}$ times (15), we have:

$$\alpha^{(t)}\theta^{(t)} - \theta_{ag}^{(t)} = \alpha^{(t)}\theta^{(t-1)} - \theta_{md}^{(t)}. \tag{17}$$

Then, substituting $\theta_{md}^{(t)}$ in (17) by (14), we have:

$$\theta^{(t)} = \frac{\theta_{ag}^{(t)} + (1-\alpha^{(t)})\theta_{ag}^{(t-1)}}{\alpha^{(t)}}. \tag{18}$$

Plugging (18) back into (14), we get:

$$\theta_{md}^{(t)} = (1-\alpha^{(t)})\theta_{ag}^{(t-1)} + \alpha^{(t)}\theta^{(t-1)} \tag{19}$$

$$= (1-\alpha^{(t)})\theta_{ag}^{(t-1)} + \alpha^{(t)}\frac{\theta_{ag}^{(t-1)} + (1-\alpha^{(t-1)})\theta_{ag}^{(t-2)}}{\alpha^{(t-1)}} \tag{20}$$

$$= \theta_{ag}^{(t-1)} + \frac{\alpha^{(t)}(1-\alpha^{(t-1)})}{\alpha^{(t-1)}}(\theta_{ag}^{(t-1)} - \theta_{ag}^{(t-2)}). \tag{21}$$

So by (14) and (21), we could simplify Algorithm 3 into a two variables update:

$$\theta_{md}^{(t)} = \theta_{ag}^{(t-1)} + \frac{\alpha^{(t)}(1-\alpha^{(t-1)})}{\alpha^{(t-1)}}(\theta_{ag}^{(t-1)} - \theta_{ag}^{(t-2)}) \tag{22}$$

$$\theta_{ag}^{(t)} = \theta_{md}^{(t)} - \beta^{(t)}\nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\theta_{md}^{(t)}} \tag{23}$$

Finally, by plugging $\alpha^{(t)} = \frac{2}{t+1}$ and $\beta^{(t)} = \eta^{(t)}$, we reach our desired result:

$$\theta_{md}^{(t)} = \theta_{ag}^{(t-1)} + \frac{t-2}{t+1}(\theta_{ag}^{(t-1)} - \theta_{ag}^{(t-2)}) \tag{24}$$

$$\theta_{ag}^{(t)} = \theta_{md}^{(t)} - \eta^{(t)} \nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\theta_{md}^{(t)}} \tag{25}$$

Note that we've rearranged the ordering of (24) and (25) to reach our Algorithm 2. In summary, θ_{md}, θ_{ag} in (24) and (25) corresponds to ω, θ in Algorithm 2 respectively. And also we've turned the first step (24) into initializing ω in Algorithm 2 and follow the residual update.

□

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

Input: Learning rate $s = \frac{1}{L}$, where L is the Lipschitz constant of the objective function f.
Initialize: x_0 and $y_0 = x_0$.
for $t = 1$ to T **do**

$$x_k = y_{k-1} - s\nabla f(y_{k-1}) \tag{26}$$

$$y_k = x_k + \frac{k-1}{k+2}(x_k - x_{k-1}) \tag{27}$$

end for

B. Supporting Lemmas

B.1. Useful Properties

For ease of notation, we use $\nabla_{s,a}^{(t)}$ as the shorthand for $\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\big|_{\theta=\omega^{(t)}}$. Moreover, in the sequel, for ease of exposition, for any pair of positive integers (j,t), we define

$$G(j,t) := \begin{cases} 1 & \text{, if } t=j, \\ 1+\frac{j}{j+3} & \text{, if } t=j+1, \\ 1+\frac{j}{j+3}+\frac{(j+1)j}{(j+4)(j+3)} & \text{, if } t=j+2, \\ 1+\frac{j}{j+3}+\frac{(j+1)j}{(j+4)(j+3)}+\frac{(j+2)(j+1)j}{(j+5)(j+4)(j+3)} & \text{, if } t=j+3, \\ 1+\frac{j}{j+3}+\frac{(j+1)j}{(j+4)(j+3)}+\frac{(j+2)(j+1)j}{(j+5)(j+4)(j+3)}+\sum_{k=4}^{t-j}\frac{(j+2)(j+1)j}{(j+k+2)(j+k+1)(j+k)} & \text{, if } t\geq j+4 \\ 0 & \text{, otherwise.} \end{cases} \tag{28}$$

Lemma 5. *Under APG, we could express the policy parameter as follows:*

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

$$\theta_{s,a}^{(1)} = \eta^{(1)}\nabla_{s,a}^{(0)} + \theta_{s,a}^{(0)}, \tag{29}$$

$$\theta_{s,a}^{(2)} = \eta^{(2)}\nabla_{s,a}^{(1)} + \eta^{(1)}\nabla_{s,a}^{(0)} + \theta_{s,a}^{(0)} \tag{30}$$

$$\theta_{s,a}^{(3)} = \eta^{(3)}\nabla_{s,a}^{(2)} + \eta^{(2)}(1+\frac{1}{4})\nabla_{s,a}^{(1)} + \eta^{(1)}\nabla_{s,a}^{(0)} + \theta_{s,a}^{(0)} \tag{31}$$

$$\theta_{s,a}^{(4)} = \eta^{(4)}\nabla_{s,a}^{(3)} + \eta^{(3)}(1+\frac{2}{5})\nabla_{s,a}^{(2)} + \eta^{(2)}(1+\frac{1}{4}+\frac{2\cdot 1}{5\cdot 4})\nabla_{s,a}^{(1)} + \eta^{(1)}\nabla_{s,a}^{(0)} + \theta_{s,a}^{(0)} \tag{32}$$

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

$$\theta_{s,a}^{(t+1)} = \eta^{(t+1)}\nabla_{s,a}^{(t)} + \eta^{(t)}\Big(1+\frac{t-1}{t+2}\Big)\nabla_{s,a}^{(t-1)} + \eta^{(t-1)}\Big(1+\frac{t-2}{t+1}+\frac{(t-1)(t-2)}{(t+2)(t+1)}\Big)\nabla_{s,a}^{(t-2)} \tag{33}$$

$$+\sum_{j=1}^{t-3}\eta^{(j+1)}\Big(1+\frac{j}{j+3}+\frac{(j+1)j}{(j+4)(j+3)}+\frac{(j+2)(j+1)j}{(j+5)(j+4)(j+3)}+\sum_{k=4}^{t-j}\frac{(j+2)(j+1)j}{(j+k+2)(j+k+1)(j+k)}\Big)\nabla_{s,a}^{(j)} \tag{34}$$

$$+\eta^{(1)}\nabla_{s,a}^{(0)} + \theta_{s,a}^{(0)} \tag{35}$$

$$=\sum_{j=1}^{t} G(j,t)\cdot\eta^{(j+1)}\nabla_{s,a}^{(j)} + \eta^{(1)}\nabla_{s,a}^{(0)} + \theta_{s,a}^{(0)}. \tag{36}$$

Proof of Lemma 5. Regarding a), one could verify (29)-(32) by directly using the APG update in Algorithm 2. Regarding b), we prove this by induction. Specifically, suppose (33)-(36) hold for all iterations up to t. By the APG update, we know

$$\theta_{s,a}^{(t+1)} = \theta_{s,a}^{(t)} + \eta^{(t+1)}\nabla_{s,a}^{(t)} + \frac{t-1}{t+2}(\theta_{s,a}^{(t)} - \theta_{s,a}^{(t-1)}). \tag{37}$$

By plugging into (37) the expressions of $\theta_{s,a}^{(t)}$ and $\theta_{s,a}^{(t-1)}$ as suggested by (33)-(36), we could verify that (33)-(36) into hold for iteration $t+1$. □

Lemma 6. *(Vector Composition Lemma in (Boyd et al., 2004)). Let $f = h\left(g_1(x), g_2(x), \cdots, g_k(x)\right)$ where $h:\mathbb{R}^k\to\mathbb{R}$, $g_i:\mathbb{R}^n\to\mathbb{R}$. Then, f is concave in x if h is concave and non-decreasing in each argument and g_i are concave.*

Lemma 7 (Performance Difference Lemma in (Kakade & Langford, 2002))**.** *For each state s_0, the difference in the value of s_0 between two policies π and π' can be characterized as:*

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{\pi}}\mathbb{E}_{a\sim\pi(\cdot|s)}\left[A^{\pi'}(s,a)\right]. \tag{38}$$

Lemma 8. *(Lemma 1. in (Mei et al., 2020)). Softmax policy gradient w.r.t. θ is*

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\Big|_{\theta=\theta} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot A^{\pi_\theta}(s,a). \tag{39}$$

Lemma 9. *(Lemma 2. in (Mei et al., 2020)).* $\forall r \in [0,1]^{|\mathcal{A}|}, \theta \to \pi_\theta^\top r$ *is* $5/2$*-smooth.*

B.2. Convergence to First-Order Stationary Points Under APG

For ease of exposition, we restate several theoretical results stated in (Ghadimi & Lan, 2016) as follows. Note that we have made a minor modification to ensure that Theorem 5 and Theorem 6 from the convex regime can be easily applied to the concave regime without any loss of generality. This modification also allows for the use of a unified symbol across both regimes, providing a streamlined and consistent approach.

Theorem 5. *(Theorem 1 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 and* Γ_t *be defined by:*

$$\Gamma^{(t)} \coloneqq \begin{cases} 1, & t = 1 \\ (1-\alpha^{(t)})\Gamma^{(t-1)}, & t \geq 2 \end{cases} \tag{40}$$

Given a convex set $\mathcal{X}$ *such that* $V^{\pi_\theta}(\rho)$ *is concave in* $\mathcal{X}$*. Suppose* $\left\{\theta_{md}^{(t)}, \theta_{ag}^{(t)}\right\}_{t\geq 1}$ *always remain in the set* $\mathcal{X}$*, for all* t*. If* $\alpha^{(t)}, \beta^{(t)}, \lambda^{(t)}$ *are chosen such that*

$$\alpha^{(t)}\lambda^{(t)} \leq \beta^{(t)} \leq \frac{1}{L}, \tag{41}$$

$$\frac{\alpha^{(1)}}{\lambda^{(1)}\Gamma^{(1)}} \geq \frac{\alpha^{(2)}}{\lambda^{(2)}\Gamma^{(2)}} \geq \cdots, \tag{42}$$

where L *is the Lipschitz constant of the objective. Then for any* $t \geq 1$ *and any* θ^{**}*, we have*

$$V^{\pi_{\theta^{**}}}(\mu) - V^{\pi_\theta^{(t)}}(\mu) \leq \frac{\Gamma^{(t)}\left\|\theta^{(0)} - \theta^{**}\right\|^2}{2\lambda^{(1)}}. \tag{43}$$

Corollary 1. *(Corollary 1 in (Ghadimi & Lan, 2016) with a slight modification). Suppose that* $\left\{\alpha^{(t)}\right\}$*,* $\left\{\lambda^{(t)}\right\}$ *and* $\left\{\beta^{(t)}\right\}$ *in Algorithm 3 are set to*

$$\alpha^{(t)} = \frac{2}{(t+1)+c}, \quad \lambda^{(t)} = \frac{(t+1)+c}{2}\cdot\beta^{(t)}, \quad \beta^{(t)} = \frac{t+c}{(t+1)+c}\cdot\frac{1}{2L}, \quad \textit{where } c > 0. \tag{44}$$

Given a convex set $\mathcal{X}$ *such that* $V^{\pi_\theta}(\rho)$ *is concave in* $\mathcal{X}$*. Suppose* $\left\{\theta_{md}^{(t)}, \theta_{ag}^{(t)}\right\}_{t\geq 1}$ *always remain in the set* $\mathcal{X}$*, for all* t*. Then, for any* $t \geq 1$ *and any* θ^{**}*, we have*

$$V^{\pi_{\theta^{**}}}(\mu) - V^{\pi_{\theta_t}}(\mu) \leq \frac{4L(2+c)\left\|\theta^{(0)} - \theta^{**}\right\|^2}{(t+c+1)(t+c)} = O(\frac{1}{t^2}). \tag{45}$$

Remark 7. Note that we have made the following minor modifications: (i) We have introduced a constant c since our objective is not concave initially and hence the theoretical result had to be revised to account for the shifted initial learning rate. (ii) We have adjusted lambda from $\frac{t}{2}$ to $\frac{t+1}{2}$ and β from $\frac{1}{2L}$ to $\frac{t+c}{(t+1)+c}\cdot\frac{1}{2L}$ to ensure the applicability of both Lemma 4 and Theorem 5 results.

Remark 8. Note that Theorem 5 and Corollary 1 are built on the *local concavity* of the objective function. In Appendix D, we formally show that such local concavity indeed holds under APG in the multi-action bandit setting.

Proof of Corollary 1. We leverage Theorem 5 to reach our desired result. And it remains to show that the chosen of $\left\{\alpha^{(t)}, \lambda^{(t)}, \beta^{(t)}\right\}$ in (44) satisfy (41) and (42). Note that $\alpha^{(t)}\cdot\lambda^{(t)} = \beta^{(t)}$, (41) easily holds. And by the definition of $\Gamma^{(t)}$, we have:

$$\Gamma^{(t)} = \frac{(2+c)(1+c)}{((t+1)+c)(t+c)}. \tag{46}$$

Hence we have:

$$\frac{\alpha^{(t)}}{\lambda^{(t)}\Gamma^{(t)}} = \frac{\frac{2}{(t+1)+c}}{\frac{(t+1)+c}{2}\cdot\frac{t+c}{(t+1)+c}\cdot\frac{1}{2L}\cdot\frac{(2+c)(1+c)}{((t+1)+c)(t+c)}} = \frac{8L}{(2+c)(1+c)}, \tag{47}$$

which makes the condition (42) holds. And hence we reach the desired result by plugging $\Gamma^{(t)}$ and $\lambda^{(1)}$ into (43). □

Theorem 6. *(Theorem 2 in (Ghadimi & Lan, 2016) with a slight modification). Suppose that $\alpha^{(t)}, \beta^{(t)}, \lambda^{(t)}$ are chosen such that (41)-(42) hold. Then for any $t \geq 1$ and any θ^{**}, we have*

$$\min_{k=1,2,\cdots,t} \left\| \nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\theta^{(t)}} \right\| \tag{48}$$

$$\leq 2\left[\sum_{k=1}^{t} \frac{1}{\Gamma^{(k)}}\beta^{(k)}(1-L\beta^{(k)})\right]^{-1}\left[\frac{\left\|\theta^{(0)}-\theta^{**}\right\|}{2\lambda^{(1)}} + \frac{L}{\Gamma^{(t)}}(\|\theta^{**}\|^2 + \max_{k=1,2,\cdots,t}\left\|\theta^{(t)}\right\|^2) + \frac{|V^{\pi_{\theta^{(t)}}}(\mu) - V^{\pi_{\theta^{**}}}(\mu)|}{\Gamma^{(t)}}\right]. \tag{49}$$

Remark 9. Note that we have made the following minor modifications: (i) Instead of the bounded domain $\mathcal{X}$ stated in (Ghadimi & Lan, 2016), we consider an unbounded domain $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. (ii) Consequently, we replace the bounded domain constant M with the norm of the parameters without loss of generality. (iii) As the domain is unbounded, the problem reduces to an unconstrained optimization problem where $\mathcal{G}(x_k^{md}, \nabla\Psi(x_k^{md}), \beta_k) = \nabla\Psi(x_k^{md})$, which represents the gradient norm and $L_\Psi = L_f = L$ where L is the Lipschitz constant of the objective.

