OPT2023: 15th Annual Workshop on Optimization for Machine Learning

Practical Principled Policy Optimization for Finite 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), Coordinated Science Laboratory (CSL), UIUC

Sharan Vaswani VASWANI.SHARAN@GMAIL.COM
Simon Fraser University

Abstract

We consider (stochastic) softmax policy gradient (PG) methods for finite Markov Decision Processes (MDP). While the PG objective is not concave, recent research has used smoothness and gradient dominance to achieve convergence to an optimal policy. However, these results depend on having extensive knowledge of the environment, such as the optimal action or the true mean reward vector, to configure the algorithm parameters. This makes the resulting algorithms impractical in real applications. To alleviate this problem, we propose PG methods that employ an Armijo line-search in the deterministic setting and an exponentially decreasing step-size in the stochastic setting. We demonstrate that these proposed algorithms offer similar theoretical guarantees as previous works but now do not require the knowledge of oracle-like quantities. Furthermore, we apply the similar techniques to develop practical, theoretically sound entropy-regularized methods for both deterministic and stochastic settings. Finally, we empirically compare the proposed methods with previous approaches in single-state MDP environments.

Keywords: Reinforcement Learning, Policy Gradient, Non-convex Optimization

1. Introduction

Stochastic policy gradient (PG) methods have played a vital role in the achievements of deep reinforcement learning (RL) [6, 19, 20]. Zhang et al. [24] first proved convergence results for stochastic PG methods to locally-optimal policies. However, it's only been recently that these methods have been rigorously proven to demonstrate convergence to a globally optimal policy in the tabular setting [1]. Despite the non-concave nature of the PG objective, recent research has harnessed concepts like smoothness and gradient dominance to achieve convergence towards an optimal policy. Specifically, for *softmax policy gradient* methods, recent studies have established global convergence rates in both deterministic [3, 10, 11] and stochastic settings [11, 13, 14, 23]. . Furthermore, the use of regularization techniques such as entropy [5, 10] or log-barrier [1] has also been studied. These approaches have been shown to expedite the convergence rate, at the cost of additional bias in the resulting policies.

While these convergence results are notable, the methods that stem from them are impractical for real applications. This impracticality arises from the methods' dependence on oracle-like knowledge of the environment, which includes factors such as the concentrability coefficient, the optimal action,

the true mean reward vector, and even access to the full gradient in stochastic settings. The need for this oracle-like knowledge renders previous PG methods ineffective because there is already sufficient information to derive an optimal policy. Our objective is to design practical approaches while retaining similar theoretical guarantees. To this end, we make the following contributions.

Contribution 1: In Section 3, we consider the deterministic setting and present a practical PG method employing Armijo line-search [2] to determine the step-sizes, thus enabling the utilization of the objective's local smoothness. In the worst-case, the method achieves a convergence rate of $\mathcal{O}(1/T)$, and under more practical conditions, it can attain a linear convergence rate.

Contribution 2: In Section 4, we consider the stochastic setting for which previous approaches relied on various oracle-like quantities to choose the step-size [11, 14, 23]. To design a practical algorithm that adapts to the amount of stochasticity, we utilize exponentially decreasing step-sizes [9]. These step-sizes have been demonstrated to achieve desirable convergence rates for smooth, non-convex functions for which the Polyak-Lojasiewicz (PL) condition [16] is satisfied [8] but have not been analyzed for PG methods. We extend the use and analysis of these step-sizes to encompass the more general gradient domination case, which is of particular interest in the context of RL. Following [14], we make use of the strong growth condition (SGC) [18] which implies that the variance in the stochastic gradients decreases as we approach a stationary point. Under this condition, Mei et al. [14] prove faster convergence, but require using typically unknown problem-dependent quantities. By utilizing exponential step-sizes, we demonstrate that the same PG algorithm can achieve an $\tilde{\mathcal{O}}(1/T + 1/T^{1/3})$ convergence rate, which smoothly transitions between the deterministic and stochastic settings, and does not require any unknown quantities.

Contribution 3: In Section 5, we consider maximizing the entropy-regularized objective in the deterministic setting. Since the entropy regularization introduces a bias, prior work [10] decays the entropy at an appropriate rate. However, prior strategies again require (typically unavailable) information about the environment to set the algorithm parameters. We introduce a multi-stage algorithm to iteratively reduce the entropy regularization. This results in convergence to the globally optimal policy at an $\mathcal{O}(1/\epsilon)$ rate, while eliminating the reliance on unknown quantities. In Section 6, we combine the multi-stage approach with exponential step-sizes and design a practical theoretically principled algorithm to maximize the entropy-regularized objective in the stochastic setting.

Contribution 4: Finally, in Appendix G, we experimentally benchmark the proposed algorithms on synthetic single-state MDPs. Our empirical results indicate that the proposed practical algorithms have comparable performance as baselines that require oracle-like knowledge.

2. Problem Formulation and Background

An infinite-horizon discounted Markov Decision Process (MDP) [17] 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 set of transition probability functions, $\rho \in \Delta_{\mathcal{S}}$ 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. In this work, we consider the setting when the policy $\pi_\theta : \mathcal{S} \to \Delta_{\mathcal{A}}$ is parameterized with the softmax function i.e. $\pi_\theta(a|s) = {\exp(\theta(a,s))}/{\sum_{a' \in \mathcal{A}} \exp(\theta(a',s))}$ given the logits $\theta \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$. For a given policy π_θ, *action-value function* $Q^{\pi_\theta} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined as: $Q^{\pi_\theta}(s, a) := \mathbb{E}\big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\big]$, with $s_0 = s$, $a_0 = a$ and for $t \geq 1$, $s_{t+1} \sim p(\cdot|s_t, a_t)$ and $a_{t+1} \sim \pi_\theta(\cdot|s_t)$, the *value function* $V^{\pi_\theta} : \mathcal{S} \to \mathbb{R}$ is defined such that $V^{\pi_\theta}(s) = \mathbb{E}_{a \sim \pi_\theta(\cdot|s)}[Q^{\pi_\theta}(s, a)]$, and the *discounted state visitation distribution* $d_{s_0}^{\pi_\theta} \in \Delta_{\mathcal{S}}$ is defined

such that $d_{s_0}^{\pi_\theta} := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \Pr[s_t = s\,|s_0, \pi_\theta]$ where $\Pr[s_t = s\,|s_0, \pi_\theta]$ denotes the probability of encountering state s at time t under policy π_θ.

The objective is find a policy that maximizes $J(\pi_\theta) := \mathbb{E}_{s\sim\rho}[V^{\pi_\theta}(s)]$. Let us denote the special case when MDP has $|\mathcal{S}| = 1$ with stochastic rewards as the bandit setting. In the bandit setting, the objective is $f(\theta) := \mathbb{E}[\pi_\theta^\top r]$, where r is sampled from some unknown distribution and reduces to the single-state MDP setting when the rewards are deterministic. To cover all presented settings, we will abstract the objective as $f(\theta)$ and note that it is L-smooth[1] and upper-bounded by f^*. Finally, we note that $f(\theta)$ is non-concave [1], but satisfies a non-uniform Łojasiewciz condition $\|\nabla f(\theta)\|_2 \geq C(\theta)\,|f^* - f(\theta)|^{1-\xi}$ for some $\xi \in [0, 1]$ [10]. When $C(\theta)$ is constant, and $\xi = 1/2$, this condition matches the well studied PL condition [16]. Details of $C(\theta)$ for each abstract environment can be found in Appendix A. We define $\mu := \inf_{t\geq 1} C(\theta_t)$.

3. Policy Gradient

We first consider the deterministic setting and consider Softmax PG with the following update,

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

In this setting, the non-uniform Łojasiewciz condition is satisfied i.e.$\|\nabla f(\theta)\|_2 \geq C(\theta)\,|f^* - f(\theta)|$ and is required to prove global convergence guarantees [10]. In order for softmax PG to be able to adapt to the objective's local smoothness, we propose to use Armijo line-search to set the step-size. Armijo line-search [2] searches for a prospective step-size $\tilde{\eta}_t$ until it satisfies the following equation: $f(\theta_t + \tilde{\eta}_t \nabla f(\theta_t)) \geq f(\theta_t) + h\tilde{\eta}_t \|\nabla f(\theta_t)\|_2^2$ where $h \in (0, 1)$ is a hyper-parameter. The line-search is guaranteed to return a step-size η_t that satisfies $\eta_t \geq \min\{2(1-h)/L, \eta_{\max}\}$, where $\eta_{\max}$ is the maximum allowable step-size and L is the smoothness of f. In Appendix B.1, we prove that the resulting algorithm can achieve a convergence rate of $\mathcal{O}(1/T)$. This rate is consistent with the one obtained when employing a fixed step-size of $\eta_t = 1/L$ [10].

Although softmax PG with Armijo line-search results in an $\mathcal{O}(1/T)$ convergence in the worst case, it can result in faster convergence in practice. This is because the objective satisfies a non-uniform smoothness property, making it possible to iteratively increase the resulting step-size. The Geometry-Aware Normalized Policy Gradient (GNPG) approach introduced in [11] is able to exploit this non-uniform smoothness and exhibits a convergence rate of $\mathcal{O}(\exp(-\mu T))$. However, for MDPs, GNPG requires the knowledge of intricate constants like $C_\infty = \max_\pi \left\| d_\mu^\pi / \mu \right\|_\infty$ to determine the step-size. In contrast, softmax PG with Armijo line-search does not require such information but can exploit the non-uniform smoothness to achieve fast convergence towards the optimal policy. This result is presented in Theorem 1 with a non-asymptotic rate and is proved in Appendix B.2.

Theorem 1 *Using Update 1 and an increasing line-search of* $\eta_{\max_k} = r^k \eta_{\max_0}$ *where* k *is the number of times* $\eta_{\max}$ *has been returned by Armijo line search, then we obtain a convergence rate of*

$$f^* - f(\theta_T) \leq \frac{1}{\mu}\left(\frac{L}{2(1-h)}\frac{1}{T-k} + \frac{1}{\eta_{\max_0} r^{k-1}}\mathbb{1}\{k > 0\}\right) \tag{1}$$

In the worst-case, Armijo line-search requires back-tracking in each iteration meaning that $k = 0$ and we recover the $\mathcal{O}(1/T)$ convergence rate. However, in our experiments, we observe that k is large and the resulting convergence is linear.

1. For definitions of smoothness, and the specific non-uniform Łojasiewciz conditions see Appendix A

4. Stochastic Policy Gradient

In the stochastic setting, we construct a gradient estimator using importance sampling (IS) that is unbiased and has bounded variance. For illustrative purposes, we consider the bandit setting. For each iteration $t \in [1, T]$, sample an action $a_t \sim \pi_{\theta_t}$ and construct the IS reward estimate $\hat{r}_t(a) = \frac{\mathbb{1}\{a_t = a\}}{\pi_{\theta_t}(a)} r(a)$ for each $a \in \mathcal{A}$. The stochastic objective is $\widetilde{f}(\theta_t) = \pi_{\theta_t}^\top \hat{r}_t$ and the resulting gradient estimator satisfies $\mathbb{E}[\nabla \widetilde{f}(\theta)] = \nabla f(\theta)$ and $\mathbb{E}\left\|\nabla \widetilde{f}(\theta) - \nabla f(\theta)\right\|_2^2 \leq \sigma^2$ [11]. We now consider the following stochastic gradient updates,

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

Yuan et al. [23] have also considered the stochastic setting under the same assumptions, but their method required the knowledge of $\mu = \pi_\theta(a^*)$ when setting the step-size; hence, knowledge of the optimal action is required. A faster $\mathcal{O}(1/\sqrt{T})$ rate can be achieved but requires the true gradient $\|\nabla f(\theta_t)\|$ when setting the step-size [11]. By using an exponentially decaying step-size [9] we can match the $\tilde{\mathcal{O}}(1/T + \sigma^2/T^{1/3})$ rate in [23] without the knowledge of μ. The following theorem is proved in Appendix C.1.

Theorem 2 *Given* $\epsilon > 0$, *assuming* $\mu := \inf_{t \geq 1} C(\theta_t) > 0$, $T \geq 3$, *using Update 2 with* $\eta_0 = \frac{1}{L}$ *and* $\eta_t = \eta_0 \, \alpha^t$ *where* $\alpha = \left(\frac{\beta}{T}\right)^{\frac{1}{T}}$, $\beta > 0$ *results in the following convergence: If* $\mathbb{E}[f^* - f(\theta_t)] \geq \epsilon$ *for all* $t \in [1, T]$ *then,*

$$\mathbb{E}[f^* - f(\theta_t)] \leq \frac{5LC(\beta)}{e^2 \delta^2 \mu^2} \frac{\ln^2\left(\frac{T}{\beta}\right) \sigma^2}{T} + C(\beta) \exp\left(\frac{-0.69\delta\mu}{L}\left(\frac{T}{\ln \frac{T}{\beta}}\right)\right)(f^* - f(\theta_t)) \quad (2)$$

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

We note that the proof requires the assumption $\mu := \inf_{t \geq 1} C(\theta_t) > 0$ which we verify experimentally, and leave a formal proof of this claim to future work. Additionally, although β is a hyperparameter, it does not depend on any problem-dependent constants. To obtain ϵ-convergence, let $T = \max\{\mathcal{O}(\epsilon^{-3}), \mathcal{O}(\epsilon^{-1} \log \epsilon^{-1})\}$. This results in a convergence rate of $\tilde{\mathcal{O}}(1/T + \sigma^2/T^{1/3})$.

