Towards Principled, Practical Policy Gradient for Bandits and Tabular MDPs

Michael Lu
michael_lu_3@sfu.ca
Simon Fraser University

Matin Aghaei
matin_aghaei@sfu.ca
Simon Fraser University

Anant Raj
araj@inria.fr
SIERRA Project Team (Inria)

Sharan Vaswani
vaswani.sharan@gmail.com
Simon Fraser University

Abstract

We consider (stochastic) softmax policy gradient (PG) methods for bandits and tabular Markov decision processes (MDPs). While the PG objective is non-concave, recent research has used the objective's smoothness and gradient domination properties to achieve convergence to an optimal policy. However, these theoretical results require setting the algorithm parameters according to unknown problem-dependent quantities (e.g. the optimal action or the true reward vector in a bandit problem). To address this issue, we borrow ideas from the optimization literature to design practical, principled PG methods in both the exact and stochastic settings. In the exact setting, we employ an Armijo line-search to set the step-size for softmax PG and demonstrate a linear convergence rate. In the stochastic setting, we utilize exponentially decreasing step-sizes, and characterize the convergence rate of the resulting algorithm. We show that the proposed algorithm offers similar theoretical guarantees as the state-of-the art results, but does not require the knowledge of oracle-like quantities. For the multi-armed bandit setting, our techniques result in a theoretically-principled PG algorithm that does not require explicit exploration, the knowledge of the reward gap, the reward distributions, or the noise. Finally, we empirically compare the proposed methods to PG approaches that require oracle knowledge, and demonstrate competitive performance.

1 Introduction

Policy gradient (PG) methods have played a vital role in the achievements of deep reinforcement learning (RL) (Sutton et al., 1999a; Schulman et al., 2017). Recent theoretical research (Agarwal et al., 2021; Mei et al., 2020; 2021a; Bhandari & Russo, 2021; Lan, 2023; Shani et al., 2020) have analyzed PG methods in simplified settings, exploiting the objective's properties to guarantee global convergence to an optimal policy. We focus on *softmax policy gradient methods* that parameterize the policy using the softmax function, and consider the *tabular parameterization* for which the number of parameters scales with the number of states and actions. For this class of methods, recent studies

have established global convergence rates in both the exact (Mei et al., 2020; 2021a; Agarwal et al., 2021) and stochastic (inexact) settings (Mei et al., 2021a; 2022; 2023; Yuan et al., 2022).

Specifically, in the exact setting where the rewards and transition probabilities are known, Agarwal et al. (2021) proved that softmax PG can attain asymptotic convergence to an optimal policy despite the non-concave nature of the PG objective. Mei et al. (2020) improve this result and quantify the rate of convergence, proving that softmax PG requires $\mathcal{O}(1/\epsilon)$ iterations to converge to an ϵ-optimal policy. On the other hand, when using the tabular parameterization in the exact setting, natural policy gradient (NPG) (Kakade, 2001) and geometry-aware normalized policy gradient (GNPG) (Mei et al., 2021b) have been shown to achieve a linear convergence (Bhandari & Russo, 2021; Cen et al., 2022; Lan, 2023; Xiao, 2022) matching policy iteration.

In the stochastic setting where the rewards and transition probabilities are unknown and algorithms require sampling from the environment, (Zhang et al., 2020b) first proved that REINFORCE (Williams, 1992; Sutton et al., 1999b) converges to a first-order stationary point at an $\tilde{\mathcal{O}}(1/\epsilon^2)$ rate. Mei et al. (2021a; 2022) analyzed the convergence of stochastic softmax PG, proving that it requires $\mathcal{O}(1/\epsilon^2)$ iterations to converge to an ϵ-optimal policy. However, the resulting algorithm requires the full gradient (which in turn requires the knowledge of the environment) to set algorithm parameters, making it impractical in the stochastic setting. Similarly, Yuan et al. (2022) proved that stochastic softmax PG converges to an optimal policy at a slower $\tilde{\mathcal{O}}(1/\epsilon^3)$ rate. However, this result requires knowledge of the optimal action making it vacuous. More recently, Mei et al. (2023) analyzed stochastic softmax PG in the multi-armed bandit setting and proved that it converges to the optimal arm at an $\mathcal{O}(1/\epsilon)$ rate. Unfortunately, the algorithm requires knowledge of the reward gap which is typically unknown for bandit problems.

Consequently, while the above convergence results are notable, the methods that stem from them are impractical. The impracticality arises from the methods' dependence on oracle-like knowledge of the environment, which includes factors such as the optimal action (Yuan et al., 2022), reward gap (Mei et al., 2023) and even access to the full gradient (Mei et al., 2021a) in stochastic settings. The need for this oracle-like knowledge renders these methods ineffective because they assume access to information sufficient to derive an optimal policy. In this paper, our objective is to *design practical softmax PG methods while retaining theoretical convergence guarantees to the optimal policy*. We believe that this is an important first step towards developing practical but theoretically-principled PG methods in the general function approximation setting. To this end, we make the following contributions.

Contribution 1: In Section 3, we first consider the exact setting as a test bed for analyzing softmax PG. In this setting, theoretical step-sizes that enable convergence to the optimal policy are often too conservative in practice. We present a practical approach by employing an Armijo line-search (Armijo, 1966) to set the step-size for softmax PG. Armijo line-search enables adaptation to the objective's local smoothness which results in larger step-sizes and improved empirical performance. Furthermore, we design an alternative line-search condition that takes advantage of the objective's non-uniform smoothness and enables softmax PG to use larger step-sizes. The resulting algorithm achieves linear convergence matching GNPG (Mei et al., 2021b).

Contribution 2: In Section 4, we consider the stochastic setting where the policy gradient is estimated using finitely many interactions with an environment. To design a practical softmax PG algorithm that can adapt to the stochasticity, we utilize exponentially decreasing step-sizes (Li et al., 2021; Vaswani et al., 2022). The resulting algorithm matches the $\tilde{\mathcal{O}}(1/\epsilon^3)$ rate of Yuan et al. (2022) without the knowledge of oracle-like information. In order to attain faster convergence, we use the strong growth condition (SGC) (Schmidt & Roux, 2013; Vaswani et al., 2019) satisfied by the PG objective (Mei et al., 2023). We prove that the same algorithm with exponentially decreasing step-sizes is robust to unknown problem-dependent constants and can effectively interpolate between the fast $\tilde{\mathcal{O}}(1/\epsilon)$ and slow $\tilde{\mathcal{O}}(1/\epsilon^3)$ rate.

Contribution 3: Finally, in Section 5, we experimentally benchmark the proposed algorithms in the bandit setting. Our empirical results indicate that the proposed algorithms have comparable performance as baselines that require oracle-like knowledge.

Contribution 4: In Appendix D, we study the use of entropy regularization for PG methods in both the exact and stochastic settings. Entropy regularization has been successfully used in RL (Haarnoja et al., 2018; Hiraoka et al., 2022). It helps smooth the objective function, enabling PG methods to escape flat regions and allowing the use of larger step-sizes (Ahmed et al., 2019). Although entropy regularization allows for faster convergence, it results in convergence to a biased policy.

We introduce a practical multi-stage algorithm that iteratively reduces the entropy regularization and ensures convergence to the optimal policy. The resulting algorithm does not require the knowledge of any problem dependent constants such as the reward gap (as in prior work (Mei et al., 2020)). Under additional assumptions, we prove that softmax PG with entropy regularization converges to the optimal policy at an $\tilde{\mathcal{O}}(1/\epsilon)$ rate in the exact setting and at an $\tilde{\mathcal{O}}(1/\epsilon^3)$ rate in the stochastic setting. Although we do not prove a theoretical advantage of entropy regularization; in practice, we find that adding entropy enables the resulting algorithms to be more robust to "bad" initializations.

2 Problem Setup & Background

An infinite-horizon discounted Markov decision process (MDP) (Puterman, 2014) is defined by tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \rho, \gamma)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$ is the transition probability function, $\rho \in \Delta_{\mathcal{S}}$ is the initial state distribution, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function, and $\gamma \in [0, 1)$ is the discount factor. We will only consider *tabular MDPs*, assuming that the state and action spaces are finite and define $S := |\mathcal{S}|$ and $A := |\mathcal{A}|$. For policy π, the *action-value function* $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined as: $Q^\pi(s, a) := \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]$, with $s_0 = s$, $a_0 = a$ and for $t \geq 1$, $s_{t+1} \sim p(\cdot|s_t, a_t)$ and $a_{t+1} \sim \pi(\cdot|s_t)$. The corresponding *value function* $V^\pi : \mathcal{S} \to \mathbb{R}$ is defined as $V^\pi(s) := \mathbb{E}_{a\sim\pi(\cdot|s)}[Q^\pi(s, a)]$. The *advantage function* $A^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined as $A^\pi(s, a) := Q^\pi(s, a) - V^\pi(s)$. For state $s \in \mathcal{S}$, we define $\Pr^\pi[s_t = s\,|s_0]$ to be the probability of visiting state s at time t under policy π when starting at state s_0. The *discounted state visitation distribution* is denoted by $d^\pi_{s_0} \in \Delta_{\mathcal{S}}$ and defined as $d^\pi_{s_0} := (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \Pr^\pi[s_t = s\,|s_0]$.

Given a class of feasible policies Π, the policy optimization objective is: $\max_{\pi\in\Pi} J(\pi) := \mathbb{E}_{s\sim\rho}[V^\pi(s)]$. For brevity, we define $V^\pi(\rho) := \mathbb{E}_{s\sim\rho}[V^\pi(s)]$. We denote the optimal policy as $\pi^* = \arg\max_{\pi\in\Pi} J(\pi)$. Throughout this paper, we will consider both the general MDP setting and the *bandits* setting. For the bandit setting, $S = 1$ and $\gamma = 1$, and the corresponding objective is to find a policy that maximizes $\mathbb{E}[\langle \pi, r\rangle]$ where the expectation is over the stochastic rewards.

In this work, we consider policies with a *softmax tabular parameterization*, i.e. for parameters $\theta \in \mathbb{R}^{S\times A}$, the set Π consists of policies $\pi_\theta : \mathcal{S} \to \Delta_{\mathcal{A}}$ parameterized using the softmax function such that $\pi_\theta(a|s) = \exp(\theta(s,a))/\sum_{a'\in\mathcal{A}} \exp(\theta(s,a'))$. Such a tabular parameterization has been recently used to study the theoretical properties of policy gradient methods (Agarwal et al., 2021; Mei et al., 2020). Throughout, we will present our results considering $f(\theta)$ as an abstract objective with specific properties, and when required, instantiate it in the general MDP or bandits setting. In the general MDP setting, $f(\theta) := V^{\pi_\theta}(\rho)$, while in the bandits setting, $f(\theta) := \langle \pi_\theta, r\rangle$. With this abstraction, we hope that our results can be easily generalized to other settings such as constrained MDPs (Altman, 2021) or convex MDPs (Zahavy et al., 2021; Zhang et al., 2020a). Next, we specify the properties of f that will be used to analyze the convergence of PG methods.

First, we note that f is a non-concave function for both bandits and general MDPs (Mei et al., 2020, Proposition 1). However, in both cases, it is twice-differentiable and L-smooth, i.e. for all θ, there exists a constant $L \in (0, \infty)$, $\nabla^2 f(\theta) \preceq L I_{SA}$. Since this property holds for all θ and L is a constant independent of θ, we refer to this as *uniform smoothness*. For both bandits and general MDPs, f also satisfies a notion of *non-uniform smoothness*, i.e. for all θ, there exists a $L_1 \in (0, \infty)$ such that $\nabla^2 f(\theta) \preceq L_1 \|\nabla f(\theta)\| I_{SA}$. Intuitively, non-uniform smoothness states that the landscape is flatter closer to a stationary point $\tilde{\theta}$, meaning that as $\theta \to \tilde{\theta}$, $\nabla^2 f(\theta) \to \mathbf{0}$, i.e. the Hessian becomes

Setting	$f(\theta)$	$[\nabla f(\theta)]_{s,a}$	L	L_1	ν	$C(\theta)$
Bandits	$\langle \pi_\theta, r \rangle$	$\pi_\theta(a)\left[r(a) - \langle \pi_\theta, r \rangle\right]$	$5/2$	3	$\frac{\sqrt{2}}{\Delta^*}$	$\pi_\theta(a^*)$
MDP	$V^{\pi_\theta}(\rho)$	$\frac{d^{\pi_\theta}(s)\,\pi_\theta(a\|s)\,A^{\pi_\theta}(s,a)}{1-\gamma}$	$\frac{8}{(1-\gamma)^3}$	$\left[3 + \frac{2\,C_\infty - (1-\gamma)}{(1-\gamma)\gamma}\right]\sqrt{S}$	$\frac{\sqrt{2}}{(1-\gamma)\,\Delta^*}$	$\frac{\min_s \pi_\theta(a^*(s)\|s)}{\sqrt{S}\left\|\frac{d_\rho^{\pi^*}}{d_\rho^{\pi_\theta}}\right\|_\infty}$

Table 1: Function and gradient expressions, (non)-uniform smoothness, non-uniform and reversed Łojasiewciz properties for bandits and general tabular MDPs with $\xi = 0$ (Mei et al., 2020). Here, a^* is index of the optimal arm in the bandit problem , $C_\infty := \max_\pi \left\| \frac{d_\rho^\pi}{\rho} \right\|_\infty$ is the distribution mismatch ratio (Agarwal et al., 2021), and $\Delta^* := \min_s Q^*(s, a^*(s)) - \max_{a(s) \neq a^*(s)} Q^*(s, a(s))$ is the reward gap corresponding to the optimal policy.

degenerate. Together, the uniform and non-uniform smoothness properties are related to the (L_0, L_1) smoothness recently used to study the optimization of transformer models (Zhang et al., 2019).

Since the rewards are bounded, $f(\theta)$ is upper-bounded by a value $f^* := \max_\theta f(\theta)$. Furthermore, f satisfies a *non-uniform Łojasiewciz condition*, i.e. for all θ, there exists a $C(\theta) \in (0, \infty)$ and $\xi \in [0, 1]$ such that $\|\nabla f(\theta)\|_2 \geq C(\theta)\,|f^* - f(\theta)|^{1-\xi}$ (Mei et al., 2020). For the special case where $C(\theta)$ is an absolute constant and $\xi = 1/2$, this condition matches the well studied Polyak Łojasiewciz (PŁ) condition (Polyak, 1963; Karimi et al., 2016). The Łojasiewciz condition states that every stationary point $\tilde{\theta}$ (s.t. $\nabla f(\tilde{\theta}) = 0$) is also a global maximum s.t. $f(\tilde{\theta}) = f^*$. This condition enables the convergence of local ascent methods such as PG to an optimal solution $\theta^* := \arg\max_\theta f(\theta)$ despite the problem's non-concavity (Karimi et al., 2016; Mei et al., 2020; Agarwal et al., 2021). Finally, f satisfies a *reversed Łojasiewciz* condition, i.e. for all θ, there exists a $\nu > 0$ such that $\|\nabla f(\theta_t)\| \leq \nu\,(f^* - f(\theta))$ (Mei et al., 2020). This condition bounds how quickly the gradient norm vanishes near the optimal solution. Table 1 summarizes both the uniform and non-uniform smoothness and Łojasiewciz properties for bandits and general MDPs.

Similar to Mei et al. (2020), we assume a uniform starting state distribution, i.e. $\forall s \in \mathcal{S}$, $\rho(s) = 1/S$ and hence $C_\infty \leq \frac{1}{\min_s \rho(s)} < \infty$. This is a common assumption in the policy gradient literature that obviates the need for exploration in the general MDP setting and allows us to exclusively focus on the optimization aspects. We note that for both these settings, the optimal policy is deterministic (Puterman, 2014) i.e. in the general MDP setting, for each state $s \in \mathcal{S}$, there is an action $a^*(s) \in \mathcal{A}$ such that $\pi^*(a^*(s)|s) = 1$ and for all $a \neq a^*(s)$, $\pi^*(a|s) = 0$. This implies that when using the softmax tabular parameterization, $\theta^*(s, a^*(s)) \to \infty$ and for all $a \neq a^*(s)$, $\theta^*(s, a) \to -\infty$. This property is similar to that for logistic regression for classification on linearly separable data (Ji & Telgarsky, 2018).

In the next section, we will use the above properties of f and study the convergence of PG methods in the exact setting.

3 Policy Gradient in the Exact Setting

We first consider the exact setting that assumes complete knowledge of the rewards and transition probabilities, and consequently enables the exact calculation of the policy gradient. This setting has been used as a test bed to study the convergence properties of PG methods (Bhandari & Russo, 2021; Agarwal et al., 2021; Mei et al., 2020).

Softmax policy gradient (softmax PG) uses gradient ascent to iteratively maximize $f(\theta)$. In particular, at iteration $t \in [T]$, softmax PG uses a step-size of η_t and has the following update:

Update 1. *(Softmax PG, True Gradient)* $\theta_{t+1} = \theta_t + \eta_t \nabla f(\theta_t)$.

Refer to Table 1 for the gradient expressions of the policy gradient $\nabla f(\theta)$ in both the bandits and general MDP cases.

In this setting, Mei et al. (2020) prove that softmax PG converges to an optimal solution at an $\mathcal{O}(1/T)$ rate, implying that the algorithm requires $\mathcal{O}(1/\epsilon)$ iterations to guarantee that $f^* - f(\theta_{T+1}) \leq \epsilon$. From a policy optimization perspective, this implies that softmax PG can return a stochastic policy whose value function is ϵ close to the optimal policy's value function. In order to achieve this convergence, Mei et al. (2020) requires using a constant step-size $\eta_t = \eta = 1/L$. Furthermore, for any $\eta_t \in (0, 1]$, Mei et al. (2020, Theorem 9) proves an $\Omega(1/\epsilon)$ lower-bound showing that this rate is tight.

In most scenarios, we can only obtain a loose upper-bound on the smoothness L. This over-estimation of L implies that the resulting step-size is typically smaller than necessary, often resulting in worse empirical performance. In practice, when doing gradient ascent with access to the exact gradient, it is standard to employ a *line-search* (Armijo, 1966; Nocedal & Wright) to adaptively set the step-size in each iteration. This results in faster empirical convergence while requiring minimal tuning, and preserving the rate of convergence. Hence, we propose to use a backtracking Armijo line-search (Armijo, 1966) to adaptively set the step-size for softmax PG.

At every iteration t, backtracking Armijo line-search starts from an initial guess for the step-size ($\eta_{\max}$) and backtracks until the *Armijo condition* is satisfied. In particular, the procedure thus returns the largest step-size η_t such that following condition is satisfied:

$$f(\theta_t + \eta_t \nabla f(\theta_t)) \geq f(\theta_t) + h\eta_t \|\nabla f(\theta_t)\|_2^2, \quad \text{(Armijo condition)} \tag{1}$$

where $h \in (0, 1)$ is a hyper-parameter. For smooth functions, the backtracking procedure is guaranteed to terminate and return a step-size η_t that satisfies $\eta_t \geq \min\{2(1-h)/L, \eta_{\max}\}$. Hence, Armijo line-search guarantees improvement in the function value (ensuring monotonic policy improvement at each iteration t), while selecting a step-size larger than the $1/L$ step-size used in Mei et al. (2020).

The following theorem shows that using the Armijo line-search preserves the theoretical $\mathcal{O}(1/T)$ convergence rate.

Theorem 1. Assuming f is (i) L-smooth, (ii) satisfies the non-uniform Łojasiewciz condition with $\xi = 0$, and (iii) $\mu := \inf_{t \geq 1}[C(\theta_t)]^2 > 0$, using Update 1 and Armijo line-search to set the step-size results in the following convergence:

$$f^* - f(\theta_{T+1}) \leq \max\left\{\frac{L}{2\,h\,(1-h)}, \frac{1}{h\,\eta_{\max}}\right\} \frac{1}{\mu\,T} \tag{2}$$

where $h \in (0, 1)$ and $\eta_{\max}$ is the upper-bound on the step-size.

While assumptions (i) and (ii) are satisfied for both the general MDP and bandit settings, we need to ensure that assumption (iii) also holds. We first note that this property holds for a constant step-size $\eta_t = \eta = 1/L$ (Mei et al., 2020, Lemma 5, Lemma 9). However, the proof can be extended to any varying step-size sequence that guarantees ascent ($f(\theta_{t+1}) \geq f(\theta_t)$) in every iteration. When using the Armijo line-search to set the step-size, this condition is satisfied by definition, thus guaranteeing that $\mu := \inf_{t \geq 1}[C(\theta_t)]^2 > 0$.

The Armijo condition in Equation (1) takes advantage of the objective's uniform smoothness in order to attain an $\mathcal{O}(1/T)$ convergence. In our initial experiments, we observed that for most iterations, the maximum step-size $\eta_{\max}$ satisfies the Armijo condition, and is hence returned by the line-search procedure. By using a sufficiently large $\eta_{\max}$ or by progressively increasing the maximum step-size as a function of t, the resulting algorithm converges at a linear rate. This is because the objective satisfies a non-uniform smoothness property and the optimization landscape becomes flatter as the gradient norm decreases closer to the solution. This enables the use of larger step-sizes than those returned by the Armijo line-search when using a fixed $\eta_{\max}$. In order to take advantage of the non-uniform smoothness more explicitly, we design an alternative line-search on the logarithm of the suboptimality. Formally, we use the following condition:

$$\ln(f^* - f(\theta_t + \eta_t\,\nabla f(\theta_t))) \leq \ln(f^* - f(\theta_t)) - h\,\eta_t\,\frac{\|\nabla f(\theta_t)\|_2^2}{f^* - f(\theta_t)} \quad \text{(Armijo condition for log-loss).} \tag{3}$$

When using the above condition, Lemma 2 guarantees that the backtracking line-search procedure terminates and returns $\eta_t \geq \min\left\{\eta_{\max}, \frac{2(1-h)}{L_1\,\nu\,[f^*-f(\theta_t)]}\right\}$ (refer to Table 1 for the values of L_1 and ν). Hence, the resulting line-search accepts step-sizes proportional to $\frac{1}{f^*-f(\theta_t)}$, meaning that as the optimization progresses and $f(\theta_t) \to f^*$, larger step-sizes can be used.

The following theorem (proved in Appendix B.2) characterizes the rate of convergence of softmax PG when using the Armijo condition for the log-loss in Equation (3).

Theorem 2. For a given $\epsilon \in (0,1)$, assuming f is (i) L_1 non-uniform smooth, (ii) satisfies the non-uniform Łojasiewciz condition with $\xi = 0$, (iii) $\mu := \inf_{t\geq 1}[C(\theta_t)]^2 > 0$, (iv) f satisfies a reversed Łojasiewciz condition with $\nu > 0$, using Update 1 with backtracking line-search using the Armijo condition in Equation (3) and setting $\eta_{\max} = C/\epsilon$ results in the following convergence:
If $f^* - f(\theta_t) > \epsilon$ for all $t \in [1, T]$, then,

$$f^* - f(\theta_{T+1}) \leq [f^* - f(\theta_1)]\,\exp\left(-\min\left\{C\,h, \frac{2\,h\,(1-h)}{L_1\,\nu}\right\}\,\mu\,T\right) \tag{4}$$

where $C > 0$ and $h \in (0,1)$ are hyper-parameters. Otherwise $\min_{t\in[1,T]} f^* - f(\theta_t) \leq \epsilon$.

For a target ϵ, setting $T = \mathcal{O}\left(\log\left(1/\epsilon\right)\right)$ iterations results in a linear convergence rate. In comparison to Theorem 1, using the Armijo condition in Equation (4) enables the use of larger step-sizes resulting in a faster ($\mathcal{O}(1/\epsilon)$ vs $\mathcal{O}\left(\log\left(1/\epsilon\right)\right)$) rate. However, the Armijo condition in Equation (4) requires the knowledge of f^*, making the resulting method less practical. This requirement is similar to the Polyak step-size (Polyak, 1987) used for gradient descent. For future work, we hope to remove this dependence of f^*. In comparison, the geometry-aware normalized policy gradient (GNPG) approach introduced in Mei et al. (2021a) also explicitly exploits this non-uniform smoothness and exhibits a convergence rate of $\mathcal{O}\left(\log\left(1/\epsilon\right)\right)$. However, in the general MDP setting, GNPG requires the knowledge of unknown constants such as the concentrability coefficient $C_\infty := \max_\pi \left\|d_\rho^\pi/\rho\right\|_\infty$ to determine the step-size, making it impractical. In concurrent work, Liu et al. (2024) show that softmax PG with *any* constant step-size can attain an $\Theta(1/\epsilon)$ convergence to the optimal policy. Moreover, they prove that softmax PG with a specific adaptive step-size scheme that only depends on the advantage function and the policy (PG-A) can attain a fast $\mathcal{O}\left(\log\left(1/\epsilon\right)\right)$ convergence.

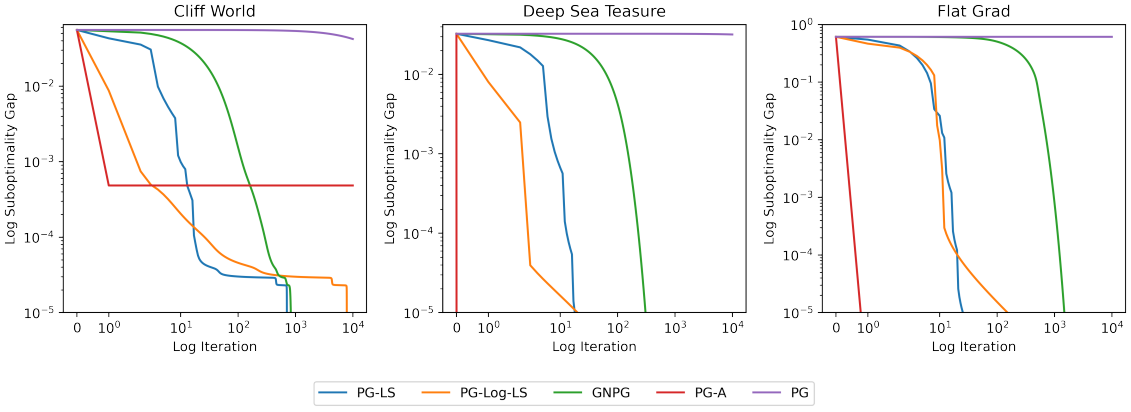

Figure 1: Comparing softmax PG that (i) uses a step-size that satisfies the Armijo condition in Equation (1) (denoted as `PG-LS`), (ii) uses a step-size that satisfies the Armijo condition in Equation (3) (`PG-Log-LS`) to GNPG (`GNPG`), PG-A (`PG-A`) and PG with a fixed step-size (`PG`) in the tabular MDP setting.

In Figure 1, we compare the presented line-search methods with the Armijo condition in Equation (1) and the Armijo condition on the log-loss in Equation (3) to GNPG, PG-A and PG with a constant step-size on three tabular MDP environments (see Appendix G for details). For the methods that use

backtracking line-search, we set $\eta_{\max} = \frac{1}{\epsilon}$ with $\epsilon = 10^{-4}$ and $h = 0.5$. For `GNPG`, we use the step-size of $\eta_t = \frac{(1-\gamma)\,\gamma}{6\,(1-\gamma)+4\,(S^{-1}-(1-\gamma))}$. Since C_∞ is unknown, we upper-bound it as: $C_\infty \leq \min_s \frac{1}{\rho(s)} = \frac{1}{S}$. For `PG-A`, we use the theoretical step-size $\eta_t = \frac{1}{\min_{s\in\hat{\mathcal{S}}_t} \max_a |\hat{A}_t(s,a)|}$ where $\hat{A}_t(s,a) := \pi_{\theta_t}(a|s)\, A^{\pi_{\theta_t}}(s,a)$ and $\hat{\mathcal{S}}_t := \{s \in \mathcal{S} \mid \hat{A}_t(s,a) > 0\}$. Finally for `PG` we use a constant step-size of $\eta_t = \frac{1}{L} = \frac{(1-\gamma)^3}{8}$. We observe that `PG-LS` is comparable to `GNPG` and `PG-A` while `PG-Log-LS` can better exploit the non-uniform smoothness, enabling larger step-sizes as the algorithm approaches the optimal policy. The performance of `PG` is negligible due to the loose upper-bound of L, resulting in a conservative step-size. In Appendix G, we plot the wall-clock time to justify the performance gains of the proposed methods.

In the next section, we study the more realistic stochastic setting where the rewards and transition probabilities are unknown, and the policy gradients need to be estimated via interactions with the environment. Although GNPG and NPG can obtain faster convergence rates in the exact setting, they are not guaranteed to converge to the optimal policy in the stochastic setting (Mei et al., 2021a). This is because these methods are too aggressive and can quickly commit to sub-optimal actions. Consequently, we restrict ourselves to softmax PG in the stochastic setting.

4 Policy Gradient in the Stochastic Setting

In this section, we analyze softmax PG with an estimated (stochastic) policy gradient. In Section 4.1, we construct PG estimators that are unbiased and have bounded variance. We design a PG algorithm that uses the stochastic policy gradient along with exponentially decreasing step-sizes (Li et al., 2021; Vaswani et al., 2022). In Section 4.2, we prove that the resulting algorithm can obtain convergence rates comparable to the state-of-the-art, but do not require oracle-like knowledge of the environment. Finally, in Section 4.2.1, we exploit the fact that the variance in the stochastic gradients decreases as the algorithm approaches a stationary point, and prove that the same stochastic softmax PG algorithm can obtain a faster convergence rate.

4.1 Stochastic Softmax Policy Gradient

For illustrative purposes, we mainly focus on the bandit setting in the main paper. In the stochastic multi-armed bandit setting (Lattimore & Szepesvári, 2020), each action (arm) has an underlying unknown reward distribution. In every iteration t, the algorithm chooses an action to pull and receives a stochastic reward sampled from the distribution of the corresponding arm. The stochastic softmax PG algorithm maintains a distribution $\pi_{\theta_t} \in \Delta_{\mathcal{A}}$ over the actions. In each iteration $t \in [1, T]$, the algorithm samples an action $a_t \sim \pi_{\theta_t}$ and receives reward $R_t \sim P_{a_t}$ where P_{a_t} is the reward distribution of arm a_t. The reward R_t is used to construct the on-policy importance sampling (IS) reward estimate $\hat{r}_t(a) = \frac{\mathbb{1}\{a_t=a\}}{\pi_{\theta_t}(a)}\, R_t$ for each $a \in \mathcal{A}$. The IS reward estimate is then used to form the stochastic gradient $\nabla\widetilde{f}(\theta_t)$ such that $\nabla\widetilde{f}(\theta_t)(a) = \pi_{\theta_t}(a)[\hat{r}_t(a) - \langle \pi_{\theta_t}, \hat{r}_t\rangle]$. Mei et al. (2021a, Lemma 5) showed that the resulting stochastic gradients are (i) unbiased i.e. $\mathbb{E}[\nabla\widetilde{f}(\theta)] = \nabla f(\theta)$ and have (ii) bounded variance i.e. $\mathbb{E}\left\|\nabla\widetilde{f}(\theta) - \nabla f(\theta)\right\|_2^2 \leq \sigma^2$. Similarly, we can construct gradient estimators that are unbiased and have bounded variance for general MDPs (refer to Appendix C.4). Given these estimators, the resulting stochastic softmax PG algorithm has the following update:

Update 2. *(Stochastic Softmax PG, Importance Sampling)* $\theta_{t+1} = \theta_t + \eta_t \nabla\widetilde{f}(\theta_t)$.

We note that this update has also been used in Yuan et al. (2022); Mei et al. (2021a) that attain global convergence to the optimal solution in both the bandit and general MDP settings. In order to prove theoretical convergence, Yuan et al. (2022) used the knowledge of $\mu := \inf_{t\geq 1}[C(\theta_t)]^2$ when setting the step-size. However, in both the bandit and general MDP settings (see Table 1 for details) $C(\theta)$ and consequently μ depends on the optimal action. This makes the resulting algorithm impractical. On the other hand, Mei et al. (2021a) require the full gradient to set the step-size and obtain global convergence. Since the full gradient is not available in the stochastic setting, it is not practical to use

	Convergence Rate	Knowledge required to set η
Mei et al. (2021a)	$\mathcal{O}(1/\epsilon^2)$	$\|\nabla f(\theta)\|$
Yuan et al. (2022)	$\mathcal{O}(1/\epsilon^3)$	π^*
Mei et al. (2023)	$\mathcal{O}(1/\epsilon)$	mean reward vector r
This work	Interpolates between $\tilde{\mathcal{O}}(1/\epsilon)$ & $\tilde{\mathcal{O}}(1/\epsilon^3)$	T

Table 2: Global convergence rates and knowledge required to set the step-size η for each method in the bandits setting. Our proposed method achieves comparable convergence rates to prior state-of-the-art results without any oracle-like knowledge.

their algorithm. Table 2 summarizes the global convergence rates for stochastic softmax PG and the method's step-size dependencies.

We make use of exponentially decaying step-sizes (Li et al., 2021; Vaswani et al., 2022) that have been previously used for stochastic gradient descent when minimizing smooth non-convex functions satisfying the PŁ-inequality (Polyak, 1963; Karimi et al., 2016). In this setting, the benefit of exponentially decaying step-sizes is that they can achieve (up to poly-logarithmic terms) the best known convergence rates without the knowledge of σ^2 or μ. Given the knowledge of T, the step-size in iteration t is set as: $\eta_t = \eta_0\,\alpha^t$ where η_0 is the initial step-size, $\alpha = \left(\frac{\beta}{T}\right)^{\frac{1}{T}}$ and $\beta \geq 1$. Although β is a hyper-parameter, we emphasize that it does not depend on any problem-dependent constants. We leverage these step-sizes for designing a stochastic softmax PG algorithm and characterize its convergence in the next section.

4.2 Theoretical Convergence

By using the proof techniques from Yuan et al. (2022) and Li et al. (2021), we prove the following theorem in Appendix C.1.