Proof of Theorem 6. The proof is identical to the one in (Ghadimi & Lan, 2016) until equation (2.53). However, instead of letting $x^* = x$, we choose a surrogate optimal solution $x^{**} = x$. Hence, we have:

$$\frac{\Psi(x_N^{ag}) - \Psi(x^{**})}{\Gamma_N} + \sum_{k=1}^{N}\frac{1-L_\Psi\beta_k}{2\beta_k\Gamma_k}\left\|x_k^{ag} - x_k^{md}\right\|^2 \leq \frac{\|x_0 - x^{**}\|}{2\lambda_1} + \frac{L_f}{\Gamma_N}(\|x^{**}\|^2 + M^2), \tag{50}$$

where, for ease of exposition, we continue to use the same symbols. By rearranging (50) and incorporating the modifications we have made, we obtain the desired result:

$$\min_{k=1,2,\cdots,N}\left\|\mathcal{G}(x_k^{md}, \nabla\Psi(x_k^{md}), \beta_k)\right\|^2\left(\sum_{k=1}^{N}\frac{\beta_k(1-L_\Psi\beta_k)}{2\Gamma_k}\right) \leq \sum_{k=1}^{N}\frac{\beta_k(1-L_\Psi\beta_k)}{2\Gamma_k}\left\|\mathcal{G}(x_k^{md}, \nabla\Psi(x_k^{md}), \beta_k)\right\|^2 \tag{51}$$

$$= \sum_{k=1}^{N}\frac{1-L_\Psi\beta_k}{2\beta_k\Gamma_k}\left\|x_k^{ag} - x_k^{md}\right\|^2 \tag{52}$$

$$\leq \frac{\|x_0 - x^{**}\|}{2\lambda_1} + \frac{L_f}{\Gamma_N}(\|x^{**}\|^2 + M^2) \tag{53}$$

$$+ \frac{|\Psi(x_N^{ag}) - \Psi(x^{**})|}{\Gamma_N}. \tag{54}$$

Under APG, our objective function is the value function and since we use the time index t instead of N, $\Psi(x_N^{ag})$ will be the value function under $\theta^{(t)}$, i.e., $V^{\pi_{\theta^{(t)}}}(\mu)$. Similarly, $\Psi(x^{**})$ will be $V^{\pi_{\theta^{**}}}(\mu)$. Hence, we obtain the results. □

Corollary 2. *(Corollary 2 in (Ghadimi & Lan, 2016) with a slight modification). Suppose that $\{\alpha^{(t)}\}$, $\{\lambda^{(t)}\}$ and $\{\beta^{(t)}\}$ in Algorithm 3 are set to*

$$\alpha^{(t)} = \frac{2}{t+1}, \quad \lambda^{(t)} = \frac{t+1}{2}\cdot\beta^{(t)}, \quad \beta^{(t)} = \frac{t}{t+1}\cdot\frac{1}{2L}, \tag{55}$$

Then for any $t \geq 1$ and any θ^{**}, we have

$$\min_{k=1,2,\cdots,t}\left\|\nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\theta^{(t)}}\right\| \leq 192L^2\frac{\left\|\theta^{(0)}-\theta^{**}\right\|}{t(t+1)(2t+1)} + \frac{48L^2}{2t+1}(\|\theta^{**}\|^2 + \max_{k=1,2,\cdots,t}\left\|\theta^{(t)}\right\|^2) \tag{56}$$

$$+ \frac{48L\cdot|V^{\pi_{\theta^{(t)}}}(\mu) - V^{\pi_{\theta^{**}}}(\mu)|}{2t+1}. \tag{57}$$

Proof of Corollary 2. The results directly follow by plugging the value of $\Gamma^{(t)} = \frac{2}{t(t+1)}$ defined in (40), $\beta^{(k)}, \lambda^{(1)}$ defined in (55) into (49):

$$\min_{k=1,2,\cdots,t}\left\|\nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\theta^{(t)}}\right\| \tag{58}$$

$$\leq 2\left[\sum_{k=1}^{t}\frac{1}{\Gamma^{(k)}}\beta^{(k)}(1-L\beta^{(k)})\right]^{-1}\left[\frac{\left\|\theta^{(0)}-\theta^{**}\right\|}{2\lambda^{(1)}}+\frac{L}{\Gamma^{(t)}}(\|\theta^{**}\|^2+\max_{k=1,2,\cdots,t}\left\|\theta^{(t)}\right\|^2)+\frac{|V^{\pi_{\theta^{(t)}}}(\mu)-V^{\pi_{\theta^{**}}}(\mu)|}{\Gamma^{(t)}}\right] \tag{59}$$

$$=2\left[\sum_{k=1}^{t}\frac{k(k+1)}{2}\cdot\frac{k}{2L(k+1)}\cdot(1-L\cdot\frac{k}{2L(k+1)})\right]^{-1} \tag{60}$$

$$\cdot\left[2L\left\|\theta^{(0)}-\theta^{**}\right\|+\frac{L\cdot t(t+1)}{2}(\|\theta^{**}\|^2+\max_{k=1,2,\cdots,t}\left\|\theta^{(t)}\right\|^2)+\frac{t(t+1)\cdot|V^{\pi_{\theta^{(t)}}}(\mu)-V^{\pi_{\theta^{**}}}(\mu)|}{2}\right] \tag{61}$$

$$\leq 2\left[\sum_{k=1}^{t}\frac{k^2}{4L}\cdot\frac{1}{2}\right]^{-1} \tag{62}$$

$$\cdot\left[2L\left\|\theta^{(0)}-\theta^{**}\right\|+\frac{L\cdot t(t+1)}{2}(\|\theta^{**}\|^2+\max_{k=1,2,\cdots,t}\left\|\theta^{(t)}\right\|^2)+\frac{t(t+1)\cdot|V^{\pi_{\theta^{(t)}}}(\mu)-V^{\pi_{\theta^{**}}}(\mu)|}{2}\right] \tag{63}$$

$$=192L^2\frac{\left\|\theta^{(0)}-\theta^{**}\right\|}{t(t+1)(2t+1)}+\frac{48L^2}{2t+1}(\|\theta^{**}\|^2+\max_{k=1,2,\cdots,t}\left\|\theta^{(t)}\right\|^2)+\frac{48L\cdot|V^{\pi_{\theta^{(t)}}}(\mu)-V^{\pi_{\theta^{**}}}(\mu)|}{2t+1}. \tag{64}$$

□

C. Asymptotic Convergence

For ease of exposition, we restate the bandit case setting as follows:

Bandit Case. A bandit case is a special case of the reinforcement learning problem in which there is only one single state. Given a bandit case with $|\mathcal{A}|$ actions $\left[a^*, a_2, \cdots, a_{|\mathcal{A}|}\right]$ and rewards $r = \left[r(a^*), r(a_2), \cdots, r(a_{|\mathcal{A}|})\right]$. We parametrized the policy under softmax, (i.e. we define $\pi_\theta = \text{softmax}(\theta)$ where $\theta = [\theta_{a^*}, \theta_{a_2}, \cdots, \theta_{a_{|\mathcal{A}|}}]$. Without loss of generality and for simplicity, we assume that the optimal action a^* is unique and $r(a^*) > r(a_2) > \cdots > r(a_{|\mathcal{A}|})$. The uniqueness assumption can be lifted with a little extra work.

C.1. Existence of Limiting Value Functions

As mentioned in Section 5.1, one fundamental challenge of the convergence analysis of APG lies in the lack of monotonic improvement. As a result, it remains unknown if the limiting value functions even exist. Despite this, we are able to show that the limiting value functions indeed exist in the general multi-action bandit setting.

Lemma 10. *Under APG, in the bandit setting, 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 state* $s \in \mathcal{S}$.

Proof of Lemma 10.

Claim 1. *The proof can be completed by making the following claims:*

a) *If the gradient vector and the momentum vector share identical signs, then we have* $V^{\pi_\theta^{(t+1)}}(\mu) \geq V^{\pi_\theta^{(t)}}(\mu)$.

b) *In the bandit setting, if* $r(a_i) > V^{\pi_\theta^{(T_{a_i}-1)}}(s) > r(a_{i+1})$ *and* $r(a_{i-1}) > V^{\pi_\theta^{(T_{a_i})}}(s) > r(a_i)$ *for some* T_a, *then we have* $V^{\pi_\theta^{(t)}}(s) > r(a_i)$ *for all* $t \geq T_{a_i}$ *where* $i = 2, 3, \cdots, |\,\mathcal{A}\,|$.

c) *Under APG, in the bandit setting, there exists a finite time* T_0 *after which* $V^{\pi_\theta}(\mu)$ *either consistently increases or converges at a stationary point. Hence, there may be two distinct situations:*

- $V^{\pi_\theta^{(t+1)}}(\mu) \geq V^{\pi_\theta^{(t)}}(\mu)$ *for all* $t > T_0$.
- $V^{\pi_\theta^{(t)}}(\mu) \to r(a')$ *as* $t \to \infty$ *for some* $a' \in \mathcal{A}$.

d) *In both situations, since* $V^{\pi_\theta}(\mu)$ *is bounded above, by the monotone convergence theorem, we conclude that 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 states* $s \in \mathcal{S}$.

Claim a). We show the desired result by leveraging Lemma 7. According to Performance Difference Lemma, in order to show that $V^{\pi_\theta^{(t+1)}}(\mu) \geq V^{\pi_\theta^{(t)}}(\mu)$, it is sufficient to show that:

$$\sum_{a\in\mathcal{A}} \pi_\theta^{(t+1)}(a|s) A^{\pi_\theta^{(t)}}(s,a) > 0, \quad \forall s \in \mathcal{S}. \tag{65}$$

To reach (65), we have $\forall s \in \mathcal{S}$:

$$\begin{aligned}
\sum_{a\in\mathcal{A}} \pi_\theta^{(t+1)}(a|s) A^{\pi_\theta^{(t)}}(s,a) &= \sum_{a\in\mathcal{A}} \frac{\exp(\theta_{s,a}^{(t+1)})}{\sum_{a\in\mathcal{A}} \theta_{s,a}^{(t+1)}} A^{\pi_\theta^{(t)}}(s,a) && (66)\\
&= \frac{\sum_{a\in\mathcal{A}} \theta_{s,a}^{(t)}}{\sum_{a\in\mathcal{A}} \theta_{s,a}^{(t+1)}} \sum_{a\in\mathcal{A}} \frac{\exp(\theta_{s,a}^{(t+1)})}{\sum_{a\in\mathcal{A}} \theta_{s,a}^{(t)}} A^{\pi_\theta^{(t)}}(s,a) && (67)\\
&> \frac{\sum_{a\in\mathcal{A}} \theta_{s,a}^{(t)}}{\sum_{a\in\mathcal{A}} \theta_{s,a}^{(t+1)}} \sum_{a\in\mathcal{A}} \frac{\exp(\theta_{s,a}^{(t)})}{\sum_{a\in\mathcal{A}} \theta_{s,a}^{(t)}} A^{\pi_\theta^{(t)}}(s,a) && (68)\\
&= \frac{\sum_{a\in\mathcal{A}} \theta_{s,a}^{(t)}}{\sum_{a\in\mathcal{A}} \theta_{s,a}^{(t+1)}} \sum_{a\in\mathcal{A}} \pi_m(a|s) A^{\pi_\theta^{(t)}}(s,a) && (69)\\
&= 0, && (70)
\end{aligned}$$

where (68) holds based on the assumption that the gradient vector and the momentum vector share identical signs and the fact that $\theta_{s,a}^{(t+1)} \geq \theta_{s,a}^{(t)}$ if $A^{\pi_\theta^{(t)}}(s,a) \geq 0$ and $\theta_{s,a}^{(t+1)} \leq \theta_{s,a}^{(t)}$ if $A^{\pi_\theta^{(t)}}(s,a) \leq 0$ as shown in Lemma 8.

Claim b). By Algorithm 2, the momentum vector at time t is $\frac{t-1}{t+2}(\theta^{(t)} - \theta^{(t-1)})$. Hence we have that the gradient and the momentum can exhibit opposite signs exclusively when there is a change in the sign of the gradient. So given the fact that $r(a_i) > V^{\pi_\theta^{(T_{a_i}-1)}}(s) > r(a_{i+1})$ and $r(a_{i-1}) > V^{\pi_\theta^{(T_{a_i})}}(s) > r(a_i)$, we have that the gradient $\frac{\partial V^{\pi_\theta^{(t)}}(\mu)}{\partial \theta_{s,a_i}}$ and the momentum $\frac{t-1}{t+2}(\theta_{s,a_i}^{(t)} - \theta_{s,a_i}^{(t-1)})$ might exhibit opposite signs while other element in the gradient vector must be identical to the sign with the corresponding element in the momentum.

Hence, by Lemma 7, we have for all $T(a_{i-1}) > t > T(a_i)$:

$$
\begin{aligned}
V^{\pi_\theta^{(t+1)}}(s) - V^{\pi_\theta^{(t)}}(s) &= \sum_{a\in\mathcal{A}} \pi_\theta^{(t+1)}(a|s) A^{\pi_\theta^{(t)}}(s,a) && (71)\\
&= \sum_{a\in\mathcal{A}} \frac{\exp(\theta_{s,a}^{(t+1)})}{\sum_{a\in\mathcal{A}}\exp(\theta_{s,a}^{(t+1)})} A^{\pi_\theta^{(t)}}(s,a) && (72)\\
&= \frac{\sum_{a\in\mathcal{A}}\exp(\theta_{s,a}^{(t)})}{\sum_{a\in\mathcal{A}}\exp(\theta_{s,a}^{(t+1)})} \sum_{a\in\mathcal{A}} \frac{\exp(\theta_{s,a}^{(t+1)})}{\sum_{a\in\mathcal{A}}\exp(\theta_{s,a}^{(t)})} A^{\pi_\theta^{(t)}}(s,a) && (73)\\
&\geq \frac{\sum_{a\in\mathcal{A}}\exp(\theta_{s,a}^{(t)})}{\sum_{a\in\mathcal{A}}\exp(\theta_{s,a}^{(t+1)})} \sum_{a\in\mathcal{A}} \frac{\exp(\theta_{s,a}^{(t)})}{\sum_{a\in\mathcal{A}}\exp(\theta_{s,a}^{(t)})} A^{\pi_\theta^{(t)}}(s,a) + \frac{\exp(\theta_{s,a_i}^{(t+1)}) - \exp(\theta_{s,a_i}^{(t)})}{\sum_{a\in\mathcal{A}}\exp(\theta_{s,a}^{(t+1)})} A^{\pi_\theta^{(t)}}(s,a_i) && (74)\\
&\geq \frac{\sum_{a\in\mathcal{A}}\exp(\theta_{s,a}^{(t)})}{\sum_{a\in\mathcal{A}}\exp(\theta_{s,a}^{(t+1)})} \sum_{a\in\mathcal{A}} \pi_m(a|s) A^{\pi_\theta^{(t)}}(s,a) + \frac{\exp(\theta_{s,a_i}^{(t+1)})}{\sum_{a\in\mathcal{A}}\exp(\theta_{s,a}^{(t+1)})} A^{\pi_\theta^{(t)}}(s,a_i) && (75)\\
&\geq 0 + A^{\pi_\theta^{(t)}}(s,a_i) && (76)\\
&= A^{\pi_\theta^{(t)}}(s,a_i), && (77)
\end{aligned}
$$

where (74) holds based on the assumption that the gradient vector and the momentum vector share identical signs for all $a \neq a_i$ and the fact that $\theta_{s,a}^{(t+1)} \geq \theta_{s,a}^{(t)}$ if $A^{\pi_\theta^{(t)}}(s,a) \geq 0$ and $\theta_{s,a}^{(t+1)} \leq \theta_{s,a}^{(t)}$ if $A^{\pi_\theta^{(t)}}(s,a) \leq 0$ as shown in Lemma 8.