The previous result assumes that the variance σ^2 is a constant. However, it has been observed that the noise decreases as we get closer to a stationary point and the policy become more deterministic, meaning that $\sigma_t^2 \to 0$ as $t \to \infty$. Using the same gradient estimator with IS, Mei et al. [13] has shown that the SGC [18] holds, implying that $\mathbb{E}_t\left\|\nabla \widetilde{f}(\theta_t)\right\|_2^2 \leq \rho \|\nabla f(\theta_t)\|$ for a problem-dependent $\rho > 0$. Using Update 4 in the bandit setting and knowledge of ρ, a faster convergence rate of $\mathcal{O}(1/T)$ rate can be achieved [14]. However, the resulting algorithm requires setting the step-size proportional to ρ that has a dependence on the *reward gap* $\Delta := \min_{i \neq j} |r(i) - r(j)|$. This implies that the true mean reward vector is required when setting the step-size of the algorithm, rendering it ineffective in most practical cases. In Appendix C.2, we show that the Δ dependence on ρ is necessary, meaning that it is unlikely to achieve fast convergence rates by exploiting SGC, while still being practical. Hence, we aim to develop a practical algorithm that can automatically adapt to both ρ and σ_t^2. We use the same exponentially decreasing step-sizes for this and analyze its convergence in Appendix C.4.

Theorem 3 *Under the same assumptions as Theorem 2 and SGC, Update 2 with exponential step-sizes has the following convergence: if* $\mathbb{E}[f^* - f(\theta_t)] \geq \epsilon$ *for all* $t \in [1, T]$ *then*

$$\mathbb{E}[f^* - f(\theta_T)] \leq C_1 \exp\left(-\frac{\alpha\, T}{\kappa\, \ln(T)}\right) \mathbb{E}[f^* - f(\theta_1)] + \frac{4\,\rho\, C_2\, L}{\epsilon^2} \frac{\sum_{t=1}^{T_0-1} \mathbb{E}[f^* - f(\theta_t)]}{T^2} \tag{3}$$

where $\kappa = \frac{2}{\mu\,\epsilon\,\eta_0}$, $C_1 := \frac{2\beta}{\kappa \ln(T/\beta)}$, $C_2 := \exp\left(\frac{2\beta}{\kappa \ln(T/\beta)}\right) \frac{4\kappa^2}{e^2\alpha^2} \ln^2(T/\beta)$, $T_0 := T \max\left\{\frac{\ln(4\rho\eta_0)}{\ln(T/\beta)}, 0\right\}$, $\eta_0 = \frac{1}{18}$ *and* ρ *and* L *are known problem dependent constants. Otherwise,* $\min_{t\in[1,T]} \mathbb{E}[f^* - f(\theta_t)] \leq \epsilon$

In comparison to Theorem 2, the algorithm achieves an $\tilde{\mathcal{O}}\left(1/T + T_0^{1/3}/T^{2/3}\right)$ convergence rate. If $\eta_0 \leq \frac{1}{4\rho}$, $T_0 = 0$ and we obtain a "fast" $\mathcal{O}(1/T)$ rate. In the worst-case, if we choose η_0 to be large, $T_0 = \mathcal{O}(T)$, resulting in the "slow" $\tilde{\mathcal{O}}(1/T^{1/3})$ rate. Hence, the resulting algorithm can is robust to ρ and can interpolate between the "slow" and "fast" rates. In Appendix C.3, we show that the SGC property is not limited to bandits and also holds for the general MDP setting.

5. Policy Gradient With Entropy Regularization

For the following sections, we restrict our analysis to the single-state MDP setting, but note that the results can also be extended to the general MDP setting. The entropy regularized objective for single-state MDP is defined as: $f^\tau(\theta) := \pi_\theta^\top (r - \tau \log \pi_\theta)$ where $\tau \geq 0$ denotes the strength of the entropy regularization. In this case, $\nabla f^\tau(\theta) := H(\pi_\theta)(r - \tau \log \pi_\theta)$ where $H(\pi) := \text{diag}(\pi_\theta) - \pi_\theta \pi_\theta^\top \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{A}|}$. The entropy regularized objective satisfies following non-uniform Lojasiewciz condition [10]: $\|\nabla f^\tau(\theta)\|_2 \geq C(\theta)\, |f^{*\tau} - f^\tau(\theta)|^{1/2}$ where $C(\theta) := \sqrt{2\tau}\, c$ and $c := \min_a \pi_\theta(a)$. This property can be interpreted as the non-uniform variant of the classical PL condition [7]. In the subsequent sections, we assume that $\mu := \inf_{t\geq 1} C(\theta_t) > 0$ across the iterations and $\tau \leq 1$. The update for the entropy regularized objective is given as,

Update 3 *(Softmax PG with Entropy Regularization, True Gradient)*

$$\theta_{t+1} = \theta_t + \nabla f^\tau(\theta_t) = \theta_t + \eta_t H(\pi_\theta)(r - \tau \log \pi_\theta)$$

In Appendix D.1, we prove softmax PG with entropy regularization and Armijo line-search will converge to a biased optimal policy with an $\mathcal{O}(\exp(-\mu\, T))$ convergence rate which also matches the rate when employing a fixed step-size $\eta_t = 1/L^\tau$ [10]. The resulting policy is biased due to the presence of entropy regularization. In order for entropy regularized objectives to converge to the globally optimal policy, $\tau \to 0$. Using Update 3, prior work [10] used a two-stage approach to decay τ, but the resulting algorithm requires knowledge of Δ, rendering the method ineffective. Consequently, in Theorem 4 we establish a rate of convergence to an ϵ-neighbourhood of the optimal policy by choosing a sufficiently small τ that does not depend on Δ and prove the following result in Appendix D.2.

Theorem 4 *Setting* $\eta_t = 1/L^\tau$, $\tau = \epsilon / \left(2\, W\left((|\mathcal{A}|-1)/e\right)\right)$ *where* $W(x)$ *is the Lambert function, Update 3 achieves* ϵ*-suboptimality after* $\mathcal{O}\left(1/\epsilon \log 1/\epsilon\right)$ *iterations.*

We note that in the proof we treat $c := \min_a \pi_\theta(a)$ and hence μ as a constant. We conjecture that there is a hidden strong dependence on ϵ in c, and that this dependence can be weakened by

proving a tighter bound on μ. Unfortunately, using a fixed τ leads to a $\mathcal{O}\left(1/\epsilon \log 1/\epsilon\right)$ convergence rate that is slower than PG methods without the presence of entropy. Thus, we present Algorithm 1 that decays the entropy in stages and achieves a $\mathcal{O}\left(1/\epsilon\right)$ rate, matching the two-stage approach in [10], but without requiring the prior knowledge of Δ. Additionally, in comparison to Theorem 4, the algorithm do not require ϵ to be set advance. We analyze Algorithm 1 and prove it's convergence rate in Appendix D.3.

6. Stochastic Policy Gradient With Entropy Regularization

Following Section 4, we construct a gradient estimator using IS for the entropy regularized objective. This estimator $\nabla \widetilde{f}^{\tau}(\theta_t)$ has been shown to be unbiased and has bounded variance [5], i.e. $\mathbb{E}\left\|\nabla \widetilde{f}^{\tau}(\theta_t) - \mathbb{E}[\nabla \widetilde{f}^{\tau}(\theta_t)]\right\|_2^2 \leq b := 8(1 + (\tau \log |\mathcal{A}|)^2)$. We consider the following update,

Update 4 *(Stochastic Softmax PG with Entropy, Importance Sampling)*

$$\theta_{t+1} = \theta_t + \nabla \widetilde{f}^{\tau}(\theta_t) = \theta_t + \eta_t H(\pi_\theta)(\hat{r} - \tau \log \pi_\theta)$$

Under the same setting, prior work [4] proposes a two-stage approach that converges to a globally optimal policy by modifying the batch size to counteract the variance at a $\tilde{\mathcal{O}}(1/\epsilon^2)$ convergence rate, but requires knowledge of the optimal policy to set the algorithm parameters. Using the same rationale as in Section 4, using Update 4 with fixed τ and in conjunction with an exponentially decaying step-size to mitigate the variance, we can converge to an ϵ-neighbourhood of a globally optimal policy after $\tilde{\mathcal{O}}\left(1/\epsilon + b/\epsilon^3\right)$ iterations. We prove the following theorem in Appendix E.1.

Theorem 5 *Assuming* $c := \inf_{t \geq 0} \min_a \pi_{\theta_t}(a) > 0$. *Using Update 4 with* $\tau = \epsilon / \left(2\left(W\left(|\mathcal{A}|-1/e\right) + \log |\mathcal{A}|\right)\right)$ *and using exponential decreasing step-size* $\eta_t = \eta_0\, \alpha^t$, *where* $\eta_0 = 1/L^{\tau}$, *achieves* ϵ*-suboptimality after* $\tilde{\mathcal{O}}\left(1/\epsilon + b/\epsilon^3\right)$ *iterations.*

Finally, we extend the multi-stage approach of Algorithm 1 to the stochastic setting, resulting in Algorithm 3 in Appendix E.2. In Theorem 19 in Appendix E.2, we prove an $\tilde{\mathcal{O}}\left(1/\epsilon + b/\epsilon^3\right)$ convergence rate for the resulting algorithm. This matches the corresponding rate in Theorem 2.

7. Discussion

We propose and analyze (stochastic) PG methods that can be used *in the wild* since they do not require oracle-like knowledge to set algorithmic parameters. In comparison to prior work, we achieve similar theoretical guarantees, but we are able to use our proposed algorithms in practice. We additionally propose practical multi-stage methods to decay PG methods with entropy regularization that allows the algorithms to converge to a globally optimal policy. We experimentally validate our results and display competitive results compared to prior work. For future work, we aim to remove the assumption in the stochastic setting that the non-uniform Łojasiewciz condition $\inf_{t \geq 1} C(\theta_t) > 0$ and tighten the non-condition for the entropy regularized setting.

References

[1] 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.

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

[3] Jalaj Bhandari and Daniel Russo. On the linear convergence of policy gradient methods for finite mdps. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2386–2394. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/bhandari21a.html.

[4] Yuhao Ding, Junzi Zhang, and Javad Lavaei. Beyond exact gradients: Convergence of stochastic soft-max policy gradient methods with entropy regularization. *arXiv preprint arXiv:2110.10117*, 2021.

[5] Yuhao Ding, Junzi Zhang, and Javad Lavaei. Local analysis of entropy-regularized stochastic soft-max policy gradient methods. In *2023 European Control Conference (ECC)*, pages 1–8. IEEE, 2023.

[6] 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.

[7] 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*, pages 795–811. Springer, 2016.

[8] Jonathan Wilder Lavington, Sharan Vaswani, Reza Babanezhad Harikandeh, Mark Schmidt, and Nicolas Le Roux. Target-based surrogates for stochastic optimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 18614–18651. PMLR, 2023. URL https://proceedings.mlr.press/v202/lavington23a.html.

[9] 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*, pages 6553–6564. PMLR, 2021.

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

[11] 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, 2021.

[12] 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*, pages 7555–7564. PMLR, 2021.

[13] 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.

[14] 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, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 24325–24360. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/mei23a.html.

[15] 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, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 24325–24360. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/mei23a.html.

[16] 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.

[17] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[18] Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013.

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

[20] 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, page 1057–1063, Cambridge, MA, USA, 1999. MIT Press.

[21] Sharan Vaswani, Aaron Mishkin, Issam Laradji, Mark Schmidt, Gauthier Gidel, and Simon Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. *Advances in neural information processing systems*, 32, 2019.

[22] Sharan Vaswani, Benjamin Dubois-Taine, and Reza Babanezhad. Towards noise-adaptive, problem-adaptive (accelerated) stochastic gradient descent. In *International Conference on Machine Learning*, pages 22015–22059. PMLR, 2022.

[23] Rui Yuan, Robert M. Gower, and Alessandro Lazaric. A general sample complexity analysis of vanilla policy gradient, 2022.

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

Supplementary Material

Organization of the Appendix

Appendix A. Definitions

A function f is L-smooth if for all v and w

$$f(v) \leq f(w) + \langle \nabla f(w), v - w \rangle + \frac{L}{2}\|v - w\|_2^2 \tag{4}$$

The non-uniform Łojasiewciz condtion of degree ξ for $\xi \in [0, 1]$ is defined as

$$\|\nabla f(\theta_t)\| \geq C(\theta)|f^* - f(\theta)|^{1-\xi} \tag{5}$$

Setting	$f(\theta)$	$C(\theta)$	ξ	L
Single-State MDP	$\pi_\theta^\top r$	$\pi_\theta(a^*)$	0	$5/2$
General MDP	$V^{\pi_\theta}(\rho)$	$\frac{\min_s \pi_\theta(a^*(s)\|s)}{\sqrt{\|\mathcal{S}\|}\left\|d_\rho^{\pi^*}/d_\mu^{\pi_\theta}\right\|_\infty}$	0	$8/(1-\gamma)^3$
Single-State MDP with Entropy Regularization	$\pi_\theta^\top(r - \tau \log \pi_\theta)$	$\sqrt{2\tau}\min_a \pi_\theta(a)$	$1/2$	$5/2 + 5\tau(1 + \log K)$

Table 1: Summary of abstracted objectives, smoothness, and non-uniform Łojasiewciz condition

Appendix B. Policy Gradient Proofs

B.1. Policy Gradient with Armijo line-search

In the following theoerm, we show that PG with Armijo line-search obtains the same $\mathcal{O}(1/T)$ rate as using a constant fixed step-size of $\eta_t = 1/L$ [10].