Theorem 3. For a given $\epsilon \in (0, 1)$, assuming f is (i) L-smooth, (ii) satisfies the non-uniform Łojasiewciz condition with $\xi = 0$, (iii) $\mu := \left[\mathbb{E}\left[\inf_{t\geq 1}[C(\theta_t)]^{-2}\right]\right]^{-1} > 0$, using Update 2 with (a) unbiased stochastic gradients whose variance is bounded by σ^2 and (b) exponentially decreasing step-sizes $\eta_t = \eta_0\,\alpha^t$ where $\eta_0 = \frac{1}{L}$ and $\alpha = \left(\frac{\beta}{T}\right)^{\frac{1}{T}}$, $\beta \geq 1$ results in the following convergence: If $\mathbb{E}[f^* - f(\theta_t)] > \epsilon$ for all $t \in [1, T]$, then,

$$\mathbb{E}[f^* - f(\theta_{T+1})] \leq \mathbb{E}[f^* - f(\theta_1)]\, C_1\, \exp\left(-\frac{\alpha\,\epsilon\,T}{\kappa\,\ln(T/\beta)}\right) + \frac{C_1\,C_2}{2\,L}\frac{\ln^2\left(\frac{T}{\beta}\right)\sigma^2}{\epsilon^2\,T} \tag{5}$$

where $\kappa := \frac{2\,L}{\mu}$, $C_1 := \exp\left(\frac{2\,\beta}{\kappa\,\ln(T/\beta)}\right)$ and $C_2 := \frac{4\,\kappa^2}{e^2\,\alpha^2}$. Otherwise $\min_{t\in[1,T]} \mathbb{E}[f^* - f(\theta_t)] \leq \epsilon$.

In order to ensure that assumption (iii) holds, let us consider the bandit setting where $C(\theta) = \pi_\theta(a^*)$. To guarantee that $\mu := \left[\mathbb{E}\left[\inf_{t\geq 1}[C(\theta_t)]^{-2}\right]\right]^{-1} > 0$, we must ensure that $\pi_{\theta_0}(a^*) > 0$. Since T is finite and θ_0, η_t and the stochastic gradients are bounded (refer to Lemmas 10 and 11 in Appendix C), no parameter including $\theta(a^*)$ can diverge to $-\infty$, guaranteeing that $\pi_\theta(a^*) > 0$.

To determine the resulting convergence rate, let us first analyze the case when $\sigma^2 = 0$. In this case, given a target ϵ, we set $T = \mathcal{O}(1/\epsilon\,\log(1/\epsilon))$ iterations to make the first term $\mathcal{O}(\epsilon)$. On the other hand, when $\sigma^2 > 0$ and the second term of $\tilde{\mathcal{O}}\left(\sigma^2/\epsilon^2 T\right)$ dominates, we set $T = \tilde{\mathcal{O}}(1/\epsilon^3)$ iterations to make the second term $\mathcal{O}(\epsilon)$. Putting both cases together, in order to make the sub-optimality $\mathcal{O}(\epsilon)$, we can set $T = \max\{\tilde{\mathcal{O}}\left(1/\epsilon, \sigma^2/\epsilon^3\right)\}$. This convergence rate matches that in Yuan et al. (2022) without requiring the knowledge of μ. We emphasize that the above convergence rate holds without the knowledge of any oracle-like information.

The previous result assumes that the variance σ^2 is constant w.r.t. θ. However, it has been observed that the noise depends on θ, and decreases as the algorithm gets closer to a stationary point since the policy become more deterministic. Next, we leverage this property to prove faster rates.

4.2.1 Faster Rates

In the bandit setting, Mei et al. (2023) formalized the above intuition, and proved that the stochastic gradient $\nabla\widetilde{f}(\theta)$ satisfies the strong growth condition (SGC) (Schmidt & Roux, 2013; Vaswani et al., 2019) implying that $\mathbb{E}\left\|\nabla\widetilde{f}(\theta)\right\|_2^2 \leq \varrho\,\|\nabla f(\theta)\|$ for a problem-dependent $\varrho > 1$. This implies that the variance decreases as the algorithm approaches a stationary point and $\|\nabla f(\theta)\| \to 0$. For the bandit setting, using Update 2 and the knowledge of ϱ to set the step-size, Mei et al. (2023) can attain a faster $\mathcal{O}(1/\epsilon)$ convergence rate. We generalize the above SGC result to the general MDP setting in Theorem 6 (proved in Appendix C.4).

Theorem 4. Using Update 2, we have for all θ, $\mathbb{E}\left\|\nabla\widetilde{f}(\theta)\right\|_2^2 \leq \varrho\,\|\nabla f(\theta)\|_2$, where $\varrho := \frac{8\,A^{3/2}}{\Delta^2}$ in the bandit setting with $\Delta := \min_{a\neq a'}|r(a) - r(a')|$ and $\varrho = \frac{4\,A^{3/2}\,S^{1/2}}{(1-\gamma)^4\,\Delta^2}$ in the tabular MDP setting with $\Delta := \min_s \min_{a\neq a'}|Q^{\pi_\theta}(s,a) - Q^{\pi_\theta}(s,a')|$.

However, in the bandit setting, ϱ depends on the unknown *reward gap* $\Delta := \min_{a\neq a'}|r(a) - r(a')|$ and we prove that this dependence is necessary (Proposition 1 in Appendix C). This makes the resulting algorithm ineffective in most practical cases. Hence, we aim to develop a practical algorithm that can automatically adapt to ϱ and result in a faster convergence. In Theorem 5, proved in Appendix C.2, we show that the same stochastic softmax PG algorithm (with exponentially decreasing step-sizes) can attain such fast convergence. In addition to the properties in Theorem 3, we exploit the function's non-uniform smoothness, the SGC and the boundedness of stochastic gradients to prove this result.

Theorem 5. For a given $\epsilon \in (0,1)$, assuming f is (i) L_1 non-uniform smooth, (ii) satisfies the non-uniform Łojasiewciz condition with $\xi = 0$, (iii) $\mu := \left[\mathbb{E}\left[\inf_{t\geq 1}[C(\theta_t)]^{-2}\right]\right]^{-1} > 0$, using Update 2 with unbiased stochastic gradients that are (a) bounded, i.e. $\|\nabla\widetilde{f}(\theta)\| \leq B$ and satisfy the strong growth condition with ϱ and (b) exponentially decreasing step-sizes $\eta_t = \eta_0\,\alpha^t$ where $\eta_0 < \frac{1}{L_1^2 B}$ and $\alpha = \left(\frac{\beta}{T}\right)^{\frac{1}{T}}$, $\beta \geq 1$, results in the following convergence:
If $\mathbb{E}[f^* - f(\theta_t)] > \epsilon$ for all $t \in [1,T]$, then,

$$\mathbb{E}[f^* - f(\theta_{T+1})] \leq \mathbb{E}[f^* - f(\theta_1)]\,C_1\,\exp\left(-\frac{\alpha\,\epsilon\,T}{\kappa\,\ln(T)}\right) + \frac{C_2\,\sum_{t=1}^{T_0-1}\mathbb{E}[f^* - f(\theta_t)]}{\epsilon^2\,T^2} \tag{6}$$

where $\kappa := \frac{2}{\mu\,\eta_0}$, $C_1 := \exp\left(\frac{2\,\beta}{\kappa\,\ln(T/\beta)}\right)$, $C_2 := \exp\left(\frac{2\,\beta}{\kappa\,\ln(T/\beta)}\right)\frac{16\,\varrho\,L\,\kappa^2}{e^2\,\alpha^2}\ln^2(T/\beta)$, $T_0 := T\,\max\left\{\frac{\ln(\varrho\,\eta_0)}{\ln(T/\beta)}, 0\right\}$. Otherwise $\min_{t\in[1,T]}\mathbb{E}[f^* - f(\theta_t)] \leq \epsilon$.

Similar to Theorem 3, assumption (iii) is true when $\pi_{\theta_0}(a^*) > 0$ and T is finite. In Lemmas 10 and 11 (proved in Appendix C), we prove that the stochastic gradients are bounded in both the bandit and MDP settings. In the above result, T_0 represents the iteration when the step-size is small enough to take advantage of the SGC. Given the knowledge of ϱ, we can set set $\eta_0 \leq 1/\varrho$ in which case $T_0 = 0$. In this case, setting $T = \tilde{\mathcal{O}}(1/\epsilon)$ iterations enables us to obtain a "fast" $\mathcal{O}(1/\epsilon)$ rate. Since ϱ is unknown in general, setting η_0 to be large can result in $T_0 = \mathcal{O}(T)$ in the worst case. In this case, the second term of order $\tilde{\mathcal{O}}(1/\epsilon^2 T)$ dominates. In this case, setting $T = \mathcal{O}(1/\epsilon^3)$ iterations results in a "slow" $\tilde{\mathcal{O}}(1/\epsilon^3)$ rate. Hence, the resulting algorithm is robust to ϱ and depending on how η_0 is set, it can interpolate between the "slow" and "fast" rates.

Below, we instantiate Theorem 5 in the bandit setting.

Corollary 1. In the bandit setting, for a given $\epsilon \in (0,1)$, using Update 2 with exponentially decreasing step-sizes $\eta_t = \eta_0\,\alpha^t$ where $\eta_0 \leq \frac{1}{18}$, $\alpha = \left(\frac{\beta}{T}\right)^{\frac{1}{T}}$, $\beta \geq 1$ results in the following convergence:
If $\mathbb{E}[(\pi^* - \pi_{\theta_t})^\top r] \geq \epsilon$ for all $t \in [1,T]$, then,

$$\mathbb{E}[(\pi^* - \pi_{\theta_{T+1}})^\top r] \leq \mathbb{E}[(\pi^* - \pi_{\theta_1})^\top r]\, C_1 \exp\left(-\frac{\alpha\,\epsilon\,T}{\kappa\,\ln(T/\beta)}\right) + \frac{C_2\,\sum_{t=1}^{T_0-1}\mathbb{E}[(\pi^* - \pi_{\theta_t})^\top r]}{\epsilon^2\,T^2} \tag{7}$$

where $\kappa := \frac{2}{\mu\,\eta_0}$, $C_1 := \exp\left(\frac{2\,\beta}{\kappa\,\ln(T/\beta)}\right)$, $C_2 := \exp\left(\frac{2\beta}{\kappa\,\ln(T/\beta)}\right)\frac{32\,\varrho\,\kappa^2}{5\,e^2\,\alpha^2}\,\ln^2(T/\beta)$, $T_0 := T\,\max\left\{\frac{\ln(4\,\varrho\,\eta_0)}{\ln(T/\beta)}, 0\right\}$, $\rho = \frac{8\,A^{3/2}}{\Delta^2}$ and $\mu := \left[\mathbb{E}\left[\min_{t\in[1,T]}[\pi_{\theta_t}(a^*)]^{-2}\right]\right]^{-1} > 0$. Otherwise $\min_{t\in[1,T]}\mathbb{E}[(\pi^* - \pi_{\theta_t})^\top r] \leq \epsilon$.

In the multi-armed bandit setting, using stochastic softmax PG with exponentially decreasing step-sizes allows for implicit automatic exploration without requiring the knowledge of any problem-dependent constants such as the reward gap. Unlike Mei et al. (2023), we note that the above result does not imply asymptotic convergence to the optimal arm. This difference stems from the fact that Mei et al. (2023) uses a constant step-size, while the above result requires a decreasing step-size that asymptotically goes to zero. Compared to the standard algorithms for multi-armed bandits such as upper confidence bound (UCB) (Auer et al., 2002) which requires the knowledge of the noise magnitude to design confidence intervals or Thompson sampling (TS) (Agrawal & Goyal, 2012) which requires knowledge of the reward distribution, stochastic softmax PG does not require such information.

In the next section, we empirically validate our theoretical results and compare the proposed methods to prior algorithms in the bandits setting.

5 Experimental Evaluation [1]

We evaluate the methods in multi-armed bandit environments with $A = 10$. For each environment, we compare the various algorithms on the basis of their expected sub-optimality gap $\mathbb{E}[(\pi^* - \pi_{\theta_t})^\top r]$. For each instance of an environment, we run an algorithm 5 times to account for the stochasicity of each algorithm. We plot the average and 95% confidence interval of the expected sub-optimality gap across 25 instances over $T = 10^6$ iterations. For each run, the initial policy is uniform, i.e. $\pi_{\theta_0}(a) = 1/A$ for all $a \in \mathcal{A}$.

Environment Details: Each environment's underlying reward distribution is either a Bernoulli, Gaussian, or Beta distribution with a fixed mean reward vector $r \in \mathbb{R}^A$ and support $[0,1]$. The difficulty of the environment is determined by the maximum reward gap $\bar{\Delta} := \min_{a^*\neq a} r(a^*) - r(a)$. In easy environments $\bar{\Delta} = 0.5$ and in the hard environments $\bar{\Delta} = 0.1$. For each environment, r is randomly generated for each run.

Methods: We compare stochastic softmax PG with exponentially decreasing step-size (`SPG-ESS`) to prior work that uses the full gradient (`SPG-O-G`) (Mei et al., 2021a) and the reward gap (`SPG-O-R`) (Mei et al., 2023) when setting the step-size. For `SPG-ESS`, we select $\beta = 1$ and $\eta_0 = \frac{1}{18}$ for all experiments. For `SPG-O-R` and `SPG-O-G`, we use the corresponding theoretical step-size of $\eta_t = \frac{\Delta^2}{(40)\,10^{3/2}}$ and $\eta_t = \frac{1}{12}\left\|\frac{d\langle \pi_{\theta_t}, r\rangle}{d\theta_t}\right\|$ respectively. We emphasize that both these step-sizes depend on the unknown mean reward vector, making the resulting methods impractical.

In our experiments, we observed that `SPG-ESS` slows down and stops making progress because of overly conservative step-sizes. To counteract this, we additionally try a "doubling trick" (`SPG-ESS [D]`). This is a common trick when adapting algorithms that depend on a fixed number of iterations

[1] The code to reproduce results is available here

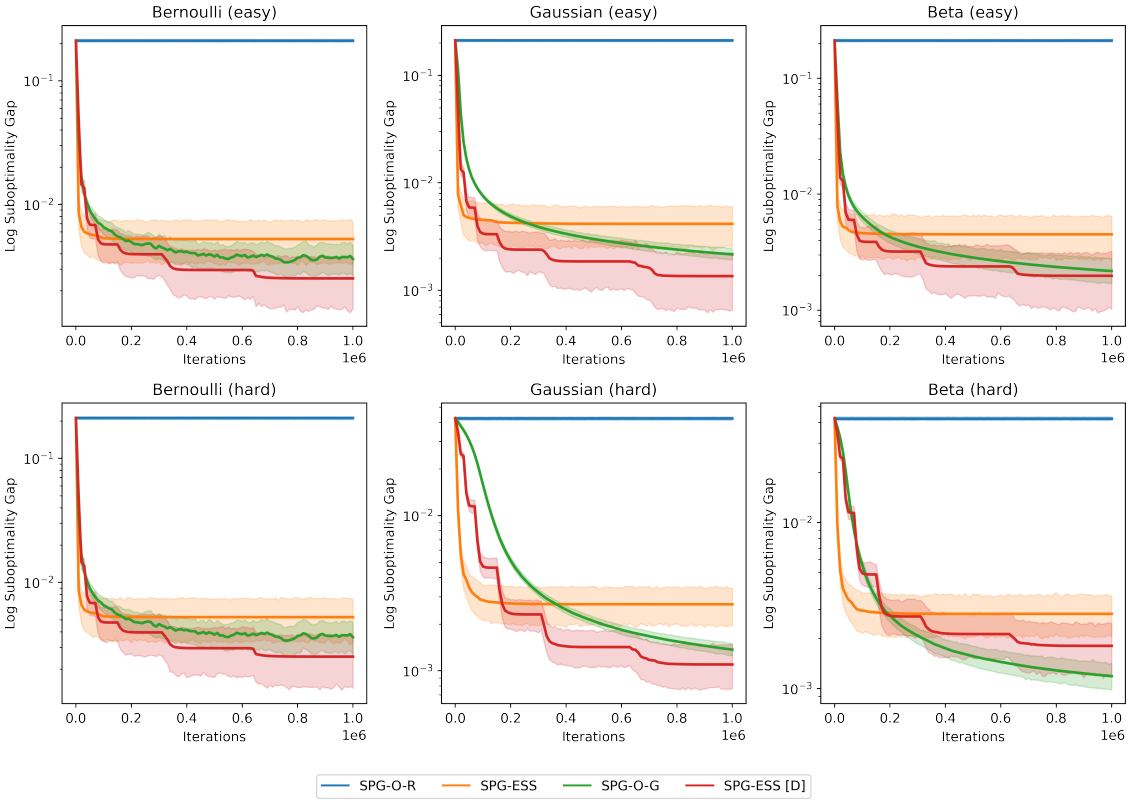

Figure 2: Expected sub-optimality gap across various environments. `SPG-ESS` and `SPG-ESS [D]` is comparable to `SPG-O-G` and `SPG-O-R` without using any oracle-like knowledge of the environment.

(Auer et al., 1995; Hazan & Kale, 2014). For this "doubling trick", we first start with a smaller time horizon $\mathcal{T}_0 << T$ when setting the step-size, i.e. for $t \leq \mathcal{T}_0$, $\eta_t = \eta_0 \left(\frac{\beta}{\mathcal{T}_0}\right)^{\frac{t}{\mathcal{T}_0}}$. After $\mathcal{T}_0$ iterations, we restart the step-size schedule, double the length of the next time horizon i.e. $\mathcal{T}_1 = 2\,\mathcal{T}_0$ and set η_t with the time horizon equal to $\mathcal{T}_1$. This process repeats until the desired number of iterations is reached. For `SPG-ESS [D]` we select $\beta = 1$, $\eta_0 = \frac{1}{18}$ and $\mathcal{T}_0 = 5000$ for all environments.

Results: From Figure 2, we conclude that `SPG-ESS` and `SPG-ESS [D]` are consistently comparable to `SPG-O-G` and `SPG-O-R` without access to any oracle-like knowledge. While `SPG-O-R` has the best theoretical convergence rate, its step-size is proportional to the reward gap. When the reward gap is small, so is the resulting step-size which results in its poor empirical performance.

6 Discussion

We designed (stochastic) softmax policy gradient (PG) methods for bandits and tabular Markov decision processes (MDPs). Throughout, we demonstrated that the proposed methods offer similar theoretical guarantees as the state-of-the art results, but do not require the knowledge of oracle-like quantities. Concretely, in the exact setting, we empirically demonstrated that using softmax PG with Armijo line-search to set the step-size is competitive to GNPG without requiring knowledge of the concentrability coefficient to set the step-size. In the stochastic setting, we used exponentially decreasing step-sizes and showed that the resulting algorithm is robust to problem-dependent constants and can interpolate between slow and fast rates. For future work, we hope to analyze the convergence rate when using the "doubling trick" with exponentially decreasing step-sizes. Finally, we aim to generalize our results to support complex (non)-linear policy parameterization.

References

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *J. Mach. Learn. Res.*, 22 (98):1–76, 2021. (cited on 1, 2, 3, 4, 64)

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pp. 39–1. JMLR Workshop and Conference Proceedings, 2012. (cited on 10)

Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the impact of entropy on policy optimization. In *International conference on machine learning*, pp. 151–160. PMLR, 2019. (cited on 3, 41)

Eitan Altman. *Constrained Markov decision processes*. Routledge, 2021. (cited on 3)

Larry Armijo. Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16(1):1 – 3, 1966. (cited on 2, 5, 16)

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th annual foundations of computer science*, pp. 322–331. IEEE, 1995. (cited on 11)

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002. (cited on 10)

Jalaj Bhandari and Daniel Russo. On the linear convergence of policy gradient methods for finite mdps. In *International Conference on Artificial Intelligence and Statistics*, pp. 2386–2394. PMLR, 2021. (cited on 1, 2, 4)

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4):2563–2578, 2022. (cited on 2, 41, 56)

Yuhao Ding, Junzi Zhang, Hyunin Lee, and Javad Lavaei. Beyond exact gradients: Convergence of stochastic soft-max policy gradient methods with entropy regularization. (cited on 41, 45, 67)

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. PMLR, 2018. (cited on 3, 41)

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1):2489–2512, 2014. (cited on 11)

Takuya Hiraoka, Takahisa Imagawa, Taisei Hashimoto, Takashi Onishi, and Yoshimasa Tsuruoka. Dropout q-functions for doubly efficient reinforcement learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xCVJMsPv3RT. (cited on 3)

Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018. (cited on 4)

Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001. (cited on 2)

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, pp. 795–811. Springer, 2016. (cited on 4, 8)

Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, 198(1):1059–1106, 2023. (cited on 1, 2)

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020. (cited on 7)

Xiaoyu Li, Zhenxun Zhuang, and Francesco Orabona. A second look at exponential and cosine step sizes: Simplicity, adaptivity, and performance. In *International Conference on Machine Learning*, pp. 6553–6564. PMLR, 2021. (cited on 2, 7, 8, 45, 63)

Jiacai Liu, Wenye Li, and Ke Wei. Elementary analysis of policy gradient methods. *arXiv preprint arXiv:2404.03372*, 2024. (cited on 6)

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pp. 6820–6829. PMLR, 2020. (cited on 1, 2, 3, 4, 5, 17, 19, 20, 21, 41, 42, 43, 65, 66)

Jincheng Mei, Bo Dai, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. Understanding the effect of stochasticity in policy optimization. *Advances in Neural Information Processing Systems*, 34:19339–19351, 2021a. (cited on 1, 2, 6, 7, 8, 10, 33, 46, 66, 67)

Jincheng Mei, Yue Gao, Bo Dai, Csaba Szepesvari, and Dale Schuurmans. Leveraging non-uniformity in first-order non-convex optimization. In *International Conference on Machine Learning*, pp. 7555–7564. PMLR, 2021b. (cited on 2, 65, 66)

Jincheng Mei, Wesley Chung, Valentin Thomas, Bo Dai, Csaba Szepesvari, and Dale Schuurmans. The role of baselines in policy gradient optimization. *Advances in Neural Information Processing Systems*, 35:17818–17830, 2022. (cited on 2)

Jincheng Mei, Zixin Zhong, Bo Dai, Alekh Agarwal, Csaba Szepesvari, and Dale Schuurmans. Stochastic gradient succeeds for bandits. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24325–24360. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/mei23a.html. (cited on 2, 8, 9, 10, 31, 37, 40, 41, 46)

Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer. (cited on 5)

Ian Osband, Yotam Doron, Matteo Hessel, John Aslanides, Eren Sezener, Andre Saraiva, Katrina McKinney, Tor Lattimore, Csaba Szepesvari, Satinder Singh, et al. Behaviour suite for reinforcement learning. *arXiv preprint arXiv:1908.03568*, 2019. (cited on 64)

Boris T Polyak. Introduction to optimization. 1987. (cited on 6)

B.T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(63)90382-3. URL https://www.sciencedirect.com/science/article/pii/0041555363903823. (cited on 4, 8)

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014. (cited on 3, 4)

Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013. (cited on 2, 9)

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. (cited on 1)

Lior Shani, Yonathan Efroni, and Shie Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5668–5675, 2020. (cited on 1)

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018. (cited on 64)

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, pp. 1057–1063, Cambridge, MA, USA, 1999a. MIT Press. (cited on 1)

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999b. (cited on 2, 33)

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1195–1204. PMLR, 2019. (cited on 2, 9)

Sharan Vaswani, Benjamin Dubois-Taine, and Reza Babanezhad. Towards noise-adaptive, problem-adaptive (accelerated) stochastic gradient descent. In *International Conference on Machine Learning*, pp. 22015–22059. PMLR, 2022. (cited on 2, 7, 8, 39, 40, 45)

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992. (cited on 2, 41)

Lin Xiao. On the convergence rates of policy gradient methods. *arXiv preprint arXiv:2201.07443*, 2022. (cited on 2)

Rui Yuan, Robert M. Gower, and Alessandro Lazaric. A general sample complexity analysis of vanilla policy gradient, 2022. (cited on 2, 7, 8)

Tom Zahavy, Brendan O'Donoghue, Guillaume Desjardins, and Satinder Singh. Reward is enough for convex mdps. *Advances in Neural Information Processing Systems*, 34:25746–25759, 2021. (cited on 3)

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*, 2019. (cited on 4)

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. *Advances in Neural Information Processing Systems*, 33:4572–4583, 2020a. (cited on 3)

Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Basar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6): 3586–3612, 2020b. (cited on 2)

Supplementary Material

Organization of the Appendix

A Definitions

A function f is L-smooth if for all θ and θ'

$$|f(\theta) - f(\theta') - \langle \nabla f(\theta'), \theta - \theta' \rangle| \leq \frac{L}{2} \|\theta - \theta'\|_2^2. \tag{8}$$

A function f is L_1-non-uniform smooth if for all θ and θ'

$$|f(\theta) - f(\theta') - \langle \nabla f(\theta'), \theta - \theta' \rangle| \leq \frac{L_1 \|\nabla f(\theta')\|}{2} \|\theta - \theta'\|_2^2. \tag{9}$$

A function f satisfies the non-uniform Łojasiewciz condition of degree ξ for $\xi \in [0, 1]$ is defined as

$$\|\nabla f(\theta)\| \geq C(\theta) |f^* - f(\theta)|^{1-\xi} \qquad (f^* := \sup_\theta f(\theta))$$

where $C : \theta \to \mathbb{R} > 0$.

A function f satisfies the reversed Łojasiewciz condition if for all θ

$$\|\nabla f(\theta)\| \leq \nu \left[f^* - f(\theta)\right] \tag{10}$$

where $\nu > 0$.

B Proofs in Section 3

B.1 Proof Of Theorem 1

Theorem 1. Assuming f is (i) L-smooth, (ii) satisfies the non-uniform Łojasiewciz condition with $\xi = 0$, and (iii) $\mu := \inf_{t\geq 1}[C(\theta_t)]^2 > 0$, using Update 1 and Armijo line-search to set the step-size results in the following convergence:

$$f^* - f(\theta_{T+1}) \leq \max\left\{\frac{L}{2\,h\,(1-h)}, \frac{1}{h\,\eta_{\max}}\right\}\frac{1}{\mu\,T} \tag{2}$$

where $h \in (0,1)$ and $\eta_{\max}$ is the upper-bound on the step-size.

Proof. From Equation (1), Armijo line-search selects a step-size that satisfies the following condition where $h \in (0,1)$ is a hyper-parameter

$$f(\theta_t + \eta_t\,\nabla f(\theta_t)) \geq f(\theta_t) + h\,\eta_t\,\|\nabla f(\theta_t)\|_2^2. \tag{11}$$

For any L-smooth function the step-size η_t returned by the Armijo line-search is guaranteed to satisfy $\eta_{\max} \geq \eta_t \geq \min\left\{\frac{2\,(1-h)}{L}, \eta_{\max}\right\}$ (Armijo, 1966) which implies that

$$f(\theta_{t+1}) \geq f(\theta_t) + \min\left\{\frac{2\,h\,(1-h)}{L}, h\,\eta_{\max}\right\}\|\nabla f(\theta_t)\|_2^2 \tag{12}$$

Adding f^* to both sides and multiplying by -1

$$f^* - f(\theta_{t+1}) \leq f^* - f(\theta_t) - \min\left\{\frac{2\,h\,(1-h)}{L}, h\,\eta_{\max}\right\}\|\nabla f(\theta_t)\|_2^2 \tag{13}$$

Let $\delta(\theta_t) := f^* - f(\theta_t)$

$$\delta(\theta_{t+1}) \leq \delta(\theta_t) - \min\left\{\frac{2\,h\,(1-h)}{L}, h\,\eta_{\max}\right\}\|\nabla f(\theta_t)\|_2^2 \tag{14}$$

Since f satisfies the non-uniform Łojasiewciz condition with $\xi = 0$

$$\leq \delta(\theta_t) - \min\left\{\frac{2\,h\,(1-h)}{L}, h\,\eta_{\max}\right\}[C(\theta_t)]^2\,[\delta(\theta_t)]^2 \tag{15}$$

Assuming $\mu := \inf_{t\geq 1}[C(\theta_t)]^2 > 0$

$$\leq \delta(\theta_t) - \underbrace{\mu\,\min\left\{\frac{2\,h(1-h)}{L}, h\,\eta_{\max}\right\}}_{:=\frac{1}{C}}[\delta(\theta_t)]^2 \tag{16}$$

Dividing by $\delta(\theta_t)\,\delta(\theta_{t+1})$

$$\implies \frac{1}{\delta(\theta_t)} \leq \frac{1}{\delta(\theta_{t+1})} - \frac{1}{C}\,\frac{\delta(\theta_t)}{\delta(\theta_{t+1})} \tag{17}$$

Using Equation (17) and recursing from $t = 1$ to T

$$\frac{1}{\delta(\theta_1)} \leq \frac{1}{\delta(\theta_{T+1})} - \frac{1}{C}\sum_{t=1}^{T}\frac{\delta(\theta_t)}{\delta(\theta_{t+1})} \tag{18}$$

$$\leq \frac{1}{\delta(\theta_{T+1})} - \frac{T}{C} \qquad \left(\tfrac{\delta(\theta_t)}{\delta(\theta_{t+1})} \geq 1\right)$$

$$\implies \frac{T}{C} \leq \frac{1}{\delta(\theta_{T+1})}. \tag{19}$$

Therefore

$$f^* - f(\theta_{T+1}) \leq \max\left\{\frac{L}{2\,h\,(1-h)}\,\frac{1}{h\,\eta_{\max}}\right\}\frac{1}{\mu}. \tag{20}$$

□

Corollary 2. In the bandit setting, using Update 1 with Armijo line-search to set the step-size results in the following convergence:

$$(\pi^* - \pi_{\theta_{T+1}})^\top r \leq \max\left\{\frac{5}{4\,h\,(1-h)}, \frac{1}{h\,\eta_{\max}}\right\}\frac{1}{\mu\,T} \tag{21}$$

where $h \in (0,1)$, $\eta_{\max}$ is the upper-bound on the step-size, and $\mu := \inf_{t\geq 1}[\pi_{\theta_t}(a^*)]^2 > 0$.

Proof. We can extend Theorem 1 to the bandit setting since:

- by Lemma 24, f is $\frac{5}{2}$-smooth
- by Lemma 31, f is non-uniform Łojsiewciz with $\xi = 0$ and $C(\theta) = \pi_\theta(a^*)$
- we observe that (Mei et al., 2020, Lemma 5) works for any step-size sequence guaranteeing monotonic improvement $\mu := \inf_{t\geq 1}[C(\theta_t)]^2 > 0$

□

Corollary 3. Assuming $\min_{s\in\mathcal{S}} \rho(s) > 0$, in the tabular MDP setting, using Update 1 with Armijo line-search to set the step-size results in the following convergence:

$$V^*(\rho) - V^{\pi_{\theta_{T+1}}}(\rho) \leq \max\left\{\frac{8}{2\,h\,(1-h)\,(1-\gamma)^3}\,\frac{1}{\eta_{\max}\,h}\right\}\frac{1}{\mu\,T} \tag{22}$$

where $h \in (0,1)$, $\eta_{\max}$ is the upper-bound on the step-size, and $\mu := \inf_{t\geq 1}\left(\frac{\min_s \pi_{\theta_t}(a^*(s)|s)}{\sqrt{S}\left\|d_\rho^{\pi^*}/d_\rho^{\pi_{\theta_t}}\right\|_\infty}\right)^2 > 0$.

Proof. We can extend Theorem 1 to the tabular MDP setting since:

- by Lemma 27, f is $\frac{8}{(1-\gamma)^3}$-smooth,
- by Lemma 32, f is non-uniform Łojsiewciz with $\xi = 0$ and $C(\theta) = \frac{\min_s \pi_\theta(a^*(s)|s)}{\sqrt{S}\left\|d_\rho^{\pi^*}/d_\rho^{\pi_\theta}\right\|_\infty}$
- we observe that (Mei et al., 2020, Lemma 9) works for any step-size sequence guaranteeing monotonic improvement $\mu := \inf_{t\geq 1}[C(\theta_t)]^2 > 0$

□

B.2 Proof Of Theorem 2

Theorem 2. For a given $\epsilon \in (0,1)$, assuming f is (i) L_1 non-uniform smooth, (ii) satisfies the non-uniform Łojasiewciz condition with $\xi = 0$, (iii) $\mu := \inf_{t\geq 1}[C(\theta_t)]^2 > 0$, (iv) f satisfies a reversed Łojasiewciz condition with $\nu > 0$, using Update 1 with backtracking line-search using the Armijo condition in Equation (3) and setting $\eta_{\max} = C/\epsilon$ results in the following convergence:
If $f^* - f(\theta_t) > \epsilon$ for all $t \in [1,T]$, then,

$$f^* - f(\theta_{T+1}) \leq [f^* - f(\theta_1)]\,\exp\left(-\min\left\{C\,h, \frac{2\,h\,(1-h)}{L_1\,\nu}\right\}\mu\,T\right) \tag{4}$$

where $C > 0$ and $h \in (0,1)$ are hyper-parameters. Otherwise $\min_{t\in[1,T]} f^* - f(\theta_t) \leq \epsilon$.