Claim c). By Algorithm 2, the momentum vector at time t is $\frac{t-1}{t+2}(\theta^{(t)} - \theta^{(t-1)})$. Hence we have that the gradient and the momentum can exhibit opposite signs exclusively when there is a change in the sign of the gradient. That said, the gradient and the momentum can exhibit opposite signs at time t exclusively when there exists a time $T_a < t$ such that $V^{\pi_\theta^{(T_a)}}(s) > r(a) > V^{\pi_\theta^{(T_a-1)}}(s)$ for some $a \in \mathcal{A}$ and also the gradient and the momentum exhibit opposite signs for all $t' \in [T_a, t]$.

Additionally, by **Claim (b)**, we have that there exist at most $|\mathcal{A}| - 1$ time such that $V^{\pi_\theta^{(T_a)}}(s) > r(a) > V^{\pi_\theta^{(T_a-1)}}(s)$ for some $a \in \mathcal{A}$. Hence we discuss two possible case:

- **Case 1.** For all T_a such that once $V^{\pi_\theta^{(T_a)}}(s) > r(a) > V^{\pi_\theta^{(T_a-1)}}(s)$, there exists a finite time ${T_a}' > T_a$ such that the gradient vector and the momentum vector share identical signs at $t = {T_a}'$:

  Since the gradient vector and the momentum vector share identical signs at $t = {T_a}'$, we have that $V^{\pi_\theta^{(t)}}(\mu)$ enjoys monotonic improvement if $T_{a'} > t > {T_a}'$ where $T_{a'}$ is the next time step such that the sign of the gradient has changed again. Moreover, since there only exist at most $|\mathcal{A}| - 1$ time such that $V^{\pi_\theta^{(T_a)}}(s) > r(a) > V^{\pi_\theta^{(T_a-1)}}(s)$ for some $a \in \mathcal{A}$, we have that after time $T_0 = \max_a {T_a}'$, the gradient vector and the momentum vector share identical signs. And so by **Claim (a)**, we have that $V^{\pi_\theta^{(t+1)}}(\mu) \geq V^{\pi_\theta^{(t)}}(\mu)$ for all $t \geq T_0$.

- **Case 2.** There exists some T_a such that once $V^{\pi_\theta^{(T_a)}}(s) > r(a) > V^{\pi_\theta^{(T_a-1)}}(s)$, the gradient and the momentum exhibit opposite signs for *all* $t \geq T_a$:

Since the magnitude of the momentum is bounded above, we have that the gradient and the momentum exhibit opposite signs for *all* $t \geq T_a$ if and only if $\frac{\partial \pi_\theta^\top r}{\partial \theta_a}\Big|_{\theta=\omega^{(t)}} \to 0$. Based on the converging gradient and the fact that the gradient and the momentum exhibit opposite signs, we can conclude that $\pi_\theta^\top r \to r(a)$, which also leading to the stationary point of the objective.

□

C.2. Supporting Lemmas for Asymptotic Convergence of APG

In the sequel, we use $A^{(t)}(s,a)$, $Q^{(t)}(s,a)$, and $V^{(t)}(s)$ as the shorthand of $A^{\pi_\omega^{(t)}}(s,a)$, $Q^{\pi_\omega^{(t)}}(s,a)$, and $V^{\pi_\omega^{(t)}}(s)$, respectively. For ease of exposition, we divide the action space into the following subsets based on the advantage function:

$$I_s^+ := \{a \in \mathcal{A} : \lim_{t\to\infty} A^{(t)}(s,a) > 0\} \tag{78}$$

$$I_s^- := \{a \in \mathcal{A} : \lim_{t\to\infty} A^{(t)}(s,a) < 0\} \tag{79}$$

$$I_s^0 := \{a \in \mathcal{A} : \lim_{t\to\infty} A^{(t)}(s,a) = 0\} \tag{80}$$

Note that the above action sets are well-defined as the limiting value functions exist by Lemma 10. Moreover, we would like to highlight that the theoretical results in Appendix C.2 and C.3 are directly applicable to the general MDP case as long as the limiting value functions exist.

For ease of notation, for each state s, we define

$$\Delta_s := \min_{a\in I_s^+\cup I_s^-} |A^{(t)}(s,a)|. \tag{81}$$

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 $a \in I_s^+$, we have
$$A^{(t)}(s,a) \geq +\frac{\Delta_s}{4}, \quad \text{for all } t \geq \bar{T}_s, \tag{82}$$
- (ii) For all $a \in I_s^-$, we have
$$A^{(t)}(s,a) \leq -\frac{\Delta_s}{4}, \quad \text{for all } t \geq \bar{T}_s. \tag{83}$$
- (iii) For all $a \in I_s^0$, we have
$$|A^{(t)}(s,a)| \leq \frac{\Delta_s}{4}, \quad \text{for all } t \geq \bar{T}_s. \tag{84}$$

Lemma 11. *For any state $s \in \mathcal{S}$, we have $\sum_{a\in I_s^+\cup I_s^-} \pi_\theta^{(t)}(a|s) \to 0$, as $t \to \infty$. As a result, we also have $\sum_{a\in I_s^0} \pi_\theta^{(t)}(a|s) \to 1$, as $t \to \infty$.*

Proof of Lemma 11. Given that the limiting value functions exist as well as the fact that $d_\mu^{\pi_\theta}(s) \geq \frac{\mu(s)}{1-\gamma} > 0$, we know that for any state-action pair (s,a),

$$\pi_\theta^{(t)}(a|s) A^{\pi_\theta^{(t)}}(s,a) \to 0, \quad \text{as } t \to \infty. \tag{85}$$

- For any $a \in I_s^+$, by definition we have $\lim_{t\to\infty} A^{\pi_\theta^{(t)}}(s,a) > 0$. By (85), this implies that $\pi_\theta^{(t)}(a|s) \to 0$, as $t \to \infty$.
- Similarly, for any $a \in I_s^-$, by definition we have $\lim_{t\to\infty} A^{\pi_\theta^{(t)}}(s,a) < 0$. Again, by (85), this property implies that $\pi_\theta^{(t)}(a|s) \to 0$, as $t \to \infty$.

Hence, we have $\sum_{a\in I_s^+\cup I_s^-} \pi_\theta^{(t)}(a|s) \to 0$, as $t \to \infty$. □

Lemma 12. *Under APG, for any iteration k and any state-action pair (s,a), we have*

$$\sum_{a\in\mathcal{A}} \theta_{s,a}^{(k)} = \sum_{a\in\mathcal{A}} \theta_{s,a}^{(0)}. \tag{86}$$

Proof of Lemma 12. We prove this by induction and show 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

$$\theta_{s,a}^{(1)} = \omega_{s,a}^{(0)} + \eta^{(1)} \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\Big|_{\theta=\omega^{(0)}} = \theta_{s,a}^{(0)} + \eta^{(1)} \cdot \frac{1}{1-\gamma} d^{\pi_\theta^{(0)}}(s) \pi_\theta^{(0)}(a|s) A^{\pi_\theta^{(0)}}(s,a), \tag{87}$$

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 8. By taking the sum of (87) over all the actions, we have $\sum_{a\in\mathcal{A}} \theta_{s,a}^{(1)} = \sum_{a\in\mathcal{A}} \theta_{s,a}^{(0)}$ due to the fact that $\sum_{a\in\mathcal{A}} \pi_\theta(a|s) A^{\pi_\theta}(s,a) = 0$, for any θ. Similarly, we have

$$\theta_{s,a}^{(2)} = \omega_{s,a}^{(1)} + \eta^{(2)} \cdot \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\Big|_{\theta=\omega^{(1)}} \tag{88}$$

$$= \theta_{s,a}^{(1)} + \frac{0}{3} \cdot (\theta_{s,a}^{(1)} - \theta_{s,a}^{(0)}) + \eta^{(2)} \cdot \frac{1}{1-\gamma} d^{\pi_\omega^{(1)}}(s) \pi_\omega^{(1)}(a|s) A^{\pi_\omega^{(1)}}(s,a). \tag{89}$$

By taking the sum of (89) 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 the fact that$\sum_{a\in\mathcal{A}} \pi_\theta(a|s) A^{\pi_\theta}(s,a) = 0$, for all θ.

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 (89). That is,

$$\theta_{s,a}^{(M+1)} = \omega_{s,a}^{(M)} + \eta^{(M+1)} \cdot \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\Big|_{\theta=\omega^{(M)}} \tag{90}$$

$$= \theta_{s,a}^{(M)} + \frac{M-1}{M+2} \cdot (\theta_{s,a}^{(M)} - \theta_{s,a}^{(M-1)}) + \eta^{(M+1)} \cdot \frac{1}{1-\gamma} d^{\pi_\omega^{(M)}}(s) \pi_\omega^{(M)}(a|s) A^{\pi_\omega^{(M)}}(s,a). \tag{91}$$

By taking the sum of (91) over all the actions, we could verify that $\sum_{a\in\mathcal{A}} \theta_{s,a}^{(M+1)} = \sum_{a\in\mathcal{A}} \theta_{s,a}^{(0)}$. □

Lemma 13. *Let a be an action in I_s^+. Under APG, $\theta_{s,a}^{(t)}$ and $\omega_{s,a}^{(t)}$ must be bounded from below, for all t.*

Proof of Lemma 13. Recall that we define $\Delta_s := \min_{a\in I_s^+ \cup I_s^-} |A^{\pi_\omega^{(t)}}(s,a)|$. Then, there must exist $T_0 \in \mathbb{N}$ such that $A^{\pi_\omega^{(t)}}(s,a) \geq \frac{\Delta_s}{4}$, for all $t \geq T_0$.

For ease of notation, we let $\delta_{T_0} := \theta_{s,a}^{(T_0)} - \theta_{s,a}^{(T_0-1)}$. By a similar argument, for any $M \in \mathbb{N}$, we have

$$\theta_{s,a}^{(T_0+M)} = \omega_{s,a}^{(T_0+M-1)} + \eta^{(T_0+M)} \cdot \underbrace{\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\Big|_{\theta=\omega^{(T_0+M-1)}}}_{\geq 0} \tag{92}$$

$$\geq \theta_{s,a}^{(T_0+M-1)} + \frac{T_0+M-2}{T_0+M+1}(\theta_{s,a}^{(T_0+M-1)} - \theta_{s,a}^{(T_0+M-2)}) \tag{93}$$

$$\geq \theta_{s,a}^{(T_0)} + \frac{T_0-1}{T_0+2}\delta_{T_0} + \frac{T_0(T_0-1)}{(T_0+3)(T_0+2)}\delta_{T_0} + \cdots + \frac{(T_0+M-2)\cdots(T_0-1)}{(T_0+M+1)\cdots(T_0+2)}\delta_{T_0} \tag{94}$$

$$= \theta_{s,a}^{(T_0)} + \Big[\frac{T_0-1}{T_0+2} + \frac{T_0(T_0-1)}{(T_0+3)(T_0+2)} + \sum_{\tau=2}^{M} \frac{(T_0+1)T_0(T_0-1)}{(T_0+\tau+2)(T_0+\tau+1)(T_0+\tau)}\Big]\delta_{T_0}. \tag{95}$$

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

$$\sum_{\tau=2}^{M} \frac{(T_0+1)T_0(T_0-1)}{(T_0+\tau+2)(T_0+\tau+1)(T_0+\tau)} \tag{96}$$

$$= (T_0+1)T_0(T_0-1) \sum_{\tau=2}^{M} \frac{1}{2}\Big(\frac{1}{(T_0+\tau)(T_0+\tau+1)} - \frac{1}{(T_0+\tau+1)(T_0+\tau+2)}\Big) \tag{97}$$

$$= (T_0+1)T_0(T_0-1) \cdot \frac{1}{2}\Big(\frac{1}{(T_0+2)(T_0+3)} - \frac{1}{(T_0+M+1)(T_0+M+2)}\Big) \tag{98}$$

$$\leq \frac{T_0}{2}. \tag{99}$$

Therefore, we know that for any $M \in \mathbb{N}$,

$$\theta_{s,a}^{(T_0+M)} \geq \theta_{s,a}^{(T_0)} - (2+\frac{T_0}{2})|\delta_{T_0}|. \tag{100}$$

Hence, $\theta_{s,a}^{(t)} \geq \theta_{s,a}^{(T_0)} - (2+\frac{T_0}{2})|\delta_{T_0}|$, for all $t \geq T_0$. As the gradient under softmax parameterization is always bounded, this also implies that $\omega_{s,a}^{(t)}$ is bounded from below, for all t. □

Lemma 14. *Let a be an action in I_s^-. Under APG, $\theta_{s,a}^{(t)}$ and $\omega_{s,a}^{(t)}$ must be bounded from above, for all t.*

Proof of Lemma 14. To prove this, we could follow the same procedure as that in Lemma 13. Again, for ease of notation, we define $\Delta_s := \min_{a \in I_s^+ \cup I_s^-} |A^{\pi_\omega^{(t)}}(s,a)|$ and define $\delta_{T_0} := \theta_{s,a}^{(T_0)} - \theta_{s,a}^{(T_0-1)}$. Accordingly, there must exist $T_0 \in \mathbb{N}$ such that $A^{\pi_\omega^{(t)}}(s,a) \leq -\frac{\Delta_s}{4}$, for all $t \geq T_0$. Moreover, by the update scheme of APG, we have

$$\theta_{s,a}^{(T_0+1)} = \omega_{s,a}^{(T_0)} + \eta^{(T_0+1)} \cdot \underbrace{\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\Big|_{\theta=\omega^{(T_0)}}}_{\leq 0} \leq \omega_{s,a}^{(T_0)} = \theta_{s,a}^{(T_0)} + \frac{T_0-1}{T_0+2}(\theta_{s,a}^{(T_0)} - \theta_{s,a}^{(T_0-1)}). \tag{101}$$

Similarly, for any $M \in \mathbb{N}$, we have

$$\theta_{s,a}^{(T_0+M)} = \omega_{s,a}^{(T_0+M-1)} + \eta^{(T_0+M)} \cdot \underbrace{\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\Big|_{\theta=\omega^{(T_0+M-1)}}}_{\leq 0} \tag{102}$$

$$\leq \theta_{s,a}^{(T_0+M-1)} + \frac{T_0+M-2}{T_0+M+1}(\theta_{s,a}^{(T_0+M-1)} - \theta_{s,a}^{(T_0+M-2)}) \tag{103}$$

$$\leq \theta_{s,a}^{(T_0)} + \frac{T_0-1}{T_0+2}\delta_{T_0} + \frac{T_0(T_0-1)}{(T_0+3)(T_0+2)}\delta_{T_0} + \cdots + \frac{(T_0+M-2)\cdots(T_0-1)}{(T_0+M+1)\cdots(T_0+2)}\delta_{T_0} \tag{104}$$

$$= \theta_{s,a}^{(T_0)} + \Big[\frac{T_0-1}{T_0+2} + \frac{T_0(T_0-1)}{(T_0+3)(T_0+2)} + \sum_{\tau=2}^{M} \frac{(T_0+1)T_0(T_0-1)}{(T_0+\tau+2)(T_0+\tau+1)(T_0+\tau)}\Big]\delta_{T_0}. \tag{105}$$

By (96)-(99), we know $\sum_{\tau=2}^{M} \frac{(T_0+1)T_0(T_0-1)}{(T_0+\tau+2)(T_0+\tau+1)(T_0+\tau)} \leq \frac{T_0}{2}$. As a result, for any $M \in \mathbb{N}$,

$$\theta_{s,a}^{(T_0+M)} \leq \theta_{s,a}^{(T_0)} + (2+\frac{T_0}{2})|\delta_{T_0}|. \tag{106}$$

Hence, $\theta_{s,a}^{(t)} \leq \theta_{s,a}^{(T_0)} + (2+\frac{T_0}{2})|\delta_{T_0}|$, for all $t \geq T_0$. As the gradient under softmax parameterization is always bounded, this also implies that $\omega_{s,a}^{(t)}$ is bounded from above, for all t. □

Lemma 15. *Under APG, if I_s^+ is non-empty, then we have $\max_{a \in I_s^0} \theta_{s,a}^{(t)} \to \infty$, as $t \to \infty$.*

Proof. By Lemma 11, we know $\sum_{a \in I_s^0} \pi_\theta^{(t)}(a|s) \to 1$, as $t \to \infty$. Moreover, by Lemma 13, we know $\theta_{s,a}^{(t)}$ is bounded from below, for all $a \in I_s^+$. Therefore, under the softmax policy parameterization, we must have $\max_{a \in I_s^0} \theta_{s,a}^{(t)} \to \infty$. □

Recall from (83) that for all $a \in I_s^-$, we have $A^{(t)}(s,a) \leq -\frac{\Delta_s}{4}$ for all $t \geq \bar{T}_s$.