Theorem 6 . *Using Update 1, softmax policy gradient and Armijo line-search converges to a globally optimal solution with a rate of* $\mathcal{O}(\frac{1}{T})$.

$$f^* - f(\theta_T) \leq \max\left\{\frac{L}{2(1-h)}, \frac{1}{\eta_{\max}}\right\}\frac{1}{T\,\mu} \tag{6}$$

where L *is the smoothness of* f*,* μ *is the non-uniform PŁ-constant,* $\eta_{\max}$ *is the maximum allowable step-size, and* $h \in (0,1)$ *is a hyperparameter from Armijo line-search. In the single-state MDP setting,* $L = \frac{5}{2}$ *and* $\mu = \min_{1\leq t\leq T} \pi_{\theta_t}(a^*)$*. In the general MDP setting,* $L = \frac{8}{(1-\gamma)^3}$ *and* $\mu = \min_{1\leq t\leq T} \min_{s\in\mathcal{S}} \pi_{\theta_t}(a^*|s) \left\|\frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1} |\mathcal{S}|^{-2}$.

Proof For simplicity, we will present the proof for the single-state MDP setting, but note that this proof can easily extend to the general MDP setting.

For any L-smooth function the step-size η_t returned by the Armijo line-search is guaranteed to satisfy $\eta_{\max} \geq \eta_t \geq \min\left\{\frac{2(1-h)}{L}, \eta_{\max}\right\}$ (Lemma 1 in [21]) which implies that

$$f(\theta_{t+1}) \geq f(\theta_t) + \min\left\{\frac{2(1-h)}{L}, \eta_{\max}\right\}\|\nabla f(\theta_t)\|_2^2 \tag{7}$$

By Theorem 21, $L = \frac{5}{2}$

$$f(\theta_{t+1}) \geq f(\theta_t) + \min\left\{\frac{4(1-h)}{5}, \eta_{\max}\right\}\|\nabla f(\theta_t)\|_2^2 \tag{8}$$

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

$$f^* - f(\theta_{t+1}) \leq f^* - f(\theta_t) - \min\left\{\frac{4(1-h)}{5}, \eta_{\max}\right\}\|\nabla f(\theta_t)\|_2^2 \tag{9}$$

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

$$\delta(\theta_{t+1}) \leq \delta(\theta_t) - \min\left\{\frac{4(1-h)}{5}, \eta_{\max}\right\}\|\nabla f(\theta_t)\|_2^2 \tag{10}$$

By Theorem 29

$$\leq \delta(\theta_t) - \min\left\{\frac{4(1-h)}{5}, \eta_{\max}\right\}(\pi_{\theta_t}(a^*)\,\delta(\theta_t))^2 \tag{11}$$

Let $\mu := \inf_{t\geq 1} \pi_{\theta_t}(a^*)$

$$\leq \delta(\theta_t) - \underbrace{\mu\,\min\left\{\frac{4(1-h)}{5}, \eta_{\max}\right\}}_{\frac{1}{C}}\delta(\theta_t)^2 \tag{12}$$

Therefore, we have

$$\frac{1}{\delta(\theta_T)} = \frac{1}{\delta(\theta_0)} + \sum_{t=0}^{T-1}\left[\frac{1}{\delta(\theta_{t+1})} - \frac{1}{\delta(\theta_t)}\right] \tag{13}$$

$$= \frac{1}{\delta(\theta_0)} + \sum_{t=0}^{T-1}\left[\frac{1}{\delta(\theta_{t+1})\,\delta(\theta_t)}\left(\delta(\theta_t) - \delta(\theta_{t+1})\right)\right] \tag{14}$$

By Eq. (12)

$$\geq \frac{1}{\delta(\theta_0)} + \sum_{t=0}^{T-1}\left[\frac{1}{\delta(\theta_{t+1})\,\delta(\theta_t)}\,\frac{1}{C}\,\delta(\theta_t)^2\right] \tag{15}$$

Since $\delta(\theta_t) \geq \delta(\theta_{t+1})$

$$\geq \frac{1}{\delta(\theta_0)} + \sum_{t=0}^{T-1}\frac{1}{C} \tag{16}$$

$$= \frac{1}{\delta(\theta_0)} + \frac{T}{C} \tag{17}$$

$$\geq \frac{T}{C} \tag{18}$$

Now let us show that using Update 1 with Armijo line-search guarantees that $\mu := \inf_{t\geq 1}\pi_{\theta_t}(a^*) > 0$. We will now present an extended proof of Lemma 5 of [10] where the authors originally show that using a fixed step size of $\eta = \frac{2}{5}$ guarantees that $\inf_{t\geq 1}\pi_{\theta_t}(a^*) > 0$. We modify the proof to work for any step size returned by Armijo line-search. Let

$$c = \frac{K}{2\Delta}\left(1 - \frac{\Delta}{K}\right) \tag{19}$$

and

$$\Delta = r(a^*) - \max_{a\neq a^*} r(a) > 0 \tag{20}$$

denote the reward gap of r. We will prove that $\inf_{t\geq 1}\pi_{\theta_t}(a^*) = \min_{1\leq t\leq t_0}\pi_{\theta_t}(a^*)$, where $t_0 = \min\{t : \pi_{\theta_t}(a^* \geq \frac{d}{d+1}\}$. Note that t_0 depends only θ_1 and d. Define the following regions

$$\mathcal{R}_1 = \left\{\theta : \frac{\pi_\theta^T r}{d\theta(a^*)} \geq \frac{\pi_\theta^T r}{d\theta(a)}, \forall a \neq a^*\right\} \tag{21}$$

$$\mathcal{R}_2 = \{\theta : \pi_\theta(a^*) \geq \pi_\theta(a), \forall a \neq a^*\} \tag{22}$$

$$\mathcal{N}_c = \left\{\theta : \pi_\theta(a^*) \geq \frac{c}{c+1}\right\} \tag{23}$$

We will make the following three-part claim. **Claim 1:**

a $\mathcal{R}_1$ is a "nice" region, that if $\theta_t \in \mathcal{R}_1$ then, with any $\eta > 0$, following a gradient update (i) $\theta_{t+1} \in \mathcal{R}_1$ and (ii) $\pi_{\theta_{i+1}}(a^*) \geq \pi_{\theta_t}(a^*)$.

b We have $\mathcal{R}_2 \subset \mathcal{R}_1$ and $\mathcal{N}_c \subset \mathcal{R}_1$.

c For $\eta_t = \min\left\{\frac{2(1-h)}{L}, \eta_{\max}\right\}$, there exists a finite time $t_0 \geq 1$ such that $\theta_{t_0} \in \mathcal{N}_c$ and thus $\theta_{t_0} \in \mathcal{R}_1$, which implies that $\inf_{t\geq 1} \pi_{\theta_t}(a^*) = \min_{1\leq t\leq t_0} \pi_{\theta_t}(a^*)$

We note that **Claim 1: a)** holds for any $\eta > 0$ and **Claim 1: b)** is independent of η. Thus we will only prove **Claim 1: c)**. To show that $\pi_{\theta_t}(a^*) \to 1$ as $t \to \infty$, we will use the same convergence result as in [1], but extend it to work for any arbitrary $\eta > 0$ as long as it satisfies Theorem 20 (see Lemma C.2 in [1]). Let η_t be returned by Armijo line-search and thus satisfies

$$f(\theta_{t+1}) \geq f(\theta_t) - \min\left\{\frac{2(1-h)}{L}, \eta_{\max}\right\} \|\nabla f(\theta_t)\|_2^2 \tag{24}$$

and therefore

$$f(\theta_{t+1}) \geq f(\theta_t) \tag{25}$$

Since we have monotonic improvement and $f(\theta)$ is bounded above, by monotone convergence theorem, it must converge to some limit point. The rest of the proof follows summarily as in the proof of Theorem 5.1 in [1]. Continuing from Lemma 5 in [10], there exists $t_0 \geq 1$, such that $\pi_{\theta_{t_0}}(a^*) \geq \frac{d}{d+1}$, which implies $\theta_{t_0} \in \mathcal{N}_c \subset \mathcal{R}_1$. Therefore $\pi_{\theta_t}(a^*)$ in increasing in $\mathcal{R}_1$ and we have $\inf_t \pi_{\theta_t}(a^*) = \min_{1\leq t\leq t_0} \pi_{\theta_t}(a^*)$, where t_0 depends on initialization and c depends on the problem only. ■

B.2. Proof Of Theorem 1

Theorem 1 *Using Update 1 and an increasing line-search of $\eta_{\max_k} = r^k \eta_{\max_0}$ where k is the number of times $\eta_{\max}$ has been returned by Armijo line search, then we obtain a convergence rate of*

$$f^* - f(\theta_T) \leq \frac{1}{\mu}\left(\frac{L}{2(1-h)}\frac{1}{T-k} + \frac{1}{\eta_{\max_0} r^{k-1}} \mathbb{1}\{k > 0\}\right) \tag{1}$$

Proof The analysis is similar to the proof in Theorem 6. Suppose that we satisfy Armijo's condition for k iterations, then we will have the following update for $i \in [k]$ where the Armijo's condition is satisfied:

$$\delta(\theta_{t+1}) \leq \delta(\theta_t) - \underbrace{\mu \eta_{\max_i}}_{\frac{1}{C_i}} \delta(\theta_t)^2 \tag{26}$$

and otherwise

$$\delta(\theta_{t+1}) \leq \delta(\theta_t) - \underbrace{\mu \frac{2(1-h)}{L}, \eta_{\max}}_{\frac{1}{C}} \delta(\theta_t)^2 \tag{27}$$

for remaining $T - k$ updates. Following similar steps in Theorem 6 we have

$$\frac{1}{\delta(\theta_T)} \geq \frac{1}{\delta(\theta_0)} + \sum_{i=0}^{T-k-1} \frac{1}{C} + \sum_{i=1}^{k} \frac{1}{C_i} \tag{28}$$

$$\geq \frac{T-k}{C} + \mu\, \eta_{\max_0} \sum_{i=1}^{k-1} r^i \tag{29}$$

Since $r > 0$, we can throw away smaller terms

$$\geq \frac{T-k}{C} + \mu\, \eta_{\max_0} r^{k-1} \tag{30}$$

■

Appendix C. Stochastic Policy Gradient Proofs

C.1. Proof Of Theorem 2

Theorem 2 *Given $\epsilon > 0$, assuming $\mu := \inf_{t\geq 1} C(\theta_t) > 0$, $T \geq 3$, using Update 2 with $\eta_0 = \frac{1}{L}$ and $\eta_t = \eta_0\, \alpha^t$ where $\alpha = \left(\frac{\beta}{T}\right)^{\frac{1}{T}}$, $\beta > 0$ results in the following convergence: If $\mathbb{E}[f^* - f(\theta_t)] \geq \epsilon$ for all $t \in [1, T]$ then,*

$$\mathbb{E}[f^* - f(\theta_t)] \leq \frac{5LC(\beta)}{e^2\delta^2\mu^2} \frac{\ln^2\left(\frac{T}{\beta}\right)\sigma^2}{T} + C(\beta)\exp\left(\frac{-0.69\delta\mu}{L}\left(\frac{T}{\ln\frac{T}{\beta}}\right)\right)(f^* - f(\theta_t)) \tag{2}$$

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

Proof

Starting with the smoothness of f

$$|f(\theta_{t+1}) - f(\theta_t) - \langle \nabla f(\theta_t), \theta_{t+1} - \theta_t\rangle| \leq \frac{L}{2}\|\theta_t - \theta_t\|_2^2 \tag{31}$$

$$f(\theta_{t+1}) - f(\theta_t) - \langle \nabla f(\theta_t), \theta_{t+1} - \theta_t\rangle \geq -\frac{L}{2}\|\theta_t - \theta_t\|_2^2 \tag{32}$$

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

$$f(\theta_{t+1}) - f(\theta_t) - \eta_t\left\langle \nabla f(\theta_t), \nabla\widetilde{f}(\theta_t)\right\rangle \geq -\frac{L}{2}\eta^2\left\|\nabla\widetilde{f}(\theta_t)\right\|_2^2 \tag{33}$$

$$f(\theta_{t+1}) \geq f(\theta_t) + \eta_t\left\langle \nabla f(\theta_t), \nabla\widetilde{f}(\theta_t)\right\rangle - \frac{L}{2}{\eta_t}^2\left\|\nabla\widetilde{f}(\theta_t)\right\|_2^2 \tag{34}$$

Multiplying both sides by -1 and adding f^*

$$f^* - f(\theta_{t+1}) \leq f^* - f(\theta_t) - \eta_t\left\langle \nabla f(\theta_t), \nabla\widetilde{f}(\theta_t)\right\rangle + \frac{L}{2}{\eta_t}^2\left\|\nabla\widetilde{f}(\theta_t)\right\|_2^2 \tag{35}$$

(36)

Taking expectation with respect to the gradients on both sides

$$\underbrace{\mathbb{E}[f^* - f(\theta_{t+1})]}_{\delta(\theta_{t+1})} \leq \underbrace{\mathbb{E}[f^* - f(\theta_t)]}_{\delta(\theta_t)} - \eta_t \Big\langle \nabla f(\theta_t), \mathbb{E}\Big[\nabla \widetilde{f}(\theta_t)\Big] \Big\rangle + \frac{L\eta^2}{2} \mathbb{E}\left[\left\| \nabla \widetilde{f}(\theta_t) \right\|_2^2 \right] \tag{37}$$