Proof. Since the rewards are bounded, we will overload the notation and let $f^* - f(\theta_t)$ denote the normalized sub-optimality gap. This implies that $f^* - f(\theta_t) \leq 1$. Using backtracking line-search

using Armijo condition in Equation (3) selects a step-size that satisfies the following condition where $h \in (0,1)$ is a hyper-parameter:

$$\ln(f^* - f(\theta_t + \eta_t \nabla f(\theta_t))) \leq \ln(f^* - f(\theta_t)) - h\,\eta_t \frac{\|\nabla f(\theta_t)\|_2^2}{f^* - f(\theta_t)} \tag{23}$$

Applying $\exp(\cdot)$ to both sides

$$f^* - f(\theta_t + \eta_t\,\nabla f(\theta_t)) \leq [f^* - f(\theta_t)]\,\exp\left(-h\,\eta_t \frac{\|\nabla f(\theta_t)\|_2^2}{f^* - f(\theta_t)}\right) \tag{24}$$

By Lemma 2, we can guarantee that the backtracking line-search is guaranteed to satisfy $\eta_t \geq \min\left\{\eta_{\max}, \frac{2(1-h)}{L_1\,\nu\,[f^* - f(\theta_t)]}\right\}$ which implies that

$$f^* - f(\theta_{t+1}) \leq [f^* - f(\theta_t)]\,\exp\left(-\min\left\{\eta_{\max}\,h, \frac{2\,h\,(1-h)}{L_1\,\nu\,[f^* - f(\theta_t)]}\right\}\frac{\|\nabla f(\theta_t)\|_2^2}{f^* - f(\theta_t)}\right) \tag{25}$$

Assuming that for a target $\epsilon \in (0,1)$, $\epsilon < f^* - f(\theta_t)$ for $t \in [1,T]$, selecting $\eta_{\max} = \frac{C}{\epsilon}$ for $C > 0$ implies $\eta_{\max} > \frac{C}{f^* - f(\theta_t)}$

$$\leq [f^* - f(\theta_t)]\,\exp\left(-\min\left\{C\,h, \frac{2\,h\,(1-h)}{L_1\,\nu}\right\}\frac{\|\nabla f(\theta_t)\|_2^2}{(f^* - f(\theta_t))^2}\right) \tag{26}$$

Since f satisfies the non-uniform Łojasieciz condition with $\xi = 0$

$$\leq [f^* - f(\theta_t)]\,\exp\left(-\min\left\{C\,h, \frac{2\,h\,(1-h)}{L_1\,\nu}\right\}[C(\theta_t)]^2\right) \tag{27}$$

Assuming $\mu := \inf_{t\geq 1}[C(\theta_t)]^2 > 0$

$$\implies f^* - f(\theta_{t+1}) \leq [f^* - f(\theta_t)]\,\exp\left(-\min\left\{C\,h, \frac{2\,h\,(1-h)}{L_1\,\nu}\right\}\mu\right). \tag{28}$$

Using Equation (28) and recursing from $t = 1$ to T we have

$$f^* - f(\theta_{T+1}) \leq [f^* - f(\theta_1)]\,\exp\left(-\min\left\{C\,h, \frac{2\,h\,(1-h)}{L_1\,\nu}\right\}\mu\,T\right). \tag{29}$$

□

Corollary 4. In the bandit setting, for a given $\epsilon \in (0,1)$, using Update 1 with backtracking line-search using the Armijo condition in Equation (3) and setting $\eta_{\max} = C/\epsilon$ results in the following convergence:
If $(\pi^* - \pi_{\theta_t})^\top r > \epsilon$ for all $t \in [1,T]$, then,

$$(\pi^* - \pi_{\theta_{T+1}})^\top r \leq (\pi^* - \pi_{\theta_1})^\top r\,\exp\left(-\min\left\{C\,h, \frac{2\,h\,(1-h)\,\Delta^*}{3\sqrt{2}}\right\}\mu\,T\right) \tag{30}$$

where $C > 0$ and $h \in (0,1)$ are hyper-parameters, $\Delta^* := r(a^*) - \max_{a\neq a^*} r(a)$, and $\mu := \inf_{t\geq 1}[\pi_{\theta_t}(a^*)]^2 > 0$. Otherwise $\min_{t\in[1,T]}(\pi^* - \pi_{\theta_t})^\top r \leq \epsilon$.

Proof. We can extend Theorem 2 to the bandit setting since:

- by Lemma 29, f is 3-non-uniform smooth

- by Lemma 31, f is non-uniform Łojsiewciz with $\xi = 0$ and $C(\theta) = \pi_\theta(a^*)$
- by Lemma 3, f satisfies the reverse Łojasiewciz condition with $\nu = \frac{\sqrt{2}}{\Delta^*}$
- since we observe that Lemma 5 in (Mei et al., 2020) works for any step-size sequence guaranteeing monotonic improvement, $\mu := \inf_{t\geq 1}[C(\theta_t)]^2 > 0$

□

Corollary 5. Assuming $\min_{s\in\mathcal{S}} \rho(s) > 0$, in the tabular MDP setting, for a given $\epsilon \in (0,1)$, using Update 1 with backtracking line-search using the Armijo condition in Equation (3) and setting $\eta_{\max} = C/\epsilon$ results in the following convergence:
If $V^*(\rho) - V^{\pi_{\theta_t}}(\rho) > \epsilon$ for all $t \in [1,T]$, then,

$$V^*(\rho) - V^{\pi_{\theta_{T+1}}}(\rho) \leq [V^*(\rho) - V^{\pi_{\theta_1}}(\rho)] \exp\left(-\min\left\{C\,h, \frac{2\,h\,(1-h)\,(1-\gamma)\,\Delta^*}{D\sqrt{2}}\right\}\mu\,T\right) \tag{31}$$

where $C > 0$ and $h \in (0,1)$ are hyper-parameters, $D := \left[3 + \frac{2\,C_\infty - (1-\gamma)}{(1-\gamma)\,\gamma}\right]\sqrt{S}$, $C_\infty := \max_\pi \left\|\frac{d_\rho^\pi}{\rho}\right\|_\infty \leq \frac{1}{\min_s \rho(s)} < \infty$, $\Delta^* := \min_{s\in\mathcal{S}}\left\{Q^*(s,a^*(s)) - \max_{a(s)\neq a^*(s)} Q^*(s,a)\right\}$, and $\mu := \inf_{t\geq 1}\left(\frac{\min_s \pi_{\theta_t}(a^*(s)|s)}{\sqrt{S}\left\|d_\rho^{\pi^*}/d_\rho^{\pi_{\theta_t}}\right\|_\infty}\right)^2 > 0$. Otherwise $\min_{t\in[1,T]} V^*(\rho) - V^{\pi_{\theta_t}}(\rho) \leq \epsilon$.

Proof. We can extend Theorem 2 to the tabular MDP setting since:

- by Lemma 30, f is D-non-uniform smooth where $D := \left[3 + \frac{2\,C_\infty - (1-\gamma)}{(1-\gamma)\,\gamma}\right]\sqrt{S}$ and $C_\infty := \max_\pi \left\|\frac{d_\rho^\pi}{\rho}\right\|_\infty \leq \frac{1}{\min_s \rho(s)} < \infty$
- by Lemma 32, f is non-uniform Łojsiewciz with $\xi = 0$ and $C(\theta) = \frac{\min_s \pi_\theta(a^*(s)|s)}{\sqrt{S}\left\|d_\rho^{\pi^*}/d_\rho^{\pi_\theta}\right\|_\infty}$
- by Lemma 4 f satisfies the reverse Łojsiewciz condition with $\nu = \frac{\sqrt{2}}{(1-\gamma)\,\Delta^*}$ and $\Delta^* := \min_{s\in\mathcal{S}}\left\{Q^*(s,a^*(s)) - \max_{a(s)\neq a^*(s)} Q^*(s,a)\right\}$
- since we observe that Lemma 9 in (Mei et al., 2020) works for any step-size sequence guaranteeing monotonic improvement $\mu := \inf_{t\geq 1}[C(\theta_t)]^2 > 0$

□

B.3 Additional Lemmas

Lemma 1. *Suppose that (i) f is L_1-non-uniform smooth and (ii) satisfies a reversed Łojasiewciz inequality then $\theta \to \ln(f^* - f(\theta))$ is $L_1\,\nu$-smooth.*

Proof. Let $g(\theta) := \ln(f^* - f(\theta))$. By Taylor's theorem it suffices to show that the Hessian is bounded by $L_1\,\nu$

$$\nabla^2 g(\theta) = \frac{-\nabla^2 f(\theta)\,(f^* - f(\theta)) - [\nabla f(\theta)]\,[\nabla f(\theta)]^\top}{(f^* - f(\theta))^2} \tag{32}$$

Since for any $x \in \mathbb{R}^{SA}$ $x\,x^\top \succeq 0$

$$\preceq \frac{\nabla^2 f(\theta)}{f^* - f(\theta)} \tag{33}$$

Since f is L_1-non-uniform smooth,

$$\preceq \frac{L_1 \|\nabla f(\theta)\|}{f^* - f(\theta)} \tag{34}$$

Since f satisfies the reverse Łojsaiewciz inequality

$$\preceq L_1 \, \nu \, I_{SA}. \tag{35}$$

□

Lemma 2. *The (exact) backtracking procedure with the following Armijo condition on the log-loss:*

$$\ln(f^* - f(\theta_t + \eta_t \nabla f(\theta_t))) \leq \ln(f^* - f(\theta_t)) - h \, \eta_t \, \frac{\|\nabla f(\theta_t)\|_2^2}{f^* - f(\theta_t)} \tag{36}$$

terminates and returns

$$\eta_t \geq \min \left\{ \eta_{\max}, \frac{2(1-h)}{L_1 \, \nu \, [f^* - f(\theta_t)]} \right\} \tag{37}$$

where $h \in (0, 1)$ is a hyper-parameter.

Proof. Let $g(\theta) = \ln(f^* - f(\theta))$. By Lemma 1, g is $L_1 \, \nu$-smooth. Starting with the quadratic bound using the smoothness of g:

$$g(\theta_{t+1}) \leq g(\theta_t) - \eta_t \left\langle \frac{\nabla f(\theta_t)}{f^* - f(\theta_t)}, \nabla f(\theta_t) \right\rangle + \frac{L_1 \, \nu \, {\eta_t}^2}{2} \|\nabla f(\theta_t)\|_2^2 \tag{38}$$

$$\leq \underbrace{g(\theta_t) - \|\nabla f(\theta_t)\|_2^2 \left(\frac{\eta_t}{f^* - f(\theta_t)} - \frac{L_1 \, \nu \, {\eta_t}^2}{2} \right)}_{:= h_1(\eta_t)} \tag{39}$$

From Equation (3)

$$g(\theta_t + \eta_t \nabla f(\theta_t)) \leq \underbrace{g(\theta_t) - h \, \eta_t \, \frac{\|\nabla f(\theta_t)\|_2^2}{f^* - f(\theta_t)}}_{:= h_2(\eta_t)} \tag{40}$$

If Equation (3) is satisfied, the backtracking line-search procedure terminates. If $\eta_{\max} \leq \frac{2(1-h)}{L_1 \, \nu \, [f^* - f(\theta_t)]}$ then $g(\theta_{t+1}) \leq h_1(\eta_{\max}) \leq h_2(\eta_{\max})$ implying the line-search terminates and $\eta_t = \eta_{\max}$. Otherwise, if $\eta_{\max} > \frac{2(1-h)}{L_1 \, \nu \, [f^* - f(\theta_t)]}$ and Equation (3) is satisfied for step-size η_t then

$$\ln(\theta_t + \eta_t \nabla f(\theta_t)) \leq h_2(\eta_t) \leq h_1(\eta_t) \tag{41}$$

$$\implies \frac{h \eta_t}{f^* - f(\theta_t)} \geq \frac{\eta_t}{f^* - f(\theta_t)} - \frac{L_1 \, \nu {\eta_t}^2}{2} \tag{42}$$

$$\implies \eta_t \geq \frac{2(1-h)}{L_1 \, \nu \, [f^* - f(\theta_t)]} \tag{43}$$

Putting the above conditions together, we have:

$$\eta_t \geq \min \left\{ \eta_{\max}, \frac{2(1-h)}{L_1 \, \nu \, [f^* - f(\theta_t)]} \right\}. \tag{44}$$

□

Lemma 3 (Lemma 17 in (Mei et al., 2020))**.** *For any $r \in [0, 1]^A$. Denote $\Delta^* := r(a^*) - \max_{a \neq a^*} r(a)$. Then,*

$$\left\| \frac{d\langle \pi_\theta, r \rangle}{d\theta} \right\| \leq \frac{\sqrt{2}}{\Delta^*} \langle \pi^* - \pi_\theta, r \rangle. \tag{45}$$

Lemma 4 (Lemma 28 in (Mei et al., 2020)). *Denote* $\Delta^*(s) := Q^*(s, a^*(s)) - \max_{a \neq a^*(s)} Q^*(s, a)$ *as the optimal value gap of state* s*, where* $a^*(s)$ *is the action that the optimal policy selects under state* s*, and* $\Delta^* := \min_{s \in \mathcal{S}} \Delta^*(s) > 0$ *as the optimal value gap of the MDP. Then we have*

$$\left\| \frac{\partial V^{\pi_\theta}(\rho)}{\partial \theta} \right\| \leq \frac{1}{1-\gamma} \frac{\sqrt{2}}{\Delta^*} [V^*(\rho) - V^{\pi_\theta}(\rho)]. \tag{46}$$

C Proofs in Section 4

C.1 Proof Of Theorem 3

Theorem 3. For a given $\epsilon \in (0,1)$, assuming f is (i) L-smooth, (ii) satisfies the non-uniform Łojasiewciz condition with $\xi = 0$, (iii) $\mu := \left[\mathbb{E}\left[\inf_{t\geq 1}[C(\theta_t)]^{-2}\right]\right]^{-1} > 0$, using Update 2 with (a) unbiased stochastic gradients whose variance is bounded by σ^2 and (b) exponentially decreasing step-sizes $\eta_t = \eta_0\,\alpha^t$ where $\eta_0 = \frac{1}{L}$ and $\alpha = \left(\frac{\beta}{T}\right)^{\frac{1}{T}}$, $\beta \geq 1$ results in the following convergence: If $\mathbb{E}[f^* - f(\theta_t)] > \epsilon$ for all $t \in [1, T]$, then,

$$\mathbb{E}[f^* - f(\theta_{T+1})] \leq \mathbb{E}[f^* - f(\theta_1)]\, C_1\, \exp\left(-\frac{\alpha\,\epsilon\,T}{\kappa\,\ln(T/\beta)}\right) + \frac{C_1\,C_2}{2\,L}\frac{\ln^2\left(\frac{T}{\beta}\right)\sigma^2}{\epsilon^2\,T} \tag{5}$$

where $\kappa := \frac{2\,L}{\mu}$, $C_1 := \exp\left(\frac{2\,\beta}{\kappa\,\ln(T/\beta)}\right)$ and $C_2 := \frac{4\,\kappa^2}{e^2\,\alpha^2}$. Otherwise $\min_{t\in[1,T]} \mathbb{E}[f^* - f(\theta_t)] \leq \epsilon$.

Proof.

Starting with the smoothness of f

$$|f(\theta_{t+1}) - f(\theta_t) - \langle \nabla f(\theta_t), \theta_{t+1} - \theta_t\rangle| \leq \frac{L}{2}\,\|\theta_t - \theta_t\|_2^2 \tag{47}$$

$$f(\theta_{t+1}) - f(\theta_t) - \langle \nabla f(\theta_t), \theta_{t+1} - \theta_t\rangle \geq -\frac{L}{2}\,\|\theta_t - \theta_t\|_2^2 \tag{48}$$

Using Update 2, $\theta_{t+1} = \theta_t + \eta_t\,\nabla\widetilde{f}(\theta_t)$

$$f(\theta_{t+1}) - f(\theta_t) - \eta_t\left\langle \nabla f(\theta_t), \nabla\widetilde{f}(\theta_t)\right\rangle \geq -\frac{L}{2}\,{\eta_t}^2\left\|\nabla\widetilde{f}(\theta_t)\right\|_2^2 \tag{49}$$

$$\implies f(\theta_{t+1}) \geq f(\theta_t) + \eta_t\left\langle \nabla f(\theta_t), \nabla\widetilde{f}(\theta_t)\right\rangle - \frac{L}{2}\,{\eta_t}^2\left\|\nabla\widetilde{f}(\theta_t)\right\|_2^2 \tag{50}$$

Multiplying both sides by -1 and adding f^*

$$f^* - f(\theta_{t+1}) \leq f^* - f(\theta_t) - \eta_t\left\langle \nabla f(\theta_t), \nabla\widetilde{f}(\theta_t)\right\rangle + \frac{L}{2}\,{\eta_t}^2\left\|\nabla\widetilde{f}(\theta_t)\right\|_2^2 \tag{51}$$

Taking expectation with respect to the randomness in iteration t on both sides

$$\underbrace{\mathbb{E}[f^* - f(\theta_{t+1})]}_{:=\delta(\theta_{t+1})} \leq \underbrace{\mathbb{E}[f^* - f(\theta_t)]}_{:=\delta(\theta_t)} - \eta_t\left\langle \nabla f(\theta_t), \mathbb{E}\left[\nabla\widetilde{f}(\theta_t)\right]\right\rangle + \frac{L\,{\eta_t}^2}{2}\,\mathbb{E}\left[\left\|\nabla\widetilde{f}(\theta_t)\right\|_2^2\right] \tag{52}$$

Assuming that the gradient is unbiased

$$\implies \delta(\theta_{t+1}) = \delta(\theta_t) - \eta_t\,\|\nabla f(\theta_t)\|_2^2 + \frac{L\,{\eta_t}^2}{2}\,\mathbb{E}\left[\left\|\nabla\widetilde{f}(\theta_t)\right\|_2^2\right] \tag{53}$$

$$\leq \delta(\theta_t) - \eta_t\,\|\nabla f(\theta_t)\|_2^2 + \frac{L\,{\eta_t}^2}{2}\,\mathbb{E}\left[\left\|\nabla\widetilde{f}(\theta_t) - \nabla f(\theta_t) + \nabla f(\theta_t)\right\|_2^2\right] \tag{54}$$

Expanding the square and since $\mathbb{E}\left[\left\langle \nabla f(\theta_t), \nabla\widetilde{f}(\theta_t) - \nabla f(\theta_t)\right\rangle\right] = 0$

$$\leq \delta(\theta_t) - \eta_t\,\|\nabla f(\theta_t)\|_2^2 + \frac{L\,{\eta_t}^2}{2}\mathbb{E}\left[\left\|\nabla\widetilde{f}(\theta_t) - \nabla f(\theta_t)\right\|_2^2\right] + \frac{L\,{\eta_t}^2}{2}\,\mathbb{E}\left[\|\nabla f(\theta_t)\|_2^2\right] \tag{55}$$

Assuming that the variance is bounded by σ^2

$$\leq \delta(\theta_t) - \eta_t \left\|\nabla f(\theta_t)\right\|_2^2 + \frac{L\,{\eta_t}^2}{2}\left(\sigma^2 + \mathbb{E}\Big[\left\|\nabla f(\theta_t)\right\|_2^2\Big]\right) \tag{56}$$

$$\leq \delta(\theta_t) - \frac{\eta_t}{2}\left\|\nabla f(\theta_t)\right\|_2^2 + \frac{L\,{\eta_t}^2}{2}\,\sigma^2 \qquad (\eta_t \leq \tfrac{1}{L}) \tag{57}$$

Since f satisfies the non-uniform Łojsaiewciz condition with $\xi = 0$

$$\leq \delta(\theta_t) - \frac{\eta_t}{2}\left[\delta(\theta_t)\right]^2\left[C(\theta_t)\right]^2 + \frac{L\,{\eta_t}^2}{2}\,\sigma^2 \tag{58}$$

Assuming $m := \inf_{t\geq 1}[C(\theta_t)]^2 > 0$

$$\leq \delta(\theta_t)\left(1 - \frac{\eta_t\, m}{2}\,\delta(\theta_t)\right) + \frac{L\,{\eta_t}^2}{2}\,\sigma^2. \tag{59}$$

Taking expectation with respect to all previous iterations $t \geq 1$ on both sides

$$\implies \mathbb{E}[\delta(\theta_{t+1})] \leq \mathbb{E}[\delta(\theta_t)] - \frac{\eta_t}{2}\mathbb{E}[m\left[\delta(\theta_t)\right]^2] + \frac{L\,{\eta_t}^2}{2}\,\sigma^2 \tag{60}$$

To lower-bound $\mathbb{E}[m\left[\delta(\theta_t)\right]^2]$

$$\mathbb{E}[\delta(\theta_t)] = \mathbb{E}\left[\frac{1}{\sqrt{m}}\,\sqrt{m}\,\delta(\theta_t)\right] \tag{61}$$

Using Cauchy-Schwarz since $m > 0$ and $\delta(\theta_t) > 0$

$$\leq \sqrt{\mathbb{E}\left[\frac{1}{m}\right]}\,\sqrt{\mathbb{E}\left[m\left[\delta(\theta_t)\right]^2\right]} \tag{62}$$

$$\implies \underbrace{\left[\mathbb{E}\left[\frac{1}{m}\right]\right]^{-1}}_{:=\mu}\mathbb{E}[\delta(\theta_t)]^2 \leq \mathbb{E}\left[m\left[\delta(\theta_t)\right]^2\right]$$

Hence

$$\mathbb{E}[\delta(\theta_{t+1})] \leq \mathbb{E}[\delta(\theta_t)]\left(1 - \frac{\eta_t\,\mu}{2}\,\mathbb{E}[\delta(\theta_t)]\right) + \frac{L\,{\eta_t}^2}{2}\,\sigma^2 \tag{63}$$

If for some $t \in [1, T]$ we have $\mathbb{E}[\delta(\theta_t)] < \epsilon$ then we are done and have converged to a ϵ-neighbourhood within T iterations and have achieved

$$\min_{t\in[1,T]}\mathbb{E}[f^* - f(\theta_t)] \leq \epsilon. \tag{64}$$

Otherwise, we have $\mathbb{E}[\delta(\theta_t)] \geq \epsilon$ and thus

$$\mathbb{E}[\delta(\theta_{t+1})] \leq \mathbb{E}[\delta(\theta_t)]\left(1 - \frac{\eta_t\,\mu\,\epsilon}{2}\,\eta_t\right) + \frac{L\,\sigma^2}{2}\,{\eta_t}^2 \tag{65}$$

$$= \mathbb{E}[\delta(\theta_t)]\left(1 - \frac{\eta_0\,\mu\,\epsilon}{2}\,\alpha^t\right) + \frac{\alpha^{2t}\,L\,\eta_0^2\,\sigma^2}{2} \qquad (\eta_t = \eta_0\,\alpha^t)$$

Define $\frac{1}{\kappa} := \frac{\eta_0\,\mu\,\epsilon}{2}$ and since $\eta_0 = \frac{1}{L}$

$$\leq \mathbb{E}[\delta(\theta_t)]\left(1 - \frac{1}{\kappa}\,\alpha^t\right) + \frac{\alpha^{2t}\,\sigma^2}{2\,L}. \tag{66}$$

Using Equation (66) and recursing from $t = 1$ to T we have

$$\mathbb{E}[\delta(\theta_{T+1})] \leq \mathbb{E}[\delta(\theta_1)] \prod_{t=1}^{T} \left(1 - \frac{1}{\kappa}\,\alpha^t\right) + \frac{\sigma^2}{2\,L} \sum_{t=1}^{T} \alpha^{2t} \prod_{i=t+1}^{T} \left(1 - \frac{1}{\kappa}\,\alpha^i\right) \tag{67}$$

Using $1 - x \leq \exp(-x)$ and by summing up the geometric series

$$\leq \mathbb{E}[\delta(\theta_1)] \exp\!\left(-\frac{1}{\kappa}\frac{\alpha - \alpha^{T+1}}{1-\alpha}\right) + \frac{\sigma^2}{2\,L} \sum_{t=1}^{T} \alpha^{2t} \exp\!\left(-\frac{1}{\kappa}\,\frac{\alpha^{t+1} - \alpha^{T+1}}{1-\alpha}\right). \tag{68}$$

Let us now bound the second term on the RHS

$$\frac{\sigma^2}{2\,L} \sum_{t=1}^{T} \alpha^{2t} \exp\left(-\frac{1}{\kappa}\frac{\alpha^{t+1} - \alpha^{T+1}}{1-\alpha}\right) = \frac{\sigma^2}{2\,L} \exp\left(\frac{\alpha^{T+1}}{\kappa\,(1-\alpha)}\right) \sum_{t=1}^{T} \alpha^{2t} \exp\left(-\frac{\alpha^{t+1}}{\kappa\,(1-\alpha)}\right) \tag{69}$$

By Lemma 8, $\exp(-x) \leq \left(\frac{2}{e\,x}\right)^2$

$$\leq \frac{\sigma^2}{2\,L} \exp\left(\frac{\alpha^{T+1}}{\kappa\,(1-\alpha)}\right) \sum_{t=1}^{T} \alpha^{2t} \left(\frac{2\,(1-\alpha)\,\kappa}{e\,\alpha^{t+1}}\right)^2 \tag{70}$$

$$= \frac{\sigma^2}{2\,L} \exp\left(\frac{\alpha^{T+1}}{\kappa\,(1-\alpha)}\right) \frac{4\,(1-\alpha)^2\,\kappa^2}{e^2\,\alpha^2}\,T \tag{71}$$

Since $1 - x \leq \ln\!\left(\frac{1}{x}\right)$ and using it to bound $(1-\alpha)^2$ where $\alpha = \left(\frac{\beta}{T}\right)^{\frac{1}{T}}$

$$\leq \frac{\sigma^2}{2\,L} \exp\left(\frac{\alpha^{T+1}}{\kappa\,(1-\alpha)}\right) \frac{4\,\kappa^2}{e^2\,\alpha^2} \frac{\ln^2\!\left(\frac{T}{\beta}\right)}{T}. \tag{72}$$

Putting everything together

$$\mathbb{E}[\delta(\theta_{T+1})] \leq \mathbb{E}[\delta(\theta_1)] \exp\!\left(-\frac{1}{\kappa}\frac{\alpha - \alpha^{T+1}}{1-\alpha}\right) + \frac{\sigma^2}{2\,L} \exp\left(\frac{\alpha^{T+1}}{\kappa\,(1-\alpha)}\right) \frac{4\,\kappa^2}{e^2\,\alpha^2} \frac{\ln^2\!\left(\frac{T}{\beta}\right)}{T} \tag{73}$$

$$= \mathbb{E}[\delta(\theta_1)] \exp\!\left(\frac{\alpha^{T+1}}{\kappa\,(1-\alpha)}\right) \exp\!\left(-\frac{\alpha}{\kappa\,(1-\alpha)}\right) + \frac{\sigma^2}{2\,L} \exp\left(\frac{\alpha^{T+1}}{\kappa\,(1-\alpha)}\right) \frac{4\,\kappa^2}{e^2\,\alpha^2} \frac{\ln^2\!\left(\frac{T}{\beta}\right)}{T} \tag{74}$$

By Lemma 6, $\frac{\alpha^{T+1}}{1-\alpha} \leq \frac{2\beta}{\ln(T/\beta)}$

$$\leq \mathbb{E}[\delta(\theta_1)] \exp\!\left(\frac{2\,\beta}{\kappa\,\ln(T/\beta)}\right) \exp\!\left(-\frac{\alpha}{\kappa\,(1-\alpha)}\right) + \frac{\sigma^2}{2\,L} \exp\!\left(\frac{2\,\beta}{\kappa\,\ln(T/\beta)}\right) \frac{4\,\kappa^2}{e^2\,\alpha^2} \frac{\ln^2\!\left(\frac{T}{\beta}\right)}{T} \tag{75}$$

Since $1 - x \leq \ln\!\left(\frac{1}{x}\right)$, $\frac{\alpha}{(1-\alpha)} \geq \frac{\alpha\,T}{\ln(T/\beta)}$

$$\leq \mathbb{E}[\delta(\theta_1)] \exp\!\left(\frac{2\,\beta}{\kappa\,\ln(T/\beta)}\right) \exp\!\left(-\frac{\alpha\,T}{\kappa\,\ln(T/\beta)}\right) + \frac{\sigma^2}{2\,L} \exp\!\left(\frac{2\,\beta}{\kappa\,\ln(T/\beta)}\right) \frac{4\,\kappa^2}{e^2\,\alpha^2} \frac{\ln^2\!\left(\frac{T}{\beta}\right)}{T}. \tag{76}$$

Making the dependence on the constants explicit

$$\implies \mathbb{E}[f^* - f(\theta_{T+1})]$$

$$\leq \mathbb{E}[f^* - f(\theta_1)] \exp\!\left(\frac{2\,\beta}{\kappa\,\ln(T/\beta)}\right) \exp\!\left(-\frac{\alpha\,T}{\kappa\,\ln(T/\beta)}\right) + \frac{\sigma^2}{2\,L} \exp\!\left(\frac{2\,\beta}{\kappa\,\ln(T/\beta)}\right) \frac{4\,\kappa^2}{e^2\,\alpha^2} \frac{\ln^2\!\left(\frac{T}{\beta}\right)}{T} \tag{77}$$

Since $\epsilon < 1$

$$= \mathbb{E}[f^* - f(\theta_1)] \exp\left(\frac{\mu\,\beta}{L\,\ln(T/\beta)}\right) \exp\left(-\frac{\mu\,\epsilon\,\alpha\,T}{2\,L\,\ln(T/\beta)}\right) + \exp\left(\frac{\mu\,\beta}{L\,\ln(T/\beta)}\right) \frac{32\,L\,\sigma^2\,\ln^2\left(\frac{T}{\beta}\right)}{e^2\,\alpha^2\,\mu^2\,\epsilon^2\,T}. \tag{78}$$

□

Corollary 6. In the bandit setting, for a given $\epsilon \in (0,1)$, using Update 2 with exponentially decreasing step-sizes $\eta_t = \eta_0\,\alpha^t$ where $\eta_0 = \frac{5}{2}$ and $\alpha = \left(\frac{\beta}{T}\right)^{\frac{1}{T}}$, $\beta \geq 1$ results in the following convergence:
If $\mathbb{E}[(\pi^* - \pi_{\theta_t})^\top r] \geq \epsilon$ for all $t \in [1,T]$, then,

$$\mathbb{E}[(\pi^* - \pi_{\theta_{T+1}})^\top r] \leq \mathbb{E}[(\pi^* - \pi_{\theta_1})^\top r]\,C_1 \exp\left(-\frac{\alpha\,\epsilon\,T}{\kappa\,\ln(T/\beta)}\right) + \frac{C_1\,C_2\,\ln^2\left(\frac{T}{\beta}\right)}{\epsilon^2\,T} \tag{79}$$

where $\mu := \left[\mathbb{E}\left[\min_{t\in[1,T]}[\pi_{\theta_t}(a^*)]\right]^{-2}\right]^{-1} > 0$, $\kappa := \frac{5}{\mu}$, $C_1 := \exp\left(\frac{2\,\beta}{\kappa\,\ln(T/\beta)}\right)$ and $C_2 := \frac{4\kappa^2}{5\,e^2\alpha^2}$.
Otherwise, $\min_{t\in[1,T]} \mathbb{E}[f^* - f(\theta_t)] \leq \epsilon$.

Proof. We can extend Theorem 3 to the bandit setting since:

- by Lemma 24, f is $\frac{5}{2}$-smooth
- by Lemma 31, f is non-uniform Łojsiewciz with $\xi = 0$ and $C(\theta) = \pi_\theta(a^*)$
- since T is finite and the updates are bounded, $\mu := \left[\mathbb{E}\left[\min_{t\in[1,T]}[\pi_{\theta_t}(a^*)]^{-2}\right]\right]^{-1} > 0$
- by Lemma 35, the stochastic gradient is unbiased and $\sigma^2 \leq 2$

□

Corollary 7. In the tabular MDP setting, for a given $\epsilon \in (0,1)$, using Update 2 with exponentially decreasing step-sizes $\eta_t = \eta_0\,\alpha^t$ where $\eta_0 = \frac{(1-\gamma)^3}{8}$ and $\alpha = \left(\frac{\beta}{T}\right)^{\frac{1}{T}}$, $\beta \geq 1$ results in the following convergence:
If $\mathbb{E}[V^*(\rho) - V^{\pi_{\theta_t}}(\rho)] \geq \epsilon$ for all $t \in [1,T]$, then,

$$\mathbb{E}[V^*(\rho) - V^{\pi_{\theta_{T+1}}}(\rho)] \leq \mathbb{E}[V^*(\rho) - V^{\pi_{\theta_1}}(\rho)]\,C_1 \exp\left(-\frac{\alpha\,\epsilon\,T}{\kappa\,\ln(T/\beta)}\right) + \frac{C_1\,C_2\,\ln^2\left(\frac{T}{\beta}\right)}{\epsilon^2\,T} \tag{80}$$

where $\mu := \left[\mathbb{E}\left[\min_{t\in[1,T]}\left(\frac{\min_s \pi_\theta(a^*(s)|s)}{\sqrt{S}\left\|d_\rho^{\pi^*}/d_\rho^{\pi_\theta}\right\|_\infty}\right)\right]^{-2}\right]^{-1} > 0$, $\kappa := \frac{16}{\mu\,(1-\gamma)^3}$, $C_1 := \exp\left(\frac{2\,\beta}{\kappa\,\ln(T/\beta)}\right)$ and $C_2 := \frac{A\,\kappa^2}{4\,(1-\gamma)\,e^2\,\alpha^2}$. Otherwise, $\min_{t\in[1,T]} \mathbb{E}[V^*(\rho) - V^{\pi_{\theta_t}}(\rho)] \leq \epsilon$.

Proof. We can extend Theorem 3 to the tabular MDP setting since:

- by Lemma 27, f is $\frac{8}{(1-\gamma)^3}$-smooth
- by Lemma 32, f is non-uniform Łojsiewciz with $\xi = 0$ and $C(\theta) = \frac{\min_s \pi_\theta(a^*(s)|s)}{\sqrt{S}\left\|\frac{d_\rho^{\pi^*}}{d_\rho^{\pi_\theta}}\right\|_\infty}$
- since T is finite and the updates are bounded, $\mu := \left[\mathbb{E}\left[\min_{t\in[1,T]}\left(\frac{\min_s \pi_\theta(a^*(s)|s)}{\sqrt{S}\left\|d_\rho^{\pi^*}/d_\rho^{\pi_\theta}\right\|_\infty}\right)^{-2}\right]\right]^{-1} > 0$
- by Lemma 36, the stochastic gradient is unbiased and $\sigma^2 \leq \frac{2\,S}{(1-\gamma)^4}$

□

C.2 Proof of Theorem 5

Theorem 5. For a given $\epsilon \in (0,1)$, assuming f is (i) L_1 non-uniform smooth, (ii) satisfies the non-uniform Łojasiewciz condition with $\xi = 0$, (iii) $\mu := \left[\mathbb{E}\left[\inf_{t\geq 1}[C(\theta_t)]^{-2}\right]\right]^{-1} > 0$, using Update 2 with unbiased stochastic gradients that are (a) bounded, i.e. $\|\nabla \widetilde{f}(\theta)\| \leq B$ and satisfy the strong growth condition with ϱ and (b) exponentially decreasing step-sizes $\eta_t = \eta_0\,\alpha^t$ where $\eta_0 < \frac{1}{L_1^2 B}$ and $\alpha = \left(\frac{\beta}{T}\right)^{\frac{1}{T}}$, $\beta \geq 1$, results in the following convergence:
If $\mathbb{E}[f^* - f(\theta_t)] > \epsilon$ for all $t \in [1, T]$, then,

$$\mathbb{E}[f^* - f(\theta_{T+1})] \leq \mathbb{E}[f^* - f(\theta_1)]\, C_1\, \exp\left(-\frac{\alpha\,\epsilon\,T}{\kappa\,\ln(T)}\right) + \frac{C_2\,\sum_{t=1}^{T_0-1}\,\mathbb{E}[f^* - f(\theta_t)]}{\epsilon^2\,T^2} \tag{6}$$

where $\kappa := \frac{2}{\mu\,\eta_0}$, $C_1 := \exp\left(\frac{2\,\beta}{\kappa\,\ln(T/\beta)}\right)$, $C_2 := \exp\left(\frac{2\,\beta}{\kappa\,\ln(T/\beta)}\right)\frac{16\,\varrho\,L\,\kappa^2}{e^2\,\alpha^2}\ln^2(T/\beta)$, $T_0 := T\,\max\left\{\frac{\ln(\varrho\,\eta_0)}{\ln(T/\beta)}, 0\right\}$. Otherwise $\min_{t\in[1,T]}\mathbb{E}[f^* - f(\theta_t)] \leq \epsilon$.