Lemma 16. *Under APG, if I_s^+ is non-empty, then for any $a \in I_s^-$, we have $\theta_{s,a}^{(t)} \to -\infty$, as $t \to \infty$.*

Proof of Lemma 16. We prove this contradiction. Motivated by the proof of Lemma C11 in (Agarwal et al., 2021), our proof here extends the argument to the case with the momentum by considering the cumulative effect of all the gradient terms on the policy parameter.

Specifically, given an action $a \in I_s^-$, suppose that there exists ϑ such that $\theta_{s,a}^{(t)} > \vartheta$, for all $t \geq \bar{T}_s$. Then, by Lemma 12 and Lemma 15, 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 $\theta_{s,a}^{(\bar{T}_s)} \geq \vartheta - \delta$. For each $t \geq \bar{T}_s$, define

$$\nu(t) := \max\{\tau : \theta_{s,a'}^{(\tau)} \geq \vartheta - \delta, \bar{T}_s \leq \tau \leq t\}, \tag{107}$$

which is essentially the latest iteration at which $\theta_{s,a'}^{(\tau)}$ crosses $\vartheta - \delta$ from the above. Moreover, we define an index set

$$\mathcal{J}^{(t)} := \Big\{\tau : \frac{\partial V^{(\tau)}(\mu)}{\partial \theta_{s,a'}} < 0, \nu(t) < \tau < t\Big\}. \tag{108}$$

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

$$Z^{(t)} := \sum_{t' \in \mathcal{J}^{(t)}} \eta^{(t'+1)} \cdot \frac{\partial V^{(t')}(\mu)}{\partial \theta_{s,a'}} \cdot G(t', t), \tag{109}$$

where $G(t, t')$ is the function defined in (28). Note that if $\mathcal{J}^{(t)} = \varnothing$, we define $Z^{(t)} = 0$. Accordingly, we know that for any $t > \bar{T}_s$, we have

$$Z^{(t)} \leq \sum_{t' \in \mathcal{J}^{(t)}} \eta^{(t'+1)} \cdot \frac{\partial V^{(t')}(\mu)}{\partial \theta_{s,a'}} \cdot G(t', t) + \sum_{t': t' \notin \mathcal{J}^{(t)}, \nu(t) < t' < t} \underbrace{\eta^{(t'+1)} \cdot \frac{\partial V^{(t')}(\mu)}{\partial \theta_{s,a'}} \cdot G(t', t)}_{\geq 0, \text{by the definition of } \mathcal{J}^{(t)}} \tag{110}$$

$$+ \sum_{t' \leq \nu(t)} \underbrace{\eta^{(t'+1)} \cdot \Big(\frac{\partial V^{(t')}(\mu)}{\partial \theta_{s,a'}} + \frac{1}{(1-\gamma)^2}\Big) \cdot G(t', t)}_{\geq 0, \text{by the fact that } |\partial V^{(t')}(\mu)/\partial \theta_{s,a'}| \leq 1/(1-\gamma)^2} \tag{111}$$

$$\leq (\theta_{s,a'}^{(t)} - \theta_{s,a}^{(1)}) + \sum_{t' \leq \nu(t)} \eta^{(t'+1)} \frac{1}{(1-\gamma)^2} G(t', t), \tag{112}$$

where (112) holds by the update scheme of APG as in Algorithm 2. Note that as $\liminf_{t\to\infty} \theta_{s,a'}^{(t)} = -\infty$, then $\nu(t)$ must be finite, for all t. This also implies that $\sum_{t' \leq \nu(t)} \eta^{(t'+1)} \frac{1}{(1-\gamma)^2} G(t', t)$ is finite, for all t. Therefore, by taking the limit infimum on both sides of (112), we know

$$\liminf_{t\to\infty} Z^{(t)} = -\infty. \tag{113}$$

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 must have

$$\frac{|\partial V^{(t')}(\mu)/\partial \theta_{s,a}|}{|\partial V^{(t')}(\mu)/\partial \theta_{s,a'}|} = \left|\frac{\pi^{(t')}(a|s) A^{(t')(s,a)}}{\pi^{(t')}(a'|s) A^{(t')(s,a')}}\right| \geq \exp(\vartheta - \theta_{s,a'}^{(t')}) \cdot \frac{(1-\gamma)\Delta_s}{4} \geq \exp(\delta) \cdot \frac{(1-\gamma)\Delta_s}{4}, \tag{114}$$

where the first inequality follows from that $|A^{(t')}(s,a)| \leq 1/(1-\gamma)$ and that $A^{(t')}(s,a) \leq -\Delta_s/4$, and the second equality holds by the definition of $\nu(t)$. For any $\mathcal{J}^{(t)} \neq \varnothing$, we have

$$\theta_{s,a}^{(t)} - \theta_{s,a}^{(1)} = \sum_{t': 1 \leq t' < \bar{T}_s} \eta^{(t'+1)} \cdot \frac{\partial V^{(t')}(\mu)}{\partial \theta_{s,a}} \cdot G(t', t) + \sum_{t': t' \geq \bar{T}_s} \eta^{(t'+1)} \cdot \frac{\partial V^{(t')}(\mu)}{\partial \theta_{s,a}} \cdot G(t', t) \tag{115}$$

$$\leq \sum_{t': 1 \leq t' < \bar{T}_s} \eta^{(t'+1)} \cdot \frac{\partial V^{(t')}(\mu)}{\partial \theta_{s,a}} \cdot G(t', t) + \sum_{t': t' \in \mathcal{J}^{(t)}} \eta^{(t'+1)} \cdot \frac{\partial V^{(t')}(\mu)}{\partial \theta_{s,a}} \cdot G(t', t) \tag{116}$$

$$\leq \underbrace{\sum_{t': 1 \leq t' < \bar{T}_s} \eta^{(t'+1)} \cdot \frac{\partial V^{(t')}(\mu)}{\partial \theta_{s,a}} \cdot G(t', t)}_{< \infty \text{ and does not depend on } t} + \exp(\delta) \cdot \frac{(1-\gamma)\Delta_s}{4} \underbrace{\sum_{t': t' \in \mathcal{J}^{(t)}} \eta^{(t'+1)} \cdot \frac{\partial V^{(t')}(\mu)}{\partial \theta_{s,a'}} \cdot G(t', t)}_{\equiv Z^{(t)}}, \tag{117}$$

where (116) holds by the fact that $A^{(t)}(s,a) < 0$ for all $t \geq \bar{T}_s$ and (117) is a direct result of (114). Therefore, by taking the limit infimum on both sides of (117), we have $\liminf_{t\to\infty} \theta_{s,a}^{(t)} = -\infty$, which leads to contradiction. □

For ease of notation, we define $\Delta\theta_{s,a}^{(t)} := \theta_{s,a}^{(t)} - \theta_{s,a}^{(t-1)}$, for each state-action pair s, a and $t \in \mathbb{N}$.

Lemma 17. *Consider any state s with non-empty I_s^+. Let $a_+ \in I_s^+$ and $a \in I_s^0$. Suppose $\theta_{s,a_+}^{(\tau)} > \theta_{s,a}^{(\tau)}$ and $\Delta\theta_{s,a_+}^{(\tau)} > \Delta\theta_{s,a}^{(\tau)}$, then we also have $\theta_{s,a_+}^{(t)} > \theta_{s,a}^{(t)}$ and $\Delta\theta_{s,a_+}^{(t)} > \Delta\theta_{s,a}^{(t)}$, for all $t > \tau$.*

Proof of Lemma 17. We prove this by induction. Suppose at some time $\tau > \bar{T}_s$, we have $\theta_{s,a_+}^{(\tau)} > \theta_{s,a}^{(\tau)}$ and $\Delta\theta_{s,a_+}^{(\tau)} > \Delta\theta_{s,a}^{(\tau)}$. Then, we have

$$\omega_{s,a_+}^{(\tau)} = \theta_{s,a_+}^{(\tau)} + \frac{\tau-1}{\tau+2}(\theta_{s,a_+}^{(\tau)} - \theta_{s,a_+}^{(\tau-1)}) > \theta_{s,a}^{(\tau)} + \frac{\tau-1}{\tau+2}(\theta_{s,a}^{(\tau)} - \theta_{s,a}^{(\tau-1)}) = \omega_{s,a}^{(\tau)} \tag{118}$$

Recall that we use $A^{(t)}(s,a)$, $Q^{(t)}(s,a)$ and $V^{(t)}(s)$ as the shorthand of $A^{\pi_\omega^{(t)}}(s,a)$, $Q^{\pi_\omega^{(t)}}(s,a)$ and $V^{\pi_\omega^{(t)}}(s)$, respectively. Note that

$$\frac{\partial V^{(t)}(s)}{\partial \theta_{s,a_+}} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_{\omega^{(t)}}}(s) \cdot \pi_\omega^{(t)}(a_+|s) \cdot (Q^{(t)}(s,a_+) - V^{(t)}(s)) \tag{119}$$

$$> \frac{1}{1-\gamma} \cdot d_\mu^{\pi_{\omega^{(t)}}}(s) \cdot \pi_\omega^{(t)}(a|s) \cdot (Q^{(t)}(s,a) - V^{(t)}(s)) \tag{120}$$

$$= \frac{\partial V^{(t)}(s)}{\partial \theta_{s,a}}, \tag{121}$$

where (120) holds by (118) and the fact that $\tau > \bar{T}_s$ implies $A^{(t)}(s,a_+) \geq A^{(t)}(s,a)$. Therefore, by (118)-(121), we must have

$$\theta_{s,a_+}^{(\tau+1)} = \omega_{s,a_+}^{(\tau)} + \eta^{(t+1)}\frac{\partial V^{(t)}(s)}{\partial \theta_{s,a_+}} > \omega_{s,a}^{(\tau)} + \eta^{(t+1)}\frac{\partial V^{(t)}(s)}{\partial \theta_{s,a}} = \theta_{s,a}^{(\tau+1)}, \tag{122}$$

$$\Delta\theta_{s,a_+}^{(\tau+1)} = \frac{\tau-1}{\tau+2}(\theta_{s,a_+}^{(\tau)} - \theta_{s,a_+}^{(\tau-1)}) + \frac{\partial V^{(t)}(s)}{\partial \theta_{s,a_+}} > \frac{\tau-1}{\tau+2}(\theta_{s,a}^{(\tau)} - \theta_{s,a}^{(\tau-1)}) + \frac{\partial V^{(t)}(s)}{\partial \theta_{s,a}} = \Delta\theta_{s,a}^{(\tau+1)}. \tag{123}$$

By repeating the above argument, we know $\theta_{s,a_+}^t > \theta_{s,a}^t$ 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: For any state with non-empty I_s^+, for any $a \in I_s^+$, we define

$$B_s^0(a_+) := \left\{ a \in I_s^0 : \text{For any } t \geq \bar{T}_s, \text{either } \theta_{s,a}^{(t)} \geq \theta_{s,a_+}^{(t)} \text{ or } \Delta\theta_{s,a}^{(t)} \geq \Delta\theta_{s,a_+}^{(t)} \right\} \tag{124}$$

We use $\bar{B}_s^0(a_+)$ to denote the complement of $B_s^0(a_+)$. As a result, we could write $\bar{B}_s^0(a_+)$ as

$$\bar{B}_s^0(a_+) := \left\{ a \in I_s^0 : \theta_{s,a}^{(t)} < \theta_{s,a_+}^{(t)} \text{ and } \Delta\theta_{s,a}^{(t)} < \Delta\theta_{s,a_+}^{(t)}, \text{for some } t \geq \bar{T}_s \right\}. \tag{125}$$

Lemma 18. *Under APG, if I_s^+ is not empty, then:*

a) *For any $a_+ \in I_s^+$, we have*

$$\sum_{a \in B_s^0(a_+)} \pi_\theta^{(t)}(a|s) \to 1, \quad \text{as } t \to \infty. \tag{126}$$

b) *For any $a_+ \in I_s^+$, we have*

$$\max_{a \in B_s^0(a_+)} \theta_{s,a}^{(t)} \to \infty, \quad \text{as } t \to \infty. \tag{127}$$

c) *For any $a_+ \in I_s^+$, we have*

$$\sum_{a \in B_s^0(a_+)} \theta_{s,a}^{(t)} \to \infty, \quad \text{as } t \to \infty. \tag{128}$$

Proof of Lemma 18. Regarding **(a)**, by the definition of $\bar{B}_s^0(a_+)$, for each $a \in B_s^0(a_+)$, there must exist some $T' \geq \bar{T}_s$ such that $\theta_{s,a_+}^{(T')} > \theta_{s,a}^{(T')}$ and $\Delta\theta_{s,a_+}^{(T')} > \Delta\theta_{s,a}^{(T')}$. Then, by Lemma 17, we know

$$\theta_{s,a_+}^{(t)} > \theta_{s,a}^{(t)} \text{ and } \Delta\theta_{s,a_+}^{(t)} > \Delta\theta_{s,a}^{(t)}, \quad \text{for all } t \geq T'. \tag{129}$$

Moreover, by Lemma 11 and that $a_+ \in I_s^+$, we have $\pi^{(t)}(a_+|s) \to 0$ as $t \to \infty$. Based on (129), this shall further imply that $\pi^{(t)}(a_+|s) \to 0$ as $t \to \infty$, for any $a \in \bar{B}_s^0(a_+)$. Hence, we conclude that

$$\sum_{a \in \bar{B}_s^0(a_+)} \pi^{(t)}(a|s) \to 1, \quad \text{as } t \to \infty. \tag{130}$$

Regarding **(b)**, based on the result in **(a)**, we could leverage exactly the same argument as that of Lemma 15 and obtain that $\max_{a \in B_s^0(a_+)} \theta_{s,a}^{(t)} \to \infty$, as $t \to \infty$.