By Theorem 25, the gradient is unbiased

$$= \delta(\theta_t) - \eta_t \|\nabla f(\theta_t)\|_2^2 + \frac{L\eta^2}{2} \mathbb{E}\left[\left\| \nabla \widetilde{f}(\theta_t) \right\|_2^2 \right] \tag{38}$$

$$= \delta(\theta_t) - \eta \|\nabla f(\theta_t)\|_2^2 + \frac{L\eta^2}{2} \mathbb{E}\left[\left\| \nabla \widetilde{f}(\theta_t) - \nabla f(\theta_t) + \nabla f(\theta_t) \right\|_2^2 \right] \tag{39}$$

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

$$\leq \delta(\theta_t) - \eta \|\nabla f(\theta_t)\|_2^2 + \frac{L\eta^2}{2} \mathbb{E}\left[\left\| \nabla \widetilde{f}(\theta_t) - \nabla f(\theta_t) \right\|_2^2 \right] + \frac{L\eta^2}{2} \mathbb{E}\Big[\|\nabla f(\theta_t)\|_2^2 \Big] \tag{40}$$

By Theorem 25, the variance is bounded

$$\leq \delta(\theta_t) - \eta \|\nabla f(\theta_t)\|_2^2 + \frac{L\eta^2}{2} \Big(\sigma(\theta_t)^2 + \mathbb{E}\Big[\|\nabla f(\theta_t)\|_2^2 \Big] \Big) \tag{41}$$

Since $\eta_t \leq \frac{1}{L}$

$$\leq \delta(\theta_t) - \frac{\eta}{2} \|\nabla f(\theta_t)\|_2^2 + \frac{L\eta^2}{2} \sigma(\theta_t)^2 \tag{42}$$

By Theorem 29 and the assumption that $\mu := \inf_{t \geq 1} C(\theta_t) > 0$

$$\leq \delta(\theta_t) - \frac{\eta_t}{2} \delta(\theta_t)^2 \, \mu + \frac{L\eta^2}{2} \sigma(\theta_t)^2 \tag{43}$$

For clarity, let $\mu = \mu/2$ and $\sigma = \max_t \sigma(\theta_t)$

$$\leq \delta(\theta_t) \Big(1 - \frac{\eta}{2} \delta(\theta_t) \, \mu' \Big) + \frac{L\eta^2}{2} \sigma^2 \tag{44}$$

Hence, we have the following recursion,

$$\delta(\theta_{t+1}) \leq (1 - \eta_t \mu \delta(\theta_t)) \delta(\theta_t) + \frac{L}{2} \eta^2 \sigma^2,$$

To turn this inequality into the structure as the PL-condition, let's consider when the sub-optimality gap is small for some constant $\delta > 0$ to be chosen later. If for some $t \in [0, T-1]$ we have $\delta(\theta_t) < \delta$ then we are done and have converged to a δ-neighbourhood within T iterations and have achieved

$$\min_{t \in [1,T]} \mathbb{E}[f^* - f(\theta_t)] \leq \delta \tag{45}$$

Otherwise, we have $\delta(\theta_t) \geq \delta$ and thus

$$\delta(\theta_{t+1}) \leq (1 - \delta\mu\eta_t)\delta(\theta_t) + \frac{L\sigma^2}{2}\eta_t^2 \tag{46}$$

By recursion on (46), and using $1 - x \leq \exp(-x)$ and Lemma 2 in [9] we have

$$\delta(\theta_{T+1}) \leq \exp\left(-\delta\mu \sum_{t=1}^{T} \eta_t\right)\delta(\theta_1) + \frac{L\sigma^2}{2}\sum_{t=1}^{T}\exp\left(-\delta\mu \sum_{i=t+1}^{T} \eta_i\right)\eta_t^2 \tag{47}$$

Let's start by lower bounding $\sum_{t=1}^{T} \eta_t$

$$\sum_{t=1}^{T} \eta_t = \eta_0 \frac{\alpha - \alpha^{T+1}}{1 - \alpha} \tag{48}$$

By Lemma 4 in [9]

$$\geq \frac{\eta_0\alpha}{1-\alpha} - \frac{2\eta_0\beta}{\ln\frac{T}{\beta}} \tag{49}$$

By Lemma 5 in [9]

$$= T\,\frac{0.69\eta_0}{\ln\frac{T}{\beta}} - \frac{2\eta_0\beta}{\ln\frac{T}{\beta}} \tag{50}$$

Now let's upper bound $\sum_{t=1}^{T}\exp\left(-\delta\mu\sum_{i=t+1}^{T}\eta_i\right)\eta_t^2$

$$\sum_{t=1}^{T}\exp\left(-\delta\mu\sum_{i=t+1}^{T}\eta_i\right)\eta_t^2 = \eta_0^2\sum_{t=1}^{T}\exp\left(-\mu\eta_0\frac{\alpha^{t+1} - \alpha^{T+1}}{1-\alpha}\right)\alpha^{2t} \tag{51}$$

Using (50)

$$\leq \eta_0^2 C(\beta)\sum_{t=1}^{T}\exp\left(-\frac{\mu\eta_0\alpha^{t+1}}{1-\alpha}\right)\alpha^{2t} \tag{52}$$

$$\leq \eta_0^2 C(\beta)\sum_{t=1}^{T}\left(\frac{e}{2}\frac{\delta\mu\eta_0\alpha^{t+1}}{1-\alpha}\right)^{-2}\alpha^{2t} \tag{53}$$

Using the fact that $\exp(-x) \leq \left(\frac{\gamma}{ex}\right)^{\gamma}, \forall x > 0, \gamma > 0$, with $\gamma = 2$ and $x = \alpha^{t+1}$

$$\leq \eta_0^2 C(\beta)\sum_{t=1}^{T}\left(\frac{e}{2}\frac{\delta\mu\alpha^{t+1}}{L(1+a)(1-\alpha)}\right)^{-2}\alpha^{2t} \tag{54}$$

$$\leq \frac{4L^2}{e^2(\delta\mu)^2}\sum_{t=1}^{T}\frac{1}{\alpha^2}\ln^2\left(\frac{1}{\alpha}\right) \tag{55}$$

$$\leq \frac{10L^2\ln^2\frac{T}{\beta}}{e^2(\delta\mu)^2 T} \tag{56}$$

Combining the bounds we have (47):

$$\mathbb{E}[f^* - f(\theta_t)] \leq \frac{5LC(\beta)}{e^2\delta^2\mu^2}\frac{\ln^2 \frac{T}{\beta}\sigma^2}{T} + C(\beta)\exp\left(\frac{-0.69\delta\mu}{L}\left(\frac{T}{\ln\frac{T}{\beta}}\right)\right)(f^* - f(\theta_t)) \tag{57}$$

Now to pick δ. The dominating term in the above equation is $\frac{1}{\delta^2 c^2 T}$. In order to converge $\epsilon > 0$ close to the optimal solution. Let $\delta = \epsilon$ and $T = \mathcal{O}(\delta^{-2}\epsilon^{-1})$. Picking $\delta = \epsilon$, we have T = $\mathcal{O}(\epsilon^{-3})$ ■

C.2. 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.

Theorem 7 *The dependence of Δ in [14, Lemma 4.3] 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 [14, Lemma 4.3], 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 $a \in \mathcal{A}$

$$\mathbb{E}\left[\left\|\frac{d[\langle \pi_\theta, \hat{r}_t\rangle]}{d\theta}\right\|_2^2\right] \leq \rho \left\|\frac{d[\langle \pi_\theta, r\rangle]}{d\theta}\right\| \tag{58}$$

Calculating the LHS,

$$\frac{d[\langle \pi_\theta, \hat{r}_t\rangle]}{d\theta(a)} = \left[\mathcal{I}\left\{a_t = a\right\} - \pi_\theta(a)\right] r(a_t) \implies \left\|\frac{d[\langle \pi_\theta, \hat{r}_t\rangle]}{d\theta}\right\|_2^2 = \sum_a \left[\left[\mathcal{I}\left\{a_t = a\right\} - \pi_\theta(a)\right] r(a_t)\right]^2 \tag{59}$$

Let $p := \pi_\theta(a_1)$ as the probability of pulling the optimal arm.

$$= \left[\left[\mathcal{I}\left\{a_t = a_1\right\} - p\right] r(a_t)\right]^2 + \left[\left[\mathcal{I}\left\{a_t = a_2\right\} - (1-p)\right] r(a_t)\right]^2 \tag{60}$$

$$\mathbb{E}\left[\left\|\frac{d[\langle \pi_\theta, \hat{r}_t\rangle]}{d\theta}\right\|_2^2\right] = \mathbb{E}\left[\left\|\frac{d[\langle \pi_\theta, \hat{r}_t\rangle]}{d\theta}\right\|_2^2 \middle| a_t = a_1\right] \Pr[a_t = a_1] + \mathbb{E}\left[\left\|\frac{d[\langle \pi_\theta, \hat{r}_t\rangle]}{d\theta}\right\|_2^2 \middle| a_t \neq a_1\right] \Pr[a_t \neq a_1] \tag{61}$$

$$= \left((1-p)^2 r_1^2 + (1-p)^2 r_1^2\right) p + \left(p^2 r_2^2 + p^2 r_2^2\right)(1-p) \tag{62}$$

$$\implies \text{LHS} = 2p(1-p)^2 r_1^2 + 2(1-p)p^2 r_2^2 = 2p(1-p)\left[(1-p)r_1^2 + p r_2^2\right] \tag{63}$$

Calculating the RHS,

$$\frac{d[\langle \pi_\theta, r\rangle]}{d\theta(a)} = \pi_\theta(a)\left[r_a - \langle \pi_\theta, r\rangle\right] \tag{64}$$

$$\implies \left\|\frac{d[\langle \pi_\theta, r\rangle]}{d\theta}\right\|_2^2 = \sum_a \pi_\theta(a)^2 \left[r_a - \langle \pi_\theta, r\rangle\right]^2 = p^2\left[r_1 - \langle \pi_\theta, r\rangle\right]^2 + (1-p)^2\left[r_2 - \langle \pi_\theta, r\rangle\right]^2 \tag{65}$$

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

$$= p^2 \left[r_1 - [p\, r_1 + (1-p)\, r_2] \right]^2 + (1-p)^2 \left[r_2 - [p\, r_1 + (1-p)\, r_2] \right]^2 \quad (66)$$

$$= p^2 (1-p)^2 \Delta^2 + (1-p)^2 p^2 \Delta^2 = 2\, p^2 (1-p)^2 \Delta^2 \quad (67)$$

$$\implies \text{RHS} = \left\| \frac{d[\langle \pi_\theta, r \rangle]}{d\theta} \right\| = \sqrt{2}\, p\, (1-p)\, \Delta \quad (68)$$

Hence,

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

For rewards $r_1 > r_2 > 0$, the numerator is a constant independent of Δ irrespective of p, while the denominator is Δ. Since we have derived an equality, the dependence on $\frac{1}{\Delta}$ in ρ is necessary. ■

C.3. Strong Growth Condition - General MDP Setting

Theorem 8 *Using Update 2, we have for all $t \geq 1$*

$$\mathbb{E}_t \left[\left\| \nabla \widetilde{f}(\theta_t) \right\|_2^2 \right] \leq \rho \|\nabla f(\theta_t)\|_2 \quad (69)$$

where $\rho := \frac{8\, R_{\max}^3 K^{2/3}}{\Delta^2}$ in the bandit setting with $\Delta := \min_{a \neq a'} |r(a) - r(a')|$ and $\rho = 4 \left(\sum_{s \in \mathcal{S}'} \frac{d_\mu^{\pi_{\theta_t}}(s)^2}{(1-\gamma)^4} \frac{|\mathcal{A}|}{\Delta_s^2} \sqrt{\frac{|A|}{\delta}} \right)$ with $\Delta_s := \min_{a \neq a'} |Q^{\pi_\theta}(s,a) - Q^{\pi_\theta}(s,a')|$, $\mathcal{S}' := \{ s : d^{\pi_\theta}(s) > 0 \}$ and $\delta = \min_{s \in \mathcal{S}'} d^{\pi_\theta}(s)$ in the general MDP setting.