Proof. Assuming f is $L_1\,\|\nabla f(\theta)\|$ non-uniform smooth and the stochastic gradients are bounded, i.e. $\|\nabla \widetilde{f}(\theta)\| \leq B$, by Lemma 5, using Update 2 with $\eta_t \in \left(0, \frac{1}{L_1 B}\right)$

$$|f(\theta_{t+1}) - f(\theta_t) - \langle \nabla f(\theta_t), \theta_{t+1} - \theta_t\rangle| \leq \frac{1}{2}\,\frac{L_1\,\|\nabla f(\theta_t)\|}{1 - L_1\,B\,\eta_t}\,\|\theta_{t+1} - \theta_t\|_2^2 \tag{81}$$

Then following the initial proof of Theorem 3 we obtain

$$\underbrace{\mathbb{E}[f^* - f(\theta_{t+1})]}_{:=\delta(\theta_{t+1})} \leq \underbrace{\mathbb{E}[f^* - f(\theta_t)]}_{:=\delta(\theta_t)} - \eta_t\,\|\nabla f(\theta_t)\|_2^2 + \frac{{\eta_t}^2}{2}\,\frac{L_1\,\|\nabla f(\theta_t)\|}{1 - L_1\,B\,\eta_t}\,\mathbb{E}\left[\left\|\nabla \widetilde{f}(\theta_t)\right\|_2^2\right] \tag{82}$$

Assuming f satisfies the strong growth condition, $\mathbb{E}\left\|\nabla \widetilde{f}(\theta_t)\right\|_2^2 \leq \varrho\,\|\nabla f(\theta_t)\|$

$$\leq \delta(\theta_t) - \eta_t\,\|\nabla f(\theta_t)\|_2^2 + \frac{\varrho\,{\eta_t}^2}{2}\,\frac{L_1}{1 - L_1\,B\,\eta_t}\,\|\nabla f(\theta_t)\|_2^2 \tag{83}$$

Since for all $t \geq 1$, $\eta_t \leq \eta_0$, $\frac{1}{1-L_1\,B\,\eta_t} \leq \frac{1}{1-L_1\,B\,\eta_0}$

$$\leq \delta(\theta_t) - \eta_t\,\|\nabla f(\theta_t)\|_2^2 + \frac{\varrho\,{\eta_t}^2}{2}\,\frac{L_1}{1 - L_1\,B\,\eta_0}\,\|\nabla f(\theta_t)\|_2^2 \tag{84}$$

Picking η_0 such that $\frac{L_1}{1-L_1\,B\eta_0} < 1 \implies \eta_0 < \frac{1}{L_1^2 B}$

$$\implies \delta(\theta_{t+1}) \leq \delta(\theta_t) - \eta_t\,\|\nabla f(\theta_t)\|_2^2 + \frac{\varrho\,{\eta_t}^2}{2}\,\|\nabla f(\theta_t)\|_2^2. \tag{85}$$

Since η_t is decreasing, we will now consider the following phases:

Phase 1 : When η_t is "large", i.e. $\eta_t > \frac{1}{\varrho}$

Phase 2 : When η_t is "small", i.e. $\eta_t \leq \frac{1}{\varrho}$.

For $\eta_t \leq \frac{1}{\varrho}$ we require that

$$\eta_0\,\left(\frac{\beta}{T}\right)^{\frac{t}{T}} \leq \frac{1}{\varrho} \implies t \geq T_0 := T\,\frac{\ln(\varrho\,\eta_0)}{\ln\left(\frac{T}{\beta}\right)}. \tag{86}$$

Hence, when $t \geq T_0$, the step-size is small enough to be in Phase 2. Let us first analyze Phase 1.

Phase 1: In Phase 1 we have $\eta_t > \frac{1}{\varrho}$. Starting with Equation (85),

$$\delta(\theta_{t+1}) \leq \delta(\theta_t) - \eta_t \left\|\nabla f(\theta_t)\right\|_2^2 + \frac{\varrho\, {\eta_t}^2}{2} \left\|\nabla f(\theta_t)\right\|_2^2. \tag{87}$$

To simplify $\left\|\nabla f(\theta_t)\right\|_2^2$, since f is L-smooth for any θ and θ'

$$f(\theta') \geq f(\theta) + \langle \nabla f(\theta), \theta' - \theta \rangle - \frac{L}{2} \left\|\theta' - \theta\right\|_2^2 \tag{88}$$

Setting $\theta' = \theta + \frac{1}{L} \nabla f(\theta)$

$$\geq f(\theta) + \frac{1}{L} \left\|\nabla f(\theta)\right\|_2^2 \tag{89}$$

$$\implies \left\|\nabla f(\theta)\right\|_2^2 \leq 2L \left[f(\theta') - f(\theta)\right] \leq 2L \left[f^* - f(\theta)\right] \tag{90}$$

$$\implies \frac{\varrho}{2} \|\nabla f(\theta_t)\|_2^2 \leq \varrho\, L \left[f^* - f(\theta)\right] = \varrho\, L\, \delta(\theta_t). \tag{91}$$

Hence,

$$\delta(\theta_{t+1}) \leq \delta(\theta_t) - \frac{\eta_t}{2} \left\|\nabla f(\theta_t)\right\|_2^2 + L\, \varrho\, {\eta_t}^2\, \delta(\theta_t) \tag{92}$$

Since f satisfies the non-uniform Łojsaiewciz condition with $\xi = 0$

$$\leq \delta(\theta_t) - \frac{\eta_t \left[C(\theta_t)\right]^2}{2} \left[\delta(\theta_t)\right]^2 + L\, \varrho\, {\eta_t}^2\, \delta(\theta_t) \tag{93}$$

Since $m := \inf_{t \geq 1} [C(\theta_t)]^2 > 0$

$$\leq \delta(\theta_t) - \frac{\eta_t\, m}{2} \left[\delta(\theta_t)\right]^2 + {\eta_t}^2 \underbrace{L\, \varrho\, \delta(\theta_t)}_{:=\Gamma_t} \tag{94}$$

Taking expectation with respect to all previous iterations $t \geq 1$ on both sides

$$\implies \mathbb{E}[\delta(\theta_{t+1})] \leq \mathbb{E}[\delta(\theta_t)] - \frac{\eta_t}{2} \mathbb{E}[m\, \delta(\theta_t)^2] + {\eta_t}^2 \underbrace{L\, \varrho\, \mathbb{E}[\delta(\theta_t)]}_{:=\Gamma_t} \tag{95}$$

Using Cauchy-Schwarz to lower-bound $\mathbb{E}[m\, [\delta(\theta_t)]^2]$

$$\leq \mathbb{E}[\delta(\theta_t)] - \frac{\eta_t}{2\, \mathbb{E}[m^{-1}]} \mathbb{E}[\delta(\theta_t)] + {\eta_t}^2\, \Gamma_t \tag{96}$$

Define $\mu := \frac{1}{\mathbb{E}[m^{-1}]}$

$$\leq \mathbb{E}[\delta(\theta_t)] - \frac{\eta_t\, \mu}{2} \mathbb{E}[\delta(\theta_t)] + {\eta_t}^2\, \Gamma_t \tag{97}$$

If $\mathbb{E}[\delta(\theta_t)] \leq \epsilon$ for some $t \in \{1, \ldots, T\}$, then we are done. Else for all $t \in \{1, \ldots, T\}$, $\mathbb{E}[\delta(\theta_t)] > \epsilon$. Hence,

$$\mathbb{E}[\delta(\theta_{t+1})] \leq \mathbb{E}[\delta(\theta_t)] \left(1 - \frac{\eta_t\, \mu\, \epsilon}{2}\right) + {\eta_t}^2\, \Gamma_t \tag{98}$$

$$= \mathbb{E}[\delta(\theta_t)] \left(1 - \frac{\eta_0\, \mu\, \epsilon}{2}\, \alpha^t\right) + \eta_0^2\, \alpha^{2t}\, \Gamma_t \tag{99}$$

Define $\frac{1}{\kappa} := \frac{\eta_0\, \mu\, \epsilon}{2}$

$$\implies \mathbb{E}[\delta(\theta_{t+1})] = \mathbb{E}[\delta(\theta_t)] \left(1 - \frac{1}{\kappa}\, \alpha^t\right) + \eta_0^2\, \alpha^{2t}\, \Gamma_t. \tag{100}$$

Recall we are in Phase 1 when $t < T_0$. Using Equation (100) and recursing from $t = 1$ to $T_0 - 1$

$$\mathbb{E}[\delta(\theta_{T_0})] \leq \mathbb{E}[\delta(\theta_1)] \prod_{t=1}^{T_0-1} \left(1 - \frac{1}{\kappa}\,\alpha^t\right) + \eta_0^2 \sum_{t=1}^{T_0-1} \alpha^{2t}\,\Gamma_t \prod_{i=t+1}^{T_0-1} \left(1 - \frac{1}{\kappa}\,\alpha^i\right) \tag{101}$$

Using $1 - x \leq \exp(-x)$ and by summing up the geometric series

$$\implies \mathbb{E}[\delta(\theta_{T_0})] \leq \mathbb{E}[\delta(\theta_1)] \exp\left(-\frac{1}{\kappa}\,\frac{\alpha - \alpha^{T_0}}{1-\alpha}\right) + \eta_0^2 \sum_{t=1}^{T_0-1} \alpha^{2t}\,\Gamma_t \exp\left(-\frac{1}{\kappa}\,\frac{\alpha^{t+1} - \alpha^{T_0}}{1-\alpha}\right). \tag{102}$$

Let us now bound the second term on the RHS

$$\eta_0^2 \sum_{t=1}^{T_0-1} \alpha^{2t}\,\Gamma_t \exp\left(-\frac{1}{\kappa}\,\frac{\alpha^{t+1} - \alpha^{T_0}}{1-\alpha}\right) = \eta_0^2 \exp\left(\frac{\alpha^{T_0}}{\kappa\,(1-\alpha)}\right) \sum_{t=1}^{T_0-1} \alpha^{2t}\,\Gamma_t \exp\left(-\frac{\alpha^{t+1}}{\kappa\,(1-\alpha)}\right) \tag{103}$$

By Lemma 8, $\exp(-x) \leq \left(\frac{2}{ex}\right)^2$

$$\leq \eta_0^2 \exp\left(\frac{\alpha^{T_0}}{\kappa\,(1-\alpha)}\right) \sum_{t=1}^{T_0-1} \alpha^{2t}\,\Gamma_t \left(\frac{2\,(1-\alpha)\,\kappa}{e\,\alpha^{t+1}}\right)^2 \tag{104}$$

$$= \exp\left(\frac{\alpha^{T_0}}{\kappa\,(1-\alpha)}\right) \frac{4\,\eta_0^2\,(1-\alpha)^2\,\kappa^2}{e^2\,\alpha^2} \sum_{t=1}^{T_0-1} \Gamma_t \tag{105}$$

Since $1 - x \leq \ln\left(\frac{1}{x}\right)$ and using it to bound $(1-\alpha)^2$ where $\alpha = \left(\frac{\beta}{T}\right)^{\frac{1}{T}}$

$$\leq \exp\left(\frac{\alpha^{T_0}}{\kappa\,(1-\alpha)}\right) \frac{4\,\eta_0^2\,\kappa^2}{e^2\,\alpha^2} \frac{\ln^2\left(\frac{T}{\beta}\right) \sum_{t=1}^{T_0-1} \Gamma_t}{T^2}. \tag{106}$$

Putting everything together,

$$\implies \mathbb{E}[\delta(\theta_{T_0})] \leq \mathbb{E}[\delta(\theta_1)] \exp\left(-\frac{1}{\kappa}\,\frac{\alpha - \alpha^{T_0}}{1-\alpha}\right) + \exp\left(\frac{\alpha^{T_0}}{\kappa\,(1-\alpha)}\right) \frac{4\,\eta_0^2\,\kappa^2}{e^2\,\alpha^2} \frac{\ln^2\left(\frac{T}{\beta}\right) \sum_{t=1}^{T_0-1} \Gamma_t}{T^2}. \tag{107}$$

Now let us consider Phase 2.

Phase 2: We are in Phase 2 when $\eta_t \leq \frac{1}{\varrho}$. Starting with Equation (85),

$$\delta(\theta_{t+1}) \leq \delta(\theta_t) - \eta_t\,\|\nabla f(\theta_t)\|_2^2 + \frac{\varrho\,{\eta_t}^2}{2}\,\|\nabla f(\theta_t)\|_2^2 \tag{108}$$

Since f satisfies the non-uniform Łojsaiewciz condition with $\xi = 0$

$$\leq \delta(\theta_t) - \frac{\eta_t\,[C(\theta_t)]^2}{2}\,[\delta(\theta_t)]^2 \tag{109}$$

Since $m := \inf_{t \geq 1} [C(\theta_t)]^2 > 0$

$$\leq \delta(\theta_t) - \frac{\eta_t\,m}{2}\,[\delta(\theta_t)]^2 \tag{110}$$

Taking expectation with respect to all previous iterations $t \geq 1$ on both sides

$$\mathbb{E}[\delta(\theta_{t+1})] \leq \mathbb{E}[\delta(\theta_t)] - \frac{\eta_t}{2}\,\mathbb{E}[m[\delta(\theta_t)]^2] \tag{111}$$

Using Cauchy-Schwarz to lower-bound $\mathbb{E}[m\,[\delta(\theta_t)]^2]$

$$\mathbb{E}[\delta(\theta_{t+1})] \leq \mathbb{E}[\delta(\theta_t)] - \frac{\eta_t}{2\,\mathbb{E}[m^{-1}]}\,\mathbb{E}[\delta(\theta_t)] \tag{112}$$

Define $\mu := \frac{1}{\mathbb{E}[m^{-1}]}$

$$\mathbb{E}[\delta(\theta_{t+1})] \leq \mathbb{E}[\delta(\theta_t)] - \frac{\eta_t\,\mu}{2}\,\mathbb{E}[\delta(\theta_t)] \tag{113}$$

If $\mathbb{E}[\delta(\theta_t)] \leq \epsilon$ for some $t \in \{1, \ldots, T\}$, then we are done. Else for all $t \in \{1, \ldots, T\}$, $\mathbb{E}[\delta(\theta_t)] > \epsilon$. Hence,

$$\mathbb{E}[\delta(\theta_{t+1})] \leq \mathbb{E}[\delta(\theta_t)]\left(1 - \frac{\eta_t\,\mu\,\epsilon}{2}\right). \tag{114}$$

Recall we are in Phase 2 when $t \geq T_0$. Using Equation (114) and recursing from $t = T_0$ to T

$$\mathbb{E}[\delta(\theta_{T+1})] \leq \prod_{t=T_0}^{T}\left(1 - \frac{\eta_t\,\mu\,\epsilon}{2}\right)\mathbb{E}[\delta(\theta_{T_0})] \tag{115}$$

Using $1 - x \leq \exp(-x)$

$$\mathbb{E}[\delta(\theta_{T+1})] \leq \exp\left(-\frac{\mu\,\epsilon}{2}\sum_{t=T_0}^{T}\eta_t\right)\mathbb{E}[\delta(\theta_{T_0})] \tag{116}$$

Since $\eta_t = \eta_0\,\alpha^t$ and summing up the geometric series

$$\implies \mathbb{E}[\delta(\theta_{T+1})] \leq \exp\left(-\frac{\eta_0\,\mu\,\epsilon}{2}\,\frac{\alpha^{T_0} - \alpha^{T+1}}{1-\alpha}\right)\mathbb{E}[\delta(\theta_{T_0})] \tag{117}$$

Since $\frac{1}{\kappa} = \frac{\eta_0\,\mu\,\epsilon}{2}$

$$= \exp\left(-\frac{1}{\kappa}\,\frac{\alpha^{T_0} - \alpha^{T+1}}{1-\alpha}\right)\mathbb{E}[\delta(\theta_{T_0})]. \tag{118}$$

$$\tag{119}$$

Combining the results of Phase 1 (Equation (107)) and Phase 2 (Equation (118))

$$\mathbb{E}[\delta(\theta_{T+1})] \leq \exp\left(-\frac{1}{\kappa}\,\frac{\alpha^{T_0} - \alpha^{T+1}}{1-\alpha}\right)$$
$$\left[\mathbb{E}[\delta(\theta_1)]\,\exp\left(-\frac{1}{\kappa}\,\frac{\alpha - \alpha^{T_0}}{1-\alpha}\right) + \exp\left(\frac{\alpha^{T_0}}{\kappa\,(1-\alpha)}\right)\frac{4\,\eta_0^2\,\kappa^2}{e^2\,\alpha^2}\,\frac{\ln^2\left(\frac{T}{\beta}\right)\sum_{t=1}^{T_0-1}\Gamma_t}{T^2}\right] \tag{120}$$

$$= \mathbb{E}[\delta(\theta_1)]\,\exp\left(-\frac{1}{\kappa}\,\frac{\alpha - \alpha^{T+1}}{1-\alpha}\right) + \exp\left(\frac{\alpha^{T+1}}{\kappa\,(1-\alpha)}\right)\frac{4\,\eta_0^2\,\kappa^2}{e^2\,\alpha^2}\,\frac{\ln^2\left(\frac{T}{\beta}\right)\sum_{t=1}^{T_0-1}\Gamma_t}{T^2} \tag{121}$$

$$= \mathbb{E}[\delta(\theta_1)]\,\exp\left(\frac{\alpha^{T+1}}{\kappa\,(1-\alpha)}\right)\exp\left(-\frac{\alpha}{\kappa\,(1-\alpha)}\right)$$
$$+ \exp\left(\frac{\alpha^{T+1}}{\kappa\,(1-\alpha)}\right)\frac{4\,\eta_0^2\,\kappa^2}{e^2\,\alpha^2}\,\frac{\ln^2\left(\frac{T}{\beta}\right)\sum_{t=1}^{T_0-1}\Gamma_t}{T^2} \tag{122}$$

By Lemma 6, $\frac{\alpha^{T+1}}{(1-\alpha)} \leq \frac{2\beta}{\ln(T/\beta)}$

$$\leq \mathbb{E}[\delta(\theta_1)] \exp\left(\frac{2\beta}{\kappa \ln(T/\beta)}\right) \exp\left(-\frac{\alpha}{\kappa(1-\alpha)}\right) + \exp\left(\frac{2\beta}{\kappa \ln(T/\beta)}\right) \frac{4\eta_0^2 \kappa^2}{e^2 \alpha^2} \frac{\ln^2\left(\frac{T}{\beta}\right) \sum_{t=1}^{T_0-1} \Gamma_t}{T^2} \tag{123}$$

Since $1 - x \leq \ln\left(\frac{1}{x}\right)$, $\frac{\alpha}{1-\alpha} \geq \frac{\alpha T}{\ln(T/\beta)}$

$$\leq \mathbb{E}[\delta(\theta_1)] \underbrace{\exp\left(\frac{2\beta}{\kappa \ln(T/\beta)}\right)}_{:=C_1} \exp\left(-\frac{\alpha T}{\kappa \ln(T/\beta)}\right) + \underbrace{\exp\left(\frac{2\beta}{\kappa \ln(T/\beta)}\right) \frac{4\eta_0^2 \kappa^2}{e^2 \alpha^2} \ln^2\left(\frac{T}{\beta}\right)}_{:=C_2} \frac{\sum_{t=1}^{T_0-1} \Gamma_t}{T^2} \tag{124}$$

$$\implies \mathbb{E}[\delta(\theta_{T+1})] \leq C_1 \mathbb{E}[\delta(\theta_1)] \exp\left(-\frac{\alpha T}{\kappa \ln(T/\beta)}\right) + C_2 \frac{\sum_{t=1}^{T_0-1} \Gamma_t}{T^2}. \tag{125}$$

Making the dependence on the constants explicit

$$\implies \mathbb{E}[\delta(\theta_{T+1})] \leq \mathbb{E}[\delta(\theta_1)] \exp\left(\frac{\mu \epsilon \eta_0 \beta}{\ln(T/\beta)}\right) \exp\left(\frac{-\mu \epsilon \eta_0 \alpha T}{2 \ln(T/\beta)}\right) + \exp\left(\frac{\mu \epsilon \eta_0 \beta}{\ln(T/\beta)}\right) \frac{16 L \varrho \ln^2(T/\beta)}{e^2 \alpha^2 \mu^2 \epsilon^2} \frac{\sum_{t=1}^{T_0-1} \mathbb{E}[\delta(\theta_t)]}{T^2} \tag{126}$$

Since $\epsilon < 1$

$$\leq \mathbb{E}[\delta(\theta_1)] \exp\left(\frac{\mu \eta_0 \beta}{\ln(T/\beta)}\right) \exp\left(\frac{-\mu \epsilon \eta_0 \alpha T}{2 \ln(T/\beta)}\right) + \exp\left(\frac{\mu \eta_0 \beta}{\ln(T/\beta)}\right) \frac{16 L \varrho \ln^2(T/\beta)}{e^2 \alpha^2 \mu^2 \epsilon^2} \frac{\sum_{t=1}^{T_0-1} \mathbb{E}[\delta(\theta_t)]}{T^2} \tag{127}$$

□

Corollary 1. In the bandit setting, for a given $\epsilon \in (0,1)$, using Update 2 with exponentially decreasing step-sizes $\eta_t = \eta_0 \alpha^t$ where $\eta_0 \leq \frac{1}{18}$, $\alpha = \left(\frac{\beta}{T}\right)^{\frac{1}{T}}$, $\beta \geq 1$ results in the following convergence:
If $\mathbb{E}[(\pi^* - \pi_{\theta_t})^\top r] \geq \epsilon$ for all $t \in [1, T]$, then,

$$\mathbb{E}[(\pi^* - \pi_{\theta_{T+1}})^\top r] \leq \mathbb{E}[(\pi^* - \pi_{\theta_1})^\top r] C_1 \exp\left(-\frac{\alpha \epsilon T}{\kappa \ln(T/\beta)}\right) + \frac{C_2 \sum_{t=1}^{T_0-1} \mathbb{E}[(\pi^* - \pi_{\theta_t})^\top r]}{\epsilon^2 T^2} \tag{7}$$

where $\kappa := \frac{2}{\mu \eta_0}$, $C_1 := \exp\left(\frac{2\beta}{\kappa \ln(T/\beta)}\right)$, $C_2 := \exp\left(\frac{2\beta}{\kappa \ln(T/\beta)}\right) \frac{32 \varrho \kappa^2}{5 e^2 \alpha^2} \ln^2(T/\beta)$, $T_0 := T \max\left\{\frac{\ln(4 \varrho \eta_0)}{\ln(T/\beta)}, 0\right\}$, $\rho = \frac{8 A^{3/2}}{\Delta^2}$ and $\mu := \left[\mathbb{E}\left[\min_{t \in [1,T]} [\pi_{\theta_t}(a^*)]^{-2}\right]\right]^{-1} > 0$. Otherwise $\min_{t \in [1,T]} \mathbb{E}[(\pi^* - \pi_{\theta_t})^\top r] \leq \epsilon$.

Proof. We can extend Theorem 5 to the bandit setting since:

- by Lemma 24, f is $\frac{5}{2}$-smooth
- by Lemma 29, f is 3-non-uniform smooth
- by Lemma 31, f is non-uniform Łojsiewciz with $\xi = 0$ and $C(\theta) = \pi_\theta(a^*)$

- since T is finite and the updates are bounded, $\mu := \left[\mathbb{E}\left[\min_{t\in[1,T]} \pi_{\theta_t}(a^*)^{-2}\right]\right]^{-1} > 0$
- by Lemma 35, the stochastic gradient is unbiased
- by Lemma 7, the stochastic gradient satisfies the strong growth condition with $\varrho = \frac{8\,A^{3/2}}{\Delta^2}$ where $\Delta := \min_{a\neq a'} |r(a) - r(a')|$
- by Mei et al. (2023, Equation 52) $\|\frac{d\langle \pi_\theta, r\rangle}{d\theta}\| \leq \sqrt{2}$ and $\eta_0 := \frac{1}{18} < \frac{1}{L_1^2 B} = \frac{1}{9\sqrt{2}}$.

□

Corollary 8. Assuming $\min_{s\in\mathcal{S}} \rho(s) > 0$, in the tabular MDP setting, for a given $\epsilon \in (0,1)$, using Update 2 with exponentially decreasing step-sizes $\eta_t = \eta_0\,\alpha^t$ where $\eta_0 < \frac{1}{C^2 B}$ and $\alpha = \left(\frac{\beta}{T}\right)^{\frac{1}{T}}$, $\beta \geq 1$ results in the following convergence:
If $\mathbb{E}[V^{\pi^*}(\rho) - V^{\pi_{\theta_t}}(\rho)] \geq \epsilon$ for all $t \in [1,T]$, then,

$$\mathbb{E}[V^{\pi^*}(\rho) - V^{\pi_{\theta_{T+1}}}(\rho)] \leq \mathbb{E}[V^{\pi^*}(\rho) - V^{\pi_{\theta_1}}(\rho)]\, C_1\, \exp\left(-\frac{\alpha\,\epsilon\,T}{\kappa\,\ln(T/\beta)}\right) + \frac{C_2\,\sum_{t=1}^{T_0-1} \mathbb{E}[V^{\pi^*}(\rho) - V^{\pi_{\theta_t}}(\rho)]}{\epsilon^2\,T^2} \tag{128}$$

where $C := \left[3 + \frac{2\,C_\infty - (1-\gamma)}{(1-\gamma)\,\gamma}\right]\sqrt{S}$, $C_\infty := \max_\pi \left\|\frac{d_\rho^\pi}{\rho}\right\|_\infty \leq \frac{1}{\min_s \rho(s)} < \infty$, $B := \frac{\sqrt{2\,S}}{(1-\gamma)^4}$, $\kappa := \frac{2}{\mu\,\eta_0}$, $C_1 := \exp\left(\frac{2\,\beta}{\kappa\,\ln(T/\beta)}\right)$, $C_2 := \exp\left(\frac{2\,\beta}{\kappa\,\ln(T/\beta)}\right)\frac{128\,\varrho\,\kappa^2}{(1-\gamma)^3\,e^2\,\alpha^2}\,\ln^2(T/\beta)$, $T_0 := T\,\max\left\{\frac{\ln(\varrho\,\eta_0)}{\ln(T/\beta)}, 0\right\}$ and $\mu := \left[\mathbb{E}\left[\min_{t\in[1,T]}\left(\frac{\min_s \pi_\theta(a^*(s)|s)}{\sqrt{S}\left\|d_\rho^{\pi^*}/d_\rho^{\pi_\theta}\right\|_\infty}\right)^{-2}\right]\right]^{-1} > 0$. Otherwise, $\min_{t\in[1,T]} \mathbb{E}[V^{\pi^*}(\rho) - V^{\pi_{\theta_t}}(\rho)] \leq \epsilon$.

Proof. We can extend Theorem 5 to the tabular MDP setting since:

- by Lemma 27, f is $\frac{8}{(1-\gamma)^3}$-smooth
- by Lemma 30, f is C-non-uniform smooth where $C := \left[3 + \frac{2\,C_\infty - (1-\gamma)}{(1-\gamma)\gamma}\right]\sqrt{S}$ and $C_\infty := \max_\pi \left\|\frac{d_\rho^\pi}{\rho}\right\|_\infty \leq \frac{1}{\min_s \rho(s)} < \infty$,
- by Lemma 32, f is non-uniform Łojsiewciz with $\xi = 0$ and $C(\theta) = \frac{\min_s \pi_\theta(a^*(s)|s)}{\sqrt{S}\left\|d_\rho^{\pi^*}/d_\rho^{\pi_\theta}\right\|_\infty}$
- since T is finite and the update is bounded, $\mu := \left[\mathbb{E}\left[\min_{t\in[1,T]}\left(\frac{\min_s \pi_\theta(a^*(s)|s)}{\sqrt{S}\left\|d_\rho^{\pi^*}/d_\rho^{\pi_\theta}\right\|_\infty}\right)^{-2}\right]\right]^{-1} > 0$
- by Lemma 36, the stochastic gradient is unbiased
- by Theorem 6, the stochastic gradient satisfies the strong growth condition with $\varrho = \frac{4\,A^{3/2}\,S^{1/2}}{(1-\gamma)^4\,\Delta^2}$ where $\Delta := \min_s \min_{a\neq a'} |Q^{\pi_\theta}(s,a) - Q^{\pi_\theta}(s,a')|$
- by Equation (153), $\|\nabla \widetilde{f}(\theta_t)\| \leq B := \frac{\sqrt{2\,S}}{(1-\gamma)^2}$

□

Corollary 9. Assuming $\rho(s) = \frac{1}{S}$ for all $s \in \mathcal{S}$, in the tabular MDP setting, for a given $\epsilon \in (0,1)$, using Update 2 with exponentially decreasing step-sizes $\eta_t = \eta_0\,\alpha^t$ where $\eta_0 < \frac{1}{C^2 B}$ and $\alpha = \left(\frac{\beta}{T}\right)^{\frac{1}{T}}$, $\beta \geq 1$ results in the following convergence:
If $\mathbb{E}[V^{\pi^*}(\rho) - V^{\pi_{\theta_t}}(\rho)] \geq \epsilon$ for all $t \in [1,T]$, then,

$$\mathbb{E}[V^{\pi^*}(\rho) - V^{\pi_{\theta_{T+1}}}(\rho)] \leq \mathbb{E}[V^{\pi^*}(\rho) - V^{\pi_{\theta_1}}(\rho)]\, C_1\, \exp\left(-\frac{\alpha\,\epsilon\,T}{\kappa\,\ln(T/\beta)}\right) + \frac{C_2 \sum_{t=1}^{T_0-1} \mathbb{E}[V^{\pi^*}(\rho) - V^{\pi_{\theta_t}}(\rho)]}{\epsilon^2\,T^2} \tag{129}$$

where $C := \left[3 + \frac{2\,A^{-1} - (1-\gamma)}{(1-\gamma)\,\gamma}\right]\sqrt{S}$, $B := \frac{\sqrt{2\,S}}{(1-\gamma)^2}$, $\kappa := \frac{2}{\mu\,\eta_0}$, $C_1 := \exp\left(\frac{2\,\beta}{\kappa\,\ln(T/\beta)}\right)$, $C_2 := \exp\left(\frac{2\,,\beta}{\kappa\,\ln(T/\beta)}\right)\frac{128\,\varrho\,\kappa^2}{(1-\gamma)^3\,e^2\,\alpha^2}\,\ln^2(T/\beta)$, $T_0 := T\,\max\left\{\frac{\ln(\varrho\,\eta_0)}{\ln(T/\beta)}, 0\right\}$ and $\mu := \left[\mathbb{E}\left[\min_{t\in[1,T]}\left(\frac{\min_s \pi_\theta(a^*(s)|s)}{\sqrt{S}\left\|d_\rho^{\pi^*}/d_\rho^{\pi_\theta}\right\|_\infty}\right)^{-2}\right]\right]^{-1} > 0$. Otherwise, $\min_{t\in[1,T]} \mathbb{E}[V^{\pi^*}(\rho) - V^{\pi_{\theta_t}}(\rho)] \leq \epsilon$.

Proof. Follows from Corollary 8. □

C.3 Strong Growth Condition - Dependence of Reward Gap

We first show that the dependence of the reward gap Δ in the SGC constant ϱ cannot be removed.

Proposition 1. *The dependence of Δ in the strong growth condition in Lemma 7 is necessary.*

Proof. Consider a 2-arm bandit problem with deterministic rewards: $r_1 := r(1)$ and $r_2 := r(2)$. Assume that $\Delta := r_1 - r_2 > 0$, and hence arm 1 is the optimal arm. We will show that in SGC in Lemma 7, the dependence of Δ in the SGC constant ϱ is necessary. Let $\hat{r}(a) := \frac{\mathbb{1}\{a_t = a\}}{\pi_{\theta_t}(a)}\, r(a)$ for all $a \in \mathcal{A}$. The stochastic gradient estimate satisfies the following SGC:

$$\mathbb{E}_t\left[\left\|\frac{d[\langle \pi_{\theta_t}, \hat{r}_t\rangle]}{d\theta_t}\right\|_2^2\right] \leq \varrho\left\|\frac{d[\langle \pi_{\theta_t}, r\rangle]}{d\theta_t}\right\|. \tag{130}$$

Calculating the LHS

$$\frac{d\langle \pi_{\theta_t}, \hat{r}_t\rangle}{d\theta_t(a)} = \left[\mathbb{1}\{a_t = a\} - \pi_{\theta_t}(a)\right]\, r(a_t) \tag{131}$$

$$\implies \left\|\frac{d\langle \pi_{\theta_t}, \hat{r}_t\rangle}{d\theta_t}\right\|_2^2 = \sum_a \left[\left[\mathbb{1}\{a_t = a\} - \pi_{\theta_t}(a)\right]\, r(a_t)\right]^2 \tag{132}$$

Let $p := \pi_{\theta_t}(a_1)$ as the probability of pulling the optimal arm

$$= \left[\left[\mathbb{1}\{a_t = a_1\} - p\right]\, r(a_t)\right]^2 + \left[\left[\mathbb{1}\{a_t = a_2\} - (1-p)\right]\, r(a_t)\right]^2. \tag{133}$$

$$\begin{aligned}
\mathbb{E}_t\left[\left\|\frac{d\langle \pi_{\theta_t}, \hat{r}_t\rangle}{d\theta_t}\right\|_2^2\right] &= \mathbb{E}_t\left[\left\|\frac{d\langle \pi_{\theta_t}, \hat{r}_t\rangle}{d\theta_t}\right\|_2^2 |\; a_t = a_1\right]\, \Pr[a_t = a_1] \\
&\quad + \mathbb{E}_t\left[\left\|\frac{d\langle \pi_{\theta_t}, \hat{r}_t\rangle}{d\theta_t}\right\|_2^2 |\; a_t \neq a_1\right]\, \Pr[a_t \neq a_1] && (134)\\
&= \left((1-p)^2\, r_1^2 + (1-p)^2\, r_1^2\right)\, p + \left(p^2\, r_2^2 + p^2\, r_2^2\right)\, (1-p) && (135)\\
\implies \text{LHS} &= 2p\,(1-p)^2\, r_1^2 + 2(1-p)\, p^2\, r_2^2 = 2p\,(1-p)\left[(1-p)\, r_1^2 + p\, r_2^2\right]. && (136)
\end{aligned}$$

Calculating the RHS

$$\frac{d\langle \pi_{\theta_t}, r\rangle}{d\theta_t(a)} = \pi_{\theta_t}(a)\ [r_a - \langle \pi_{\theta_t}, r\rangle] \tag{137}$$

$$\implies \left\| \frac{d\langle \pi_{\theta_t}, r\rangle}{d\theta_t} \right\|_2^2 = \sum_a \pi_{\theta_t}(a)^2\ [r_a - \langle \pi_{\theta_t}, r\rangle]^2 \tag{138}$$

$$= p^2\ [r_1 - \langle \pi_{\theta_t}, r\rangle]^2 + (1-p)^2\ [r_2 - \langle \pi_{\theta_t}, r\rangle]^2 \tag{139}$$

Since $\langle \pi_{\theta_t}, r\rangle = p\, r_1 + (1-p)\, r_2$

$$= p^2\ [r_1 - [p\, r_1 + (1-p)\, r_2]]^2 + (1-p)^2\ [r_2 - [p\, r_1 + (1-p)\, r_2]]^2 \tag{140}$$

$$= p^2\,(1-p)^2\,\Delta^2 + (1-p)^2\,p^2\,\Delta^2 = 2\,p^2\,(1-p)^2\,\Delta^2 \tag{141}$$

$$\implies \text{RHS} = \left\| \frac{d\langle \pi_{\theta_t}, r\rangle}{d\theta_t} \right\| = \sqrt{2}\,p\,(1-p)\,\Delta. \tag{142}$$

Hence,

$$\text{LHS} = \frac{\sqrt{2}\ \left[(1-p)\, r_1^2 + p\, r_2^2\right]}{\Delta}\ \text{RHS} \implies \varrho = \frac{\sqrt{2}\ \left[(1-p)\, r_1^2 + p\, r_2^2\right]}{\Delta}.$$

For rewards $r_1 > r_2 > 0$, the numerator depends on the magnitude of the rewards, while the denominator depends on their gap. Since we have derived an equality, the dependence on $\frac{1}{\Delta}$ in ϱ is necessary. □

C.4 Strong Growth Condition - Tabular MDP Setting, IS Parallel Estimator

Following (Mei et al., 2021a, Definition 3), we first consider stochastic gradients using the on-policy parallel IS estimator.