Regarding **(c)**, let us consider any action $a \in \bar{B}_s^0(a_+)$. By the definition of $B_s^0(a_+)$, at each iteration $t \geq \bar{T}_s$, either $\theta_{s,a}^{(t)} \geq \theta_{s,a_+}^{(t)}$ or $\Delta\theta_{s,a}^{(t)} \geq \Delta\theta_{s,a_+}^{(t)}$ holds. As a result, by Lemma 13, we know that $\theta_{s,a}^{(t)}$ must also be bounded from below, for all t. Therefore, based on the result in **(b)**, we know $\sum_{a \in B_s^0(a_+)} \theta_{s,a}^{(t)} \to \infty$, as $t \to \infty$. □

Lemma 19. *Under APG, for any $a_+ \in I_s^+$, the following two properties about $\bar{B}_s^0(a_+)$ shall hold:*

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

$$\pi^{(t)}(a_+|s) > \pi^{(t)}(a|s) \quad \textit{for all } t > T_{a_+}. \tag{131}$$

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

$$|A^{(t)}(s,a)| < \frac{\pi^{(t)}(a_+|s)}{\pi^{(t)}(a|s)} \cdot \frac{\Delta_s}{16|\mathcal{A}|}, \quad \textit{for all } t > T_{a_+}^\dagger. \tag{132}$$

Moreover, this also implies that

$$\sum_{a \in \bar{B}_s^0(a_+)} \pi^{(t)}(a|s)|A^{(t)}(s,a)| < \pi^{(t)}(a_+|s) \cdot \frac{\Delta_s}{16}, \quad \textit{for all } t > T_{a_+}^\dagger. \tag{133}$$

Proof of Lemma 19. Regarding **(a)**, for each $a \in \bar{B}_s^0(a_+)$, we define

$$u_a(a_+) := \inf\{\tau \geq \bar{T}_s : \theta_{s,a_+}^{(\tau)} > \theta_{s,a}^{(\tau)} \text{ and } \Delta\theta_{s,a_+}^{(\tau)} > \Delta\theta_{s,a}^{(\tau)}\}. \tag{134}$$

By the definition of $\bar{B}_s^0(a_+)$, we know the following two facts: (i) $u_a(a_+)$ is finite, for any $a \in \bar{B}_s^0(a_+)$. (ii) By Lemma 17, for all $t \geq u_a(a_+)$, we must have $\theta_{s,a_+}^{(t)} > \theta_{s,a}^{(t)}$ 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 $\pi^{(t)}(a_+|s) > \pi^{(t)}(a|s)$, for all $t > T_{a_+}$.

Regarding **(b)**, as $\frac{\pi^{(t)}(a_+|s)}{\pi^{(t)}(a|s)} > 1$ for all $t > T_{a_+}$ (this is a direct result of **(a)**), we know that for each $a \in \bar{B}_s^0(a_+) \subseteq I_s^0$, there must exist some finite $t'_a > T_{a_+}$ such that

$$|A^{(t)}(s,a)| < \frac{\pi^{(t)}(a_+|s)}{\pi^{(t)}(a|s)} \cdot \frac{\Delta_s}{16|\mathcal{A}|}, \quad \text{for all } t \geq t'_a. \tag{135}$$

As a result, by choosing $T_{a_+}^\dagger := \max_{a \in \bar{B}_s^0(a_+)} t'_a$, we conclude that (132)-(133) indeed hold. □

Lemma 20. *If I_s^+ is non-empty, then for any $a_+ \in I_s^+$, there exists some finite $\tilde{T}_{a_+}$ such that*

$$\sum_{a \in I_s^-} \pi^{(t)}(a|s)A^{(t)}(s,a) > -\pi^{(t)}(a_+|s)\frac{\Delta_s}{16}, \quad \textit{for all } t \geq \tilde{T}_{a_+}. \tag{136}$$

Proof. Let $a_+ \in I_s^+$ and $a_- \in I_s^-$. By Lemma 13 and Lemma 16, we know $\theta_{s,a_+}^{(t)}$ is always bounded from below and $\theta_{s,a_-}^{(t)} \to \infty$, as $t \to \infty$. This implies that $\pi^{(t)}(a_-|s)/\pi^{(t)}(a_+|s) \to 0$, as $t \to \infty$. Therefore, there must exist some finite t'_{a_-} such that

$$\frac{\pi^{(t)}(a_-|s)}{\pi^{(t)}(a_+|s)} < \frac{\Delta_s(1-\gamma)}{16|\mathcal{A}|}, \quad \text{for all } t \geq t'_{a_-}. \tag{137}$$

By choosing $\tilde{T}_{a_+} := \max_{a_- \in I_s^-} t'_{a_-}$, we know (136) holds for all $t \geq \tilde{T}_{a_+}$. □

C.3. Putting Everything Together: Asymptotic Convergence of APG

Now we are ready to put everything together and prove Theorem 1. For ease of exposition, we restate Theorem 1 as follows.

Theorem 1. ***(Global convergence under softmax parameterization)*** *Consider a tabular softmax parameterized policy π_θ. Under APG with $\eta^{(t)} = \frac{t}{t+1} \cdot \frac{1}{5}$, we have $V^{\pi_\theta^{(t)}}(s) \to V^*(s)$ as $t \to \infty$, for all $s \in \mathcal{S}$.*

Proof of Theorem 1. 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 (82)-(84), Lemma 19, and Lemma 20, respectively. We define $T_{\max} := \max\{\bar{T}_s, T_{a_+}, T_{a_+}^\dagger, \tilde{T}_{a_+}\}$. Note that for all $t > T_{\max}$, we have

$$0 = \sum_{a \in B_s^0(a_+)} \pi^{(t)}(a|s) A^{(t)}(s,a) + \underbrace{\sum_{a \in \bar{B}_s^0(a_+)} \pi^{(t)}(a|s) A^{(t)}(s,a)}_{> -\pi^{(t)}(a_+|s)\frac{\Delta_s}{16} \text{ by Lemma 19}} + \underbrace{\sum_{a \in I_s^+} \pi^{(t)}(a|s) A^{(t)}(s,a)}_{\geq \pi^{(t)}(a_+|s)\frac{\Delta_s}{4}} + \underbrace{\sum_{a \in I_s^-} \pi^{(t)}(a|s) A^{(t)}(s,a)}_{> -\pi^{(t)}(a_+|s)\frac{\Delta_s}{16} \text{ by Lemma 20}} \tag{138}$$

$$> \sum_{a \in B_s^0(a_+)} \pi^{(t)}(a|s) A^{(t)}(s,a) + \frac{1}{8} \cdot \pi^{(t)}(a_+|s) \Delta_s \tag{139}$$

$$> \sum_{a \in B_s^0(a_+)} \pi^{(t)}(a|s) A^{(t)}(s,a). \tag{140}$$

Note that (140) implies $\sum_{a \in B_s^0(a_+)} \frac{\partial V^{(t)}(\mu)}{\partial \theta_{s,a}} < 0$, for all $t > T_{\max}$. Moreover, we have

$$\sum_{a \in B_s^0(a_+)} \theta_{s,a}^{(t)} - \theta_{s,a}^{(1)} \tag{141}$$

$$= \sum_{a \in B_s^0(a_+)} \sum_{t'=1}^{t} \eta^{(t'+1)} \cdot \frac{\partial V^{(t')}(\mu)}{\partial \theta_{s,a'}} \cdot G(t',t) \tag{142}$$

$$= \sum_{t'=1}^{t} \eta^{(t'+1)} G(t',t) \cdot \Big(\sum_{a \in B_s^0(a_+)} \frac{\partial V^{(t')}(\mu)}{\partial \theta_{s,a'}} \Big) \tag{143}$$

$$= \underbrace{\sum_{t'=1}^{T_{\max}} \eta^{(t'+1)} G(t',t) \cdot \Big(\sum_{a \in B_s^0(a_+)} \frac{\partial V^{(t')}(\mu)}{\partial \theta_{s,a'}} \Big)}_{< \infty \text{ and does not depend on } t} + \sum_{t'=T_{\max}+1}^{t} \eta^{(t'+1)} G(t',t) \cdot \underbrace{\Big(\sum_{a \in B_s^0(a_+)} \frac{\partial V^{(t')}(\mu)}{\partial \theta_{s,a'}} \Big)}_{< 0 \text{ by (140)}}. \tag{144}$$

By taking the limit of the both sides of (144), we know that the left-hand side of (144) shall go to positive infinity by Lemma 18, but the right-hand side of (144) is bounded from above. This leads to contradiction and hence completes the proof. □

D. Convergence Rates of APG: The Multi-Action Bandit Case

D.1. $\tilde{O}(1/t^2)$ Convergence Rate of APG

Theorem 2. *Consider a tabular softmax parameterized policy π_θ. Under APG with $\eta^{(t)} = \frac{t}{t+1} \cdot \frac{1}{5}$, there exists a finite time T such that for all $t > T$, we have:*

$$\left(\pi^* - \pi_\theta^{(t)}\right)^\top r \leq \frac{|\mathcal{A}| - 1}{(t-T)^2 + |\mathcal{A}| - 1} \tag{10}$$

$$+ \frac{10(2+T)\left(\left\|\theta^{(T)}\right\| + 2\ln(t-T)\right)^2}{t(t+1)}. \tag{11}$$

Proof of Theorem 2.

Claim 2. *The proof can be completed by making the following claims:*

a) *By Lemma 1, the function $\theta \to \pi_\theta^\top r$ is concave if $\theta_{a^*} - \theta_a > \delta$ for all $a \neq a^*$ where $\delta = \ln \frac{r(a_2)(|\mathcal{A}|-1)}{r(a^*)-r(a_2)} + \ln(1 + \binom{|\mathcal{A}|-1}{2}) - \ln(d(a^*) - d(a_i))$.*

b) *By Lemma 2, given such δ, there exists a finite time T such that $\theta_{a^*}^{(t)} - \theta_a^{(t)} > \delta$ holds for all $t > T$, $a \neq a^*$.*

c) *APG enjoys the convergence rate of $\tilde{O}(\frac{1}{t^2})$ after time T.*

Claim c). As the objective $\pi_\theta^\top r$ enters a locally concave region after time T, it is necessary to account for a shift in the initial learning rate due to the passage of time T. In other words, if we divide the update process into two phases based on time T, the latter phase will commence with a modified learning rate. Hence, by Corollary 1 and Lemma 9 with $c = T$ and $\theta^{**} = \theta^{(t)**}$, we have that for all $t > T$:

$$\pi_{\theta^{**}}^{(t)\top} r - \pi_\theta^{(t)\top} r \leq \frac{4L(2+T)\left\|\theta^{(0)} - \theta^{(t)**}\right\|^2}{(t+T+1)(t+T)} \tag{145}$$

$$= \frac{10(2+T)\left\|\theta^{(T)} - \theta^{(t)**}\right\|^2}{(t+T+1)(t+T)} \tag{146}$$

$$= \frac{10(2+T)\left(\left\|\theta^{(T)}\right\| + 2\ln(t-T)\right)^2}{(t+T+1)(t+T)}, \tag{147}$$

where $\theta^{(t)**} := [2\ln(t-T), 0, 0, \cdots, 0]$ is a chosen surrogate optimal solution at time t. Additionally, we have the sup-optimality gap between the original optimal solution and the surrogate optimal solution can be bounded as:

$$\pi^{*\top} r - \pi_{\theta^{**}}^{(t)\top} r = r(a^*) - \pi_{\theta^{**}}^{(t)\top} r \tag{148}$$

$$\leq r(a^*)\left(1 - \pi_{\theta^{**}}^{(t)}(a^*)\right) \tag{149}$$

$$= r(a^*)\left(1 - \frac{\exp(2\ln(t-T))}{\exp(2\ln(t-T)) + \exp(0)\cdot(|\mathcal{A}|-1)}\right) \tag{150}$$

$$= r(a^*)\left(\frac{|\mathcal{A}|-1}{(t-T)^2 + |\mathcal{A}| - 1}\right) \tag{151}$$

$$\leq \frac{|\mathcal{A}|-1}{(t-T)^2 + |\mathcal{A}| - 1}, \tag{152}$$

where (149) holds since we ignore the terms $\pi_{\theta^{**}}^{(t)}(a) \cdot r(a) \geq 0$ for every $a \neq a^*$ .Combining (145-152), we get our desired result.

Remark 10. It is crucial to highlight that we consider the surrogate optimal solution $\theta^{(t)**}$ instead of the original optimal solution θ^* due to the unbounded optimal solution of softmax parameterization.

□

Lemma 1. ***(Local Concavity; Informal).*** *The function $\theta \to \pi_\theta^\top r$ is concave if $\theta_{a^*} - \theta_a > \delta$ for some $\delta > 0$, for all $a \neq a^*$.*

Proof of Lemma 1.

Claim 3. *The proof can be completed by making the following claims:*

a) *The function $\theta \to \pi_\theta(a^*)$ is concave and the functions $\theta \to \pi_\theta(a)$ are convex if $\theta_{a^*} - \theta_a > \delta$ for all $a \neq a^*$ where $\delta = \ln \frac{r(a_2)(|\mathcal{A}|-1)}{r(a^*)-r(a_2)} + \ln(1 + \binom{|\mathcal{A}|-1}{2}) - \ln(d(a^*) - d(a_i))$.*

b) *The function $\theta \to \pi_\theta^\top r'$ is concave if $\theta_{a^*} - \theta_a > \delta$ for all $a \neq a^*$, where $r' = [r(a^*) - r(a_2), r(a_2) - r(a_2), \cdots, r(a_{|\mathcal{A}|}) - r(a_2)]$ is the shifted reward function.*

c) *The update of the objective $\pi_\theta^\top r$ under the original reward r is equivalent to the update of the objective $\pi_\theta^\top r'$ under any shifted reward function $r' = \big[r(a^*) - c, r(a_2) - c, \cdots, r(a_{|\mathcal{A}|}) - c\big]$, where $c \in \mathbb{R}$ is a constant.*

d) *Combining **(a)-(c)**,we can conclude that the objective function $\pi_\theta^\top r$ will undergo an update identical to that of the concave function $\pi_\theta^\top r'$ if $\theta_a - \theta_a > \delta$ for all $a \neq a$.*

Claim a). We establish the convexity of the function $\theta \to \pi_\theta(a)$ for all $a \neq a^*$ by demonstrating that if $\theta_{a^*} - \theta_a > \ln \frac{r(a_2)(|\mathcal{A}|-1)}{r(a^*)-r(a_2)} + \ln(1 + \binom{|\mathcal{A}|-1}{2}) - \ln(d(a^*) - d(a_i))$ for all $a \neq a^*$, then the function is convex. Following that, since $\pi_\theta(a^*) = 1 - \sum_{a \neq a^*} \pi_\theta(a)$ can be viewed as a summation of concave functions, it follows that $\theta \to \pi_\theta(a^*)$ is concave.