Proof The proof in the bandits setting can be found in Lemma 4.3 in [13]. For the general MDP setting first recall that

$$\frac{V^{\pi_{\theta_t}}(\mu)}{\partial \theta(s,a)} = \frac{1}{1-\gamma}\, d_\mu^{\pi_{\theta_t}}(s)\, \pi_{\theta_t}(a|s)\, A^{\pi_{\theta_t}}(s,a) \quad (70)$$

$$\nabla V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta(\,|s)} [A^{\pi_\theta}(s,a) \nabla \log \pi_\theta(a|s)] \quad (71)$$

and let

$$\nabla \widetilde{f}(\theta_t)_s := \frac{1}{1-\gamma}\, d_\mu^{\pi_{\theta_t}}(s)\, \pi_{\theta_t}(a|s) \left(\hat{Q}^{\pi_{\theta_t}}(s,a) - \pi_{\theta_t}(\,|s)^\top \hat{Q}^{\pi_{\theta_t}}(s,\,) \right) \quad (72)$$

and given $t \geq 1$, denote $k_t(s) := \arg\max_{a \in \mathcal{A}} \pi_{\theta_t}(a|s)$ as the action with the largest probability at state s. From the second part of the proof of Lemma 11 in [11], for all $s \in \mathcal{S}$,

$$\sum_{a\in\mathcal{A}} \pi_{\theta_t}(a|s) \sum_{a\in\mathcal{A}} \pi_{\theta_t}(a|s)^2 \left(\hat{Q}^{\pi_{\theta_t}}(s,a) - \pi_{\theta_t}(\,|s)^\top \hat{Q}^{\pi_{\theta_t}}(s,\,)\right)^2$$

$$= \sum_{a\in\mathcal{A}} \pi_{\theta_t}(a|s)^2 \left[\frac{\mathbb{1}\{a(s)=a\}}{\pi_{\theta_t}(a|s)^2} Q^{\pi_{\theta_t}}(s,a)^2 \right.$$

$$\left. -2\,\frac{\mathbb{1}\{a(s)=a\}}{\pi_{\theta_t}(a|s)} Q^{\pi_{\theta_t}}(s,a)\,\pi_{\theta_t}(\,|s)^\top \hat{Q}^{\pi_{\theta_t}}(s,\,) + \left(\pi_{\theta_t}(\,|s)^\top \hat{Q}^{\pi_{\theta_t}}(s,\,)\right)^2\right] \tag{73}$$

$$= Q^{\pi_{\theta_t}}(s,a(s))^2 - 2\,\pi_{\theta_t}(a(s)|s)\,Q^{\pi_{\theta_t}}(s,a(s))^2 + \sum_{a\neq a(s)} \pi_{\theta_t}(a|s)^2\,Q^{\pi_{\theta_t}}(s,a(s))^2 \tag{74}$$

$$= (1-\pi_{\theta_t}(a(s)|s))^2\,Q^{\pi_{\theta_t}}(s,a(s))^2 + \sum_{a\neq a(s)} \pi_{\theta_t}(a|s)^2\,Q^{\pi_{\theta_t}}(s,a(s))^2 \tag{75}$$

Taking expectation over $a(s)\sim\pi_{\theta_t}(\,|s)$

$$\mathbb{E}_{a(s)\sim\pi_{\theta_t}(\,|s)}\left[\sum_{a\in\mathcal{A}} \pi_{\theta_t}(a|s)^2 \left(\hat{Q}^{\pi_{\theta_t}}(s,a) - \pi_{\theta_t}(\,|s)^\top \hat{Q}^{\pi_{\theta_t}}(s,\,)\right)^2\right]$$

$$= \sum_{a(s)\in\pi_{\theta_t}(\,|s)} \pi_{\theta_t}(a(s)|s)\,(1-\pi_{\theta_t}(a(s)|s))^2\,Q^{\pi_{\theta_t}}(s,a(s))^2$$

$$+ \sum_{a(s)\in\pi_{\theta_t}(\,|s)} \pi_{\theta_t}(a(s)|s) \sum_{a\neq a(s)} \pi_{\theta_t}(a|s)^2\,Q^{\pi_{\theta_t}}(s,a(s))^2 \tag{76}$$

Since $\|x\|_2 \le \|x\|_1$

$$\le 2 \sum_{a(s)\in\pi_{\theta_t}(\,|s)} \pi_{\theta_t}(a(s)|s)\,(1-\pi_{\theta_t}(a(s)|s))^2\,Q^{\pi_{\theta_t}}(s,a(s))^2 \tag{77}$$

Since $Q^{\pi_{\theta_t}}(s,a) \le \frac{1}{1-\gamma}$

$$\le \frac{2}{(1-\gamma)^2}\left(\pi_{\theta_t}(k_t(s)|s)\,(1-\pi_{\theta_t}(k_t(s)|s))^2 + \sum_{a(s)\neq k_t(s)} \pi_{\theta_t}(a(s)|s)\,(1-\pi_{\theta_t}(a(s)|s))^2\right) \tag{78}$$

Since $\pi_{\theta_t}(a|s)\in(0,1)$

$$\le \frac{2}{(1-\gamma)^2}\left((1-\pi_{\theta_t}(k_t(s)|s)) + \sum_{a(s)\neq k_t(s)} \pi_{\theta_t}(a(s)|s)\right) \tag{79}$$

$$= \frac{4}{(1-\gamma)^2}\,(1-\pi_{\theta_t}(k_t(s)|s)) \tag{80}$$

Now to lower bound

$$\|\nabla f(\theta)\|_2^2 = \sum_{s\in\mathcal{S}} \frac{d_\mu^{\pi_{\theta_t}}(s)^2}{(1-\gamma)^2} \sum_{a\in\mathcal{A}} \pi_{\theta_t}(a|s)^2 (A^{\pi_{\theta_t}}(s,a))^2 \tag{81}$$

$$= \sum_{s\in\mathcal{S}} \frac{d_\mu^{\pi_{\theta_t}}(s)^2}{(1-\gamma)^2} \left(\sum_{a'\in\mathcal{A}} A^{\pi_{\theta_t}}(s,a')^2 \sum_{a\in\mathcal{A}} \pi_{\theta_t}(a|s)^2 \frac{A^{\pi_{\theta_t}}(s,a)^2}{\sum_{a'\in\mathcal{A}} A^{\pi_{\theta_t}}(s,a')^2} \right) \tag{82}$$

Using Jensen's inequatliy

$$\geq \sum_{s\in\mathcal{S}} \frac{d_\mu^{\pi_{\theta_t}}(s)^2}{(1-\gamma)^2} \left(\sum_{a'\in\mathcal{A}} A^{\pi_{\theta_t}}(s,a')^2 \left[\sum_{a\in\mathcal{A}} \pi_{\theta_t}(a|s) \frac{A^{\pi_{\theta_t}}(s,a)^2}{\sum_{a'\in\mathcal{A}} A^{\pi_{\theta_t}}(s,a')^2} \right]^2 \right) \tag{83}$$

$$\geq \sum_{s\in\mathcal{S}} \frac{d_\mu^{\pi_{\theta_t}}(s)^2}{(1-\gamma)^2} \left(\frac{1}{\sum_{a'\in\mathcal{A}} A^{\pi_{\theta_t}}(s,a')^2} \left[\sum_{a\in\mathcal{A}} \pi_{\theta_t}(a|s) A^{\pi_{\theta_t}}(s,a)^2 \right]^2 \right) \tag{84}$$

Since $A^{\pi_{\theta_t}}(s,a) \leq (1-\gamma)^{-1}$

$$\geq \frac{1}{|A|} \sum_{s\in\mathcal{S}} d_\mu^{\pi_{\theta_t}}(s)^2 \left[\sum_{a\in\mathcal{A}} \pi_{\theta_t}(a|s) A^{\pi_{\theta_t}}(s,a)^2 \right]^2 \tag{85}$$

Let $\delta = \min_{s\in\mathcal{S}'} d_\mu^{\pi_{\theta_t}}(s)$ and since all terms are non-negative

$$\geq \frac{\delta}{|A|} \left[\sum_{a\in\mathcal{A}} \pi_{\theta_t}(a|s) A^{\pi_{\theta_t}}(s,a)^2 \right]^2 \tag{86}$$

Taking the square roots of both sides gives,

$$\sum_{a\in\mathcal{A}} \pi_{\theta_t}(a|s) A^{\pi_{\theta_t}}(s,a)^2 \leq \sqrt{\frac{|A|}{\delta}} \|\nabla f(\theta_t)\| \tag{87}$$

To connect (76) and (87) let's label actions $a \in \mathcal{A}$ as $a \in [K]$.

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

$$= \sum_{i=1}^{K} \pi_{\theta_t}(i)\,Q^{\pi_{\theta_t}}(s,i)^2 - \left[\sum_{i=1}^{K} \pi_{\theta_t}(i)\,Q^{\pi_{\theta_t}}(s,i)\right]^2 \tag{88}$$

$$= \sum_{i=1}^{K} \pi_{\theta_t}(i)\,Q^{\pi_{\theta_t}}(s,i)^2 - \sum_{i=1}^{K} \pi_{\theta_t}(i)^2\,Q^{\pi_{\theta_t}}(s,i)^2 - 2\sum_{i=1}^{K-1} \pi_{\theta_t}(i)\,Q^{\pi_{\theta_t}}(s,i) \sum_{j=i+1}^{K} \pi_{\theta_t}(j)\,Q^{\pi_{\theta_t}}(s,j) \tag{89}$$

$$= \sum_{i=1}^{K} \pi_{\theta_t}(i)\,Q^{\pi_{\theta_t}}(s,i)^2(1-\pi_{\theta_t}(i)) - 2\sum_{i=1}^{K-1} \pi_{\theta_t}(i)\,Q^{\pi_{\theta_t}}(s,i) \sum_{j=i+1}^{K} \pi_{\theta_t}(j)\,Q^{\pi_{\theta_t}}(s,j) \tag{90}$$

$$= \sum_{i=1}^{K} \pi_{\theta_t}(i)\,Q^{\pi_{\theta_t}}(s,i)^2 \sum_{j\neq i} \pi_{\theta_t}(j) - 2\sum_{i=1}^{K-1} \pi_{\theta_t}(i)\,Q^{\pi_{\theta_t}}(s,i) \sum_{j=i+1}^{K} \pi_{\theta_t}(j)\,Q^{\pi_{\theta_t}}(s,j) \tag{91}$$

$$= \sum_{i=1}^{K-1} \pi_{\theta_t}(i|s) \sum_{j=i+1}^{K} \pi_{\theta_t}(j|s)\,(Q^{\pi_{\theta_t}}(s,i)^2 + Q^{\pi_{\theta_t}}(s,j)^2)$$
$$- 2\sum_{i=1}^{K-1} \pi_{\theta_t}(i)\,Q^{\pi_{\theta_t}}(s,i) \sum_{j=i+1}^{K} \pi_{\theta_t}(j|s)\,Q^{\pi_{\theta_t}}(s,j) \tag{92}$$

$$= \sum_{i=1}^{K-1} \pi_{\theta_t}(i|s) \sum_{j=i+1}^{K} \pi_{\theta_t}(j|s)\,(Q^{\pi_{\theta_t}}(s,i) - Q^{\pi_{\theta_t}}(s,j))^2 \tag{93}$$

This implies

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

$$\geq \sum_{i=1}^{k_t(s)} \pi_{\theta_t}(i|s) \sum_{j=i+1}^{K} \pi_{\theta_t}(j|s)\,(Q^{\pi_{\theta_t}}(s,i) - Q^{\pi_{\theta_t}}(s,j))^2 \tag{94}$$

$$\geq \sum_{i=1}^{k_t(s)-1} \pi_{\theta_t}(i|s)\,\pi_{\theta_t}(k_t(s)|s)\,(Q^{\pi_{\theta_t}}(s,i) - Q^{\pi_{\theta_t}}(s,k_t(s))^2$$
$$+ \pi_{\theta_t}(k_t(s)|s) \sum_{j=k_t(s)+1}^{K} \pi_{\theta_t}(i|s)\,\pi_{\theta_t}(k_t(s)|s)\,(Q^{\pi_{\theta_t}}(s,k_t(s)) - Q^{\pi_{\theta_t}}(s,j)^2 \tag{95}$$

$$= \pi_{\theta_t}(k_t(s)|s) \sum_{a\neq k_t(s)} \pi_{\theta_t}(a|s)\,(Q^{\pi_{\theta_t}}(s,a) - Q^{\pi_{\theta_t}}(s,k_t(s))^2 \tag{96}$$

Let $\Delta_s = \min_{a\neq a'} |Q^{\pi_{\theta_t}}(s,a) - Q^{\pi_{\theta_t}}(s,a'|$

$$\geq (1-\pi_{\theta_t}(k_t(s)|s))\,\frac{\Delta_s^2}{|\mathcal{A}|} \tag{97}$$

$$\tag{98}$$

This implies that,

$$(1 - \pi_{\theta_t}(k_t(s)|s)) \leq \frac{|\mathcal{A}|}{\Delta_s^2} \sum_{a \in \mathcal{A}} \pi_{\theta_t}(a|s) \left(Q^{\pi_{\theta_t}}(s,a) - V^{\pi_{\theta_t}}(s)\right)^2 \tag{99}$$

Therefore the upper bound is,

$$\mathbb{E}_t\left[\left\|\nabla \widetilde{f}(\theta_t)\right\|_2^2\right] = \mathbb{E}_{a(s)\sim\pi_{\theta_t}(\,|s)}\left[\sum_{(s,a)} \frac{1}{(1-\gamma)^2} d_\mu^{\pi_{\theta_t}}(s)^2\, \pi_{\theta_t}(a|s)^2 \left(\hat{Q}^{\pi_{\theta_t}}(s,a) - \pi_{\theta_t}(\,|s)^\top \hat{Q}^{\pi_{\theta_t}}(s,\,)\right)^2\right] \tag{100}$$

By (73)

$$\leq 4 \sum_{s \in \mathcal{S}} \frac{d_\mu^{\pi_{\theta_t}}(s)^2}{(1-\gamma)^4} \left(1 - \pi_{\theta_t}(k_t(s)|s)\right) \tag{101}$$

By (94)

$$\leq 4 \sum_{s \in \mathcal{S}} \frac{d_\mu^{\pi_{\theta_t}}(s)^2}{(1-\gamma)^4} \frac{|\mathcal{A}|}{\Delta_s^2} \sum_{a \in \mathcal{A}} \pi_{\theta_t}(a|s) \left(A^{\pi_{\theta_t}}(s,a)\right)^2 \tag{102}$$