Definition 1 (On-policy parallel IS estimator)**.** *In the tabular MDP setting, at iteration t, under each state $s \in \mathcal{S}$ sample one action $a_t(s) \sim \pi_{\theta_t}(\cdot|s)$. The IS state-action value estimator $\hat{Q}^{\pi_{\theta_t}}$ is constructed as $\hat{Q}^{\pi_{\theta_t}}(s,a) = \frac{\mathbb{1}\{a_t(s)=a\}}{\pi_{\theta_t}(a|s)}\, Q^{\pi_{\theta_t}}(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.*

Using this parallel IS parallel estimator, the following PG estimator constructed in Algorithm 1 satisfies the SGC.

Algorithm 1: Softmax PG, on-policy stochastic gradient

Input: Learning rate $\eta > 0$.
Output: Policy $\pi_{\theta_t} = \text{softmax}(\theta_t)$.
Initialize parameters $\theta_1(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$
while $t \geq 1$ **do**
  Sample $a_t(s) \sim \pi_{\theta_t}(\cdot|s)$ for all $s \in \mathcal{S}$
  $\hat{Q}^{\pi_{\theta_t}}(s,a) \leftarrow \frac{\mathbb{I}\{a_t(s)=a\}}{\pi_{\theta_t}(a|s)}\, Q^{\pi_{\theta_t}}(s,a)$
  $\hat{g}_t(s,\cdot) \leftarrow \frac{1}{1-\gamma}\, d_\rho^{\pi_{\theta_t}}(s) \left[\sum_a \frac{\partial \pi_{\theta_t}(a|s)}{\partial \theta_t(s,\cdot)}\, \hat{Q}^{\pi_{\theta_t}}(s,a)\right]$
  $\theta_{t+1} \leftarrow \theta_t + \eta\, \hat{g}_t$
end

Recall in the tabular MDP setting, the PG theorem (Sutton et al., 1999b) states

$$\frac{\partial V^{\pi_{\theta_t}}(\rho)}{\partial \theta} = \frac{1}{1-\gamma} \mathbb{E}_{s' \sim d_\rho^{\pi_\theta}} \left[\sum_{a' \in \mathcal{A}} \frac{\partial \pi_\theta(a'|s')}{\partial \theta}\, Q^{\pi_\theta}(s', a') \right]. \tag{143}$$

For tabular softmax policy for any $s' \neq s$ and any $a \in \mathcal{A}$, $\frac{\partial \pi_\theta(a|s')}{\partial \theta(s,\cdot)} = \mathbf{0}$. Hence,

$$\frac{V^{\pi_\theta}(\rho)}{\partial \theta(s,a)} = \frac{1}{1-\gamma} d_\rho^{\pi_\theta}(s)\, \pi_\theta(a|s)\, (Q^{\pi_\theta}(s,a) - \langle \pi_\theta(\cdot), Q^{\pi_\theta}(s,\cdot)\rangle). \tag{144}$$

In contrast, in Algorithm 1 the stochastic gradient is

$$\hat{g}(s,a) = \frac{1}{1-\gamma} d_\rho^{\pi_\theta}(s)\, \pi_\theta(a|s) \left(\hat{Q}^{\pi_\theta}(s,a) - \left\langle \pi_\theta(\cdot), \hat{Q}^{\pi_\theta}(s,\cdot)\right\rangle\right). \tag{145}$$

Theorem 6. In the tabular MDP setting, using Update 2 with the on-policy parallel IS estimator, we have for any θ,

$$\mathbb{E}\left[\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \frac{d_\rho^{\pi_\theta}(s)^2}{(1-\gamma)^2} \pi_\theta(a\,|\,s)^2 \left(\hat{Q}^{\pi_\theta}(s,a) - \left\langle \pi_\theta(\cdot\,|\,s), \hat{Q}^{\pi_\theta}(s,\cdot)\right\rangle\right)^2\right] \leq \frac{4\, A^{3/2}\, S^{1/2}}{(1-\gamma)^4\, \Delta^2} \left\| \frac{\partial V^{\pi_\theta}(\rho)}{\partial \theta}\right\|_2 \tag{146}$$

where $\Delta := \min_s \min_{a\neq a'} |Q^{\pi_\theta}(s,a) - Q^{\pi_\theta}(s,a')|$.

Proof. In the tabular MDP setting we have

$$\left\|\nabla \widetilde{f}(\theta)\right\|_2^2 = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \frac{d_\rho^{\pi_\theta}(s)^2}{(1-\gamma)^2}\, \pi_\theta(a\,|\,s)^2 \left(\hat{Q}^{\pi_\theta}(s,a) - \left\langle \pi_\theta(\cdot|s), \hat{Q}^{\pi_\theta}(s,\cdot)\right\rangle\right)^2. \tag{147}$$

Let us first bound the RHS. For a fixed $s \in \mathcal{S}$.

$$\sum_{a\in\mathcal{A}} \pi_\theta(a\,|\,s)^2 \left(\hat{Q}^{\pi_\theta}(s,a) - \left\langle \pi_\theta(\cdot\,|\,s), \hat{Q}^{\pi_\theta}(s,\cdot)\right\rangle\right)^2 \tag{148}$$

$$= \sum_{a\in\mathcal{A}} \pi_\theta(a\,|\,s)^2 \left[\frac{\mathbb{1}\{a(s)=a\}}{\pi_\theta(a\,|\,s)^2} Q^{\pi_\theta}(s,a)^2 - 2\,\frac{\mathbb{1}\{a(s)=a\}}{\pi_{\theta_t}(a\,|\,s)} Q^{\pi_\theta}(s,a) \left\langle \pi_\theta(\cdot|s), \hat{Q}^{\pi_\theta}(s,\cdot)\right\rangle \right.$$
$$\left. + \left(\left\langle \pi_\theta(\cdot|s), \hat{Q}^{\pi_\theta}(s,\cdot)\right\rangle\right)^2\right] \tag{149}$$

$$= Q^{\pi_\theta}(s,a(s))^2 - 2\,\pi_\theta(a(s)|s)\, Q^{\pi_\theta}(s,a(s))^2 + Q^{\pi_\theta}(s,a(s))^2 \sum_{a\in\mathcal{A}} \pi_\theta(a|s)^2 \tag{150}$$

$$= (1-\pi_\theta(a(s)|s))^2\, Q^{\pi_\theta}(s,a(s))^2 + Q^{\pi_\theta}(s,a(s))^2 \sum_{a\neq a(s)} \pi_\theta(a|s)^2 \tag{151}$$

$$= \frac{1}{(1-\gamma)^2}(1-\pi_\theta(a(s)|s))^2 + \sum_{a\neq a(s)} \pi_\theta(a|s)^2 \qquad (Q^{\pi_\theta}(s,a) \leq \tfrac{1}{1-\gamma})$$

$$\leq \frac{1}{(1-\gamma)^2}\left((1-\pi_\theta(a(s)|s))^2 + \left(\sum_{a\neq a(s)} \pi_\theta(a|s)\right)^2\right) \qquad (\|x\|_2 \leq \|x\|_1)$$

$$= \frac{2}{(1-\gamma)^2}(1-\pi_\theta(a(s)|s))^2 \tag{152}$$

Accounting for every $s \in \mathcal{S}$,

$$\implies \left\|\nabla \widetilde{f}(\theta)\right\|_2^2 \leq \frac{2}{(1-\gamma)^4} \sum_{s\in\mathcal{S}} \left[d_\rho^{\pi_\theta}(s)\right]^2 (1-\pi_\theta(a(s)|s))^2 \tag{153}$$

In Algorithm 1, the only source of stochasticy is from sampling $a(s) \sim \pi_\theta(\cdot|s)$ for each $s \in \mathcal{S}$. Therefore

$$\mathbb{E}\left[\left\|\nabla \widetilde{f}(\theta)\right\|_2\right] = \mathbb{E}_{a_1\sim\pi_\theta(\cdot|s_1)}\left[\mathbb{E}_{a_2\sim\pi_\theta(\cdot|s_2)}\left[\ldots \mathbb{E}_{a_S\sim\pi_\theta(\cdot|s_S)}\left[\left\|\nabla \widetilde{f}(\theta)\right\|_2^2\right]\right]\right]. \tag{154}$$

Let us first consider $\mathbb{E}_{a_1\sim\pi_\theta(\cdot|s_1)}\left[\left\|\nabla\widetilde{f}(\theta)\right\|_2^2\right]$. By Equation (153)

$$\mathbb{E}_{a_1\sim\pi_\theta(\cdot|s_1)}\left[\left\|\nabla\widetilde{f}(\theta_t)\right\|_2^2\right] \tag{155}$$

$$\leq \frac{2}{(1-\gamma)^4}\sum_{a_1\in\mathcal{A}}\pi_\theta(a_1|s_1)\left[\left[d_\rho^{\pi_\theta}(s_1)\right]^2(1-\pi_\theta(a_1|s_1))^2+\sum_{s\neq s_1}\left[d_\rho^{\pi_\theta}(s)\right]^2(1-\pi_\theta(a(s)|s))^2\right] \tag{156}$$

$$= \frac{2}{(1-\gamma)^4} \tag{157}$$

$$\left[\underbrace{\left[d_\rho^{\pi_\theta}(s_1)\right]^2\sum_{a_1\in\mathcal{A}}\pi_\theta(a_1|s_1)\left(1-\pi_\theta(a_1|s_1)\right)^2}_{:=C_{s_1}}+\underbrace{\sum_{a_1\in\mathcal{A}}\pi_\theta(a_1|s_1)}_{=1}\sum_{s\neq s_1}\left[d_\rho^{\pi_\theta}(s)\right]^2(1-\pi_\theta(a(s)|s))^2\right] \tag{158}$$

$$= \frac{2}{(1-\gamma)^4}\left[C_{s_1}+\sum_{s\neq s_1}\left[d_\rho^{\pi_\theta}(s)\right]^2(1-\pi_\theta(a(s)|s))^2\right]. \tag{159}$$

Next let us consider $\mathbb{E}_{a_2\sim\pi_\theta(\cdot|s_2)}\mathbb{E}_{a_1\sim\pi_\theta(\cdot|s_1)}\left[\left\|\nabla\widetilde{f}(\theta)\right\|_2^2\right]$ and by the same argument

$$\mathbb{E}_{a_2\sim\pi_\theta(\cdot|s_2)}\mathbb{E}_{a_1\sim\pi_\theta(\cdot|s_1)}\left[\left\|\nabla\widetilde{f}(\theta)\right\|_2^2\right] \leq \frac{2}{(1-\gamma)^4}\left[C_{s_1}+C_{s_2}+\sum_{\substack{s\neq s_1\\ s\neq s_2}}\left[d_\rho^{\pi_\theta}(s)\right]^2(1-\pi_\theta(a(s)|s))^2\right] \tag{160}$$

Continuing in the same way for the remaining $s\in\mathcal{S}$ we have

$$\mathbb{E}\left[\left\|\nabla\widetilde{f}(\theta)\right\|_2^2\right] \leq \frac{2}{(1-\gamma)^4}\sum_{s\in\mathcal{S}}C_s \tag{161}$$

$$= \frac{2}{(1-\gamma)^4}\sum_{s\in\mathcal{S}}\left[d_\rho^{\pi_\theta}(s)\right]^2\sum_{a\in\mathcal{A}}\pi_\theta(a|s)\left(1-\pi_\theta(a|s)\right)^2 \tag{162}$$

Denote $k(s):=\arg\max_{a\in\mathcal{A}}\pi_\theta(a|s)$ as the action with the largest probability at state s

$$= \frac{2}{(1-\gamma)^4}\sum_{s\in\mathcal{S}}\left[d_\rho^{\pi_\theta}(s)\right]^2\left[\pi_\theta(k(s)|s)\left(1-\pi_\theta(k(s)|s)\right)^2+\sum_{a\neq k(s)}\pi_\theta(a|s)\left(1-\pi_\theta(a|s)\right)^2\right] \tag{163}$$

$$\leq \frac{2}{(1-\gamma)^4}\sum_{s\in\mathcal{S}}\left[d_\rho^{\pi_\theta}(s)\right]^2\left[\left(1-\pi_\theta(k(s)|s)\right)+\sum_{a\neq k_t(s)}\pi_\theta(a(s|s)\right] \tag{164}$$

$$= \frac{4}{(1-\gamma)^4}\sum_{s\in\mathcal{S}}\left[d_\rho^{\pi_\theta}(s)\right]^2\left(1-\pi_\theta(k(s)|s)\right) \qquad (\pi_\theta(a|s)\in[0,1])$$

Since $d_\rho^{\pi_\theta}(s)\leq 1$ for all $s\in\mathcal{S}$

$$\leq \frac{4}{(1-\gamma)^4}\sum_{s\in\mathcal{S}}d_\rho^{\pi_\theta}(s)\left(1-\pi_\theta(k(s)|s)\right) \tag{165}$$

$$\implies \mathbb{E}\left[\left\|\nabla \widetilde{f}(\theta)\right\|_2^2\right] \leq \frac{4}{(1-\gamma)^4} \sum_{s\in\mathcal{S}} d_\rho^{\pi_\theta}(s)\,(1-\pi_{\theta_t}(k(s)|s)). \tag{166}$$

Now we lower bound $\left\|\frac{V^{\pi_\theta}(\rho)}{\partial\theta}\right\|_2^2$

$$\left\|\frac{V^{\pi_\theta}(\rho)}{\partial\theta}\right\|_2^2 \tag{167}$$

$$= \frac{1}{(1-\gamma)^2}\left(\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_\rho^{\pi_\theta}(s)^2\,\pi_\theta(a|s)^2\,A^{\pi_\theta}(s,a)^2\right) \tag{168}$$

Multiplying and dividing by $\sum_{(s',a')} A^{\pi_\theta}(s,a)^2$

$$= \frac{1}{(1-\gamma)^2}\left(\sum_{s'\in\mathcal{S}}\sum_{a'\in\mathcal{A}} A^{\pi_\theta}(s',a')^2 \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left(\underbrace{d_\rho^{\pi_\theta}(s)\,\pi_\theta(a|s)}_{:=w(s,a)}\right)^2 \underbrace{\frac{(A^{\pi_\theta}(s,a))^2}{\sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} A^{\pi_\theta}(s',a')^2}}_{:=p(s,a)}\right) \tag{169}$$

Since $p(s,a)\geq 0$ and $\sum_{s,a} p(s,a) = 1$, using Jensen's inequality,
$\sum_{s,a} w(s,a)^2\,p(s,a) \geq (\sum_{s,a} w(s,a)\,p(s,a))^2$

$$\geq \frac{1}{(1-\gamma)^2}\left(\sum_{s'\in\mathcal{S}}\sum_{a'\in\mathcal{A}} A^{\pi_\theta}(s',a')^2\left[\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_\rho^{\pi_\theta}(s)\,\pi_{\theta_t}(a|s)\,\frac{A^{\pi_\theta}(s,a)^2}{\sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} A^{\pi_\theta}(s',a')^2}\right]^2\right) \tag{170}$$

$$= \frac{1}{(1-\gamma)^2}\left(\frac{1}{\sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} A^{\pi_\theta}(s',a')^2}\left[\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_\rho^{\pi_\theta}(s)\,\pi_\theta(a|s)\,A^{\pi_\theta}(s,a)^2\right]^2\right) \tag{171}$$

Since $A^{\pi_\theta}(s,a) \leq \frac{1}{1-\gamma}$, $\frac{1}{\sum_{(s',a')} A^{\pi_\theta}(s',a')^2} \geq \frac{(1-\gamma)^2}{S\,A}$

$$\implies \left\|\frac{\partial V^{\pi_{\theta_t}}(\rho)}{\partial\theta}\right\|_2^2 \geq \frac{1}{S\,A}\left[\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_\rho^{\pi_\theta}(s)\,\pi_\theta(a|s)A^{\pi_\theta}(s,a)^2\right]^2 \tag{172}$$

$$\implies \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_\rho^{\pi_\theta}(s)\,\pi_\theta(a|s)\,A^{\pi_\theta}(s,a)^2 \leq \sqrt{S\,A}\left\|\frac{\partial V^{\pi_\theta}(\rho)}{\partial\theta}\right\|_2. \tag{173}$$

To connect Equation (166) and Equation (173) for a fixed $s\in\mathcal{S}$

$$\sum_{a\in\mathcal{A}} \pi_\theta(a|s)\,A^{\pi_\theta}(s,a)^2$$

$$= \sum_{a\in\mathcal{A}} \pi_\theta(a|s)\,(Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s))^2 \tag{174}$$

$$= \sum_{a\in\mathcal{A}} \pi_\theta(a|s)\left[Q^{\pi_\theta}(s,a)^2 - 2\,V^{\pi_\theta}(s)\,Q^{\pi_\theta}(s,a) + V^{\pi_\theta}(s)^2\right] \tag{175}$$

$$= \sum_{a\in\mathcal{A}} \pi_\theta(a|s)\,Q^{\pi_\theta}(s,a)^2 - 2\,V^{\pi_\theta}(s)\underbrace{\sum_{a\in\mathcal{A}} \pi_\theta(a|s)\,Q^{\pi_\theta}(s,a)}_{=V^{\pi_\theta}(s)} + V^{\pi_\theta}(s)^2\underbrace{\sum_{a\in\mathcal{A}} \pi_\theta(a|s)}_{=1} \tag{176}$$

$$= \sum_{a\in\mathcal{A}} \pi_\theta(a|s)Q^{\pi_\theta}(s,a)^2 - \left[\sum_{a\in\mathcal{A}} \pi_\theta(a|s)\,Q^{\pi_\theta}(s,a)\right]^2 \tag{177}$$

Recall $k(s) := \arg\max_{a \in \mathcal{A}} \pi_\theta(a|s)$, by Lemma 9,

$$\geq \pi_\theta(k(s)|s) \sum_{a \neq k(s)} \pi_\theta(k(s)|s) \left(Q^{\pi_\theta}(s, k(s)) - Q^{\pi_\theta}(s, a)\right)^2 \tag{178}$$

Let $\Delta_s := \min_{a \neq a'} |Q^{\pi_\theta}(s, a) - Q^{\pi_\theta}(s, a')|$ and since $\pi_\theta(k(s)|s) \geq \frac{1}{A}$,

$$\geq (1 - \pi_\theta(k(s)|s) \frac{\Delta_s^2}{A} \tag{179}$$

Let $\Delta := \min_s \Delta_s$

$$\geq (1 - \pi_\theta(k(s)|s) \frac{\Delta^2}{A} \tag{180}$$

$$\implies (1 - \pi_\theta(k(s)|s) \leq \frac{A}{\Delta^2} \sum_{a \in \mathcal{A}} \pi_\theta(a|s) A^{\pi_\theta}(s, a)^2 \tag{181}$$

Putting everything together, by Equation (166)

$$\mathbb{E}\left[\left\|\nabla \widetilde{f}(\theta)\right\|_2^2\right] \leq \frac{4}{(1-\gamma)^4} \sum_s d_\rho^{\pi_\theta}(s) \left(1 - \pi_\theta(k(s) \,|\, s)\right) \tag{182}$$

By Equation (181)

$$\leq \frac{4\,A}{(1-\gamma)^4\,\Delta^2} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} d_\rho^{\pi_\theta}(s)\, \pi_\theta(a \,|\, s)\, A^{\pi_\theta}(s, a)^2 \tag{183}$$

By Equation (173)

$$\leq \frac{4\,A^{3/2}\,S^{1/2}}{(1-\gamma)^4\,\Delta^2} \left\|\frac{\partial V^{\pi_\theta}(\rho)}{\partial \theta}\right\|_2. \tag{184}$$

□

C.5 Additional Lemmas

Lemma 5. *Assuming that f is L_1-non-uniform smooth and the stochastic gradient is bounded, i.e. $\|\nabla \widetilde{f}(\theta_t)\| \leq B$, using Update 2 with $\eta_t \in (0, \frac{1}{L_1 B})$ we have,*

$$|f(\theta_{t+1}) - f(\theta_t) - \langle \nabla f(\theta_t), \theta_{t+1} - \theta_t \rangle| \leq \frac{1}{2} \frac{L_1 \|\nabla f(\theta_t)\|}{1 - L_1 B \eta_t} \|\theta_{t+1} - \theta_t\|. \tag{185}$$

Proof. Following (Mei et al., 2023, Lemma 4.2), denote $\theta_\zeta := \theta_t + \zeta(\theta_{t+1} - \theta_t)$ for some $\zeta \in [0, 1]$. According to Taylor's theorem, we have

$$|f(\theta_{t+1}) - f(\theta_t) - \langle \nabla f(\theta_t), \theta_{t+1} - \theta_t \rangle| = \frac{1}{2} \left|(\theta_{t+1} - \theta_t)^\top \nabla^2 f(\theta_\zeta)\, (\theta_{t+1} - \theta_t)\right| \tag{186}$$

Assuming f is L_1 non-uniform smooth

$$\leq \frac{L_1 \|\nabla f(\theta_\zeta)\|}{2} \|\theta_{t+1} - \theta_t\|_2^2. \tag{187}$$

Denote $\theta_{\zeta_1} := \theta_t + \zeta_1(\theta_\zeta - \theta_t)$ for some $\zeta_1 \in [0, 1]$. By the fundamental theorem of calculus,

$$\|\nabla f(\theta_\zeta) - \nabla f(\theta_t)\| = \left\| \int_0^1 \langle \nabla^2 f(\theta_{\zeta_1}), \theta_\zeta - \theta_t \rangle d\zeta_1 \right\| \tag{188}$$

Using Cauchy-Schwarz

$$\leq \int_0^1 \left\|\nabla^2 f(\theta_{\zeta_1})\right\| \|\theta_\zeta - \theta_t\| d\zeta_1 \tag{189}$$

Since f is L_1-non-uniform smooth

$$\leq \int_0^1 L_1 \|\nabla f(\theta_{\zeta_1})\| \|\theta_\zeta - \theta_t\| d\zeta_1 \tag{190}$$

$$= \int_0^1 L_1 \|\nabla f(\theta_{\zeta_1})\| \zeta \|\theta_{t+1} - \theta_t\| d\zeta_1 \qquad (\theta_\zeta := \theta_t + \zeta(\theta_{t+1} - \theta_t))$$

Since $\zeta \in [0,1]$ and using Update 2, $\theta_{t+1} = \theta_t + \eta_t \nabla \widetilde{f}(\theta_t)$

$$\implies \|\nabla f(\theta_\zeta) - \nabla f(\theta_t)\| \leq L_1 \eta_t \|\nabla \widetilde{f}(\theta_t)\| \int_0^1 \|\nabla f(\theta_{\zeta_1})\| d\zeta_1 \tag{191}$$

Therefore, we have

$$\|\nabla f(\theta_\zeta)\| = \|\nabla f(\theta) + \nabla f(\theta_\zeta) - \nabla f(\theta)\| \tag{192}$$

Using triangle inequality

$$\leq \|\nabla f(\theta_t)\| + \|\nabla f(\theta_\zeta) - \nabla f(\theta_t)\| \tag{193}$$

By Equation (191)

$$\implies \|\nabla f(\theta_\zeta)\| \leq \|\nabla f(\theta_t)\| + L_1 \eta_t \|\nabla \widetilde{f}(\theta_t)\| \int_0^1 \|\nabla f(\theta_{\zeta_1})\| d\zeta_1 \tag{194}$$

Denote $\theta_{\zeta_1} := \theta_t + \zeta_2(\theta_{\zeta_1} - \theta_t)$ with $\theta_{\zeta_2} \in [0,1]$. Using similar calculations when deriving Equation (191),

$$\|\nabla f(\theta_{\zeta_1})\| \leq \|\nabla f(\theta_t)\| + L_1 \eta_t \|\nabla \widetilde{f}(\theta_t)\| \int_0^1 \|\nabla f(\theta_\zeta)\| d\zeta_2 \tag{195}$$

Putting Equation (194) and Equation (194) together,

$$\|\nabla f(\theta_\zeta)\| \leq \left(1 + L_1 \eta_t \|\nabla \widetilde{f}(\theta_t)\|\right) \|\nabla f(\theta_t)\| + \left(L_1 \eta_t \|\nabla \widetilde{f}(\theta_t)\|\right)^2 \int_0^1 \int_0^1 \|\nabla f(\theta_{\zeta_2})\| d\zeta_2 \, d\zeta_1 \tag{196}$$

Using Equation (196) and continuing in the same way for ζ_i as $i \to \infty$

$$\|\nabla f(\theta_\zeta)\| \leq \underbrace{\sum_{i=0}^{\infty} \left(L_1 \eta_t \|\nabla \widetilde{f}(\theta_t)\|\right)^i}_{\heartsuit} \|\nabla f(\theta_t)\|. \tag{197}$$

To ensure that $\heartsuit$ is finite, we require that $L_1 \eta_t \|\nabla \widetilde{f}(\theta_t)\| < 1$. Assuming $\|\nabla \widetilde{f}(\theta_t)\| \leq B$ for all t

$$L_1 \eta_t \|\nabla f(\theta_t)\| \leq L_1 B \eta_t < 1 \implies \eta_t < \frac{1}{L_1 B} \tag{198}$$

For $\eta_t \in \left(0, \frac{1}{L_1 B}\right)$, summing the geometric series

$$\|\nabla f(\theta_\zeta)\| \leq \frac{\|\nabla f(\theta_t)\|}{1 - L_1 B \eta_t}. \tag{199}$$

Putting Equation (187) and Equation (199) together, for $\eta_t \in \left(0, \frac{1}{L_1, B}\right)$ we have

$$|f(\theta_{t+1}) - f(\theta_t) - \langle \nabla f(\theta_t), \theta_{t+1} - \theta_t \rangle| \leq \frac{1}{2} \frac{L_1 \|\nabla f(\theta_t)\|}{1 - L_1 B \eta_t} \|\theta_{t+1} - \theta_t\|. \tag{200}$$

□

Lemma 6 (Lemma 5 in (Vaswani et al., 2022))**.**

$$\frac{\alpha^{T+1}}{1-\alpha} \leq \frac{2\beta}{\ln(T/\beta)} \tag{201}$$

Lemma 7 (Lemma 4.3 in (Mei et al., 2023)). *Using Update 2, we have for all $t \geq 1$,*

$$\mathbb{E}_t\left[\left\|\frac{d\langle \pi_{\theta_t}, \hat{r}_t\rangle}{d\theta_t}\right\|_2^2\right] \leq \frac{8\,A^{3/2}}{\Delta^2}\left\|\frac{d\langle \pi_{\theta_t}, \hat{r}_t\rangle}{d\theta_t}\right\|_2 \tag{202}$$

where $\Delta := \min_{a\neq a'}|r(a) - r(a')|$.

Lemma 8 (Lemma 17 in (Vaswani et al., 2022)). *For all $x, \gamma > 0$,*

$$\exp(-x) \leq \left(\frac{\gamma}{ex}\right)^{\gamma} \tag{203}$$

Lemma 9. *Let $p, b \in \mathbb{R}^K$ such that $p_1 \geq p_2 \geq \cdots \geq p_K \geq 0$, $\sum_{i=1}^K p_i = 1$ and $b_i \geq 0$ for all i then*

$$\sum_{i=1}^K p_i\, b_i^2 - \left[\sum_{i=1}^K p_i\, b_i\right]^2 \geq p_1 \sum_{j=2}^K p_j\,[b_i - b_j]^2 \tag{204}$$

Proof.

$$\sum_{i=1}^K p_i\, b_i^2 - \left[\sum_{i=1}^K p_i\, b_i\right]^2 = \sum_{i=1}^K p_i\, b_i^2 - \sum_{i=1}^K p_i^2\, b_i^2 - 2\sum_{i=1}^{K-1} p_i\, r_i \sum_{j=i+1}^K p_j\, r_j \tag{205}$$

$$= \sum_{i=1}^K (p_i\, b_i^2 - p_i^2\, b_i^2) - 2\sum_{i=1}^{K-1} p_i\, r_i \sum_{j=i+1}^K p_j\, r_j \tag{206}$$

$$= \sum_{i=1}^K p_i\, b_i^2\,(1 - p_i) - 2\sum_{i=1}^{K-1} p_i\, r_i \sum_{j=i+1}^K p_j\, r_j \tag{207}$$

$$= \sum_{i=1}^K \underbrace{p_i}_{x_i}\,\underbrace{b_i^2}_{y_i} \sum_{i=1, j\neq i}^K \underbrace{p_j}_{x_j} - 2\sum_{i=1}^{K-1} p_i\, r_i \sum_{j=i+1}^K p_j\, r_j \qquad (p_i = 1 - \textstyle\sum_{j\neq 1} p_j)$$

For any x_i, y_i, $\sum_{i=1}^K x_i\, y_i \sum_{j=1, j\neq i}^K x_j = \sum_{i=1}^{K-1} x_i \sum_{j=i+1}^K x_j\,[y_i + y_j]$

$$= \sum_{i=1}^{K-1} p_i \sum_{j=i+1}^K p_j\,[b_i^2 + b_j^2] - 2\sum_{i=1}^{K-1} p_i\, b_i \sum_{j=i+1}^K p_j\, b_j \tag{208}$$

$$= \sum_{i=1}^{K-1} p_i \sum_{j=i+1}^K p_j\left[b_i^2 - 2b_i\, b_j + b_j^2\right] \tag{209}$$

$$= \sum_{i=1}^{K-1} p_i \sum_{j=i+1}^K p_j\,[b_i - b_j]^2 \tag{210}$$

Discarding extra terms since $p_2 \geq \cdots \geq p_{K-1} \geq 0$,

$$\geq p_1 \sum_{j=2}^K p_j [b_i - b_j]^2. \tag{211}$$

□

Lemma 10. *In the bandit setting,*

$$\left\| \frac{d\langle \pi_\theta, \hat{r} \rangle}{d\theta} \right\| \leq \sqrt{2}. \tag{212}$$

Proof. Follows from Mei et al. (2023, Equation 55). □

Lemma 11. *In the tabular MDP setting,*

$$\left\| \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \frac{d_\rho^{\pi_\theta}(s)^2}{(1-\gamma)^2} \pi_\theta(a|s)^2 \left(\hat{Q}^{\pi_\theta}(s,a) - \left\langle \pi_\theta(\cdot|s), \hat{Q}^{\pi_\theta}(s,\cdot) \right\rangle \right)^2 \right\| \leq \frac{\sqrt{2\,S}}{(1-\gamma)^2}. \tag{213}$$

Proof. Follows from Equation (153). □

D Policy Gradient with Entropy Regularization

We will next consider adding entropy regularization to the objective in the exact and stochastic settings. Entropy regularization RL, also known as maximum entropy RL, uses entropy regularization to promote action diversity and prevent premature convergence to a deterministic policy (Williams, 1992; Haarnoja et al., 2018). While it is widely believed to help with exploration, the addition of entropy regularization results in a smoother optimization landscape, enabling PG methods to escape flat regions within the optimization landscape (Ahmed et al., 2019). For example in the bandits setting, flat regions occur when a policy commits to an arm. Mei et al. (2020) showed entropy regularization helps escaping these regions when starting from a "bad" initialization, i.e. the initial policy selects an sub-optimal arm with high probability.

In the exact setting, where the full gradient can be computed, Mei et al. (2020) showed softmax PG with entropy regularization obtains a fast $\mathcal{O}(\log(1/\epsilon))$ rate to a biased ϵ-optimal policy. The resulting optimal policy is biased since the presence of entropy prevents convergence to a deterministic policy. Additionally, in the same setting, Cen et al. (2022) showed NPG with entropy regularization achieves the same $\mathcal{O}(\log(1/\epsilon))$ convergence rate to a biased ϵ-optimal policy. To ensure that the resulting optimal policy is unbiased, the strength of the entropy regularization term must be decayed or removed. Mei et al. (2020) introduced a two-stage approach to obtain the optimal policy when using softmax PG with entropy regularization. In the first stage, entropy regularization is used to obtain fast convergence close to the optimal policy. In the second stage, the regularizer is removed to guarantee convergence to the optimal policy. Unfortunately, the final convergence rate is $\mathcal{O}(1/\epsilon)$ which matches the same rate as softmax PG. Additionally, in order to transition from the first to the second stage, the reward gap is needed making the resulting algorithm impractical.