Given an action $a_i \neq a^*$, since the concavity is determined by the behavior of a function on arbitrary line on its domain, it is sufficient to show that the following function is concave (i.e. the second derivative is non-positive) when $k \to 0$:

$$f(k) = \frac{e^{\theta_{a_i} + k \cdot d_i}}{e^{\theta_{a^*} + k \cdot d_1} + e^{\theta_{a_2} + k \cdot d_2} + \cdots + e^{\theta_{a_{|\mathcal{A}|}} + k \cdot d_{|\mathcal{A}|}}} \tag{153}$$

$$= \frac{1}{e^{(\theta_{a^*} - \theta_{a_i}) + k \cdot (d_1 - d_i)} + e^{(\theta_{a_2} - \theta_{a_i}) + k \cdot (d_2 - d_i)} + \cdots + e^{(\theta_{a_{|\mathcal{A}|}} - \theta_{a_i}) + k \cdot (d_{|\mathcal{A}|} - d_i)}} \tag{154}$$

$$:= \frac{1}{m(k)}, \tag{155}$$

where $d = [d_1, d_2, \cdots, d_{|\mathcal{A}|}]$ is any unit vector on the domain.

By taking the second derivative of $f(k)$, we have:

$$f^{''}(k) = \frac{2(m^{'}(k))^2}{m(k)^3} - \frac{m^{''}(k)}{m(k)^2}. \tag{156}$$

And so, we have the second derivative of $f(k)$ when $k \to 0$ is:

$$f^{''}(0) = \frac{1}{m(0)^2} \left(\frac{2(m^{'}(0))^2}{m(0)} - m^{''}(0) \right). \tag{157}$$

Note that since $m(k) \geq 0$ for all k, we have that $f^{''}(0) > 0$ (convex) if and only if:

$$2(m^{'}(0))^2 - m^{''}(0) \cdot m(0) > 0, \tag{158}$$

where $m'(0) = \sum_{a \neq a_i} (d(a) - d(a_i)) \exp(\theta_a - \theta_{a_i})$ and $m''(0) = \sum_{a \neq a_i} (d(a) - d(a_i))^2 \exp(\theta_a - \theta_{a_i})$.

By plugging $m(0), m'(0), m''(0)$ into (158) we have:

$$2(m'(0))^2 - m^{''}(0) \cdot m(0) = 2 \sum_{a \neq a_i} (d(a) - d(a_i))^2 \exp(2\theta_a - 2\theta_{a_i}) \tag{159}$$

$$+2\sum_{a,a'\neq a_i,a\neq a'}2(d(a)-d(a_i))(d(a')-d(a_i))\exp(\theta_a+\theta_{a'}-2\theta_{a_i}) \tag{160}$$

$$-\sum_{a\neq a_i}(d(a)-d(a_i))^2\exp(2\theta_a-2\theta_{a_i}) \tag{161}$$

$$-\sum_{a,a'\neq a_i,a\neq a'}\left((d(a)-d(a_i))^2+(d(a')-d(a_i))^2\right)\exp(\theta_a+\theta_{a'}-2\theta_{a_i}) \tag{162}$$

$$-\sum_{a\neq a_i}(d(a)-d(a_i))^2\exp(\theta_a-\theta_{a_i}) \tag{163}$$

$$=\sum_{a\neq a_i}(d(a)-d(a_i))^2\left(\exp(2\theta_a-2\theta_{a_i})-\exp(\theta_a-\theta_{a_i})\right) \tag{164}$$

$$+\sum_{a,a'\neq a_i,a\neq a'}2(d(a)-d(a_i))(d(a')-d(a_i))\exp(\theta_a+\theta_{a'}-2\theta_{a_i}) \tag{165}$$

$$-\sum_{a,a'\neq a_i,a\neq a'}(d(a)^2+d(a')^2)\exp(\theta_a+\theta_{a'}-2\theta_{a_i}) \tag{166}$$

$$\geq(d(a^*)-d(a_i))\left(\exp(2\theta_{a^*}-2\theta_{a_i})-\exp(\theta_{a^*}-\theta_{a_i})\right) \tag{167}$$

$$-\sum_{a,a'\neq a_i,a\neq a'}(d(a)^2+d(a')^2)\exp(\theta_a+\theta_{a'}-2\theta_{a_i}) \tag{168}$$

$$\geq(d(a^*)-d(a_i))\exp(2\theta_{a^*}-2\theta_{a_i}) \tag{169}$$

$$-\left(1+\binom{|\mathcal{A}|-1}{2}\right)\max_{a,a'\neq a_i,a\neq a'}\exp(\theta_a+\theta_{a'}-2\theta_{a_i}) \tag{170}$$

$$>0,\text{if }\theta_{a^*}-\theta_a>\ln(1+\binom{|\mathcal{A}|-1}{2})-\ln(d(a^*)-d(a_i))\text{ for all }a\neq a^*. \tag{171}$$

To ensure $-\ln(d(a^*)-d(a_i))$ is bounded, we can introduce a tighter bound by considering $\delta=\ln\frac{r(a_2)(|\mathcal{A}|-1)}{r(a^*)-r(a_2)}+\ln(1+\binom{|\mathcal{A}|-1}{2})-\ln(d(a^*)-d(a_i))$. This choice of δ ensures that $\pi_\theta^\top r>r(a_2)$, which guarantees that the domain for d_1 is positive, while the domains for the other d_i values, where $i=2,3,\ldots,|\mathcal{A}|$, are negative.

Claim b) Since the sum of several concave functions is itself concave, we separate our proof into 2 steps and leverages Lemma 6 to show that the function $\theta\to\pi_\theta(a)(r(a)-r(a_2))$ is concave for all $a\in\mathcal{A}$ if $\theta_{a^*}-\theta_a>\delta$, for all $a\neq a^*$:

- The function $\theta\to\pi_\theta(a^*)(r(a^*)-r(a_2))$ is concave if $\theta_{a^*}-\theta_a>\delta$, for all $a\neq a^*$:
  Since we've already shown the concavity of the function $\theta\to\pi_\theta(a^*)$, it remains to show the concavity and the non-decrease of the function $\pi_\theta(a^*)\to\pi_\theta(a^*)\cdot(r(a^*)-r(a_2))$. And this property directly hold since $\pi_\theta(a^*)\cdot(r(a^*)-r(a_2))$ is a linear function of $\pi_\theta(a^*)$ with $\frac{\partial\pi_\theta(a^*)\cdot(r(a^*)-r(a_2))}{\partial\pi_\theta(a^*)}=r(a^*)-r(a_2)>0$ and $\frac{\partial^2\pi_\theta(a^*)\cdot(r(a^*)-r(a_2))}{\partial\pi(a^*)^2}=0$.
- The function $\theta\to\pi_\theta(a_i)(r(a_i)-r(a_2))$ is concave for all $i=2,3,\cdots,|\mathcal{A}|$, if $\theta_{a^*}-\theta_a>\delta$, for all $a\neq a^*$:
  Since $r(a_i)-r(a_2)\leq 0$ for all $i=2,3,\cdots,|\mathcal{A}|$, we have $\theta\to\pi_\theta(a_i)(r(a_i)-r(a_2))=-\pi_\theta(a_i)\cdot|r(a_i)-r(a_2)|$. Also by the convexity of the function $\theta\to\pi_\theta(a)$, we have the function $\theta\to-\pi_\theta(a)$ is concave. And it remains to show the concavity and the non-decrease of the function $\pi_\theta(a_i)\to\pi_\theta(a^*)\cdot|r(a_i)-r(a_2)|$. This property directly hold since $\pi_\theta(a_i)\cdot|r(a_i)-r(a_2)|$ is a linear function of $\pi_\theta(a^*)$ with $\frac{\partial\pi_\theta(a_i)\cdot|r(a_i)-r(a_2)|}{\partial\pi_\theta(a_i)}=|r(a_i)-r(a_2)|\geq 0$ and $\frac{\partial^2\pi_\theta(a_i)\cdot|r(a_i)-r(a_2)|}{\partial\pi_\theta(a_i)^2}=0$.

Claim c) To reach the equivalent update result, it is sufficient to show that the gradient of the original objective $\pi_\theta^\top r$ is equal to the shifted objective $\pi_\theta^\top r'$ under all θ. By Lemma 8, we have for all $a\in\mathcal{A}$:

$$\left.\frac{\partial\pi_\theta^\top r}{\partial\theta_a}\right|_{\theta=\theta}=\pi_\theta(a)\cdot(r(a)-\pi_\theta^\top r)=\pi_\theta(a)\cdot((r(a)-c)-(\pi_\theta^\top r-c))=\left.\frac{\partial\pi_\theta^\top r'}{\partial\theta_a}\right|_{\theta=\theta}, \tag{172}$$

which complete our proof.

□

Lemma 2. *Consider a tabular softmax parameterized policy π_θ. Under APG with $\eta^{(t)} = \frac{t}{t+1} \cdot \frac{1}{5}$, given any $\delta > 0$, there exists a finite time T such that for all $t > T$, we have $\theta_{a^*} - \theta_a > \delta$, for all $a \neq a^*$.*

Proof of Lemma 2.

Claim 4. *The proof can be completed by making the following claims:*

a) *Under APG, the gradient norm will not converge at the sub-optimal policy, i.e. we have* $\inf_{t\geq 0, \pi_\theta^{(t)\top} r \in [0, r(a_2)]} \left\| \frac{\partial \pi_\theta^\top r}{\partial \theta}\Big|_{\theta=\theta^{(t)}} \right\| \geq \inf_{t\geq 0} \pi_\theta^{(t)}(a^*) \cdot (r(a^*) - r(a_2)) > 0.$

b) *Given any $\delta > 0$, there exist an $\epsilon > 0$ such that if $\left\| \frac{\partial \pi_\theta^\top r}{\partial \theta}\Big|_{\theta=\theta^{(t)}} \right\| < \epsilon$, then $\theta_{a^*} - \theta_a > \delta$ for all $a \neq a^*$.*

c) *Given any $\epsilon > 0$, there exist a finite time T such that $\left\| \frac{\partial \pi_\theta^\top r}{\partial \theta}\Big|_{\theta=\theta^{(T)}} \right\| < \epsilon$, leading to the fact that $\theta_{a^*}^{(T)} - \theta_a^{(T)} > \delta$ for all $a \neq a^*$ at time T.*

d) *We have $\theta_{a*}^{(t)} - \theta_a^{(t)} > \delta$ for all $a \neq a^*, t \geq T$.*

Claim a) To reach the desired result, we leverage Lemma 3 to ensure the gradient norm is bounded away from 0 before achieving optimal policy:

$$\left\| \frac{\partial \pi_\theta^\top r}{\partial \theta}\Big|_{\theta=\theta^{(t)}} \right\| = \sqrt{\sum_{a\in\mathcal{A}} \pi_\theta^{(t)}(a)^2 \cdot (r(a) - \pi_\theta^{(t)\top} r)^2} \tag{173}$$

$$\geq \sqrt{\pi_\theta^{(t)}(a^*)^2 \cdot (r(a^*) - \pi_\theta^{(t)\top} r)^2} \tag{174}$$

$$= \pi_\theta^{(t)}(a^*) \cdot (r(a^*) - \pi_\theta^{(t)\top} r) \tag{175}$$

$$\geq \inf_{t\geq 0} \pi_\theta^{(t)}(a^*) \cdot (r(a^*) - r(a_2)) \tag{176}$$

$$> 0, \quad \text{for all } \pi_\theta^{(t)\top} r \in [0, r(a_2)], \tag{177}$$

where (173) is followed by Lemma 8 and (177) is by Lemma 3.

Claim b) By choosing ϵ with:

$$\epsilon = \min \left\{ \inf_{t\geq 0, \pi_\theta^{(t)\top} r \in [0, r(a_2)]} \left\| \frac{\partial \pi_\theta^\top r}{\partial \theta}\Big|_{\theta=\theta^{(t)}} \right\|, \right. \tag{178}$$

$$\left. \frac{1}{e^\delta + |\mathcal{A}| - 1} \cdot \inf_{t\geq 0, a\neq a^*, \pi_\theta^{(t)\top} r \in (r(a_2), r(a^*))} |r(a) - \pi_\theta^{(t)\top} r| \right\} > 0, \tag{179}$$

we could ensure that the minimum must occur at $\pi_\theta^\top r \in [r(a^*), r(a_2))$. And so we have $\inf_{t\geq 0, a\neq a^*, \pi_\theta^{(t)\top} r \in (r(a_2), r(a^*))} |r(a) - \pi_\theta^{(t)\top} r| > 0$.

Hence, by Lemma 8, we have for all $a \in \mathcal{A}$:

$$\frac{\inf_{t\geq 0, a\neq a^*, \pi_\theta^{(t)\top} r \in (r(a_2), r(a^*))} |r(a) - \pi_\theta^{(t)\top} r|}{e^\delta + (|\mathcal{A}| - 1)} > \left\| \nabla_\theta \pi_\theta^\top r \right\| \geq \pi_\theta(a) \cdot |r(a) - \pi_\theta^\top r| \tag{180}$$

by rearranging (180), we get:

$$\frac{1}{e^\delta + (|\mathcal{A}| - 1)} \geq \frac{\inf_{t\geq 0, a\neq a^*, \pi_\theta^{(t)\top} r \in (r(a_2), r(a^*))} |r(a) - \pi_\theta^{(t)\top} r|}{|r(a) - \pi_\theta^\top r|} \cdot \frac{1}{e^\delta + (|\mathcal{A}| - 1)} > \pi_\theta(a) \tag{181}$$

for all $a \neq a^*$, leading to our desired result.