Let $\mathcal{S}' := \{s \in \mathcal{S} \,|\, d_\mu^{\pi_{\theta_t}}(s) > 0\}$

$$= 4 \sum_{s \in \mathcal{S}'} \frac{d_\mu^{\pi_{\theta_t}}(s)^2}{(1-\gamma)^4} \frac{|\mathcal{A}|}{\Delta_s^2} \sum_{a \in \mathcal{A}} \pi_{\theta_t}(a|s) \left(A^{\pi_{\theta_t}}(s,a)\right)^2 \tag{103}$$

By (87)

$$= 4 \sum_{s \in \mathcal{S}'} \frac{d_\mu^{\pi_{\theta_t}}(s)^2}{(1-\gamma)^4} \frac{|\mathcal{A}|}{\Delta_s^2} \sqrt{\frac{|A|}{\delta}} \left\|\nabla f(\theta_t)\right\| \tag{104}$$

$$= 4 \left(\sum_{s \in \mathcal{S}'} \frac{d_\mu^{\pi_{\theta_t}}(s)^2}{(1-\gamma)^4} \frac{|\mathcal{A}|}{\Delta_s^2} \sqrt{\frac{|A|}{\delta}}\right) \left\|\nabla f(\theta_t)\right\| \tag{105}$$

■

C.4. Proof of Theorem 3

Theorem 3 *Under the same assumptions as Theorem 2 and SGC, Update 2 with exponential step-sizes has the following convergence: if* $\mathbb{E}[f^* - f(\theta_t)] \geq \epsilon$ *for all* $t \in [1, T]$ *then*

$$\mathbb{E}[f^* - f(\theta_T)] \leq C_1 \exp\left(-\frac{\alpha\, T}{\kappa\, \ln(T)}\right) \mathbb{E}[f^* - f(\theta_1)] + \frac{4\,\rho\, C_2\, L}{\epsilon^2} \frac{\sum_{t=1}^{T_0 - 1} \mathbb{E}[f^* - f(\theta_t)]}{T^2} \tag{3}$$

where $\kappa = \frac{2}{\mu\, \epsilon\, \eta_0}$, $C_1 := \frac{2\beta}{\kappa \ln(T/\beta)}$, $C_2 := \exp\left(\frac{2\beta}{\kappa \ln(T/\beta)}\right) \frac{4\kappa^2}{e^2 \alpha^2} \ln^2(T/\beta)$, $T_0 := T \max\left\{\frac{\ln(4\rho\eta_0)}{\ln(T/\beta)}, 0\right\}$, $\eta_0 = \frac{1}{18}$ *and* ρ *and* L *are known problem dependent constants. Otherwise,* $\min_{t \in [1,T]} \mathbb{E}[f^* - f(\theta_t)] \leq \epsilon$

Proof Starting from Eq. (44), using Theorem 8, $\mathbb{E}\left\|\nabla \widetilde{f}(\theta_t)\right\|_2^2 \leq \rho \|\nabla f(\theta_t)\|$,

$$\delta(\theta_{t+1}) \leq \delta(\theta_t) - \eta_t \|\nabla f(\theta_t)\|_2^2 + \frac{L_\tau \rho {\eta_t}^2}{2} \|\nabla f(\theta_t)\| \tag{106}$$

Using non-uniform smoothness [14, Lemma 4.1], $L_\tau = 3 \|\nabla f(\theta_t{}')\|$

$$\delta(\theta_{t+1}) \leq \delta(\theta_t) - \eta_t \|\nabla f(\theta_t)\|_2^2 + \frac{3\rho {\eta_t}^2}{2} \|\nabla f(\theta_t)\| \|\nabla f(\theta_t{}')\| \tag{107}$$

Bounding $\|\nabla f(\theta_t{}')\|$ according to [14, Eq 96], $\|\nabla f(\theta_t{}')\| \leq \frac{1}{1-9\eta_t/2} \|\nabla f(\theta_t)\|$. If $\eta_t \leq \frac{1}{18}$,

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

Phase 2: Let us first consider the case when $\eta_t \leq \frac{1}{4\rho}$. In this case, starting from PL-condition

$$\delta(\theta_{t+1}) \leq \delta(\theta_t) - \frac{\eta_t}{2} \|\nabla f(\theta_t)\|_2^2 \leq \delta(\theta_t) - \frac{\eta_t \mu}{2} \delta(\theta_t)^2 \tag{109}$$

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

$$\delta(\theta_{t+1}) \leq \delta(\theta_t) - \frac{\eta_t \mu \delta}{2} \delta(\theta_t) \tag{110}$$

Phase 1: When $\eta_t > \frac{1}{4\rho}$, define $\sigma_t^2 := 2\rho \|\nabla f(\theta_t)\|_2^2$. Hence, using the P Ł-condition

$$\delta(\theta_{t+1}) \leq \delta(\theta_t) - \eta_t \|\nabla f(\theta_t)\|_2^2 + {\eta_t}^2 \sigma_t^2 \leq \delta(\theta_t) - \eta_t \mu \delta(\theta_t)^2 + {\eta_t}^2 \sigma_t^2 \tag{111}$$

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

$$\delta(\theta_{t+1}) \leq \delta(\theta_t) - \eta_t \mu \delta \delta(\theta_t) + {\eta_t}^2 \sigma_t^2 \tag{112}$$

We will use the exponential step-sizes [22] s.t. $\eta_t = \eta_0 \alpha_t$ where $\alpha_t := \alpha^t$ where $\alpha = \left(\frac{\beta}{T}\right)^{1/T}$. Since $\alpha_t < 1$, if $\eta_0 \leq \frac{1}{18}$ then the condition on the step-size is satisfied. For $\eta_t \leq \frac{1}{4\rho}$, we require that

$$\eta_0 \left(\frac{\beta}{T}\right)^{t/T} \leq \frac{1}{4\rho} \implies t \geq T_0 := T \frac{\ln(4\rho\eta)}{\ln\left(\frac{T}{\beta}\right)} \tag{113}$$

Hence, when $t \geq T_0$, the step-size is small enough so that we are in Phase 2. Using Eq. (110) and recursing from $t = T_0$ to T,

$$\delta(\theta_{t+1}) \leq \delta(\theta_t) - \frac{\eta_t \mu \delta}{2} \delta(\theta_t) \tag{114}$$

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

$$\implies \delta(\theta_{T+1}) \leq \exp\left(-\frac{\mu\eta_0 \delta}{2} \sum_{t=T_0}^{T} \eta_t\right) \delta(\theta_{T_0}) \tag{115}$$

Sum of geometric series

$$\implies \delta(\theta_{T+1}) \leq \exp\left(-\frac{\mu\delta\,\eta_0}{2}\,\frac{\alpha^{T_0}-\alpha^{T+1}}{1-\alpha}\right)\,\delta(\theta_{T_0}) \tag{116}$$

Now let us consider Phase 1 where $t < T_0$. Using Eq. (112) and recursing from $t = 1$ to $T_0 - 1$,

$$\delta(\theta_{t+1}) \leq \delta(\theta_t) - \frac{\eta_t\,\mu\,\delta}{2}\,\delta(\theta_t) + {\eta_t}^2\,\sigma_t^2 \tag{117}$$

$$\implies \delta(\theta_{T_0}) \leq \prod_{t=1}^{T_0-1}\left(1-\frac{\mu\,\delta\,\eta_0}{2}\,\alpha_t\right) + \sum_{t=1}^{T_0-1}\alpha_t^2\,\sigma_t^2\prod_{i=t+1}^{T_0-1}\left(1-\frac{\mu\,\delta\,\eta_0}{2}\,\alpha_i\right) \tag{118}$$

Using $1 - x \leq \exp(-x)$, defining $\frac{1}{\kappa} := \frac{\mu\,\delta\,\eta_0}{2}$ and by summing up the geometric series,

$$\implies \delta(\theta_{T_0}) \leq \delta(\theta_1)\exp\left(-\frac{1}{\kappa}\,\frac{\alpha-\alpha^{T_0}}{1-\alpha}\right) + \sum_{t=1}^{T_0-1}\alpha_t^2\,\sigma_t^2\prod_{i=t+1}^{T_0-1}\left(1-\frac{\mu\,\delta\,\eta_0}{2}\,\alpha_i\right) \tag{119}$$

Let us now bound the second term on the RHS.

$$\sum_{t=1}^{T_0-1}\alpha_t^2\sigma_t^2\prod_{i=t+1}^{T_0-1}\left(1-\frac{1}{\kappa}\alpha_i\right) \leq \sum_{t=1}^{T_0-1}\alpha_t^2\sigma_t^2\,\exp\left(-\frac{1}{\kappa}\sum_{i=t+1}^{T_0-1}\alpha^i\right) \tag{120}$$

$$= \sum_{t=1}^{T_0-1}\alpha_t^2\sigma_t^2\,\exp\left(-\frac{1}{\kappa}\frac{\alpha^{t+1}-\alpha^{T_0}}{1-\alpha}\right) \tag{121}$$

$$= \exp\left(\frac{\alpha^{T_0}}{\kappa(1-\alpha)}\right)\sum_{t=1}^{T_0-1}\alpha_t^2\sigma_t^2\,\exp\left(-\frac{\alpha^{t+1}}{\kappa(1-\alpha)}\right) \tag{122}$$

Using that $\exp(-x) \leq \left(\frac{2}{ex}\right)^2$

$$\leq \exp\left(\frac{\alpha^{T_0}}{\kappa(1-\alpha)}\right)\sum_{t=1}^{T_0-1}\alpha_t^2\sigma_t^2\,\left(\frac{2(1-\alpha)\kappa}{e\alpha^{t+1}}\right)^2 \tag{123}$$

$$= \exp\left(\frac{\alpha^{T_0}}{\kappa(1-\alpha)}\right)\frac{4(1-\alpha)^2\kappa^2}{e^2\alpha^2}\sum_{t=1}^{T_0-1}\sigma_t^2 \tag{124}$$

Since $1 - x \leq \ln(1/x)$

$$\leq \exp\left(\frac{\alpha^{T_0}}{\kappa(1-\alpha)}\right)\frac{4\kappa^2}{e^2\alpha^2}\,\frac{\ln^2\left(\frac{T}{\beta}\right)\sum_{t=1}^{T_0-1}\sigma_t^2}{T^2} \tag{125}$$

Putting everything together,

$$\delta(\theta_{T_0}) \leq \delta(\theta_1)\exp\left(-\frac{1}{\kappa}\,\frac{\alpha-\alpha^{T_0}}{1-\alpha}\right) + \exp\left(\frac{\alpha^{T_0}}{\kappa(1-\alpha)}\right)\frac{4\kappa^2}{e^2\alpha^2}\,\frac{\ln^2\left(\frac{T}{\beta}\right)\sum_{t=1}^{T_0-1}\sigma_t^2}{T^2} \tag{126}$$

Combining the results of Phase 1 and Phase 2,

$$\delta(\theta_{T+1}) \leq \exp\left(-\frac{1}{\kappa}\frac{\alpha^{T_0}-\alpha^{T+1}}{1-\alpha}\right)\left[\delta(\theta_1)\exp\left(-\frac{1}{\kappa}\frac{\alpha-\alpha^{T_0}}{1-\alpha}\right)+\exp\left(\frac{\alpha^{T_0}}{\kappa(1-\alpha)}\right)\frac{4\kappa^2}{e^2\alpha^2}\frac{\ln^2\left(\frac{T}{\beta}\right)\sum_{t=1}^{T_0-1}\sigma_t^2}{T^2}\right] \tag{127}$$

$$= \delta(\theta_1)\,\exp\left(-\frac{1}{\kappa}\frac{\alpha-\alpha^{T+1}}{1-\alpha}\right)+\exp\left(\frac{\alpha^{T+1}}{\kappa(1-\alpha)}\right)\frac{4\kappa^2}{e^2\alpha^2}\frac{\ln^2\left(\frac{T}{\beta}\right)\sum_{t=1}^{T_0-1}\sigma_t^2}{T^2} \tag{128}$$

Using that $\frac{\alpha^{T+1}}{\kappa(1-\alpha)} \leq \frac{2\beta}{\kappa\,\ln(T/\beta)}$ from [22, Lemma 5],

$$\leq \delta(\theta_1)\,\exp\left(\frac{2\beta}{\kappa\,\ln(T/\beta)}\right)\exp\left(-\frac{1}{\kappa}\frac{\alpha}{1-\alpha}\right)+\exp\left(\frac{2\beta}{\kappa\,\ln(T/\beta)}\right)\frac{4\kappa^2}{e^2\alpha^2}\frac{\ln^2\left(\frac{T}{\beta}\right)\sum_{t=1}^{T_0-1}\sigma_t^2}{T^2} \tag{129}$$