In the stochastic setting, where the value function must be approximated, Ding et al. introduced a two-stage approach for stochastic softmax PG with entropy regularization. Instead of modifying the strength of the entropy regularizer across stages, the batch size is modified. The resulting algorithm requires $\mathcal{O}(1/\epsilon)$ iterations at the second stage and $\tilde{\mathcal{O}}(1/\epsilon^2)$ samples to converge to an biased ϵ-optimal policy. The method allows for global convergence with arbitrary initiation. However, the strength of the entropy regularizer is not decayed, preventing convergence to the optimal policy. Additionally, the biased optimal policy to set the algorithm hyper-parameters making the resulting algorithm redundant. Moreover, in the stochastic setting with access to a generative model, using NPG with entropy regularization, Cen et al. (2022) achieved a linear rate of convergence to a biased optimal policy with a $\tilde{\mathcal{O}}(1/\epsilon^2)$ sample complexity.

In the following sections, we will present a multi-stage algorithm that iteratively reduces the strength of the entropy regularization term. This method obtains convergence to the optimal policy while eliminating the reliance on unknown quantities compared to prior work. In Appendix D.1 we first state how the objective's functional property changes when entropy regularization is added. In Appendix D.2 we present the multi-stage algorithm in the exact setting and the algorithm

achieves an $\mathcal{O}(1/\epsilon^p)$ rate. Here, p relies on the estimation of the lower bound of the non-uniform Łojsiewciz condition of the entropy regularized objective. Next in Appendix D.3, we extend the same multi-stage algorithm in the stochastic setting with exponentially decreasing step-sizes to obtain an also $\mathcal{O}(1/\epsilon^{2p+1})$ rate to the optimal policy. Finally, in Appendix D.3.1 we compare the proposed our multi-stage algorithm to prior PG methods without entropy regularization and show that the multi-stage algorithm helps escape flat regions within the optimization landscape.

D.1 Problem Setup

Following Section 2, for a policy π, the *entropy regularized action-value function* is defined as $\tilde{Q}^\pi_\tau(s,a) := \mathbb{E}[\sum_{t=0}^\infty \gamma^t (r(s,a) - \tau \log \pi)]$ and the *entropy regularized value function* is defined as $\tilde{V}^\pi_\tau(s) := \mathbb{E}_{a\sim\pi}[\tilde{Q}^\pi_\tau(s,a)](s)$. The *entropy regularized advantage function* is defined as $\tilde{A}^\pi_\tau(s,a) := \tilde{Q}^\pi_\tau(s,a) - \tau \log \pi(a|s) - \tilde{V}^\pi_\tau(s)$.

Additionally, let $f^\tau(\theta) := f(\theta) + \tau\,\Lambda(\pi_\theta)$ denote the entropy regularized objective, where $\Lambda(\pi_\theta)$ is the "discounted entropy" for a policy π_θ and $\tau \geq 0$ is the "temperature" or strength of the entropy regularization. For a fixed τ, f^τ is L^τ-uniform smooth and note that the smoothness now depends on τ. Furthermore, f^τ satisfies a non-uniform Łojasiewciz condition with $C_\tau(\theta)$ and $\xi = 1/2$. Compared to f, whose non-uniform Łojasiewciz degree is $\xi = 0$ (refer to Table 1), the increase to $\xi = 1/2$ allows for faster convergence. Table 3 summarizes the entropy regularizer, uniform smoothness and non-uniform Łojsaiewciz properties for the bandit and general MDP settings with entropy regularization. Finally, we will denote the maximum value of the regularized objective function as $f^{*_\tau} := f^\tau(\theta^*_\tau)$, where $\theta^*_\tau := \arg\max_\theta f^\tau(\theta)$.

Setting	$\Lambda(\pi_\theta)$	$[\nabla f^\tau(\theta)]_{s,a}$	L^τ	$C_\tau(\theta)$
Bandits	$-\langle \pi_\theta, \log \pi_\theta \rangle$	$\pi_\theta(a)\,[r(a) - \langle \pi_\theta,\, r - \tau \log \pi_\theta \rangle]$	$5/2 + 5\,\tau\,(1 + \log A)$	$\sqrt{2\tau}\,\min_a \pi_\theta(a)$
MDP	$\mathbb{H}(\pi_\theta)$	$\frac{d^{\pi_\theta}(s)\,\pi_\theta(a\|s)\,\tilde{A}^{\pi_\theta}(s,a)}{1-\gamma}$	$\frac{8+\tau\,(4+8\log A)}{(1-\gamma)^3}$	$\frac{\sqrt{\tau}\,\min_s \sqrt{\rho(s)}\;\min_{s,a}\pi_\theta(a\|s)}{S\left\|\frac{d_\rho^{\pi^{*_\tau}}}{d_\rho^{\pi_\theta}}\right\|_\infty^{1/2}}$

Table 3: Entropy regularizer, uniform smoothness and non-uniform Łojasiewciz condition with $\xi = 1/2$ for bandits and general tabular MDPs setting with entropy regularization. Here, $\mathbb{H}(\pi_\theta) := \mathbb{E}[\sum_{t=0}^\infty -\gamma^t \log \pi_\theta(a_t|s_t)]$.

With the above properties of f^τ, we next present how to principally decay τ for softmax PG with entropy regularization to obtain convergence to the optimal policy.

D.2 Exact Setting

We first consider the exact setting as a test bed to analyze how to decay τ to obtain convergence to the optimal policy. Recall that for a constant $\tau > 0$, softmax PG with entropy regularization is unable to converge to the optimal policy since the regularizer prevents the final policy from becoming deterministic. Softmax PG with entropy regularization has the following update:

Update 3. *(Softmax PG with Entropy Regularization, True Gradient)* $\theta_{t+1} = \theta_t + \eta_t \nabla f^\tau(\theta_t)$.

Refer to Table 3 for the entropy regularized policy gradient $\nabla f^\tau(\theta)$ in both the bandits and the general MDP cases. In this setting, Mei et al. (2020) prove that softmax PG with entropy regularization converges to a biased optimal policy at an $\mathcal{O}(\log 1/\epsilon)$ rate when using a fixed step-size of $\eta_t = \eta = \frac{1}{L^\tau}$. The optimal policy is biased since $\tau > 0$ is fixed. In order for entropy regularized objective to converge to the globally optimal policy, $\tau \to 0$ is required. In the bandits setting, Mei et al. (2020) proposed a two-stage approach to decay τ to obtain global convergence. A fixed $\tau > 0$ is used in the first stage but is then set to be 0 in the second stage. However, the resulting algorithm requires knowledge of the reward gap $\Delta := \max_{a^* \neq a} r(a^*) - r(a)$ in order to transition from the first stage to the second stage, rendering the method to be impractical. Additionally, Mei et al. (2020) proposed

an additional approach by allowing τ be a function of t and slowly decreasing τ_t over time. This approach also obtain convergence to the global optimal policy. However, it required $\tau_t \propto \Delta$ and knowledge of the reward gap were again needed. Moreover, the final convergence rate to the optimal policy could not be established since it could not be proved that $\inf_{t\geq 1} C_\tau(\theta_t) > 0$.

For example, in the bandits setting (refer to Table 3) $C_\tau(\theta_t) := \sqrt{2\tau}\min_a \pi_{\theta_t}(a)$. In order for $\pi_{\theta_t} \to \pi^*$, we must have $\min_a \pi_{\theta_t}(a) \to 0$. However, in order to guarantee convergence when $\tau > 0$, we also require $\inf_{t\geq 0} \min_a \pi_{\theta_t}(a) > 0$. We conjecture that the non-uniform Łojasiewciz condition bound is loose which results in a pessimistic bound involving $\min_a \pi_\theta(a)$. We will make the benign assumption that f^τ satisfies the following non-uniform Łojasiewciz condition with $\xi = 1/2$ such that $\mu := \inf_{t\geq 0}[C_\tau(\theta_t)]^2 = \tau^p B_1$ for constants $p \geq 1$ and $B_1 > 0$.

Assumption 1. *f^τ satisfies the non-uniform Łojasiewciz condition for some $C_\tau(\theta)$ and $\xi = \frac{1}{2}$ such that $\mu := \inf_{t\geq 1}[C_\tau(\theta_t)]^2 = \tau^p B_1$ for constants $p \geq 1$ and $B_1 > 0$.*

Here we will assume the next worst dependence, which is having a polynomial dependence of τ for $\mu = \tau^p B_1$. Recall that f has a non-uniform Łojasiewciz condition with degree $\xi = 0$ and in the bandit setting $C(\theta) = \pi_\theta(a^*)$. We conjecture that as $\tau \to 0$, we switch from the non-uniform Łojasiewciz condition with degree $\xi = 1/2$ to degree $\xi = 0$. We leave the investigate of how these two conditions interpolate as future work.

Under Assumption 1, we propose a multi-stage algorithm (Algorithm 2) to decay τ that can obtain ϵ-convergence to the globally optimal policy without knowledge of the reward gap or any other problem-dependent parameters. Algorithm 2 consists of multiple stages, where the temperature is decreased in each stage. Specifically, in stage i uses τ_i for T_i iterations and is halved i.e. $\tau_{i+1} = \frac{\tau_i}{2}$ in the following stage. To prove the method achieves global convergence, we first make the following assumptions to relate the entropy regularization objective f^τ to the unregularized objective f:

Assumption 2. *f^τ is L^τ-smooth and $L^\tau \leq L^{\max}$, where $L^{\max} = \max_{\tau\in[0,1]} L^\tau$ is a constant. Furthermore, $L^\tau \geq L^{\min}$, where $L^{\min} = \min_{\tau\in[0,1]} L^\tau > 0$ is a constant.*

Assumption 3. *$f^* - f(\theta^*_\tau) \leq \tau B_2$, for a constant $B_2 > 0$.*

Assumption 4. *For a constant $B_3 > 0$, $f(\theta^*_\tau) - f(\theta) \leq f^{*\tau} - f^\tau(\theta) + \tau B_3$.*

Assumption 5. *For $\tau_2 < \tau_1$ and a constant $B_4 > 0$, $f^{*\tau_2} - f^{\tau_2}(\theta) \leq f^{*\tau_1} - f^{\tau_1}(\theta) + \tau_1 B_4$.*

The Assumptions 2 to 5 hold for both the bandits and tabular MPD setting and are proved in Appendix E.2 and Appendix E.3 respectively.

The following theorem (proved in Appendix E) shows that Algorithm 2 converges to the unbiased optimal policy at an $\mathcal{O}(1/\epsilon^p)$ rate.

Theorem 7. Assuming f^τ and f satisfy Assumptions 1 to 5, for a given $\epsilon \in (0, 1)$, Algorithm 2 achieves ϵ-suboptimality to the globally optimal after $T_{\text{total}} = \frac{4\, L^{\max}\, C_1^p}{\epsilon^p\, B_1} \log\left(2\ (1 + B_4)\right)$ iterations, where $C_1 = \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right) + B_2 + B_3$.

The resulting $\mathcal{O}(1/\epsilon^p)$ rate depends on the constant p in Assumption 1. In the best case, when $p = 1$, we recover an $\mathcal{O}(1/\epsilon)$ convergence rate. Otherwise, if p is large, we obtain a slower rate similar to the pessimistic analysis using $C_\tau(\theta) \propto \min_a \pi_\theta(a|s)$. Compared to Mei et al. (2020), when using entropy regularization, our method is able to obtain ϵ-convergence without requiring the knowledge of the reward gap.

We compare Algorithm 2 (`PG-E-MS`) assuming $p = 1$ and $B_1 = 0.01$ to softmax PG (`PG`) with a fixed step-size of $\eta_t = \frac{1}{L} = \frac{2}{5}$ and softmax PG with entropy regularization (`PG-E`) with fixed $\tau = 0.1$ and $\eta_t = \eta = \frac{1}{L^\tau} = \frac{2}{5+10\,\tau(1+\log A)}$ in the bandits setting with $A = 10$. For `PG-E-MS`, p and B_1 were selected by using grid-search on separate set of bandit instances. We test the algorithms on bandit settings of varying difficulty based on their minimum reward gap $\underline{\Delta} := \min_{a^*\neq a} r(a^*) - r(a)$. The easy, medium and hard environments correspond to $\underline{\Delta} = 0.2, 0.1, 0.05$ respectively. The figure plots the average and 95% confidence interval of 50 random mean reward vectors.

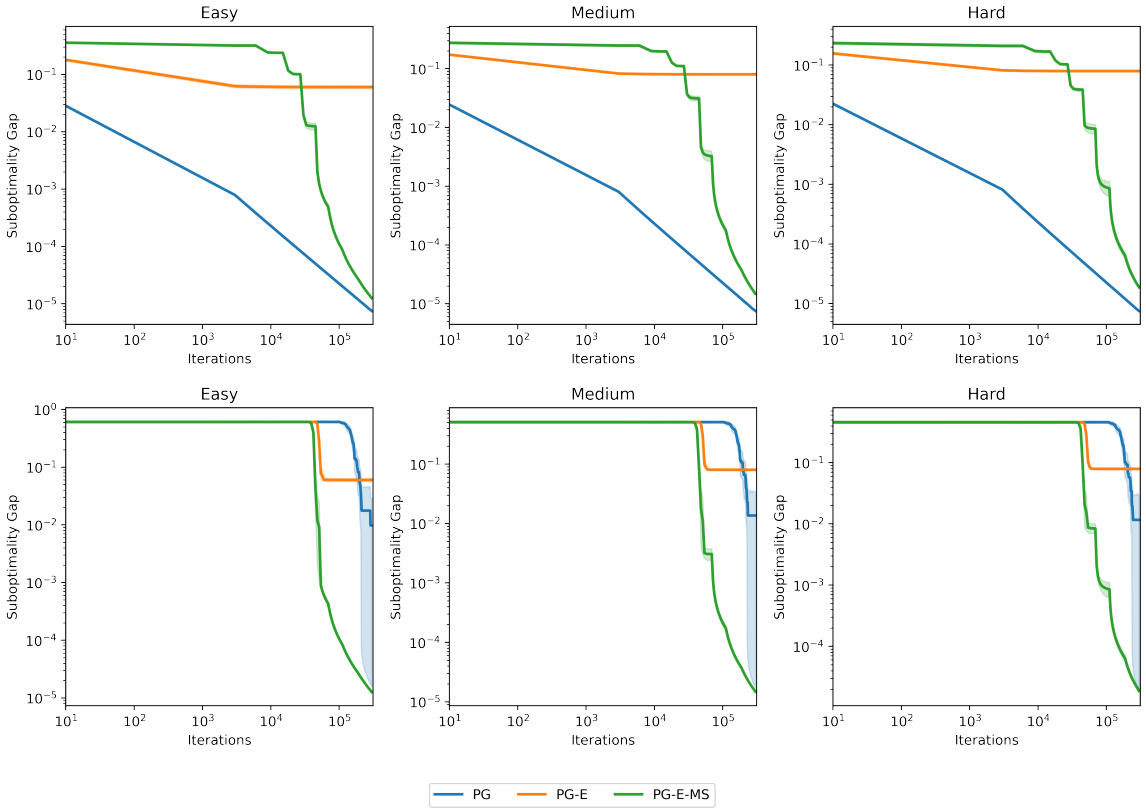

Figure 3: Sub-optimality gap across various environments and initializations. Top Row: the initial policy's parameters is uniform, i.e. $\theta_0(a) = 0 \quad \forall a$. Bottom Row: the initial policy's parameters is "bad", i.e. $\theta_0(a') = 12$ where $a' = \arg\min_a r(a)$

In Figure 3, `PG-E-MS` is able to converge to the optimal policy unlike `PG-E` since the temperature τ is decreasing. Furthermore, under "bad" initialization, `PG-E-MS` outpreforms `PG` since the addition of entropy enables the method to able to escape the initial flat region. On the other hand, `PG-E` is able to escape the initial region quickly, but is unable to converge to the optimal policy since τ is fixed.

Additionally, from our experiments, we observe that the multi-stage algorithm with $p = 1$ has a similar performance compared to softmax PG using uniform initialization. This confirms our theoretical observation that $p = 1$ results in a $\mathcal{O}(1/\epsilon)$ convergence rate. We additionally investigated how entropy regularization can help when starting with a "bad" initialization. In this case, the worst arm has a high probability of getting chosen, which results in a flat optimization landscape.

In most realistic scenarios it is difficult to calculate the exact gradient of the objective function. In the next section, we investigate how to extend the presented multi-stage algorithm to the stochastic setting.

D.3 Stochastic Setting

Following Section 4.1, we can construct an stochastic policy gradient using on-policy importance sampling (IS) reward estimates for the entropy regularized objective. Let $\nabla \widetilde{f}^{\tau}(\theta_t)$ denote the stochastic gradient with entropy regularization. By Lemma 37, the gradient estimators $\nabla \widetilde{f}^{\tau}(\theta)$ are (i) unbiased i.e. $\mathbb{E}[\nabla \widetilde{f}^{\tau}(\theta)] = \nabla f^{\tau}(\theta)$ and have (ii) bounded variance i.e. $\mathbb{E}\left\|\nabla \widetilde{f}^{\tau}(\theta) - \nabla f^{\tau}(\theta)\right\|_2^2 \leq \sigma^2$. The bound of the variance is differs compared to $\nabla \widetilde{f}(\theta)$ since σ^2 depends on the regularization strength τ. In this setting, we will consider the following update,

Update 4. *(Stochastic Softmax PG with Entropy, Importance Sampling)* $\theta_{t+1} = \theta_t + \eta_t \nabla \widetilde{f}^{\tau}(\theta_t)$.

Under the same setting when using on-policy IS reward estimates, prior work (Ding et al.) proposes a two-stage approach that converges to a biased optimal policy by modifying the batch size to counteract the variance. However, the method requires a $\tilde{\mathcal{O}}(1/\epsilon^2)$ sample complexity and knowledge of the biased optimal policy to set the algorithm hyper-parameters. Additionally, even with knowledge of the biased optimal policy, Ding et al. is unable to converge to the optimal policy.

To extend Algorithm 2 to the stochastic setting we first require an additional assumption since $\inf_{t\geq 1}[C_\tau(\theta_t)]^2$ is a now random variable in the stochastic setting.

Assumption 6. *f^τ satisfies the non-uniform Łojasiewciz condition for some $C_\tau(\theta)$ and $\xi = \frac{1}{2}$ such that $\mu := \mathbb{E}\left[\inf_{t\geq 1}[C_\tau(\theta_t)]^2\right] = \tau^p B_1$ for constants $p \geq 1$ and $B_1 > 0$.*

Under Assumption 6 and motivated by Section 4.1, we will utilize exponentially decaying step-sizes (Li et al., 2021; Vaswani et al., 2022) for each stage. At stage i, the resulting step-size at iteration t is set as: $\eta_{i,t-1} = \frac{1}{L^{\tau_i}} \alpha_i^{t-\text{last}_{i-1}}$ where $\alpha_i = \left(\frac{\beta}{T_i}\right)^{\frac{1}{T_i}}$, $\beta \geq 1$, and T_i is the length of stage i. Additionally, τ_i is the "temperature" of stage i. All together, this results in Algorithm 3.

The following theorem (proved in Appendix F.1) shows that Algorithm 3 converges to the globally optimal policy at an $\tilde{\mathcal{O}}\left(1/\epsilon^p + \sigma^2/\epsilon^{2p+1}\right)$ rate.

Theorem 8. Assuming f^τ and f satisfy Assumptions 2 to 6, for a given $\epsilon \in (0, 1)$, using Algorithm 3 with (a) unbiased stochastic gradients whose variance is bounded by σ^2 and (b) exponentially decreasing step-sizes $\eta_{i,t} = \eta_{i,\text{last}_{i-1}} \alpha_i^{t-\text{last}_{i-1}+1}$ where $\eta_{i,\text{last}_{i-1}} = \frac{1}{L^{\tau_i}}$ and $\alpha_i = \left(\frac{\beta}{T_i}\right)^{\frac{1}{T_i}}$, $\beta = 1$, achieves ϵ-sub-optimality to the globally optimal policy after $\tilde{\mathcal{O}}\left(\frac{1}{\epsilon^p} + \frac{\sigma^2}{\epsilon^{2p+1}}\right)$ iterations.

If $p = 1$, then convergence rate matches the $\tilde{\mathcal{O}}(\sigma^2/\epsilon^3)$ rate in Theorem 3. We remark that this the first stochastic softmax PG algorithm to obtain ϵ-convergece to the optimal policy while using entropy regularization. Unlike in prior work (Ding et al.), oracle-like knowledge of the environment is not necessary to obtain convergence while using entropy regularization in the stochastic setting.

In the next section, we will compare the multi-stage method with baseline methods in the bandits setting. To investigate if entropy regularization is indeed useufl, we will consider both uniform and "bad" initialization.

D.3.1 Experimental Evaluation

We evaluate the methods in multi-armed bandit environments with $A = 10$ in stochastic settings. For each environment, we compare the various algorithms based on their expected sub-optimality gap $\mathbb{E}[(\pi^* - \pi_{\theta_t})^\top r]$. We plot the average and 95% confidence interval of the expected sub-optimality gap across 25 independent bandit instances over $T = 10^6$ iterations. To counteract the randomness of each algorithm, for each bandit instance we additionally run each algorithm 5 times. In total, for each algorithm, the corresponding plot is comprised of 125 runs. To investigate if entropy regularization is helpful in escaping flat regions, we consider uniform and "bad" initialization. For experiments with uniform initialization, the initial policy is uniform, i.e. $\pi_{\theta_0}(a) = 1/A$ for all $a \in \mathcal{A}$. For experiments with bad initialization, the initial policy favours the worst arm, i.e. $\theta_0(a') = 9$ ($\pi_{\theta_0}(a') \approx 0.999$), where $a' := \arg\min_a r(a)$.

Environment Details: Each environment's underlying reward distribution is either a Bernoulli, Gaussian, or Beta distribution with a fixed mean reward vector $r \in \mathbb{R}^A$ and support $[0, 1]$. The difficulty of the environment is determined by the maximum reward gap $\bar{\Delta} := \min_{a^* \neq a} r(a^*) - r(a)$. In easy environments $\bar{\Delta} = 0.5$ and in the hard environments $\bar{\Delta} = 0.1$. For each environment, r is randomly generated for each run.

Methods: We compare the presented stochastic softmax PG multi-stage algorithm (Algorithm 3) (`SPG-E-MS`) to stochastic softmax PG (`SPG-ESS`) and stochastic softmax PG with entropy regularization (`SPG-E-ESS`) with exponentially decreasing step-sizes and when using the "doubling" trick (`SPG-ESS [D]`). We also compare with prior work that uses the full gradient (`SPG-O-G`) (Mei et al., 2021a) and the reward gap (`SPG-O-R`) (Mei et al., 2023) when setting the step-size. For `SPG-ESS` and `SPG-ESS [D]`, we select $\beta = 1$ and $\eta_0 = \frac{1}{18}$. For `SPG-E-ESS` we fix $\tau = 0.1$, and similarly select $\beta = 1$ and $\eta_0 = \frac{1}{L^\tau} = \frac{2}{5+10\,\tau\,(1+\log A)}$. Finally, for `SPG-E-MS`, we observed that the number of iterations T_i at each stage derived by Lemma 21 for the stochastic multistage algorithm are loose due to the exponentially-decreasing step-size analysis. Furthermore, we observed in the deterministic setting that when $p = 1$, the number of iterations doubles after each stage. Therefore, instead of using the theoretical number of iterations at each stage, we use the "doubling trick" (refer to Section 5). For `SPG-E-ESS` set the hyper-parameters $T_1 = 5000, \tau_0 = 0.5$, $B_1 = 1$ by employing a grid-search on a separate validation set of bandit instances. To fairly compare against `SPG-ESS` and `SPG-ESS [D]` we also select $\beta = 1$.

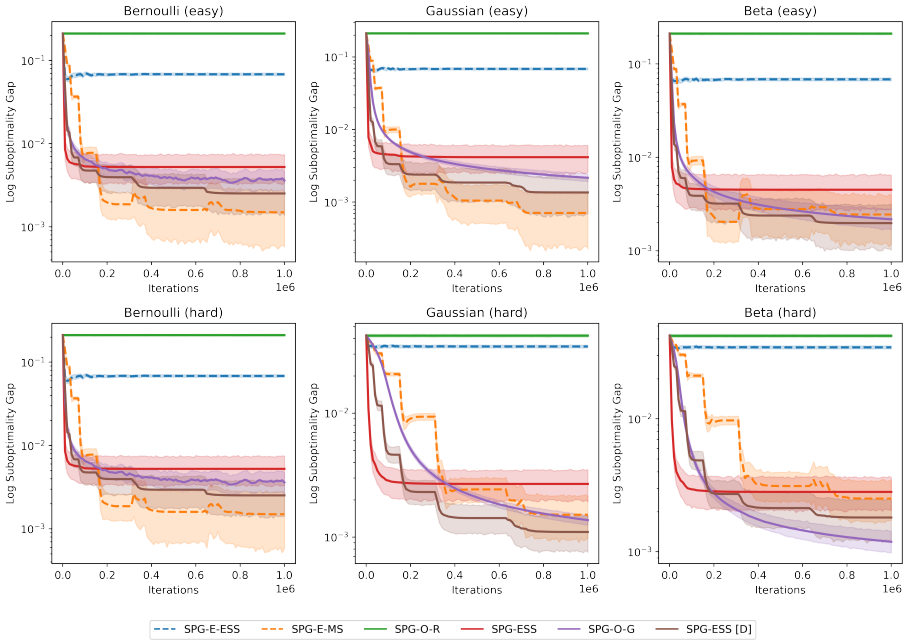

Figure 4: Expected sub-optimality gap across various environments with uniform initialization

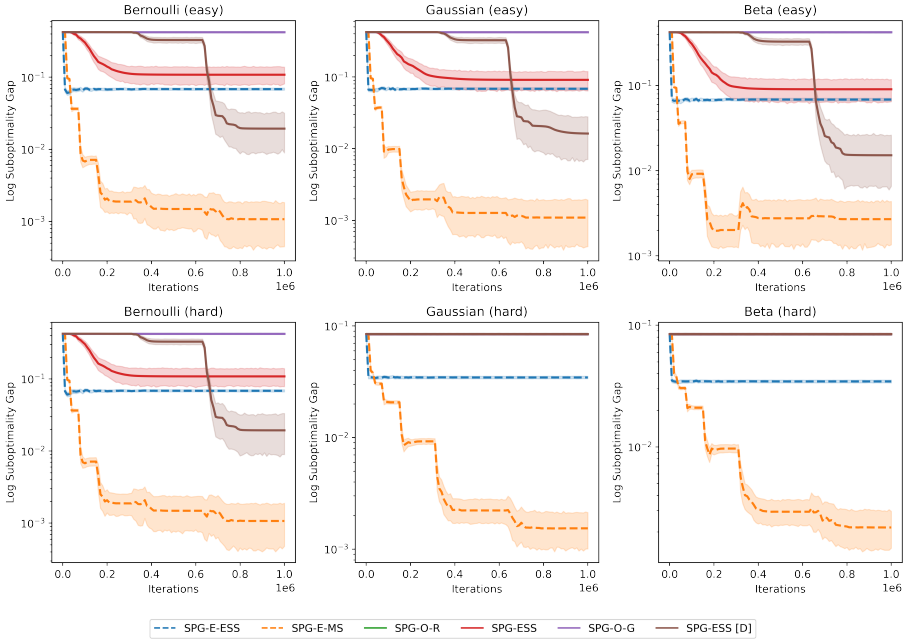

Figure 5: Expected sub-optimality gap across various environments with “bad” initialization

Results: From Figure 4, with uniform initialization, the performance of `SPG-E-MS` is comparable to `SPG-ESS`, `SPG-ESS [D]` and `SPG-O-G`. However, in the "bad" initialization settings (Figure 5), due to the presence of entropy, `SPG-E-MS` out preforms all other methods. Here we also find that entropy regularization helps escaping from flat regions in the stochastic setting. Since `SPG-E-ESS` uses a fixed entropy regularization term it is unable to converge to the optimal policy.

D.4 Discussion

We proposed a systematic method for (stochastic) softmax policy gradient (PG) to utilize the benefits of entropy regularization while guaranteeing convergence to the optimal policy. Under Assumption 1, our proposed multi-stage algorithm achieves convergence the optimal policy without any oracle-like knowledge when compared to prior methods. We empirically demonstrate that our multi-stage algorithm can escape flat regions in the exact and stochastic settings, due to entropy regularization. For future work, we aim to bridge the non-uniform Łojasiewciz conditions of f and f^τ as $\tau \to 0$.

E Proofs of Appendix D.2

Algorithm 2: Multi-Stage Softmax PG with Entropy Regularization

Output: Policy $\pi_{\theta_t} = \text{softmax}(\theta_t)$
Initialize parameters $\theta_0, \tau_0, N_{\text{stages}}$
$t \leftarrow 0$
$\text{last}_0 \leftarrow t$
$i \leftarrow 1$
while $i \leq N_{stages}$ **do**
  $\tau_i \leftarrow \tau_{i-1}/2$
  $\eta_i \leftarrow 1/L^{\tau_i}$
  $T_i \leftarrow \frac{2}{\eta_i \, \mu_i} \log\left(\frac{\tau_{i-1}}{\tau_i}\,(1+B_4)\right)$
  while $t - last_{i-1} < T_i$ **do**
    $\theta_{t+1} \leftarrow \theta_t + \eta_i \nabla f^{\tau_i}(\theta_t)$
    $t \leftarrow t+1$
  end
  $\text{last}_i \leftarrow t$
  $i \leftarrow i+1$
end

E.1 Proof of Theorem 7

Theorem 7. Assuming f^τ and f satisfy Assumptions 1 to 5, for a given $\epsilon \in (0,1)$, Algorithm 2 achieves ϵ-suboptimality to the globally optimal after $T_{\text{total}} = \frac{4\,L^{\max}\,C_1^p}{\epsilon^p\,B_1} \log\left(2\,(1+B_4)\right)$ iterations, where $C_1 = \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right) + B_2 + B_3$.

Proof. Observe that in Algorithm 2, we use τ_i and η_i at stage $i \geq 1$, which starts at iteration $\text{last}_{i-1}+1$, runs for $T_i = \frac{2}{\eta_i\,\mu_i}\log\left(\frac{\tau_{i-1}}{\tau_i}\,(1+B_4)\right)$ iterations, and ends at iteration last_i. Now, we prove by induction that $f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_i}) \leq \tau_i \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right)$ for all $i \geq 0$:
Base Case: For $i = 0$, we have

$$f^{*\tau_0} - f^{\tau_0}(\theta_0) \leq \max(\tau_0, f^{*\tau_0} - f^{\tau_0}(\theta_0)) = \tau_0 \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right). \tag{214}$$

Induction Step: Suppose $f^{*\tau_{i-1}} - f^{\tau_{i-1}}(\theta_{\text{last}_{i-1}}) \leq \tau_{i-1} \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right)$ holds.