Claim c) By Theorem 1, we have that there exist a finite time T_0 such that $\theta_{a^*}^{(t)} > \theta_a^{(t)}$ for all $a \neq a^*$ and $\theta_a^{(t)}$ will be decreasing for all $a \neq a^*$ for all $t \geq T_0$. If we run K iterations, the above statement leads to the fact that $\max_{k=1,2,\cdots,K} \left\|\theta^{(k)}\right\|^2 \leq \max_{t'\leq T_0} \left\|\theta^{(t')}\right\|^2 + |\mathcal{A}|(\theta_{a^*}^{(K)})^2$. We discuss two possible cases as follows:

- **Case 1:** $\theta_{a^*}^{(K)} \leq 2\ln(K) + \max_{t' \leq T_0, a \neq a^*}\{\theta_a^{(t')}\} + \ln(|\mathcal{A}| - 1)$:

  By Corollary 2 with the surrogate optimal solution $\theta^{**} \coloneqq \theta^{(K)**} \coloneqq [2\ln(K), 0, 0, \cdots, 0]$, we have:

$$\min_{k=1,2,\cdots,K} \left\| \frac{\partial \pi_\theta^\top r}{\partial \theta}\Big|_{\theta=\theta^{(k)}} \right\| \leq 192L^2 \frac{\left\|\theta^{(0)} - \theta^{(K)**}\right\|}{K(K+1)(2K+1)} + \frac{48L^2}{2K+1}(4|\ln(K)|^2 + \max_{k=1,2,\cdots,K} \left\|\theta^{(k)}\right\|^2) \tag{182}$$

$$+ \frac{48L \cdot |V^{\pi_{\theta^{(K)}}}(\mu) - V^{\pi_{\theta^{**}}}(\mu)|}{2K+1} \tag{183}$$

$$\leq 192L^2 \frac{\left\|\theta^{(0)}\right\| + 2|\ln(K)|}{K(K+1)(2K+1)} \tag{184}$$

$$+ \frac{48L^2}{2K+1}(4|\ln(K)|^2 + \max_{t' \leq T_0} \left\|\theta^{(t')}\right\|^2 + |\mathcal{A}|(2\ln(K) + \max_{t' \leq T_0, a \neq a^*}\{\theta_a^{(t')}\} + \ln(|\mathcal{A}| - 1))^2) \tag{185}$$

$$+ \frac{48L \cdot |V^{\pi_{\theta^{(K)}}}(\mu) - V^{\pi_{\theta^{**}}}(\mu)|}{2K+1} \tag{186}$$

$$= \tilde{O}(\frac{1}{K}). \tag{187}$$

  Hence the result directly holds by choosing (by changing variable from K to t),

$$T = \inf\Bigg\{ t : \epsilon \geq 192L^2 \frac{\left\|\theta^{(0)}\right\| + 2|\ln(t)|}{t(t+1)(2t+1)} \tag{188}$$

$$+ \frac{48L^2}{2t+1}(4|\ln(t)|^2 + \max_{t \leq T_0} \left\|\theta^{(t)}\right\|^2 + |\mathcal{A}|(2\ln(t) + \max_{t \leq T_0, a \neq a^*}\{\theta_a\} + \ln(|\mathcal{A}| - 1))^2) \tag{189}$$

$$+ \frac{48L \cdot |V^{\pi_{\theta^{(t)}}}(\mu) - V^{\pi_{\theta^{**}}}(\mu)|}{2t+1} \Bigg\}. \tag{190}$$

- **Case 2:** $\theta_{a^*}^{(K)} > 2\ln(K) + \max_{t' \leq T_0, a \neq a^*}\{\theta_a^{(t')}\} + \ln(|\mathcal{A}| - 1)$:

  Given such $\theta_{a^*}^{(K)}$, we have:

$$\pi_\theta^{(K)}(a^*) = \frac{\exp(\theta_{a^*}^{(K)})}{\sum_{a \in \mathcal{A}} \exp(\theta_a^{(K)})} > \frac{\exp(2\ln(K))}{\exp(2\ln(K)) + \sum_{a \neq \mathcal{A}} \exp(-\ln(|\mathcal{A}| - 1))} \geq \frac{K^2}{K^2+1}. \tag{191}$$

  And hence we have:

$$\left\| \frac{\partial \pi_\theta^\top r}{\partial \theta}\Big|_{\theta=\theta^{(K)}} \right\| = \sqrt{\sum_{a \in \mathcal{A}} \pi_\theta^{(K)}(a)^2 \cdot (r(a) - \pi_\theta^{(K)\top} r)^2} \tag{192}$$

$$\leq \sqrt{(r(a^*) - \pi_\theta^{(K)\top} r)^2 + \sum_{a \neq a^*} \pi_\theta^{(K)}(a)^2} \tag{193}$$

$$\leq \sqrt{\frac{1}{(K^2+1)^2} + \frac{1}{(K^2+1)^2}} \tag{194}$$

$$\leq \frac{\sqrt{2}}{K^2+1}. \tag{195}$$

  Hence the result directly holds by choosing (by changing variable from K to t),

$$T = \inf\left\{ t : \epsilon \geq \frac{\sqrt{2}}{t^2+1} \right\}. \tag{196}$$

Claim d) We elaborate on each action $a \neq a^*$ separately. For a fix $a \neq a^*$, since there exists a finite T such that $\theta_{a^*}^{(T)} - \theta_a^{(T)} > \delta$ provided by Claim (c), there is a T_a such that $T_a \leq T$, $\theta_{a^*}^{(T_a)} - \theta_a^{(T_a)} > \delta$, and $\theta_{a^*}^{(T_a-1)} - \theta_a^{(T_a-1)} \leq \delta$. Then, we have

$$\theta_{a^*}^{(T_a)} - \theta_{a^*}^{(T_a-1)} > \theta_a^{(T_a)} - \theta_a^{(T_a-1)} \tag{197}$$

We first show two useful properties:

(i) If $\theta_{a^*}^{(t)} - \theta_a^{(t)} > \delta$ and $\omega_{a^*}^{(t)} - \theta_{a^*}^{(t)} \geq \omega_a^{(t)} - \theta_a^{(t)}$, then $\frac{\partial \pi_\theta^\top r}{\partial \theta_{a^*}}\Big|_{\theta=\omega^{(t)}} \geq \frac{\partial \pi_\theta^\top r}{\partial \theta_a}\Big|_{\theta=\omega^{(t)}}$.

Since $\theta_{a^*}^{(t)} - \theta_a^{(t)} > \delta > 0$, it follows that

$$\begin{aligned}
\omega_{a^*}^{(t)} &= \theta_{a^*}^{(t)} + (\omega_{a^*}^{(t)} - \theta_{a^*}^{(t)}) && (198)\\
&> \theta_a^{(t)} + (\omega_a^{(t)} - \theta_a^{(t)}) && (199)\\
&= \omega_a^{(t)}. && (200)
\end{aligned}$$

By Lemma 8, we have

$$\begin{aligned}
\frac{\partial \pi_\theta^\top r}{\partial \theta_{a^*}}\Big|_{\theta=\omega^{(t)}} &= \pi_{\omega^{(t)}}(a^*) \cdot (r(a^*) - \pi_{\omega^{(t)}}^\top r) && (201)\\
&\geq \pi_{\omega^{(t)}}(a) \cdot (r(a) - \pi_{\omega^{(t)}}^\top r) && (202)\\
&= \frac{\partial \pi_\theta^\top r}{\partial \theta_a}\Big|_{\theta=\omega^{(t)}}, && (203)
\end{aligned}$$

where (202) holds because of that $\omega_{a^*}^{(t)} > \omega_a^{(t)}$ and thus $\pi_{\omega^{(t)}}(a^*) > \pi_{\omega^{(t)}}(a)$, and $r(a^*) > r(a)$.

(ii) If $\theta_{a^*}^{(t)} - \theta_a^{(t)} > \delta$, $\omega_{a^*}^{(t)} - \theta_{a^*}^{(t)} \geq \omega_a^{(t)} - \theta_a^{(t)}$, and $\frac{\partial \pi_\theta^\top r}{\partial \theta_{a^*}}\Big|_{\theta=\omega^{(t)}} \geq \frac{\partial \pi_\theta^\top r}{\partial \theta_a}\Big|_{\theta=\omega^{(t)}}$, then $\theta_{a^*}^{(t+1)} - \theta_a^{(t+1)} > \delta$ and $\omega_{a^*}^{(t+1)} - \theta_{a^*}^{(t+1)} \geq \omega_a^{(t+1)} - \theta_a^{(t+1)}$.

For $\theta_{a^*}^{(t+1)} - \theta_a^{(t+1)} > \delta$,

$$\theta_{a^*}^{(t+1)} - \theta_a^{(t+1)} = \left(\theta_{a^*}^{(t)} + (\omega_{a^*}^{(t)} - \theta_{a^*}^{(t)}) + \eta^{(t+1)} \cdot \frac{\partial \pi_\theta^\top r}{\partial \theta_{a^*}}\Big|_{\theta=\omega^{(t)}}\right) - \left(\theta_a^{(t)} + (\omega_a^{(t)} - \theta_a^{(t)}) + \eta^{(t+1)} \cdot \frac{\partial \pi_\theta^\top r}{\partial \theta_a}\Big|_{\theta=\omega^{(t)}}\right) \tag{204}$$

$$\geq \theta_{a^*}^{(t)} - \theta_a^{(t)} > \delta. \tag{205}$$

For $\omega_{a^*}^{(t+1)} - \theta_{a^*}^{(t+1)} \geq \omega_a^{(t+1)} - \theta_a^{(t+1)}$,

$$\begin{aligned}
\omega_{a^*}^{(t+1)} - \theta_{a^*}^{(t+1)} &= \frac{t}{t+3}(\theta_{a^*}^{(t+1)} - \theta_{a^*}^{(t)}) && (206)\\
&= \frac{t}{t+3}\left(\omega_{a^*}^{(t)} - \theta_{a^*}^{(t)} + \eta^{(t+1)} \frac{\partial \pi_\theta^\top r}{\partial \theta_{a^*}}\Big|_{\theta=\omega^{(t)}}\right) && (207)\\
&\geq \frac{t}{t+3}\left(\omega_a^{(t)} - \theta_a^{(t)} + \eta^{(t+1)} \frac{\partial \pi_\theta^\top r}{\partial \theta_a}\Big|_{\theta=\omega^{(t)}}\right) && (208)\\
&= \frac{t}{t+3}(\theta_a^{(t+1)} - \theta_a^{(t)}) && (209)\\
&= \omega_a^{(t+1)} - \theta_a^{(t+1)}, && (210)
\end{aligned}$$

where (208) is followed by the given hypotheses $\omega_{a^*}^{(t)} - \theta_{a^*}^{(t)} \geq \omega_a^{(t)} - \theta_a^{(t)}$ and $\frac{\partial \pi_\theta^\top r}{\partial \theta_{a^*}}\Big|_{\theta=\omega^{(t)}} \geq \frac{\partial \pi_\theta^\top r}{\partial \theta_a}\Big|_{\theta=\omega^{(t)}}$.

By the above two properties, it suffices to show that there is T' such that $\theta_{a^*}^{(T')} - \theta_a^{(T')} > \delta$ and $\omega_{a^*}^{(T')} - \theta_{a^*}^{(T')} \geq \omega_a^{(T')} - \theta_a^{(T')}$, then we will obtain that $\theta_{a^*}^{(t)} - \theta_a^{(t)} > \delta$ for any $t \geq T'$.

We claim that $T' = T_a$. Since T_a satisfies $\theta_{a^*}^{(T_a)} - \theta_a^{(T_a)} > \delta$, we only need to show that $\omega_{a^*}^{(T_a)} - \theta_{a^*}^{(T_a)} \geq \omega_a^{(T_a)} - \theta_a^{(T_a)}$. To check the condition, we directly expand $\omega_{a^*}^{(T_a)} - \theta_{a^*}^{(T_a)}$, we obtain

$$\omega_{a^*}^{(T_a)} - \theta_{a^*}^{(T_a)} = \frac{T_a - 1}{T_a + 2}(\theta_{a^*}^{(T_a)} - \theta_{a^*}^{(T_a-1)}) \tag{211}$$

$$> \frac{T_a - 1}{T_a + 2}(\theta_a^{(T_a)} - \theta_a^{(T_a-1)}) \tag{212}$$

$$= \omega_a^{(T_a)} - \theta_a^{(T_a)}, \tag{213}$$

where (211), (213) use the update of APG, and (212) holds by (197). Therefore, we show that $\theta_{a^*}^{(T_a)} - \theta_a^{(T_a)} > \delta$ and $\omega_{a^*}^{(T_a)} - \theta_{a^*}^{(T_a)} \geq \omega_a^{(T_a)} - \theta_a^{(T_a)}$. Hence, $\theta_{a^*}^{(t)} - \theta_a^{(t)} > \delta$ for any $t \geq T_a$.

Since the above statement holds for any action $a \neq a^*$, we obtain that $\theta_{a^*}^{(t)} - \theta_a^{(t)} > \delta$ for all $a \neq a^*, t \geq \max_{a \neq a^*} T_a$. Additionally, $T \geq \max_{a \neq a^*} T_a$, so $\theta_{a^*}^{(t)} - \theta_a^{(t)} > \delta$ for all $a \neq a^*, t \geq T$.

□

Lemma 3. *Under APG, we have* $\inf_{t \geq 0} \pi_\theta^{(t)}(a^*) > 0$.

Remark 11. Inspired by the proof of (Mei et al., 2020), we consider two "nice regions" in terms of time: one region (we call it "gradient region") characterized by a positive partial derivative with respect to the optimal action and the negative partial derivatives of all other actions, and another region (we call it "momentum region") characterized by the maximum momentum of the optimal action. If our training process enters these regions, we can ensure that the probability of selecting the optimal action does not decrease. Consequently, the infimum of $\pi_\theta^{(t)}(a^*)$ will remain greater than zero. It is important to highlight that our approach selects a more aggressive gradient region compared to the one chosen in (Mei et al., 2020). By doing so, we are able to streamline their proof to some extent without oversimplifying it.

Proof of Lemma 3. Firstly, we define two sets consisting of time indices with great properties of gradient and momentum.

$$\mathcal{R}_1 = \left\{ t : \left.\frac{\partial \pi_\theta^\top r}{\partial \theta_{a^*}}\right|_{\theta=\omega^{(t)}} \geq 0 \geq \left.\frac{\partial \pi_\theta^\top r}{\partial \theta_a}\right|_{\theta=\omega^{(t)}}, \text{ for all } a \neq a^* \right\}, \tag{214}$$

$$\mathcal{R}_2 = \left\{ t : \omega_{a^*}^{(t)} - \theta_{a^*}^{(t)} \geq \omega_a^{(t)} - \theta_a^{(t)}, \text{ for all } a \neq a^* \right\}. \tag{215}$$

Inspired by (Mei et al., 2020), we make the following claims, and then prove these claims immediately as follows.

Claim 5. *The following hold:*

a) *There exists a finite T_1 such that if $t \geq T_1$, then $\pi_\theta^{(t)}(a^*) \geq \frac{r(a_2)}{r(a^*)}$ and $\pi_\omega^{(t)}(a^*) \geq \frac{r(a_2)}{r(a^*)}$.*

b) *If $t \geq T_1$, then $t \in \mathcal{R}_1$.*

c) *There exists a finite $T_2 \geq T_1$ such that if $t \geq T_2$, then $t \in \mathcal{R}_2$.*

d) *If $t \geq T_2$, then $t \in \mathcal{R}_1 \cap \mathcal{R}_2$. Moreover, we have $\pi_\theta^{(t+1)}(a^*) \geq \pi_\theta^{(t)}(a^*)$. Hence, it follows that*

$$\inf_{t \geq 0} \pi_\theta^{(t)}(a^*) = \min_{0 \leq t \leq T_2} \pi_\theta^{(t)}(a^*). \tag{216}$$

Claim a). By the asymptotic global convergence in Theorem 1, we have that $\pi_\theta^{(t)}(a^*) \to 1$ and $\pi_\omega^{(t)}(a^*) \to 1$ as $t \to \infty$. Since a^* is unique, $\frac{r(a_2)}{r(a^*)} < 1$. According to the $\epsilon - \delta$ argument of limit, there exists a finite T_1 such that if $t \geq T_1$, then $\pi_\theta^{(t)}(a^*) \geq \frac{r(a_2)}{r(a^*)}$ and $\pi_\omega^{(t)}(a^*) \geq \frac{r(a_2)}{r(a^*)}$.

Claim b). Given $t \geq T_1$, by Lemma 8, we have

$$\left.\frac{\partial \pi_\theta^\top r}{\partial \theta_{a^*}}\right|_{\theta=\omega^{(t)}} = \pi_\omega^{(t)}(a^*)(r(a^*) - \pi_\omega^{(t)\top} r) \geq 0, \tag{217}$$

where (217) holds by $r(a^*)$ is always greater than or equal to the convex combination of rewards. Regarding the sub-optimal actions $a \neq a^*$,

$$\frac{\partial \pi_\theta^\top r}{\partial \theta_a}\bigg|_{\theta=\omega^{(t)}} = \pi_\omega^{(t)}(a)(r(a) - \pi_\omega^{(t)\top} r) \tag{218}$$

$$\leq \pi_\omega^{(t)}(a)(r(a) - \pi_\omega^{(t)}(a^*)r(a^*)) \tag{219}$$

$$\leq \pi_\omega^{(t)}(a)(r(a) - r(a_2)) \tag{220}$$

$$\leq 0, \tag{221}$$

where (219) holds since we ignore the terms $\pi_\omega^{(t)}(a)r(a) \geq 0$ for every $a \neq a^*$, and (220) holds because $t \geq T_1$, we have $\pi_\omega^{(t)}(a^*) \geq \frac{r(a_2)}{r(a^*)}$. Combining (217-221), we obtain $t \in \mathcal{R}_1$.