Using that $\frac{\alpha}{\kappa(1-\alpha)} \geq \frac{\alpha T}{\kappa\,\ln(T/\beta)}$ from [22, Lemma 5],

$$\leq \delta(\theta_1)\,\underbrace{\exp\left(\frac{2\beta}{\kappa\,\ln(T/\beta)}\right)}_{:=C_1}\exp\left(-\frac{\alpha T}{\kappa\,\ln(T/\beta)}\right)+\underbrace{\exp\left(\frac{2\beta}{\kappa\,\ln(T/\beta)}\right)\frac{4\kappa^2}{e^2\alpha^2}\ln^2\left(\frac{T}{\beta}\right)}_{:=C_2}\frac{\sum_{t=1}^{T_0-1}\sigma_t^2}{T^2} \tag{130}$$

$$\delta(\theta_{T+1}) \leq C_1\,\exp\left(-\frac{\alpha T}{\kappa\,\ln(T/\beta)}\right)\delta_1 + C_2\,\frac{\sum_{t=1}^{T_0-1}\sigma_t^2}{T^2} \tag{131}$$

$$\tag{132}$$

Let us now simplify $\sigma_t^2 = 2\rho\,\|\nabla f(\theta_t)\|_2^2$. Since f is L uniform smooth, for any u, v

$$f(v) \leq f(u) + \langle \nabla f(u), v-u\rangle + \frac{L}{2}\|u-v\|_2^2 \tag{133}$$

Setting $v = u - \frac{1}{L}\nabla f(u)$,

$$f(v) \leq f(u) - \frac{1}{2L}\|\nabla f(u)\|_2^2 \implies \|\nabla f(u)\|_2^2 \leq 2L\,[f(u)-f(v)] \leq 2L\,[f(u)-f^*] \tag{134}$$

$$\implies \sigma_t^2 \leq 4\rho L\,[f(\theta_t)-f^*] = 4\rho L\,\delta(\theta_t) \tag{135}$$

$$\delta(\theta_{T+1}) \leq C_1\,\exp\left(-\frac{\alpha T}{\kappa\,\ln(T/\beta)}\right)\delta(\theta_1) + 4\,\rho\,C_2\,L\,\frac{\sum_{t=1}^{T_0-1}\delta(\theta_t)}{T^2} \tag{136}$$

Making the dependence on the constants explicit,

$$\implies \delta(\theta_{T+1}) \leq \exp\left(\frac{\mu\delta\,\eta_0\,\beta}{\ln(T/\beta)}\right)\exp\left(\frac{-\mu\delta\eta_0\alpha}{\ln(T/\beta)}\,T\right)+4\,\rho\,L\,\exp\left(\frac{\mu\delta\,\eta_0\,\beta}{\ln(T/\beta)}\right)\frac{16\,\ln^2(T/\beta)}{e^2\alpha^2\,\mu^2\,\eta_0^2\,\delta^2}\,\frac{\sum_{t=1}^{T_0-1}\delta(\theta_t)}{T^2} \tag{137}$$

■

Appendix D. Policy Gradient with Entropy Regularization Proofs

D.1. Proof of Theorem 9

Theorem 9 (Softmax Policy Gradient with Entropy Regularization) *Using Update 3, softmax policy gradient with Armijo line-search converges to the soft globally optimal solution with a rate of* $\mathcal{O}(\exp(-T))$

$$f^{*\tau} - f^{\tau}(\theta_T) \leq f^{*\tau} - f^{\tau}(\theta_0)\, \exp\left(- \min\left\{ \frac{2(1-h)}{L^{\tau}}, \eta_{\max} \right\} \mu\, T \right) \tag{138}$$

where $\mu := \min_{1 \leq t \leq T} \min_a \pi_{\theta_t}(a|s)$, $L^{\tau} = \frac{5}{2} + 5\,\tau(1 + \log|\mathcal{A}|)$ *and* $h \in (0, 1)$.

Proof

For any L-smooth function the step-size η_t returned by the Armijo line-search is guaranteed to satisfy $\eta_{\max} \geq \eta_t \geq \min\left\{ \frac{2(1-c)}{L}, \eta_{\max} \right\}$ (Lemma 1 in [21]) which implies that

$$f^{\tau}(\theta_{t+1}) \geq f^{\tau}(\theta_t) + \min\left\{ \frac{2(1-h)}{L^{\tau}}, \eta_{\max} \right\} \|\nabla f^{\tau}(\theta_t)\|_2^2 \tag{139}$$

Subtracting $f^{*\tau}$ from both sides and multiplying by -1

$$f^{*\tau} - f^{\tau}(\theta_{t+1}) \leq f^{*\tau} - f^{\tau}(\theta_t) - \min\left\{ \frac{2(1-h)}{L^{\tau}}, \eta_{\max} \right\} \|\nabla f^{\tau}(\theta_t)\|_2^2 \tag{140}$$

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

$$\delta^{\tau}(\theta_{t+1}) \leq \delta^{\tau}(\theta_t) - \min\left\{ \frac{2(1-h)}{L^{\tau}}, \eta_{\max} \right\} \|\nabla f^{\tau}(\theta_t)\|_2^2 \tag{141}$$

By Theorem 31, with $C(\theta) := 2\tau\, \min_a \pi_\theta(a)$

$$\leq \delta^{\tau}(\theta_t) \left(1 - \min\left\{ \frac{2(1-h)}{L^{\tau}}, \eta_{\max} \right\} C(\theta_t) \right) \tag{142}$$

Recursing from $t = 0$ to $T - 1$ and for let $\mu := \min_{1 \leq t \leq T} C(\theta_t)$

$$\delta^{\tau}(\theta_T) \leq \delta^{\tau}(\theta_0)\, \exp\left(- \min\left\{ \frac{2(1-h)}{L^{\tau}}, \eta_{\max} \right\} \mu\, T \right) \tag{143}$$

■

D.2. Proof of Theorem 4

Lemma 10 *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:

$$\frac{d}{dr(\hat{a})}(\pi^* - \pi_\tau^*)^\top r = 0 \implies \frac{d}{dr(\hat{a})} \frac{\sum_{a \neq a^*} e^{\frac{r(a)}{\tau}} \Delta(a)}{\sum_{a'} e^{\frac{r(a')}{\tau}}} = 0 \tag{144}$$

$$\implies \frac{\left(\frac{e^{\frac{r(\hat{a})}{\tau}}}{\tau}[r(a^*) - r(\hat{a})] - e^{\frac{r(\hat{a})}{\tau}}\right)\left(\sum_a e^{\frac{r(a)}{\tau}}\right) - \frac{e^{\frac{r(\hat{a})}{\tau}}}{\tau}\left(\sum_a e^{\frac{r(a)}{\tau}}[r(a^*) - r(a)]\right)}{\left(\sum_{a'} e^{\frac{r(a')}{\tau}}\right)^2} = 0 \tag{145}$$

$$\implies \frac{\frac{e^{\frac{r(\hat{a})}{\tau}}}{\tau}\left(\sum_a e^{\frac{r(a)}{\tau}}[r(a) - r(\hat{a}) - \tau]\right)}{\left(\sum_{a'} e^{\frac{r(a')}{\tau}}\right)^2} = 0 \implies \sum_a e^{\frac{r(a)}{\tau}}[r(a) - r(\hat{a}) - \tau] = 0 \tag{146}$$

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

$$\implies \sum_a e^{\frac{r(a)}{\tau}}[r(a) - r(\hat{a}_i) - \tau] - \sum_a e^{\frac{r(a)}{\tau}}[r(a) - r(\hat{a}_j) - \tau] = 0 - 0 \tag{147}$$

$$\implies \sum_a e^{\frac{r(a)}{\tau}}[r(\hat{a}_j)) - r(\hat{a}_i)] = 0 \implies r(\hat{a}_j) = r(\hat{a}_i). \tag{148}$$

Therefore, all suboptimal rewards must be equal. ■

Lemma 11 *We have* $(\pi^* - \pi_\tau^*)^\top r \leq \tau W\left(\frac{|\mathcal{A}|-1}{e}\right)$*, where* $W : \mathbb{R}^+ \mapsto \mathbb{R}^+$ *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 $\pi_\tau^* = \mathrm{softmax}(r/\tau)$. Denoting $\Delta(a) = r(a^*) - r(a)$, $\Delta = \min_{a \neq a^*} \Delta(a)$, and a^* is the optimal action, we have

$$(\pi^* - \pi_\tau^*)^\top r = \sum_a \pi_\tau^*(a)\, r(a^*) - \sum_a \pi_\tau^*(a)\, r(a) = \sum_{a \neq a^*} \pi_\tau^*(a)\, \Delta(a) = \frac{\sum_{a \neq a^*} e^{\frac{r(a)}{\tau}} \Delta(a)}{\sum_{a'} e^{\frac{r(a')}{\tau}}}. \tag{149}$$

To find the upper bound, it is enough to find a reward vector $r \in \mathbb{R}^{|\mathcal{A}|}$ 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:

$$(\pi^* - \pi_\tau^*)^\top (r - c\mathbf{1}) = \frac{\sum_{a \neq a^*} e^{\frac{r(a)-c}{\tau}} \Delta(a)}{\sum_{a'} e^{\frac{r(a')-c}{\tau}}} = \frac{e^{-\frac{c}{\tau}} \sum_{a \neq a^*} e^{\frac{r(a)}{\tau}} \Delta(a)}{e^{-\frac{c}{\tau}} \sum_{a'} e^{\frac{r(a')}{\tau}}} \tag{150}$$

$$= \frac{\sum_{a \neq a^*} e^{\frac{r(a)}{\tau}} \Delta(a)}{\sum_{a'} e^{\frac{r(a')}{\tau}}} = (\pi^* - \pi_\tau^*)^\top r \tag{151}$$

Therefore, without loss of generality, we assume that the smallest reward value equals 0. Furthermore, according to Theorem 10, 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,

$$(\pi^* - \pi^*_\tau)^\top r = \frac{\sum_{a \neq a^*} e^{\frac{r(a)}{\tau}} \Delta(a)}{\sum_{a'} e^{\frac{r(a')}{\tau}}} = \frac{(|\mathcal{A}| - 1)\Delta}{e^{\frac{\Delta}{\tau}} + |\mathcal{A}| - 1}. \tag{152}$$

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

$$\frac{d}{d\Delta} \frac{(|\mathcal{A}| - 1)\Delta}{e^{\frac{\Delta}{\tau}} + |\mathcal{A}| - 1} = 0 \implies \frac{(|\mathcal{A}| - 1)\left(e^{\frac{\Delta}{\tau}} + |\mathcal{A}| - 1\right) - \frac{(|\mathcal{A}|-1)\Delta e^{\frac{\Delta}{\tau}}}{\tau}}{\left(e^{\frac{\Delta}{\tau}} + |\mathcal{A}| - 1\right)^2} = 0 \tag{153}$$

$$\implies (|\mathcal{A}| - 1)\left(e^{\frac{\Delta}{\tau}} + |\mathcal{A}| - 1\right) - \frac{(|\mathcal{A}| - 1)\Delta e^{\frac{\Delta}{\tau}}}{\tau} = 0 \implies \tau\left(e^{\frac{\Delta}{\tau}} + |\mathcal{A}| - 1\right) = \Delta e^{\frac{\Delta}{\tau}} \tag{154}$$

$$\implies \tau(|\mathcal{A}| - 1) = (\Delta - \tau)e^{\frac{\Delta}{\tau}} \implies \frac{\Delta - \tau}{\tau} e^{\frac{\Delta}{\tau}} = |\mathcal{A}| - 1 \implies \frac{\Delta - \tau}{\tau} e^{\frac{\Delta - \tau}{\tau}} = \frac{|\mathcal{A}| - 1}{e} \tag{155}$$

$$\implies W\left(\frac{|\mathcal{A}| - 1}{e}\right) = \frac{\Delta - \tau}{\tau} \implies \Delta = \tau\left(W\left(\frac{|\mathcal{A}| - 1}{e}\right) + 1\right), \tag{156}$$

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, $\Delta = \tau\left(W\left(\frac{|\mathcal{A}|-1}{e}\right) + 1\right)$ 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 Eq. (152), we get

$$(\pi^* - \pi^*_\tau)^\top r \leq \frac{(|\mathcal{A}| - 1)\tau\left(W\left(\frac{|\mathcal{A}|-1}{e}\right) + 1\right)}{e^{W\left(\frac{|\mathcal{A}|-1}{e}\right)+1} + |\mathcal{A}| - 1}. \tag{157}$$

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

$$= \frac{(|\mathcal{A}| - 1)\tau\left(W\left(\frac{|\mathcal{A}|-1}{e}\right) + 1\right)}{\frac{|\mathcal{A}|-1}{W\left(\frac{|\mathcal{A}|-1}{e}\right)} + |\mathcal{A}| - 1} \tag{158}$$

$$= \tau W\left(\frac{|\mathcal{A}| - 1}{e}\right). \tag{159}$$

■

Lemma 12 *For a fixed τ, assuming $c = \inf_{t_1 \leq t < t_2} \min_a \pi_{\theta_t}(a) > 0$, and using the update rule $\theta_{t+1} = \theta_t + \eta \nabla f^\tau(\theta_t)$, where $\eta = 1/L^\tau$, we have*

$$f^{*\tau} - f^\tau(\theta_{t_2}) \leq \exp\{-\eta\, \tau\, c^2\, (t_2 - t_1)\}[f^{*\tau} - f^\tau(\theta_{t_1})], \tag{160}$$

where $t_1 < t_2$.