Since $f^{\tau_i}(\theta)$ is L^{τ_i}-smooth and satisfies the non-uniform Łojasiewciz condition with $\mu_i := \inf_{t\geq 1} C_\tau^2(\theta_t)$, we use Lemma 12 for stage i:

$$f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_i}) \leq \exp(-\frac{\eta_i\,\mu_i}{2}\,T_i)[f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_{i-1}})] \tag{215}$$

If $T_i \geq \frac{2}{\eta_i\,\mu_i}\log\left(\frac{\tau_{i-1}}{\tau_i}\,(1+B_4)\right)$, we have

$$= \frac{f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_{i-1}})}{\exp\left(\log\left(\frac{\tau_{i-1}}{\tau_i}\,(1+B_4)\right)\right)} \tag{216}$$

Under Assumption 5

$$\leq \frac{f^{*\tau_{i-1}} - f^{\tau_{i-1}}(\theta_{\text{last}_{i-1}}) + \tau_{i-1}B_4}{\frac{\tau_{i-1}}{\tau_i}\,(1+B_4)} \tag{217}$$

Using the inductive hypothesis

$$\leq \frac{\tau_i \, \tau_{i-1} \left(\max \left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0} \right) + B_4 \right)}{\tau_{i-1} \, (1 + B_4)} \tag{218}$$

$$\leq \frac{\tau_i \max \left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0} \right) (1 + B_4)}{1 + B_4} \tag{219}$$

$$= \tau_i \max \left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0} \right). \tag{220}$$

Therefore, for all $i \geq 0$

$$f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_i}) \leq \tau_i \max \left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0} \right). \tag{221}$$

Define $\epsilon_i := f^* - f(\theta_{\text{last}_i})$ as the sub-optimality at the end of stage i. We have

$$\epsilon_i = f^* - f(\theta_{\text{last}_i}) \tag{222}$$

$$= \left[f^* - f(\theta^*_{\tau_i}) \right] + \left[f(\theta^*_{\tau_i}) - f(\theta_{\text{last}_i}) \right] \tag{223}$$

Under Assumption 4

$$\leq \left[f^* - f(\theta^*_{\tau_i}) \right] + f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_i}) + \tau_i B_3 \tag{224}$$

By Equation (221),

$$\leq \left[f^* - f(\theta^*_{\tau_i}) \right] + \tau_i \left(\max \left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0} \right) + B_3 \right) \tag{225}$$

Using Assumption 3,

$$\leq \tau_i \, B_2 + \tau_i \left(\max \left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0} \right) + B_3 \right) \tag{226}$$

$$= \tau_i \underbrace{\left(\max \left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0} \right) + B_2 + B_3 \right)}_{:= C_1} \tag{227}$$

$$= 2^{-i} \, \tau_0 \, C_1. \qquad (\tau_i = 2^{-i} \, \tau_0)$$

Therefore, the number of stages N_{stages} required to obtain an ϵ sub-optimality is given as:

$$2^{N_{\text{stages}}} \geq \frac{\tau_0 \, C_1}{\epsilon} \implies N_{\text{stages}} \geq \log_2 \left(\frac{\tau_0 \, C_1}{\epsilon} \right). \tag{228}$$

On the other hand, the sufficient number of iterations at stage i is:

$$T_i \geq \frac{2}{\eta_i \, \mu_i} \log \left(\frac{\tau_{i-1}}{\tau_i} \, (1 + B_4) \right) \tag{229}$$

Since $\eta_i = \frac{1}{L^{\tau_i}}$

$$= \frac{2 \, L^{\tau_i}}{\mu_i} \log \left(\frac{\tau_{i-1}}{\tau_i} \, (1 + B_4) \right), \tag{230}$$

Since $L^{\tau_i} \leq L^{\max}$, it is sufficient to set T_i as:

$$T_i = \frac{2 \, L^{\max}}{\mu_i} \log \left(\frac{\tau_{i-1}}{\tau_i} \, (1 + B_4) \right) \tag{231}$$

Under Assumption 1, $\mu_i = \tau_i^p B_1$

$$= \frac{2\,L^{\max}}{\tau_i^p\,B_1} \log\left(\frac{\tau_{i-1}}{\tau_i}\,(1+B_4)\right) \tag{232}$$

Since $\tau_i = 2^{-i}\,\tau_0$, we have

$$= \frac{2\,L^{\max}\,2^{ip}}{\tau_0^p\,B_1} \log\left(2\,(1+B_4)\right) \tag{233}$$

Consequently, we can calculate the sufficient total number of iterations T_{Total} in terms of ϵ:

$$T_{\text{Total}} \geq \sum_{i=1}^{N_{\text{stages}}} T_i = \sum_{i=1}^{N_{\text{stages}}} \left[\frac{2\,L^{\max}\,2^{ip}}{\tau_0^p\,B_1} \log\left(2\,(1+B_4)\right)\right] \tag{234}$$

$$= \frac{2\,L^{\max}\,\sum_{i=1}^{N_{\text{stages}}} (2^p)^i}{\tau_0^p\,B_1} \log\left(2\,(1+B_4)\right) \tag{235}$$

Since for all $x > 1, n \geq 0$, $\sum_{i=0}^{n} x^i = \frac{x^{n+1}-1}{x-1}$

$$= \frac{2\,L^{\max}\,\left[\frac{(2^p)^{N_{\text{stages}}+1}-1}{2^p-1} - 1\right]}{\tau_0^p\,B_1} \log\left(2\,(1+B_4)\right) \tag{236}$$

Therefore, it is sufficient that

$$T_{\text{Total}} \geq \frac{2\,L^{\max}\,\frac{(2^p)^{N_{\text{stages}}+1}}{2^p-1}}{\tau_0^p\,B_1} \log\left(2\,(1+B_4)\right) \tag{237}$$

$$= \frac{2\,L^{\max}\,\frac{2^p\,(2^p)^{N_{\text{stages}}}}{2^p-1}}{\tau_0^p\,B_1} \log\left(2\,(1+B_4)\right) \tag{238}$$

Since $p \geq 1$, we have $\frac{2^p}{2^p-1} \leq 2$. Hence, it is sufficient to use

$$T_{\text{Total}} = \frac{4\,L^{\max}\,(2^p)^{N_{\text{stages}}}}{\tau_0^p\,B_1} \log\left(2\,(1+B_4)\right) \tag{239}$$

$$= \frac{4\,L^{\max}\,(2^{N_{\text{stages}}})^p}{\tau_0^p\,B_1} \log\left(2\,(1+B_4)\right) \tag{240}$$

Using Equation (228),

$$\geq \frac{4\,L^{\max}\,C_1^p}{\epsilon^p\,B_1} \log\left(2\,(1+B_4)\right) \tag{241}$$

in order to guarantee $f^* - f(\theta_{T_{\text{total}}}) \leq \epsilon$. □

Corollary 10. In the bandit setting, assuming for each stage i, $\mu_i = \tau_i^p B_1$ for constants $p \geq 1$ and $B_1 > 0$, for a given $\epsilon \in (0,1)$, using Algorithm 2 with $\eta_i = \frac{2}{5+10\,\tau_i\,(1+\log A)}$ achieves ϵ-sub-optimality after $T_{\text{total}} = \frac{4\,L^{\max}\,C_1^p}{\epsilon^p\,B_1} \log\left(2\left(1+W\left(\frac{A-1}{e}\right)+\log A\right)\right)$ iterations, where $L^{\max} = \frac{5}{2} + 5\,(1+\log A)$ and $C_1 = \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right) + W\left(\frac{A-1}{e}\right) + \log A$.

Proof. Set $f(\theta) = {\pi_\theta}^\top r$ and $f^\tau(\theta) = {\pi_\theta}^\top (r - \tau \log \pi_\theta)$. We can extend Theorem 7 to the bandit setting since:

- by Lemma 26, f^τ is L^τ-smooth and since $\tau \in [0,1]$

$$\frac{5}{2} = L^{\min} \leq L^\tau = \frac{5}{2} + \tau\, 5\,(1+\log A) \leq \frac{5}{2} + 5\,(1+\log A) = L^{\max} \tag{242}$$

- by Lemma 14, we have $f^* - f(\theta^*_\tau) \leq \tau W\left(\frac{A-1}{e}\right)$
- by Lemma 15, we have for all θ, $f(\theta^*_\tau) - f(\theta) \leq f^{*\tau} - f^\tau(\theta) + \tau \log A$
- by Lemma 16, we have for all θ, $f^{*\tau_2} - f^{\tau_2}(\theta) \leq f^{*\tau_1} - f^{\tau_1}(\theta) + \tau_1 W\left(\frac{A-1}{e}\right) + \log A$

□

Corollary 11. In the tabular MDP setting, assuming for each stage i, $\mu_i = \tau_i^p B_1$ for constants $p \geq 1$ and $B_1 > 0$, for a given $\epsilon \in (0,1)$, using Algorithm 2 with $\eta_i = \frac{(1-\gamma)^3}{8+\tau_i\,(4+8\,\log A)}$ achieves ϵ-sub-optimality after $T_{\text{total}} = \frac{4\,L^{\max}\,C_1^p}{\epsilon^p\,B_1} \log\left(2\left(1+\frac{2\,\log A}{1-\gamma}\right)\right)$ iterations, where $L^{\max} = \frac{12+8\log A}{(1-\gamma)^3}$ and $C_1 = \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right) + \frac{2\,\log A}{1-\gamma}$.

Proof. Set $f(\theta) = V^{\pi_\theta}(\rho)$ and $f^\tau(\theta) = \tilde{V}_\tau^{\pi_\theta}(\rho)$. We can extend Theorem 7 to the tabular MDP setting since:

- by Lemma 28, $f^\tau(\theta)$ is L^τ-smooth and since $\tau \in [0,1]$

$$L^{\min} = \frac{8}{(1-\gamma)^3} \leq L^\tau = \frac{8+\tau(4+8\log A)}{(1-\gamma)^3} \leq \frac{12+8\log A}{(1-\gamma)^3} = L^{\max} \tag{243}$$

- by Lemma 17, we have $f^* - f(\theta^*_\tau) \leq \tau \frac{\log A}{1-\gamma}$
- by Lemma 19, we have for all θ, $f(\theta^*_\tau) - f(\theta) \leq f^{*\tau} - f^\tau(\theta) + \tau\,\frac{\log A}{1-\gamma}$
- by Lemma 20, we have for all θ, $f^{*\tau_2} - f^{\tau_2}(\theta) \leq f^{*\tau_1} - f^{\tau_1}(\theta) + \tau_1 \frac{2\log A}{1-\gamma}$

□

E.1.1 Additional Lemmas

Lemma 12. *Assuming f^τ satisfies Assumptions 1 and 2, using Update 3 with $\eta_t = \frac{1}{L^\tau}$, we have*

$$f^{*\tau} - f^\tau(\theta_{t_2}) \leq \exp\left(-\frac{\eta_t\,\mu}{2}\,(t_2 - t_1)\right)\,[f^{*\tau} - f^\tau(\theta_{t_1})] \tag{244}$$

where $t_1 < t_2$.

Proof.

Since f^τ is L^τ-smooth

$$f^\tau(\theta_{t+1}) \geq f^\tau(\theta_t) + \langle \nabla f^\tau(\theta_t),\, \theta_{t+1} - \theta_t \rangle - \frac{L^\tau}{2}\|\theta_{t+1} - \theta_t\|_2^2 \tag{245}$$

Using Update 3, $\theta_{t+1} = \theta_t + \eta_t\,\nabla f^\tau(\theta_t)$

$$= f^\tau(\theta_t) + \eta \|\nabla f^\tau(\theta_t)\|_2^2 - \frac{L^\tau\,{\eta_t}^2}{2}||\nabla f^\tau(\theta_t)||_2^2 \tag{246}$$

Using $\eta_t = \frac{1}{L^\tau}$

$$= f^\tau(\theta_t) + \frac{\eta_t}{2}\|\nabla f^\tau(\theta_t)\|_2^2 \tag{247}$$

Assuming Assumption 1 is satisfied, $\|\nabla f^\tau(\theta)\|_2^2 \geq \mu\,|f^{*\tau} - f^\tau(\theta)|$

$$\geq f^\tau(\theta_t) + \frac{\eta\,\mu}{2}\,[f^{*\tau} - f^\tau(\theta_t)] \tag{248}$$

Multiplying both sides by -1 and adding f^*

$$\implies f^{*\tau} - f^\tau(\theta_{t+1}) \leq \left(1 - \frac{\eta_t\,\mu}{2}\right)\,[f^{*\tau} - f^\tau(\theta_t)] \tag{249}$$

Using $1 - x \leq \exp(-x)$

$$\leq \exp\left(-\frac{\eta_t\,\mu}{2}\right)\,[f^{*\tau} - f^\tau(\theta_t)]. \tag{250}$$

Therefore,

$$f^{*\tau} - f^\tau(\theta_{t_2}) \leq \exp\left(-\frac{\eta_t\,\mu}{2}\,(t_2 - t_1)\right)\,[f^{*\tau} - f^\tau(\theta_{t_1})]. \tag{251}$$

□

E.2 Lemmas for the Bandit Setting

E.2.1 Verifying assumption 3

Lemma 13. *if* $\nabla_r\left[(\pi^* - \pi_\tau^*)^\top r\right] = \mathbf{0}$*, then all suboptimal rewards must be equal.*

Proof. Setting gradient of the bias of softmax optimal policy $(\pi^* - \pi_\tau^*)^\top r$ with respect to the reward vector r equal to a zero vector, the derivative of the bias with respect to an arbitrary suboptimal reward $r(\hat{a})$, where $\hat{a}$ is a suboptimal action, should be 0:

$$\frac{d}{dr(\hat{a})}(\pi^* - \pi_\tau^*)^\top r = 0 \implies \frac{d}{dr(\hat{a})}\frac{\sum_{a\neq a^*} e^{\frac{r(a)}{\tau}}\,\Delta(a)}{\sum_{a'} e^{\frac{r(a')}{\tau}}} = 0 \tag{252}$$

$$\implies \frac{\left(\frac{e^{\frac{r(\hat{a})}{\tau}}}{\tau}[r(a^*) - r(\hat{a})] - e^{\frac{r(\hat{a})}{\tau}}\right)\left(\sum_a e^{\frac{r(a)}{\tau}}\right) - \frac{e^{\frac{r(\hat{a})}{\tau}}}{\tau}\left(\sum_a e^{\frac{r(a)}{\tau}}[r(a^*) - r(a)]\right)}{\left(\sum_{a'} e^{\frac{r(a')}{\tau}}\right)^2} = 0 \tag{253}$$

$$\implies \frac{\frac{e^{\frac{r(\hat{a})}{\tau}}}{\tau}\left(\sum_a e^{\frac{r(a)}{\tau}}[r(a) - r(\hat{a}) - \tau]\right)}{\left(\sum_{a'} e^{\frac{r(a')}{\tau}}\right)^2} = 0 \implies \sum_a e^{\frac{r(a)}{\tau}}[r(a) - r(\hat{a}) - \tau] = 0 \tag{254}$$

Now, for any two suboptimal actions $\hat{a}_i$ and $\hat{a}_j$, we have

$$\implies \sum_a e^{\frac{r(a)}{\tau}}[r(a) - r(\hat{a}_i) - \tau] - \sum_a e^{\frac{r(a)}{\tau}}[r(a) - r(\hat{a}_j) - \tau] = 0 - 0 \tag{255}$$

$$\implies \sum_a e^{\frac{r(a)}{\tau}}[r(\hat{a}_j)) - r(\hat{a}_i)] = 0 \implies r(\hat{a}_j) = r(\hat{a}_i). \tag{256}$$

Therefore, all suboptimal rewards must be equal. □

Lemma 14. *We have* $(\pi^* - \pi_\tau^*)^\top r \leq \tau W\left(\frac{A-1}{e}\right)$*, where* $W\colon \mathbb{R}^+ \mapsto \mathbb{R}^+$ *is the principal branch of the Lambert W function, which is defined by* $W(x)e^{W(x)} = x \quad \forall x \geq 0$.

Proof. We want to find an upper bound on the difference between the expected reward achieved by the optimal policy π^* and the softmax optimal policy $\pi^*_\tau = \text{softmax}(r/\tau)$. Denoting $\Delta(a) = r(a^*) - r(a)$, $\Delta = \min_{a \neq a^*} \Delta(a)$, and a^* is the optimal action, we have

$$(\pi^* - \pi^*_\tau)^\top r = \sum_a \pi^*_\tau(a)\, r(a^*) - \sum_a \pi^*_\tau(a)\, r(a) = \sum_{a \neq a^*} \pi^*_\tau(a)\, \Delta(a) = \frac{\sum_{a \neq a^*} e^{\frac{r(a)}{\tau}}\, \Delta(a)}{\sum_{a'} e^{\frac{r(a')}{\tau}}}. \tag{257}$$

To find the upper bound, it is enough to find a reward vector $r \in \mathbb{R}^A$ that maximizes the bias. To do so, we find a unique stationary point and then prove that it is the reward vector with the maximum bias. First, we show that decreasing all rewards by a constant value c does not change the bias:

$$(\pi^* - \pi^*_\tau)^\top (r - c\mathbf{1}) = \frac{\sum_{a \neq a^*} e^{\frac{r(a)-c}{\tau}}\, \Delta(a)}{\sum_{a'} e^{\frac{r(a')-c}{\tau}}} = \frac{e^{-\frac{c}{\tau}} \sum_{a \neq a^*} e^{\frac{r(a)}{\tau}}\, \Delta(a)}{e^{-\frac{c}{\tau}} \sum_{a'} e^{\frac{r(a')}{\tau}}} \tag{258}$$

$$= \frac{\sum_{a \neq a^*} e^{\frac{r(a)}{\tau}}\, \Delta(a)}{\sum_{a'} e^{\frac{r(a')}{\tau}}} = (\pi^* - \pi^*_\tau)^\top r \tag{259}$$

Therefore, without loss of generality, we assume that the smallest reward value equals 0. Furthermore, according to Lemma 13, stationary reward vectors must have equal values for all non-optimal actions. Therefore, we assume that the reward vector has a value of $r_{a^*} = \Delta$ for the optimal action and 0 values for all other actions. In this case,

$$(\pi^* - \pi^*_\tau)^\top r = \frac{\sum_{a \neq a^*} e^{\frac{r(a)}{\tau}}\, \Delta(a)}{\sum_{a'} e^{\frac{r(a')}{\tau}}} = \frac{(A-1)\Delta}{e^{\frac{\Delta}{\tau}} + A - 1}. \tag{260}$$

Now, we find the reward gap Δ that makes the first derivative of the bias with respect to Δ equal to 0:

$$\frac{d}{d\Delta} \frac{(A-1)\Delta}{e^{\frac{\Delta}{\tau}} + A - 1} = 0 \implies \frac{(A-1)\left(e^{\frac{\Delta}{\tau}} + A - 1\right) - \frac{(A-1)\Delta e^{\frac{\Delta}{\tau}}}{\tau}}{\left(e^{\frac{\Delta}{\tau}} + A - 1\right)^2} = 0 \tag{261}$$

$$\implies (A-1)\left(e^{\frac{\Delta}{\tau}} + A - 1\right) - \frac{(A-1)\Delta e^{\frac{\Delta}{\tau}}}{\tau} = 0 \implies \tau\left(e^{\frac{\Delta}{\tau}} + A - 1\right) = \Delta e^{\frac{\Delta}{\tau}} \tag{262}$$

$$\implies \tau(A-1) = (\Delta - \tau)e^{\frac{\Delta}{\tau}} \implies \frac{\Delta - \tau}{\tau} e^{\frac{\Delta}{\tau}} = A - 1 \implies \frac{\Delta - \tau}{\tau} e^{\frac{\Delta - \tau}{\tau}} = \frac{A-1}{e} \tag{263}$$

$$\implies W\left(\frac{A-1}{e}\right) = \frac{\Delta - \tau}{\tau} \implies \Delta = \tau\left(W\left(\frac{A-1}{e}\right) + 1\right), \tag{264}$$

where $W : \mathbb{R} \mapsto \mathbb{R}$ is the principal branch of the Lambert W function. Since this value is the only stationary point of the bias with respect to the rewards vector, $\Delta = \tau\left(W\left(\frac{A-1}{e}\right) + 1\right)$ is either the global maximum or the global minimum point. Since π^* is the optimal policy, the bias $(\pi^* - \pi^*_\tau)^\top r$ is always non-negative. For $\Delta = 0$, the bias is equal to 0, so the unique stationary point must yield the global maximum. Substituting it in Equation (260), we get

$$(\pi^* - \pi^*_\tau)^\top r \leq \frac{(A-1)\tau\left(W\left(\frac{A-1}{e}\right) + 1\right)}{e^{W\left(\frac{A-1}{e}\right)+1} + A - 1}. \tag{265}$$

Now, since $e^{W(x)} = \frac{x}{W(x)}$,

$$= \frac{(A-1)\tau\left(W\left(\frac{A-1}{e}\right) + 1\right)}{\frac{A-1}{W\left(\frac{A-1}{e}\right)} + A - 1} \tag{266}$$

$$= \tau W\left(\frac{A-1}{e}\right). \tag{267}$$

□

E.2.2 Verifying assumption 4

Lemma 15. *For a fixed θ and τ, we have*

$$(\pi_\tau^* - \pi_\theta)^\top r \leq {\pi_\tau^*}^\top (r - \tau \log \pi_\tau^*) - {\pi_\theta}^\top (r - \tau \log \pi_\theta) + \tau \log A. \tag{268}$$

Proof.

$$(\pi_\tau^* - \pi_\theta)^\top r = {\pi_\tau^*}^\top (r - \tau \log \pi_\tau^*) - {\pi_\theta}^\top (r - \tau \log \pi_\theta) + \tau(\pi_\tau^* \log \pi_\tau^* - \pi_\theta \log \pi_\theta) \tag{269}$$

For all θ, $\log \frac{1}{A} \leq {\pi_\theta}^\top \log \pi_\theta \leq 0$

$$\leq {\pi_\tau^*}^\top (r - \tau \log \pi_\tau^*) - {\pi_\theta}^\top (r - \tau \log \pi_\theta) + \tau \left(0 - \log \frac{1}{A}\right) \tag{270}$$

$$= {\pi_\tau^*}^\top (r - \tau \log \pi_\tau^*) - {\pi_\theta}^\top (r - \tau \log \pi_\theta) + \tau \log A. \tag{271}$$

□

E.2.3 Verifying assumption 5

Lemma 16. *Set $f^\tau(\theta) = {\pi_\theta}^\top (r - \tau \log \pi_\theta)$. For a fixed θ, if $\tau_2 < \tau_1$, then*

$$f^{*\tau_2} - f^{\tau_2}(\theta) \leq f^{*\tau_1} - f^{\tau_1}(\theta) + \tau_1 W\left(\frac{A-1}{e}\right) + \tau_1 \log A. \tag{272}$$

Proof. Assuming $\tau_2 < \tau_1$, we have

$$[f^{*\tau_2} - f^{\tau_2}(\theta)] - [f^{*\tau_1} - f^{\tau_1}(\theta)] = [f^{*\tau_2} - f^{*\tau_1}] - [f^{\tau_2}(\theta) - f^{\tau_1}(\theta)] \tag{273}$$

$$= \left[{\pi_{\tau_2}^*}^\top (r - \tau_2 \log \pi_{\tau_2}^*) - {\pi_{\tau_1}^*}^\top (r - \tau_1 \log \pi_{\tau_1}^*)\right] - [{\pi_\theta}^\top (r - \tau_2 \log \pi_\theta) - {\pi_\theta}^\top (r - \tau_1 \log \pi_\theta)] \tag{274}$$

$$= (\pi_{\tau_2}^* - \pi_{\tau_1}^*)^\top r - \left[\tau_2 {\pi_{\tau_2}^*}^\top \log \pi_{\tau_2}^* - \tau_1 {\pi_{\tau_1}^*}^\top \log \pi_{\tau_1}^*\right] + (\tau_2 - \tau_1) {\pi_\theta}^\top \log \pi_\theta \tag{275}$$

For all θ, $\log \frac{1}{A} \leq {\pi_\theta}^\top \log \pi_\theta \leq 0$

$$\leq (\pi_{\tau_2}^* - \pi_{\tau_1}^*)^\top r - \left[\tau_2 \log \frac{1}{A} - \tau_1 0\right] + (\tau_2 - \tau_1) \log \frac{1}{A} \leq (\pi^* - \pi_{\tau_1}^*)^\top r + \tau_1 \log A. \tag{276}$$

By Lemma 14

$$\implies f^{*\tau_2} - f^{\tau_2}(\theta) \leq f^{*\tau_1} - f^{\tau_1}(\theta) + \tau_1 W\left(\frac{A-1}{e}\right) + \tau_1 \log A. \tag{277}$$

□

E.3 Lemmas for Tabular MDP Setting

E.3.1 Verifying assumption 3

Lemma 17 (Equation (12) in (Cen et al., 2022)). $V^*(\rho) - V^{\pi_\tau^*}(\rho) \leq \tau \frac{\log A}{1-\gamma}$.

E.3.2 Verifying assumption 4

Lemma 18. *For any π and ρ, we have*

$$\mathbb{H}(\pi) \leq \frac{\log A}{1-\gamma}, \tag{278}$$

where

$$\mathbb{H}(\pi) := \mathop{\mathbb{E}}_{\substack{s_0\sim\rho, a_t\sim\pi(\cdot|s_t),\\ s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)}} \left[\sum_{t=0}^{\infty} -\gamma^t \log \pi(a_t|s_t)\right]. \tag{279}$$

Proof.

$$\begin{aligned}
\mathbb{H}(\pi) =& \mathop{\mathbb{E}}_{\substack{s_0\sim\rho, a_t\sim\pi(\cdot|s_t),\\ s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)}} \left[\sum_{t=0}^{\infty} -\gamma^t \log \pi(a_t|s_t)\right] && (280)\\
=& \frac{1}{1-\gamma} \sum_{s,a} d_\rho^\pi(s)\, \pi(a|s) \left[-\log \pi(a|s)\right] && (281)\\
=& \frac{1}{1-\gamma} \sum_{s} d_\rho^\pi(s) \left[-\sum_a \pi(a|s)\, \log \pi(a|s)\right] && (282)
\end{aligned}$$

Since for all π, $\log \frac{1}{A} \leq \sum_a \pi(a|s)\, \log \pi(a|s) \leq 0$

$$\begin{aligned}
\leq& \frac{1}{1-\gamma} \sum_{s} d_\rho^\pi(s) \left[-\log \frac{1}{A}\right] && (283)\\
=& \frac{1}{1-\gamma} \sum_{s} d_\rho^\pi(s)\, \log A && (284)\\
=& \frac{\log A}{1-\gamma} && (285)
\end{aligned}$$

□

Lemma 19. *For a fixed θ and τ, we have*

$$V^{\pi_\tau^*}(\rho) - V^{\pi_\theta}(\rho) \leq \tilde{V}_\tau^*(\rho) - \tilde{V}_\tau^{\pi_\theta}(\rho) + \frac{\tau \log A}{1-\gamma}. \tag{286}$$

Proof.

$$\begin{aligned}
V^{\pi_\tau^*}(\rho) - V^{\pi_\theta}(\rho) =& (V^{\pi_\tau^*}(\rho) + \tau\mathbb{H}(\rho, \pi_\tau^*)) - (V^{\pi_\theta}(\rho) + \tau\mathbb{H}(\pi_\theta)) + \tau(\mathbb{H}(\pi_\theta) - \mathbb{H}(\pi_\tau^*)) && (287)\\
=& \tilde{V}_\tau^*(\rho) - \tilde{V}_\tau^{\pi_\theta}(\rho) + \tau(\mathbb{H}(\pi_\theta) - \mathbb{H}(\pi_\tau^*)) && (288)
\end{aligned}$$

Since for all π, $\mathbb{H}(\pi) \geq 0$

$$\leq \tilde{V}_\tau^*(\rho) - \tilde{V}_\tau^{\pi_\theta}(\rho) + \tau\mathbb{H}(\pi_\theta) \tag{289}$$

By Lemma 18

$$\leq \tilde{V}_\tau^*(\rho) - \tilde{V}_\tau^{\pi_\theta}(\rho) + \frac{\tau \log A}{1-\gamma} \tag{290}$$

□

E.3.3 Verifying assumption 5

Lemma 20. *For a fixed θ, if $\tau_2 < \tau_1$, then*

$$\tilde{V}^*_{\tau_2}(\rho) - \tilde{V}^{\pi_\theta}_{\tau_2}(\rho) \leq \tilde{V}^*_{\tau_1}(\rho) - \tilde{V}^{\pi_\theta}_{\tau_1}(\rho) + \frac{2\,\tau_1 \log A}{1-\gamma}. \tag{291}$$

Proof. Assuming $\tau_2 < \tau_1$, we have

$$\begin{aligned}
\tilde{V}^*_{\tau_2}(\rho) - \tilde{V}^{\pi_\theta}_{\tau_2}(\rho) - \tilde{V}^*_{\tau_1}(\rho) - \tilde{V}^{\pi_\theta}_{\tau_1}(\rho) &= [\tilde{V}^*_{\tau_2}(\rho) - \tilde{V}^*_{\tau_1}(\rho)] - [\tilde{V}^{\pi_\theta}_{\tau_2}(\rho) - \tilde{V}^{\pi_\theta}_{\tau_1}(\rho)] && (292)\\
&= \left[\left(V^{\pi^*_{\tau_2}}(\rho) + \tau_2\,\mathbb{H}(\pi^*_{\tau_2})\right) - \left(V^{\pi^*_{\tau_1}}(\rho) + \tau_1\,\mathbb{H}(\pi^*_{\tau_1})\right)\right]\\
&\quad - [(V^{\pi_\theta}(\rho) + \tau_2\,\mathbb{H}(\pi_\theta)) - (V^{\pi_\theta}(\rho) + \tau_1\,\mathbb{H}(\pi_\theta))] && (293)\\
&= \left[V^{\pi^*_{\tau_2}}(\rho) - V^{\pi^*_{\tau_1}}(\rho)\right] + \left[\tau_2\,\mathbb{H}(\pi^*_{\tau_2}) - \tau_1\,\mathbb{H}(\pi^*_{\tau_1})\right]\\
&\quad + (\tau_1 - \tau_2)\,\mathbb{H}(\rho, \pi_\theta). && (294)
\end{aligned}$$

By Lemma 18, $0 \leq \mathbb{H}(\pi) \leq \frac{\log A}{1-\gamma}$

$$\begin{aligned}
&\leq \left[V^{\pi^*_{\tau_2}}(\rho) - V^{\pi^*_{\tau_1}}(\rho)\right] + \left[\tau_2 \frac{\log A}{1-\gamma} - \tau_1\, 0\right] + (\tau_1 - \tau_2)\frac{\log A}{1-\gamma} && (295)\\
&\leq V^*(\rho) - V^{\pi^*_{\tau_1}}(\rho) + \tau_1 \frac{\log A}{1-\gamma}. && (296)
\end{aligned}$$

By Lemma 17,

$$\implies \tilde{V}^*_{\tau_2}(\rho) - \tilde{V}^{\pi_\theta}_{\tau_2}(\rho) \leq \tilde{V}^*_{\tau_1}(\rho) - \tilde{V}^{\pi_\theta}_{\tau_1}(\rho) + \frac{2\tau_1 \log A}{1-\gamma}. \tag{297}$$

□

F Proofs of Appendix D.3

Algorithm 3: Stochastic Multi-Stage Softmax PG with Entropy Regularization

Output: Policy $\pi_{\theta_t} = \text{softmax}(\theta_t)$
Initialize parameters $\theta_0, \tau_0, N_{\text{stages}}, \beta = 1$
$t \leftarrow 0$
$\text{last}_0 \leftarrow t$
$i \leftarrow 1$
while $i \leq N_{stages}$ **do**
  $\tau_i \leftarrow \frac{\tau_{i-1}}{2}$
  $X_1 \leftarrow \exp\left(\frac{\mu_i\,\beta}{L^{\tau}\,\log(T/\beta)}\right)$
  $X_2 \leftarrow \frac{0.69}{L^{\tau}}$
  $X_3 \leftarrow \frac{5\,L^{\tau}\,X_1}{e^2}$
  $T_i^{'} \leftarrow \frac{2}{X_2\,\mu_i}\log\left(\frac{2\,X_1\,\tau_{i-1}}{\tau_i}\,(1+B_4)\right)$
  $T_i^{''} \leftarrow \frac{2\,X_3\,\sigma^2}{\tau_i\,\mu_i^2}$
  $T_i \leftarrow \max(5583, 2\,T_i^{'}\,\log T_i^{'}, 4\,T_i^{''}\,\log^2 T_i^{''})$
  $\alpha_i \leftarrow \left(\frac{\beta}{T_i}\right)^{\frac{1}{T_i}}$
  $\eta_{i,t} \leftarrow \frac{\alpha_i}{L^{\tau_i}}$
  while $t - last_{i-1} < T_i$ **do**
    $\theta_{t+1} \leftarrow \theta_t + \eta_{i,t}\,\nabla\widetilde{f}^{\tau}(\theta_t)$
    $\eta_{i,t+1} \leftarrow \eta_{i,t}\,\alpha_i$
    $t \leftarrow t+1$
  end
  $\text{last}_i \leftarrow t$
  $i \leftarrow i+1$
end

F.1 Proof of Theorem 8

Theorem 8. Assuming f^{τ} and f satisfy Assumptions 2 to 6, for a given $\epsilon \in (0,1)$, using Algorithm 3 with (a) unbiased stochastic gradients whose variance is bounded by σ^2 and (b) exponentially decreasing step-sizes $\eta_{i,t} = \eta_{i,\text{last}_{i-1}}\,\alpha_i^{t-\text{last}_{i-1}+1}$ where $\eta_{i,\text{last}_{i-1}} = \frac{1}{L^{\tau_i}}$ and $\alpha_i = \left(\frac{\beta}{T_i}\right)^{\frac{1}{T_i}}$, $\beta = 1$, achieves ϵ-sub-optimality to the globally optimal policy after $\tilde{\mathcal{O}}\left(\frac{1}{\epsilon^p} + \frac{\sigma^2}{\epsilon^{2p+1}}\right)$ iterations.