Claim c). Part (i): We show that for $t \geq T_1$ and any $a \in \mathcal{A}$, if $\omega_{a^*}^{(t)} - \theta_{a^*}^{(t)} \geq \omega_a^{(t)} - \theta_a^{(t)}$, then $\omega_{a^*}^{(t+1)} - \theta_{a^*}^{(t+1)} \geq \omega_a^{(t+1)} - \theta_a^{(t+1)}$. By combining the updates (8) and (9), we have

$$\theta_a^{(t+1)} \leftarrow \theta_a^{(t)} + (\omega_a^{(t)} - \theta_a^{(t)}) + \eta^{(t+1)} \frac{\partial \pi_\theta^\top r}{\partial \theta_a}\bigg|_{\theta=\omega^{(t)}}. \tag{222}$$

By using the update (9) of APG with respect to the action a^* and (222),

$$\omega_{a^*}^{(t+1)} - \theta_{a^*}^{(t+1)} = \frac{t}{t+3}(\theta_{a^*}^{(t+1)} - \theta_{a^*}^{(t)}) \tag{223}$$

$$= \frac{t}{t+3}\left(\omega_{a^*}^{(t)} - \theta_{a^*}^{(t)} + \eta^{(t+1)} \frac{\partial \pi_\theta^\top r}{\partial \theta_{a^*}}\bigg|_{\theta=\omega^{(t)}}\right) \tag{224}$$

$$\geq \frac{t}{t+3}\left(\omega_a^{(t)} - \theta_a^{(t)} + \eta^{(t+1)} \frac{\partial \pi_\theta^\top r}{\partial \theta_a}\bigg|_{\theta=\omega^{(t)}}\right) \tag{225}$$

$$= \frac{t}{t+3}(\theta_a^{(t+1)} - \theta_a^{(t)}) \tag{226}$$

$$= \omega_a^{(t+1)} - \theta_a^{(t+1)}, \tag{227}$$

where (225) is followed by the hypothesis $\omega_{a^*}^{(t)} - \theta_{a^*}^{(t)} \geq \omega_a^{(t)} - \theta_a^{(t)}$ and $t \geq T_1$ so $\frac{\partial \pi_\theta^\top r}{\partial \theta_{a^*}}\Big|_{\theta=\omega^{(t)}} \geq \frac{\partial \pi_\theta^\top r}{\partial \theta_a}\Big|_{\theta=\omega^{(t)}}$. We finish the proof of Part (i).

Part (ii): We show that there exists a finite $T_2 \geq T_1$ such that if $t \geq T_2$, then $t \in \mathcal{R}_2$. We prove it by contradiction. Suppose that there is no such T_2. Thus, there are infinitely many $t \geq T_1$ such that there is an action a violating the condition of lying in $\mathcal{R}_2$ at time t, i.e., $\omega_{a^*}^{(t)} - \theta_{a^*}^{(t)} < \omega_a^{(t)} - \theta_a^{(t)}$, as the definition (215) of $\mathcal{R}_2$. Since our action space is finite, there is an action $\tilde{a}$ such that $\omega_{a^*}^{(t)} - \theta_{a^*}^{(t)} < \omega_{\tilde{a}}^{(t)} - \theta_{\tilde{a}}^{(t)}$ holds for infinitely many $t \geq T_1$.

We claim that there is a $\tilde{T} \geq T_1$ such that for all $t \geq \tilde{T}$, $\omega_{a^*}^{(t)} - \theta_{a^*}^{(t)} < \omega_{\tilde{a}}^{(t)} - \theta_{\tilde{a}}^{(t)}$. If not, there is a $t_0 \geq T_1$ such that $\omega_{a^*}^{(t_0)} - \theta_{a^*}^{(t_0)} \geq \omega_a^{(t_0)} - \theta_a^{(t_0)}$, then by Part (i), $\omega_{a^*}^{(t)} - \theta_{a^*}^{(t)} \geq \omega_a^{(t)} - \theta_a^{(t)}$ holds for every $t \geq t_0$, which contradicts to $\omega_{a^*}^{(t)} - \theta_{a^*}^{(t)} < \omega_{\tilde{a}}^{(t)} - \theta_{\tilde{a}}^{(t)}$ holds for infinitely many $t \geq T_1$. Thus, there is a $\tilde{T} \geq T_1$ such that for all $t \geq \tilde{T}$, $\omega_{a^*}^{(t)} - \theta_{a^*}^{(t)} < \omega_{\tilde{a}}^{(t)} - \theta_{\tilde{a}}^{(t)}$.

To meet the contradiction, for any $N > \tilde{T}$, we consider

$$\theta_{a^*}^{(N)} = \theta_{a^*}^{(\tilde{T})} + \sum_{t=\tilde{T}}^{N-1}(\theta_{a^*}^{(t+1)} - \theta_{a^*}^{(t)}) \tag{228}$$

$$= \theta_{a^*}^{(\tilde{T})} + \sum_{t=\tilde{T}}^{N-1}\frac{t+3}{t}(\omega_{a^*}^{(t+1)} - \theta_{a^*}^{(t+1)}) \tag{229}$$

$$< \theta_{a^*}^{(\tilde{T})} + \sum_{t=\tilde{T}}^{N-1}\frac{t+3}{t}(\omega_{\tilde{a}}^{(t+1)} - \theta_{\tilde{a}}^{(t+1)}) \tag{230}$$

$$= \theta_{a^*}^{(\tilde{T})} + \sum_{t=\tilde{T}}^{N-1} \frac{t+3}{t} \left[\frac{t}{t+3} (\theta_{\tilde{a}}^{(t+1)} - \theta_{\tilde{a}}^{(t)}) \right] \tag{231}$$

$$= \theta_{a^*}^{(\tilde{T})} + \sum_{t=\tilde{T}}^{N-1} (\theta_{\tilde{a}}^{(t+1)} - \theta_{\tilde{a}}^{(t)}) \tag{232}$$

$$= \theta_{a^*}^{(\tilde{T})} - \theta_{\tilde{a}}^{(\tilde{T})} + \theta_{\tilde{a}}^{(N)}, \tag{233}$$

where (228) uses a simple telescope argument, (229) uses the update (9) of APG, (230) holds since $t \geq \tilde{T}$, (231) uses the update (9).

Since (233) holds for arbitrary $N > \tilde{T}$, we obtain $\theta_{a^*}^{(N)} < \theta_{a^*}^{(\tilde{T})} - \theta_{\tilde{a}}^{(\tilde{T})} + \theta_{\tilde{a}}^{(N)}$ for any $N > \tilde{T}$, which contradicts $\pi_\theta^{(t)}(a^*) \to 1$ as $t \to \infty$, i.e., the asymptotic global convergence theorem. Hence, there is a $T_2 \geq T_1$ such that if $t \geq T_2$, then $t \in \mathcal{R}_2$.

Claim d). Given $t \geq T_2$, we have $t \in \mathcal{R}_2$. In addition, since $T_2 \geq T_1$, we also have $t \in \mathcal{R}_1$. Thus, $t \in \mathcal{R}_1 \cap \mathcal{R}_2$. Then, to show that $\pi_\theta^{(t+1)}(a^*) \geq \pi_\theta^{(t)}(a^*)$, we have

$$\pi_\theta^{(t+1)}(a^*) = \frac{\exp\{\theta_{a^*}^{(t+1)}\}}{\sum_{a \in \mathcal{A}} \exp\{\theta_a^{(t+1)}\}} \tag{234}$$

$$= \frac{\exp\left\{\theta_{a^*}^{(t)} + (\omega_{a^*}^{(t)} - \theta_{a^*}^{(t)}) + \eta^{(t+1)} \left.\frac{\partial \pi_\theta^\top r}{\partial \theta_{a^*}}\right|_{\theta=\omega^{(t)}}\right\}}{\sum_{a \in \mathcal{A}} \exp\left\{\theta_a^{(t)} + (\omega_a^{(t)} - \theta_a^{(t)}) + \eta^{(t+1)} \left.\frac{\partial \pi_\theta^\top r}{\partial \theta_a}\right|_{\theta=\omega^{(t)}}\right\}} \tag{235}$$

$$\geq \frac{\exp\left\{\theta_{a^*}^{(t)} + (\omega_{a^*}^{(t)} - \theta_{a^*}^{(t)}) + \eta^{(t+1)} \left.\frac{\partial \pi_\theta^\top r}{\partial \theta_{a^*}}\right|_{\theta=\omega^{(t)}}\right\}}{\sum_{a \in \mathcal{A}} \exp\left\{\theta_{a^*}^{(t)} + (\omega_{a^*}^{(t)} - \theta_{a^*}^{(t)}) + \eta^{(t+1)} \left.\frac{\partial \pi_\theta^\top r}{\partial \theta_{a^*}}\right|_{\theta=\omega^{(t)}}\right\}} \tag{236}$$

$$= \frac{\exp\{\theta_{a^*}^{(t)}\}}{\sum_{a \in \mathcal{A}} \exp\{\theta_a^{(t)}\}} = \pi_\theta^{(t)}(a^*), \tag{237}$$

where (235) holds by (222) and (236) leverages the properties of $\mathcal{R}_1$ and $\mathcal{R}_2$ as the definitions (214) and (215). Therefore, $\pi_\theta^{(t)}(a^*)$ is non-decreasing after T_2. Hence, it follows that

$$\inf_{t \geq 0} \pi_\theta^{(t)}(a^*) = \min_{0 \leq t \leq T_2} \pi_\theta^{(t)}(a^*) > 0. \tag{238}$$

□

D.2. Lower Bound of Sub-Optimality Under APG

Theorem 3. *Consider a simple two-armed bandit with actions a^*, a_2, reward function $r(a^*) = 1, r(a_2) = 0$, and initial policy parameters $\theta^{(0)}_{a^*} = \theta^{(0)}_{a_2} = 0$. Under APG with $\eta^{(t)} = \frac{t}{t+1} \cdot \frac{1}{5}$, for all $t > 0$, we have:*

$$\left(\pi^* - \pi^{(t)}_\theta\right)^\top r = \Omega(\frac{1}{t^2}) \tag{12}$$

Proof of Theorem 3. Given the reward function $r(a^*) = 1, r(a_2) = 0$, we know ${\pi^{(t)}_\theta}^\top r = \pi^{(t)}_\theta(a^*)$, and hence we could focus on $\theta^{(t)}_{a^*}$. Moreover, by Lemma 12 and the initialization $\theta^{(0)}_{a^*} = \theta^{(0)}_{a_2} = 0$, we know $\theta^{(t)}_{a^*} + \theta^{(t)}_{a_2} = 0$, for all t. Under APG, in this two-armed bandit case, we have

$$\theta^{(t+1)} = \theta^{(t)} + \frac{t-1}{t+2}(\theta^{(t)} - \theta^{(t-1)}) + \eta^{(t+1)} \cdot \nabla(\pi_\theta^\top r)\big|_{\theta=\omega^{(t)}}. \tag{239}$$

Under the learning rate $\eta^{(t)} = \frac{t}{t+1} \cdot \frac{1}{5}$, one could iteratively verify that $\ln(t+1) - 2 \le \theta^{(t)}_{a^*} \le \ln(t+1) - 1$, for all $t \ge 2$. This implies that the sub-optimality gap is $\Omega(\frac{1}{t^2})$.

□

E. Detailed Explanation of the Motivating Example and the Experimental Configurations

4.2 Motivating Examples of APG

Consider a simple two-action bandit with actions a^*, a_2 and reward function $r(a^*) = 1, r(a_2) = 0$. Accordingly, the objective we aim to optimize is $\mathbb{E}_{a\sim\pi_\theta}[r(a)] = \pi_\theta(a^*)$. By deriving the Hessian matrix with respect to our policy parameters θ_{a^*} and θ_{a_2}, the Hessian matrix can be written as:

$$\mathbf{H} = \begin{bmatrix} \frac{\partial^2 \pi_\theta(a^*)}{\partial \theta_{a^*} \partial \theta_{a^*}} & \frac{\partial^2 \pi_\theta(a^*)}{\partial \theta_{a^*} \partial \theta_{a_2}} \\ \frac{\partial^2 \pi_\theta(a^*)}{\partial \theta_{a_2} \partial \theta_{a^*}} & \frac{\partial^2 \pi_\theta(a^*)}{\partial \theta_{a_2} \partial \theta_{a_2}} \end{bmatrix} = \begin{bmatrix} \pi_\theta(a^*)(1-\pi_\theta(a^*))(1-2\pi_\theta(a^*)) & \pi_\theta(a^*)(1-\pi_\theta(a^*))(2\pi_\theta(a^*)-1) \\ \pi_\theta(a^*)(1-\pi_\theta(a^*))(2\pi_\theta(a^*)-1) & \pi_\theta(a^*)(1-\pi_\theta(a^*))(1-2\pi_\theta(a^*)) \end{bmatrix}, \tag{240}$$

where the eigenvalue $\lambda_1 = 2\pi_\theta(a^*)(\pi_\theta(a^*)-1)(2\pi_\theta(a^*)-1), \lambda_2 = 0$. So we have that if $\pi_\theta(a^*) \geq 0.5$, then $\lambda_1, \lambda_2 \leq 0$, leading to the fact that the Hessian is negative semi-definite. Additionally, we have that the Hessian is negative semi-definite if and only if the objective $\mathbb{E}_{a\sim\pi_\theta}[r(a)]$ is concave, which complete our proof.

4.3 Non-Monotonic Improvement Under APG

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. Notably, the weight of the optimal action in the initial policy $\pi^{(0)} \approx [0.00005, 0.00091, 0.99904]$ is exceedingly small.

6.1 Numerical Validation of the Convergence Rates of APG

(Bandit) We conduct a 3-action bandit experiment with actions $\mathcal{A} = [a^*, a_2, a_3]$, where the corresponding rewards are $r = [r(a^*), r(a_2), r(a_3)] = [1, 0.99, 0]$. We initialize the policy parameters with both a uniform initialization ($\theta^{(0)} = [0, 0, 0], \pi^{(0)} = [1/3, 1/3, 1/3]$) and a hard initialization ($\theta^{(0)} = [1, 3, 5], \pi^{(0)} = [0.01588, 0.11731, 0.86681]$ and hence the optimal action has the smallest initial probability).

(MDP) We conduct an experiment on an MDP with 5 states and 5 actions under the initial state distribution $\rho = [0.3, 0.2, 0.1, 0.15, 0.25]$. The reward, initial policy parameters, transition probability can be found in the following Table 1-8.

Table 1. Experimental settings: Reward function

$r(s,a)$	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: Hard initialization

$\theta^{(0)}_{s,a}$	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: Uniform initialization

$\theta^{(0)}_{s,a}$	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: Transition probability $P(\cdot|s_0,\cdot)$

$P(s\|s_0,a)$	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 $P(\cdot|s_1,\cdot)$

$P(s\|s_1,a)$	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 $P(\cdot|s_2,\cdot)$

$P(s\|s_2,a)$	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 $P(\cdot|s_3,\cdot)$

$P(s\|s_3,a)$	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 $P(\cdot|s_4,\cdot)$

$P(s\|s_4,a)$	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