Proof. Observe that in Algorithm 3, we use τ_i at stage $i \geq 1$, which starts at iteration $\text{last}_{i-1}+1$, ends at iteration last_i, and runs for $T_i = \max(5583, 2\,T_i^{'}\,\log T_i^{'}, 4\,T_i^{''}\,\log^2 T_i^{''})$ iterations, where

$$T_i^{'} = \frac{2\,\log\left(\frac{2\,X_1\,\tau_{i-1}(1+B_4)}{\tau_i}\right)}{X_2\,\mu_i}, \quad T_i^{''} = \frac{2\,X_3\,\sigma^2}{\tau_i\,\mu_i^2}, \tag{298}$$

where $X_1 = \exp\left(\frac{\mu_i\,\beta}{L^{\tau_i}\,\log(T/\beta)}\right)$, $X_2 = \frac{0.69}{L^{\tau_i}}$, and $X_3 = \frac{5\,L^{\tau_i}\,X_1}{e^2}$. Now, we will prove by induction that $\mathbb{E}[f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_i})] \leq \tau_i \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right)$ for all $i \geq 0$:
Base Case: For $i = 0$, we have

$$f^{*\tau_0} - f^{\tau_0}(\theta_0) \leq \max(\tau_0, f^{*\tau_0} - f^{\tau_0}(\theta_0)) = \tau_0 \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right). \tag{299}$$

Induction Step: Suppose $\mathbb{E}[f^{*\tau_{i-1}} - f^{\tau_{i-1}}(\theta_{\text{last}_{i-1}})] \leq \tau_{i-1} \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right)$ holds. At stage i, by Lemma 21, using exponentially decreasing step-size $\eta_{i,t} = \eta_{i,\text{last}_{i-1}}\, \alpha_i^{t-\text{last}_{i-1}+1}$, where $\eta_{i,\text{last}_{i-1}} = \frac{1}{L^{\tau_i}}$, $\alpha_i = \left(\frac{\beta}{T_i}\right)^{\frac{1}{T_i}}$ with $\beta = 1$, for $\mathbb{E}[f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_i})] \leq \tau_i \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right)$ to hold, it suffices that $T_i \geq \max(5583, 2\,Y_i\,\log Y_i, 4\,Y_i^{'}\,\log^2 Y_i^{'})$, where

$$Y_i = \frac{2\,\log\left(\frac{2\,X_1\,\mathbb{E}[f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_{i-1}})]}{\tau_i \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right)}\right)}{X_2\,\mu_i}, \quad Y_i^{'} = \frac{2\,X_3\,\sigma^2}{\tau_i\,\mu_i^2 \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right)}. \tag{300}$$

Under Assumption 5,

$$Y_i \leq \frac{2\,\log\left(\frac{2\,X_1\left(\mathbb{E}[f^{*\tau_{i-1}} - f^{\tau_{i-1}}(\theta_{\text{last}_{i-1}})] + \tau_{i-1} B_4\right)}{\tau_i \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right)}\right)}{X_2\,\mu_i} \tag{301}$$

Using the inductive hypothesis

$$\leq \frac{2\,\log\left(\frac{2\,X_1\left(\tau_{i-1} \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right) + \tau_{i-1} B_4\right)}{\tau_i \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right)}\right)}{X_2\,\mu_i} \tag{302}$$

$$\leq \frac{2\,\log\left(\frac{2\,X_1\,\tau_{i-1} \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right)(1 + B_4)}{\tau_i \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right)}\right)}{X_2\,\mu_i} \tag{303}$$

$$= \frac{2\,\log\left(\frac{2\,X_1\,\tau_{i-1}(1+B_4)}{\tau_i}\right)}{X_2\,\mu_i} = T_i^{'}. \tag{304}$$

On the other hand, we have

$$Y_i^{'} \leq \frac{2\,X_3\,\sigma^2}{\tau_i\,\mu_i^2} = T_i^{''}. \tag{305}$$

Therefore, $T_i = \max(5583, 2\,T_i^{'}\,\log T_i^{'}, 4\,T_i^{''}\,\log^2 T_i^{''}) \geq \max(5583, 2\,Y_i\,\log Y_i, 4\,Y_i^{'}\,\log^2 Y_i^{'})$. This implies $\mathbb{E}[f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_i})] \leq \tau_i \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right)$ holds for all $i \geq 0$. As a result, under Assumption 4, we have

$$\mathbb{E}[f(\theta_{\tau_i}^*) - f(\theta_{\text{last}_i})] \leq \mathbb{E}[f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_i})] + \tau_i\,B_3 \tag{306}$$

$$\leq \tau_i \left(\max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right) + B_3\right) \tag{307}$$

Denote $\epsilon_i := \mathbb{E}[f^* - f(\theta_{\text{last}_i})]$ as the suboptimality at the end of stage i. We have

$$\epsilon_i = \mathbb{E}[f^* - f(\theta_{\text{last}_i})] \tag{308}$$

$$= f^* - f(\theta_{\tau_i}^*) + \mathbb{E}[f(\theta_{\tau_i}^*) - f(\theta_{\text{last}_i})] \tag{309}$$

Under Assumption 3

$$\leq \tau_i\,C_1 \tag{310}$$

where $C_1 = \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right) + B_2 + B_3$. Therefore, ϵ_i has an upper bound that is proportional to τ_i. Now, since $\tau_i = 2^{-i}\,\tau_0$, the sub-optimality ϵ_i has an exponential rate in terms of the number of executed stages:

$$= 2^{-i}\,\tau_0\,C_1 \tag{311}$$

Therefore, the required number of stages N_{stages} in terms of the final sub-optimality $\epsilon := \epsilon_{N_{\text{stages}}}$ is

$$2^{N_{\text{stages}}} \geq \frac{\tau_0\,C_1}{\epsilon} \implies N_{\text{stages}} \geq \log_2\left(\frac{\tau_0\,C_1}{\epsilon}\right). \tag{312}$$

On the other hand, we have the sufficient number of iterations at stage i:

$$T_i \geq \max\left(5583, \frac{4\,\log\left(\frac{2\,X_1\,\tau_{i-1}(1+B_4)}{\tau_i}\right)}{X_2\,\mu_i}\,\log\left(\frac{\log\left(\frac{2\,X_1\,\tau_{i-1}(1+B_4)}{\tau_i}\right)}{X_2\,\mu_i}\right), \frac{8\,X_3\,\sigma^2}{\tau_i\,\mu_i^2}\,\log^2\left(\frac{2\,X_3\,\sigma^2}{\tau_i\,\mu_i^2}\right)\right) \tag{313}$$

Since $\tau_i \leq 1$, under Assumption 6, we have $\mu_i = \tau_i^p\,B_1 \leq B_1$. Furthermore, $\log\left(\frac{T_i}{\beta}\right) \geq 1$, and under Assumption 2, we have $0 < L^{\min} \leq L^{\tau_i} \leq L^{\max}$. Therefore,

$$X_1 \leq A_1 = \exp\left(\frac{B_1\,\beta}{L^{\min}}\right), \tag{314}$$

$$X_2 \geq A_2 = \frac{0.69}{L^{\max}}, \tag{315}$$

$$X_3 \leq A_3 = \frac{5\,L^{\max}\,A_1}{e^2}. \tag{316}$$

Hence, we can safely substitute variables X_1, X_2, X_3 with their corresponding constants A_1, A_2, A_3. Therefore, it is sufficient to set T_i as

$$T_i \geq \max\left(5583, \frac{4\,\log\left(\frac{2\,A_1\,\tau_{i-1}(1+B_4)}{\tau_i}\right)}{A_2\,\mu_i}\,\log\left(\frac{\log\left(\frac{2\,A_1\,\tau_{i-1}(1+B_4)}{\tau_i}\right)}{A_2\,\mu_i}\right), \frac{8\,A_3\,\sigma^2}{\tau_i\,\mu_i^2}\,\log^2\left(\frac{2\,A_3\,\sigma^2}{\tau_i\,\mu_i^2}\right)\right) \tag{317}$$

Under Assumption 6, $\mu_i = \tau_i^p\,B_1$

$$= \max\left(5583, \frac{4\,\log\left(\frac{2\,A_1\,\tau_{i-1}(1+B_4)}{\tau_i}\right)}{A_2\,\tau_i^p\,B_1}\,\log\left(\frac{\log\left(\frac{2\,A_1\,\tau_{i-1}(1+B_4)}{\tau_i}\right)}{A_2\,\tau_i^p\,B_1}\right), \frac{8\,A_3\,\sigma^2}{\tau_i^{2p+1}\,B_1^2}\,\log^2\left(\frac{2\,A_3\,\sigma^2}{\tau_i^{2p+1}\,B_1^2}\right)\right) \tag{318}$$

Since $\tau_i = 2^{-i}\,\tau_0$

$$= \max\left(5583, \frac{4\,\log(4\,A_1\,(1+B_4))\,2^{ip}}{A_2\,\tau_0^p\,B_1}\,\log\left(\frac{\log(4\,A_1\,(1+B_4))\,2^{ip}}{A_2\,\tau_0^p\,B_1}\right), \frac{8\,A_3\,\sigma^2\,2^{i(2p+1)}}{\tau_0^{2p+1}\,B_1^2}\,\log^2\left(\frac{2\,A_3\,\sigma^2\,2^{i(2p+1)}}{\tau_0^{2p+1}\,B_1^2}\right)\right) \tag{319}$$

Since $i \leq N_{\text{stages}}$, it is sufficient that

$$T_i = \max\left(5583, \frac{4\,\log(4\,A_1\,(1+B_4))\,2^{ip}}{A_2\,\tau_0^p\,B_1}\,Y_1, \frac{8\,A_3\,\sigma^2\,2^{i(2p+1)}}{\tau_0^{2p+1}\,B_1^2}\,Y_2\right) \tag{320}$$

where $Y_1 = \log\left(\frac{\log(4\,A_1\,(1+B_4))\,(2^{N_{\text{stages}}})^p}{A_2\,\tau_0^p\,B_1}\right)$ and $Y_2 = \log^2\left(\frac{2\,A_3\,\sigma^2\,(2^{N_{\text{stages}}})^{2p+1}}{\tau_0^{2p+1}\,B_1^2}\right)$. Consequently, we can calculate the sufficient total number of iterations T_{Total} in terms of ϵ:

$$T_{\text{Total}} \geq \sum_{i=1}^{N_{\text{stages}}} T_i \tag{321}$$

$$= \sum_{i=1}^{N_{\text{stages}}} \max\left(5583, \frac{4\,\log(4\,A_1\,(1+B_4))\,2^{ip}}{A_2\,\tau_0^p\,B_1}\,Y_1, \frac{8\,A_3\,\sigma^2\,2^{i(2p+1)}}{\tau_0^{2p+1}\,B_1^2}\,Y_2\right) \tag{322}$$

$$= \max\left(5583\,N_{\text{stages}}, \frac{4\,\log(4\,A_1\,(1+B_4))\,\sum_{i=1}^{N_{\text{stages}}}(2^p)^i}{A_2\,\tau_0^p\,B_1}\,Y_1, \frac{8\,A_3\,\sigma^2\,\sum_{i=1}^{N_{\text{stages}}}(2^{2p+1})^i}{\tau_0^{2p+1}\,B_1^2}\,Y_2\right) \tag{323}$$

Since $\forall x > 1, n \geq 0$, $\sum_{i=0}^{n} x^i = \frac{x^{n+1}-1}{x-1}$

$$= \max\left(5583\,N_{\text{stages}}, \frac{4\,\log(4\,A_1\,(1+B_4))\,\left[\frac{(2^p)^{N_{\text{stages}}+1}-1}{2^p-1} - 1\right]}{A_2\,\tau_0^p\,B_1}\,Y_1,\right.$$
$$\left.\frac{8\,A_3\,\sigma^2\,\left[\frac{(2^{2p+1})^{N_{\text{stages}}+1}-1}{2^{2p+1}-1} - 1\right]}{\tau_0^{2p+1}\,B_1^2}\,Y_2\right) \tag{324}$$

Therefore, it is sufficient that

$$T_{\text{Total}} \geq \max\left(5583\,N_{\text{stages}}, \frac{4\,\log(4\,A_1\,(1+B_4))\,\frac{(2^p)^{N_{\text{stages}}+1}}{2^p-1}}{A_2\,\tau_0^p\,B_1}\,Y_1, \frac{8\,A_3\,\sigma^2\,\frac{(2^{2p+1})^{N_{\text{stages}}+1}}{2^{2p+1}-1}}{\tau_0^{2p+1}\,B_1^2}\,Y_2\right) \tag{325}$$

$$= \max\left(5583\,N_{\text{stages}}, \frac{4\,\log(4\,A_1\,(1+B_4))\,\frac{2^p\,(2^p)^{N_{\text{stages}}}}{2^p-1}}{A_2\,\tau_0^p\,B_1}\,Y_1, \frac{8\,A_3\,\sigma^2\,\frac{2^{2p+1}\,(2^{2p+1})^{N_{\text{stages}}}}{2^{2p+1}-1}}{\tau_0^{2p+1}\,B_1^2}\,Y_2\right) \tag{326}$$

Since $p \geq 1$, we have $\frac{2^p}{2^p-1} \leq 2$ and $\frac{2^{2p+1}}{2^{2p+1}-1} \leq \frac{8}{7}$. Hence, it is sufficient to use

$$T_{\text{Total}} = \max\left(5583\,N_{\text{stages}}, \frac{8\,\log(4\,A_1\,(1+B_4))\,(2^p)^{N_{\text{stages}}}}{A_2\,\tau_0^p\,B_1}\,Y_1, \frac{64\,A_3\,\sigma^2\,(2^{2p+1})^{N_{\text{stages}}}}{7\,\tau_0^{2p+1}\,B_1^2}\,Y_2\right) \tag{327}$$

$$= \max\left(5583\,N_{\text{stages}}, \frac{8\,\log(4\,A_1\,(1+B_4))\,(2^{N_{\text{stages}}})^p}{A_2\,\tau_0^p\,B_1}\,Y_1, \frac{64\,A_3\,\sigma^2\,(2^{N_{\text{stages}}})^{2p+1}}{7\,\tau_0^{2p+1}\,B_1^2}\,Y_2\right) \tag{328}$$

Using Equation (312)

$$\geq \max\left(5583\,\log_2\left(\frac{\tau_0\,C_1}{\epsilon}\right), \frac{8\,\log(4\,A_1\,(1+B_4))\,C_1^p\,\log\left(\frac{\log(4\,A_1\,(1+B_4))\,C_1^p}{A_2\,B_1\,\epsilon^p}\right)}{A_2\,B_1\,\epsilon^p},\right.$$
$$\left.\frac{64\,A_3\,C_1^{2p+1}\,\log^2\left(\frac{2\,A_3\,C_1^{2p+1}\,\sigma^2}{B_1^2\,\epsilon^{2p+1}}\right)\,\sigma^2}{7\,B_1^2\,\epsilon^{2p+1}}\right) \tag{329}$$

$$\implies T_{\text{Total}} \in \tilde{\mathcal{O}}\left(\frac{1}{\epsilon^p} + \frac{\sigma^2}{\epsilon^{2p+1}}\right). \tag{330}$$

□

Corollary 12. In the bandit setting, assuming for each stage i, $\mu_i = \tau_i^p B_1$ for constants $p \geq 1$, $B_1 > 0$, for a given $\epsilon \in (0,1)$, using Algorithm 3 with exponentially decreasing step-sizes $\eta_{i,t} = \eta_{i,\text{last}_{i-1}} \alpha_i^{t-\text{last}_{i-1}+1}$ where $\eta_{i,\text{last}_{i-1}} = \frac{2}{5+10\,\tau_i\,(1+\log A)}$ and $\alpha_i = \left(\frac{\beta}{T_i}\right)^{\frac{1}{T_i}}$, $\beta = 1$, achieves ϵ-suboptimality after $T_{\text{Total}} \in \tilde{\mathcal{O}}\left(\frac{1}{\epsilon^p} + \frac{\sigma^2}{\epsilon^{2p+1}}\right)$ iterations.

Proof. Set $f(\theta) = {\pi_\theta}^\top r$ and $f^\tau(\theta) = {\pi_\theta}^\top (r - \tau \log \pi_\theta)$. We can extend Theorem 8 to the bandit setting since:

- by Lemma 26, f^τ is L^τ-smooth and $\tau \in [0,1]$

$$\frac{5}{2} = L^{\min} \leq L^\tau = \frac{5}{2} + \tau\, 5\,(1 + \log A) \leq \frac{5}{2} + 5\,(1 + \log A) = L^{\max} \tag{331}$$

- by Lemma 14, we have $f^* - f(\theta_\tau^*) \leq \tau W\left(\frac{A-1}{e}\right)$
- by Lemma 15, we have for all θ, $f(\theta_\tau^*) - f(\theta) \leq f^{*\tau} - f^\tau(\theta) + \tau \log A$
- by Lemma 16, we have for all θ, $f^{*\tau_2} - f^{\tau_2}(\theta) \leq f^{*\tau_1} - f^{\tau_1}(\theta) + \tau_1 W\left(\frac{A-1}{e}\right) + \log A$
- by Lemma 38, the gradient estimator is unbiased and have bounded variance where $\sigma^2 = 8\,(1 + (\tau \log A)^2)$.

□

Corollary 13. In the tabular MDP setting, assuming for each stage i , $\mu_i = \tau_i^p\, B_1$ for constants $p \geq 1$, $B_1 > 0$, for a given $\epsilon \in (0,1)$, using Algorithm 3 with exponentially decreasing step-sizes $\eta_{i,t} = \eta_{i,\text{last}_{i-1}} \alpha_i^{t-\text{last}_{i-1}+1}$, where $\eta_{i,\text{last}_{i-1}} = \frac{(1-\gamma)^3}{8+\tau_i(4+8\log A)}$ and $\alpha_i = \left(\frac{\beta}{T_i}\right)^{\frac{1}{T_i}}$, $\beta = 1$, achieves ϵ-sub-optimality after $T_{\text{Total}} \in \tilde{\mathcal{O}}\left(\frac{1}{\epsilon^p} + \frac{\sigma^2}{\epsilon^{2p+1}}\right)$ iterations.

Proof. Set $f(\theta) = V^{\pi_\theta}(\rho)$ and $f^\tau(\theta) = \tilde{V}_\tau^{\pi_\theta}(\rho)$. We can extend Theorem 8 to the MDP setting since:

- by Lemma 28, f^τ is L^τ-smooth and since $\tau \in [0,1]$

$$L^{\min} = \frac{8}{(1-\gamma)^3} \leq L^\tau = \frac{8 + \tau(4 + 8\log A)}{(1-\gamma)^3} \leq \frac{12 + 8\log A}{(1-\gamma)^3} = L^{\max} \tag{332}$$

- by Lemma 17, we have $f^* - f(\theta_\tau^*) \leq \tau\, \frac{\log A}{1-\gamma}$
- by Lemma 19, we have for all θ, $f(\theta_\tau^*) - f(\theta) \leq f^{*\tau} - f^\tau(\theta) + \tau\, \frac{\log A}{1-\gamma}$
- by Lemma 20, we have for all θ, $f^{*\tau_2} - f^{\tau_2}(\theta) \leq f^{*\tau_1} - f^{\tau_1}(\theta) + \tau_1\, \frac{2\log A}{1-\gamma}$
- by Lemma 37, the gradient estimators are unbiased and have bounded variance where $\sigma^2 = \frac{8}{(1-\gamma)^2}\left(\frac{1+(\tau\,\log A)^2}{(1-\gamma^{1/2})^2}\right)$.

□

F.1.1 Additional Lemmas

Lemma 21. *Assuming f^τ satisfies Assumptions 2 and 6 and the gradient estimators $\nabla \widetilde{f}^\tau(\theta_t)$ are unbiased and have bounded variance σ^2, for a given $\epsilon \in (0,1)$, using Update 4 from iteration t_1+1 to t_2 with exponentially decreasing step-sizes $\eta_t = \eta_0\, \alpha^{t-t_1+1}$, where $\eta_t = \frac{1}{L^\tau}$ and $\alpha = (\frac{\beta}{T})^{\frac{1}{T}}$, $\beta \geq 1$, and $T = t_2 - t_1 > 0$, is achieved in ϵ-sub-optimality is achieved in $\max(\beta + 1, 5583, 2\,Y_1\,\log Y_1, 4\,Y_2\,\log^2 Y_2)$ iterations, where $Y_1 = \frac{2\,\log\left(\frac{2\,X_1\,\mathbb{E}[f^{*\tau} - f^\tau(\theta_{t_1})]}{\epsilon}\right)}{X_2\,\mu}$, $Y_2 = \frac{2\,X_3\,\sigma^2}{\mu^2\,\epsilon}$,$X_1 = \exp\left(\frac{\mu\,\beta}{L^\tau\,\log(T/\beta)}\right)$, $X_2 = \frac{0.69}{L^\tau}$, and $X_3 = \frac{5\,L^\tau\,X_1}{e^2}$.*

Proof. From (Li et al., 2021, Theorem 1), using Update 4 with exponentially decreasing step-sizes results from iterations t_1+1 to t_2 results in the following convergence

$$\mathbb{E}[f^{*\tau} - f^\tau(\theta_{t_2})] \leq X_1\,\exp\left(-\frac{X_2\,\mu}{2}\,\frac{T}{\log\frac{T}{\beta}}\right)\mathbb{E}[f^{*\tau} - f^\tau(\theta_{t_1})] + \frac{X_3\,\sigma^2}{\mu^2\,\frac{T}{\log^2\frac{T}{\beta}}}, \tag{333}$$

where

$$X_1 = \exp\left(\frac{\mu\,\beta}{L^\tau\,\log\frac{T}{\beta}}\right), \quad X_2 = \frac{0.69}{L^\tau}, \quad X_3 = \frac{5\,L^\tau\,X_1}{e^2} \tag{334}$$

and $\mu := \inf_{t\geq 1} C_\tau(\theta)$ with $T = t_2 - t_1$. We show that if the inequalities $\frac{T}{\log\frac{T}{\beta}} \geq Y_1$ and $\frac{T}{\log^2\frac{T}{\beta}} \geq Y_2$ are satisfied, where

$$Y_1 = \frac{2\,\log\left(\frac{2\,X_1\,\mathbb{E}[f^{*\tau} - f^\tau(\theta_{t_1})]}{\epsilon}\right)}{X_2\,\mu}, \quad Y_2 = \frac{2\,X_3\,\sigma^2}{\mu^2\,\epsilon}, \tag{335}$$

then $\mathbb{E}[f^{*\tau} - f^\tau(\theta_{t_2})] \leq \epsilon$ holds since

$$\begin{aligned}
&\mathbb{E}[f^{*\tau} - f^\tau(\theta_{t_2})] && (336)\\
&\quad \leq X_1\,\exp\left(-\frac{X_2\,\mu}{2}\,\frac{2}{X_2\,\mu}\log\left(\frac{2\,X_1\,[f^{*\tau} - f^\tau(\theta_{t_1})]}{\epsilon}\right)\right)\mathbb{E}[f^{*\tau} - f^\tau(\theta_{t_1})] + \frac{X_3\,\sigma^2}{\mu^2\,\frac{2\,X_3\,\sigma^2}{\mu^2\,\epsilon}} && (337)\\
&\quad = \frac{\epsilon}{2} + \frac{\epsilon}{2} && (338)\\
&\quad = \epsilon. && (339)
\end{aligned}$$

By Lemma 22 and since $1 \leq \beta < T$, for $\frac{T}{\log(T/\beta)} \geq \frac{T}{\log T} \geq Y_1$ to hold, it suffices that $T \geq \max(2, 2\,Y_1\,\log Y_1)$. Furthermore, according to Lemma 23 and since $1 \leq \beta < T$, for $\frac{T}{\log^2(T/\beta)} \geq \frac{T}{\log^2 T} \geq Y_2$ to hold, it suffices that $T \geq \max(5583, 4\,Y_2\,\log^2 Y_2)$. Therefore, the required number of iterations to achieve ϵ-sub-optimality is $\max(5583, 2\,Y_1\,\log Y_1, 4\,Y_2\,\log^2 Y_2)$. □

Lemma 22. *For all $C > 0$, if $T \geq \max(2, 2\,C\,\log C)$, then $\frac{T}{\log T} \geq C$.*

Proof. If $C < 2$, knowing that $T \geq 2$, we have

$$\frac{T}{\log T} > 2 > C \tag{340}$$

Otherwise, if $C \geq 2$,

$$2\,C\,\log C = C(\log C + \log C) \tag{341}$$

Since $\forall C > 0$, $C \geq 2 \log C$,

$$\geq C(\log C + \log(2 \log C)) \tag{342}$$

$$= C \log(2\, C\, \log C) \tag{343}$$

$$\implies \frac{2\, C\, \log C}{\log(2\, C\, \log C)} \geq C. \tag{344}$$

Therefore, knowing that $T \geq 2\, C\, \log C$, since $2\, C\, \log C \geq 4 \log 2 > 2.72$, we have

$$\frac{T}{\log T} \geq \frac{2\, C\, \log C}{\log(2\, C\, \log C)} \geq C. \tag{345}$$

□

Lemma 23. *For all $C > 0$, if $T \geq \max(5583, 4\, C\, \log^2 C)$, then $\frac{T}{\log^2 T} \geq C$.*

Proof. If $C < 75$, knowing that $T \geq 5583$, we have

$$\frac{T}{\log^2 T} > 75 > C. \tag{346}$$

Otherwise, if $C \geq 75$,

$$4\, C\, \log^2 C = C(\log C + \log C)^2 \tag{347}$$

Since $C \geq 4\, \log^2 C \quad \forall C \geq 75$,

$$\geq C(\log C + \log(4 \log^2 C))^2 = C \log^2(4\, C\, \log^2 C) \tag{348}$$

$$\implies \frac{4\, C\, \log^2 C}{\log^2(4\, C\, \log^2 C)} \geq C. \tag{349}$$

Therefore, knowing that $T \geq 4\, C\, \log^2 C$, since $4\, C\, \log^2 C \geq 300 \log^2 75 > 8$, we have

$$\frac{T}{\log^2 T} \geq \frac{4\, C\, \log^2 C}{\log^2(4\, C\, \log^2 C)} \geq C. \tag{350}$$

□

G Additional Experiments

G.1 Environmental Details

In each of the following environments, we set the inital state distribution to be uniform, i.e. for all $s \in \mathcal{S}$, $\rho(s) = \frac{1}{S}$.

Cliff World (Sutton & Barto, 2018, Example 6.6): The environment consists of 21 states and 4 actions. The objective is for an agent to each the goal state while avoiding a cliff. If the agent falls into the chasm, the agent receives a reward of -100. If the agent reaches the goal, the agent receives a reward of $+1$. All other rewards are 0. In this environment $\gamma = 0.9$.

Deep Sea Treasure (Osband et al., 2019): The environment consists 25 states and 2 actions. The agent begin from the top-left corner of the grid and descends one row per each time it takes an action. The goal of the agent is to stay left in order to reach the treasure. If the agent transitions to the right, it receives a reward of -0.02. Otherwise if the agent reaches the treasure, it receives a reward of $+1$. In this environment $\gamma = 0.9$.

Flat Grad (Agarwal et al., 2021): The environment consists 22 states and 4 actions. The agent begin from the left and the objective is for the agent to reach the goal on the far right. For each state, only one action moves the agent to the right while all other actions causes the agent to remain in the same state. The agent only receives a sparse reward of $+1$ when it reaches the goal. In this environment $\gamma = \frac{22}{23}$.

G.2 Average Run-time Experiments

We additionally show the average runtime of the compared methods in Figure 1.

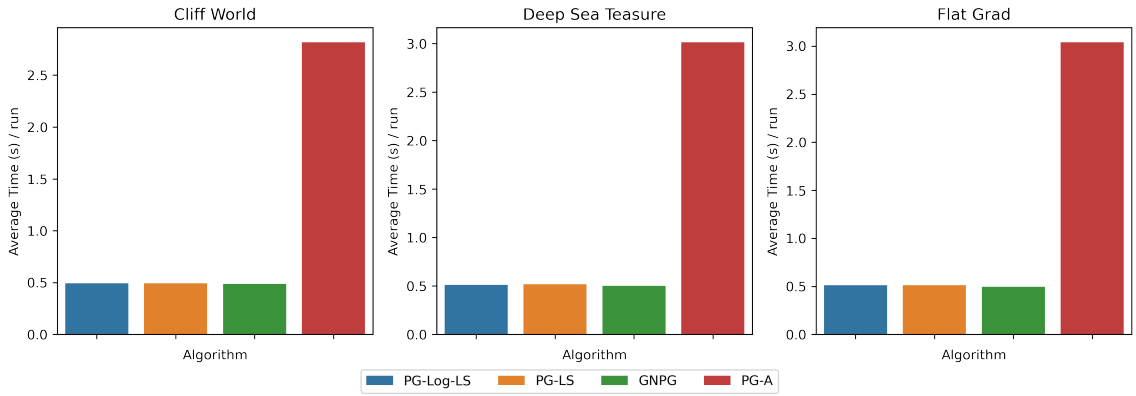

Figure 6: We compare softmax PG that (i) uses a step-size that satisfies the Armijo condition in Equation (1) (denoted as `PG-LS`), (ii) uses a step-sizes that satisfies the Armijo condition on the log-loss in Equation (3) (`PG-Log-LS`) to GNPG (`GNPG`) and PG-A (`PG-A`). The figure plots the average runtime (in seconds per run) over 50 runs for each optimization method for across all environments. Although the run time `PG-LS` and `PG-Log-LS` are longer, the methods are able to converge faster than `GNPG`. This justifies the use of line-search despite the marginal increase of runtime.

H Extra Lemmas

For completeness, we append external lemmas here.

H.1 Smoothness

Lemma 24 (Lemma 2 in Mei et al. (2020)). *$\forall r \in [0,1]^A$ $\theta \mapsto \langle \pi_\theta, r\rangle$ is $\frac{5}{2}$-smooth.*

Lemma 25 (Lemma 14 in (Mei et al., 2020)). *$\theta \to -\langle \pi_\theta, \log \pi_\theta\rangle$ is $5\,(1+\log K)$-smooth.*

Lemma 26. *$\theta \to \langle \pi_\theta, r - \tau \log \pi_\theta\rangle$ is $\frac{5}{2} + \tau\, 5\,(1+\log K)$-smooth.*

Proof. By Lemma 24 and Lemma 25. □

Lemma 27 (Lemma 7 in Mei et al. (2020)). *$\theta \to V^{\pi_\theta}(\rho)$ is $\frac{8}{(1-\gamma)^3}$-smooth.*

Lemma 28 (Lemmas 7 and 14 in (Mei et al., 2020)). *$\theta \to V^{\pi_\theta}(\rho) + \tau\, \mathbb{H}(\pi_\theta)$ is $\frac{8+\tau\,(4+8\,\log A)}{(1-\gamma)^3}$-smooth.*

Lemma 29 (Lemma 2 in (Mei et al., 2021b)). *In the bandits setting, for any $r \in [0,1]^A$, $\theta \to \langle \pi_\theta, r\rangle$ is 3-non-uniform smooth.*

Lemma 30 (Lemma 6 in (Mei et al., 2021b)). *In the tabular MDP setting, assuming* $\min_{s\in\mathcal{S}}\rho(s)>0$, $\theta\to V^{\pi_\theta}(\rho)$ *is* C*-non-uniform smooth with where* $C:=\left[3+\frac{2\,C_\infty-(1-\gamma)}{(1-\gamma)\,\gamma}\right]\sqrt{S}$ *and* $C_\infty:=\max_\pi\left\|\frac{d_\rho^\pi}{\rho}\right\|_\infty\leq\frac{1}{\min_s\rho(s)}<\infty$.

H.1.1 Non-uniform Łojasiewicz condition

Lemma 31 (Lemma 3 in Mei et al. (2020)). *Let* $\pi^*:=\max_{\pi\in\Pi}\langle\pi,r\rangle$. *Then*

$$\left\|\frac{d\langle\pi_\theta,r\rangle}{d\theta}\right\|_2\geq C(\theta)\,\langle\pi^*-\pi_\theta,r\rangle \tag{351}$$

where $C(\theta):=\pi_\theta(a^*)$.

Lemma 32 (Lemma 8 in Mei et al. (2020)). *Let* $V^*(\rho):=\max_{\pi\in\Pi}V^\pi(\rho)$. *Then*

$$\left\|\frac{\partial V^{\pi_\theta}(\rho)}{\partial\theta}\right\|_2\geq C(\theta)\,(V^*(\rho)-V^{\pi_\theta}(\rho)) \tag{352}$$

where $C(\theta):=\frac{\min_s\pi_\theta(a^*(s)\,|\,s)}{\sqrt{S}\left\|\frac{d_\rho^{\pi^*}}{d_\rho^{\pi_\theta}}\right\|_\infty}$.

Lemma 33 (Proposition 5 in (Mei et al., 2020)). *In the bandits setting, the non-uniform Łojasiewicz condition is*

$$\left\|\frac{d\langle\pi_\theta,(r-\tau\log\pi_\theta)\rangle}{d\theta}\right\|_2\geq C_\tau(\theta)\left(\mathbb{E}_{a\sim\pi_\tau^*}[r(a)-\tau\log\pi_\tau^*]-\mathbb{E}_{a\sim\pi_\theta}[r(a)-\tau\log\pi_\theta]\right)^{\frac{1}{2}} \tag{353}$$

with

$$C_\tau(\theta):=\sqrt{2\tau}\,\min_a\pi_\theta(a). \tag{354}$$

Lemma 34 (Lemma 15 in (Mei et al., 2020)). *In the tabular MDP setting, supposing* $\rho(s)>0$ *for all states* $s\in\mathcal{S}$, *the non-uniform Łojasiewicz condition is*

$$\left\|\frac{\partial\tilde{V}_\tau^{\pi_\theta}(\rho)}{\partial\theta}\right\|_2\geq C_\tau(\theta)\,\left[\tilde{V}_\tau^*(\rho)-\tilde{V}_\tau^{\pi_\theta}(\rho)\right]^{\frac{1}{2}} \tag{355}$$

with

$$C_\tau(\theta):=\frac{\sqrt{2\tau}}{\sqrt{S}}\,\min_s\sqrt{\rho(s)}\,\min_{s,a}\pi_\theta(a|s)\left\|\frac{d_\rho^{\pi_\tau^*}}{d_\rho^{\pi_\theta}}\right\|_\infty^{-\frac{1}{2}}. \tag{356}$$

H.2 Stochastic Policy Gradients

Lemma 35 (Lemma 5 from (Mei et al., 2021a)). *Let* $\hat{r}$ *be the IS estimator using on-policy sampling* $a\sim\pi_\theta(\cdot)$. *Then stochastic softmax PG estimator is:*
Unbiased*:* $\mathbb{E}_{a\sim\pi_\theta}\left[\nabla\widetilde{f}(\theta)\right]=\nabla f(\theta)$
Bounded Variance*:* $\mathbb{E}_{a\sim\pi_\theta}\left\|\nabla\widetilde{f}(\theta)\right\|_2^2\leq 2\Rightarrow\sigma^2:=\mathbb{E}_{a\sim\pi_\theta}\left[\nabla\widetilde{f}(\theta)-\nabla f(\theta)\right]=\mathbb{E}_{a\sim\pi_\theta}\left\|\nabla\widetilde{f}(\theta)\right\|_2^2-\mathbb{E}_{a\sim\pi_\theta}\|\nabla f(\theta)\|_2^2\leq 2$.

Lemma 36 (Lemma 11 from (Mei et al., 2021a)). *Let $\hat{Q}^{\pi_\theta}$ be the IS estimator using on-policy sampling $a(s) \sim \pi_\theta(\cdot|s)$. Then stochastic softmax PG estimator is:*
***Unbiased**:* $\mathbb{E}\Big[\nabla \widetilde{f}^{\tau}(\theta)\Big] = \nabla f^{\tau}(\theta)$.
***Bounded Variance**:* $\mathbb{E}\Big\|\nabla \widetilde{f}(\theta)\Big\|_2^2 \leq \frac{2\,S}{(1-\gamma)^4} \Rightarrow \sigma^2 := \mathbb{E}\Big[\nabla \widetilde{f}(\theta) - \nabla f(\theta)\Big] \leq \frac{2\,S}{(1-\gamma)^4}$.

Lemma 37 (Lemma 3 and Lemma 4 from (Ding et al.)). *Let $\hat{Q}_\tau^{\pi_\theta}$ be the entropy regularized IS estimator using on-policy sampling $a(s) \sim \pi_\theta(\cdot|s)$. Then stochastic softmax PG estimator using entropy regularization is:*
***Unbiased**:* $\mathbb{E}\Big[\nabla \widetilde{f}^{\tau}(\theta)\Big] = \nabla f^{\tau}(\theta)$.
***Bounded Variance**:* $\mathbb{E}\Big\|\nabla \widetilde{f}^{\tau}(\theta) - \mathbb{E}[\nabla \widetilde{f}^{\tau}(\theta)]\Big\|_2^2 \leq \sigma^2$, *where* $\sigma^2 = \frac{8}{(1-\gamma)^2}\left(\frac{1+(\tau\,\log A)^2}{(1-\gamma^{1/2})^2}\right)$.

Lemma 38 (Instantiation of Lemma 37 in the bandits setting). *Let $\hat{r}$ be the entropy regularized IS estimator using on-policy sampling $a \sim \pi_\theta(\cdot)$. Then stochastic softmax PG estimator using entropy regularization is:*
***Unbiased**:* $\mathbb{E}\Big[\nabla \widetilde{f}^{\tau}(\theta)\Big] = \nabla f^{\tau}(\theta)$.
***Bounded Variance**:* $\mathbb{E}\Big\|\nabla \widetilde{f}^{\tau}(\theta) - \mathbb{E}[\nabla \widetilde{f}^{\tau}(\theta)]\Big\|_2^2 \leq \sigma^2$, *where* $\sigma^2 = 8\,(1 + (\tau\,\log A)^2)$.