Proof Using the L^τ-smoothness of the entropy regularized objective function, we have

$$f^\tau(\theta_{t+1}) \geq f^\tau(\theta_t) + \langle \nabla f^\tau(\theta_t), \theta_{t+1} - \theta_t \rangle - \frac{L^\tau}{2} ||\theta_{t+1} - \theta_t||_2^2 \tag{161}$$

Using the update rule $\theta_{t+1} = \theta_t + \eta \nabla f^\tau(\theta_t)$,

$$= f^\tau(\theta_t) + \eta ||\nabla f^\tau(\theta_t)||_2^2 - \frac{L^\tau \eta^2}{2} ||\nabla f^\tau(\theta_t)||_2^2 \tag{162}$$

Using $\eta = 1/L^\tau$,

$$= f^\tau(\theta_t) + \frac{\eta}{2} ||\nabla f^\tau(\theta_t)||_2^2 \tag{163}$$

Using Theorem 31, we have

$$\geq f^\tau(\theta_t) + \eta\, \tau\, \min_a \pi_{\theta_t}(a)^2\, [f^{*\tau} - f^\tau(\theta_t)]. \tag{164}$$

$$\implies f^{*\tau} - f^\tau(\theta_{t+1}) \leq \left(1 - \eta\, \tau\, \min_a \pi_{\theta_t}(a)^2\right) [f^{*\tau} - f^\tau(\theta_t)] \tag{165}$$

Now, using $1 - x \leq \exp(-x)$,

$$\leq \exp\left(-\eta\, \tau\, \min_a \pi_{\theta_t}(a)^2\right) [f^{*\tau} - f^\tau(\theta_t)] \tag{166}$$

Assuming $c = \inf_{t_1 \leq t < t_2} \min_a \pi_{\theta_t}(a) > 0$, we have

$$\leq \exp(-\eta\, \tau\, c^2\, (t_2 - t_1))[f^{*\tau} - f^\tau(\theta_{t_1})]. \tag{167}$$

■

Lemma 13 *For a fixed θ and τ, we have*

$$(\pi_\tau^* - \pi_\theta)^\top r \leq f^{*\tau} - f^\tau(\theta) + \tau \log |\mathcal{A}|. \tag{168}$$

Proof

$$(\pi_\tau^* - \pi_\theta)^\top r = {\pi_\tau^*}^\top (r - \tau \log \pi_\tau^*) - {\pi_\theta}^\top (r - \tau \log \pi_\theta) + \tau(\pi_\tau^* \log \pi_\tau^* - \pi_\theta \log \pi_\theta) \tag{169}$$

Since $\log \frac{1}{|\mathcal{A}|} \leq {\pi_\theta}^\top \log \pi_\theta \leq 0 \quad \forall \theta$,

$$\leq f^{*\tau} - f^\tau(\theta) + \tau \left(0 - \log \frac{1}{|\mathcal{A}|}\right) \tag{170}$$

$$= f^{*\tau} - f^\tau(\theta) + \tau \log |\mathcal{A}|. \tag{171}$$

■

Theorem 4 *Setting* $\eta_t = 1/L^\tau$, $\tau = \epsilon/\left(2\,W\left((|\mathcal{A}|-1)/e\right)\right)$ *where* $W(x)$ *is the Lambert function, Update 3 achieves* ϵ*-suboptimality after* $\mathcal{O}\left(1/\epsilon \log 1/\epsilon\right)$ *iterations.*

Proof We have

$$\delta^\tau(\theta_t) = (\pi^* - \pi_{\theta_t})^\top r \tag{172}$$

$$= (\pi^* - \pi_\tau^*)^\top r + (\pi_\tau^* - \pi_{\theta_t})^\top r \tag{173}$$

Now, using Theorem 11,

$$\leq \tau W\left(\frac{|\mathcal{A}|-1}{e}\right) + (\pi_\tau^* - \pi_{\theta_t})^\top r \tag{174}$$

Furthermore, using Theorem 13, we have

$$\leq \tau\left(W\left(\frac{|\mathcal{A}|-1}{e}\right) + \log|\mathcal{A}|\right) + f^{*\tau} - f^\tau(\theta_t) \tag{175}$$

Using $\tau = \epsilon/\left(2\left(W\left(\frac{|\mathcal{A}|-1}{e}\right) + \log|\mathcal{A}|\right)\right)$,

$$\leq \frac{\epsilon}{2} + f^{*\tau} - f^\tau(\theta_t). \tag{176}$$

Therefore, to show that $\delta^\tau(\theta_t) \leq \epsilon$, it suffices that

$$f^{*\tau} - f^\tau(\theta_t) \leq \frac{\epsilon}{2}. \tag{177}$$

According to Lemma 13 from [10], we have $\min_a \pi_{\theta_t}(a) > c \;\forall t \geq 0$, where $c = 1/(|\mathcal{A}|\,\exp(\frac{1}{\tau})\,\exp(4(||\theta_0||_\infty + \frac{1}{\tau})\sqrt{|\mathcal{A}|})) > 0$, which we consider as a constant. Therefore, we can use Theorem 12:

$$f^{*\tau} - f^\tau(\theta_t) \leq \exp(-\eta\,\tau\,c^2\,t)[f^{*\tau} - f^\tau(\theta_0)] \tag{178}$$

Now, to show that $(\pi_\tau^* - \pi_{\theta_t})^\top r \leq \frac{\epsilon}{2}$, it suffices that

$$\exp(-\eta\,\tau\,c^2\,t)[f^{*\tau} - f^\tau(\theta_0)] \leq \frac{\epsilon}{2} \tag{179}$$

$$\iff \exp(\tau\,\eta\,c^2\,t) \geq \frac{2[f^{*\tau} - f^\tau(\theta_0)]}{\epsilon} \tag{180}$$

$$\iff \tau\,\eta\,c^2\,t \geq \log\left(\frac{2[f^{*\tau} - f^\tau(\theta_0)]}{\epsilon}\right) \tag{181}$$

Since $\eta = 1/L^\tau = 1/\left(\frac{5}{2} + \tau\,5\,(1+\log|\mathcal{A}|)\right)$,

$$\iff t \geq \frac{\frac{5}{2} + \tau\,5\,(1+\log|\mathcal{A}|)}{\tau\,c^2} \log\left(\frac{2[f^{*\tau} - f^\tau(\theta_0)]}{\epsilon}\right) \tag{182}$$

Again, using $\tau = \epsilon/\left(2\left(W\left(\frac{|\mathcal{A}|-1}{e}\right) + \log|\mathcal{A}|\right)\right)$,

$$\iff t \geq 5\left[\frac{W\left(\frac{|\mathcal{A}|-1}{e}\right) + \log|\mathcal{A}|}{\epsilon\,c^2} + \frac{1+\log|\mathcal{A}|}{c^2}\right] \log\left(\frac{2[f^{*\tau} - f^\tau(\theta_0)]}{\epsilon}\right). \tag{183}$$

Therefore, only $T \in \mathcal{O}\left(\frac{1}{\epsilon}\log\frac{1}{\epsilon}\right)$ iterations are required for achieving ϵ-suboptimality. ■

D.3. Proof of Theorem 15

Lemma 14 *For a fixed θ, if $\tau_2 < \tau_1$, then*

$$f^{*\tau_2} - f^{\tau_2}(\theta) \leq f^{*\tau_1} - f^{\tau_1}(\theta) + \tau_1 W\left(\frac{|\mathcal{A}|-1}{e}\right) + \tau_1 \log|\mathcal{A}|. \tag{184}$$

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

$$[f^{*\tau_2} - f^{\tau_2}(\theta)] - [f^{*\tau_1} - f^{\tau_1}(\theta)] = [f^{*\tau_2} - f^{*\tau_1}] - [f^{\tau_2}(\theta) - f^{\tau_1}(\theta)] \tag{185}$$

$$= \left[{\pi^*_{\tau_2}}^\top (r - \tau_2 \log \pi^*_{\tau_2}) - {\pi^*_{\tau_1}}^\top (r - \tau_1 \log \pi^*_{\tau_1})\right] - [{\pi_\theta}^\top (r - \tau_2 \log \pi_\theta) - {\pi_\theta}^\top (r - \tau_1 \log \pi_\theta)] \tag{186}$$

$$= (\pi^*_{\tau_2} - \pi^*_{\tau_1})^\top r - \left[\tau_2 {\pi^*_{\tau_2}}^\top \log \pi^*_{\tau_2} - \tau_1 {\pi^*_{\tau_1}}^\top \log \pi^*_{\tau_1}\right] + (\tau_2 - \tau_1) {\pi_\theta}^\top \log \pi_\theta \tag{187}$$

Since $\log \frac{1}{|\mathcal{A}|} \leq {\pi_\theta}^\top \log \pi_\theta \leq 0 \quad \forall \theta$,

$$\leq (\pi^*_{\tau_2} - \pi^*_{\tau_1})^\top r - \left[\tau_2 \log \frac{1}{|\mathcal{A}|} - \tau_1 0\right] + (\tau_2 - \tau_1) \log \frac{1}{|\mathcal{A}|} \leq (\pi^* - \pi^*_{\tau_1})^\top r + \tau_1 \log|\mathcal{A}|. \tag{188}$$

Now, using Theorem 11,

$$\implies f^{*\tau_2} - f^{\tau_2}(\theta) \leq f^{*\tau_1} - f^{\tau_1}(\theta) + \tau_1 W\left(\frac{|\mathcal{A}|-1}{e}\right) + \tau_1 \log|\mathcal{A}|. \tag{189}$$

■

Theorem 15 *Using step size $\eta_i = 1/L^{\tau_i}$, and considering $\hat{c} = \inf_i(\hat{c}_i)$ as a constant, where $\hat{c}_i = 1/(|\mathcal{A}| \exp\left(\frac{1}{\tau_i}\right) \exp\left(4(||\theta_{last_{i-1}}||_\infty + \frac{1}{\tau_i})\sqrt{|\mathcal{A}|}\right)$, Algorithm 1 achieves ϵ-suboptimality after $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ iterations.*

Proof Observe that in Algorithm 1, we use τ_i and η_i at stage $i \geq 1$, which starts at iteration $\text{last}_{i-1}+1$, runs for $T_i = \frac{1}{\eta_i \tau_i c_i^2} \log\left(\frac{\tau_{i-1}}{\tau_i}\left(1 + W\left(\frac{|\mathcal{A}|-1}{e}\right) + \log|\mathcal{A}|\right)\right)$ 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

$$f^{*\tau_0} - f^{\tau_0}(\theta_0) \leq \max(\tau_0, f^{*\tau_0} - f^{\tau_0}(\theta_0)) = \tau_0 \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right). \tag{190}$$

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 $c_i = \min_{\text{last}_{i-1} \leq t < \text{last}_i} \min_a \pi_{\theta_t}(a)$, we use Theorem 12 for stage i:

$$f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_i}) \leq \exp(-\tau_i \eta_i c_i^2 T_i)[f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_{i-1}})] \tag{191}$$

Using $T_i = \frac{1}{\eta_i \tau_i c_i^2} \log\left(\frac{\tau_{i-1}}{\tau_i}\left(1 + W\left(\frac{|\mathcal{A}|-1}{e}\right) + \log|\mathcal{A}|\right)\right)$, we have

$$\leq \frac{f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_{i-1}})}{\exp\left\{\log\left(\frac{\tau_{i-1}}{\tau_i}\left(1 + W\left(\frac{|\mathcal{A}|-1}{e}\right) + \log|\mathcal{A}|\right)\right)\right\}} \tag{192}$$

Now, using Theorem 14,

$$\leq \frac{f^{*\tau_{i-1}} - f^{\tau_{i-1}}(\theta_{\text{last}_{i-1}}) + \tau_{i-1} W\left(\frac{|\mathcal{A}|-1}{e}\right) + \tau_{i-1} \log|\mathcal{A}|}{\frac{\tau_{i-1}}{\tau_i}\left(1 + W\left(\frac{|\mathcal{A}|-1}{e}\right) + \log|\mathcal{A}|\right)} \tag{193}$$

Using the inductive hypothesis,

$$\leq \frac{\tau_i\, \tau_{i-1}\left(\max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right) + W\left(\frac{|\mathcal{A}|-1}{e}\right) + \log|\mathcal{A}|\right)}{\tau_{i-1}\left(1 + W\left(\frac{|\mathcal{A}|-1}{e}\right) + \log|\mathcal{A}|\right)} \tag{194}$$

$$\leq \frac{\tau_i \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right)\left(1 + W\left(\frac{|\mathcal{A}|-1}{e}\right) + \log|\mathcal{A}|\right)}{1 + W\left(\frac{|\mathcal{A}|-1}{e}\right) + \log|\mathcal{A}|} \tag{195}$$

$$= \tau_i \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right). \tag{196}$$

Therefore, $f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_i}) \leq \tau_i \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right)$ holds for all $i \geq 0$. As a result, we can use Theorem 13 to find an upper bound for $(\pi^*_{\tau_i} - \pi_{\theta_{\text{last}_i}})^\top r$ that is proportional to τ_i:

$$(\pi^*_{\tau_i} - \pi_{\theta_{\text{last}_i}})^\top r \leq f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_i}) + \tau_i \log|\mathcal{A}| \tag{197}$$

$$\leq \tau_i\left(\max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right) + \log|\mathcal{A}|\right) \tag{198}$$

Therefore, using Theorem 11, the total suboptimality at the end of each stage $\epsilon_i := (\pi^* - \pi_{\theta_{\text{last}_i}})^\top r$ has an upper bound that is proportional to the corresponding τ_i:

$$\epsilon_i = (\pi^* - \pi_{\theta_{\text{last}_i}})^\top r \tag{199}$$

$$= (\pi^* - \pi^*_{\tau_i})^\top r + (\pi^*_{\tau_i} - \pi_{\theta_{\text{last}_i}})^\top r \tag{200}$$

$$\leq \tau_i\, C_1 \tag{201}$$

where $C_1 = W\left(\frac{|\mathcal{A}|-1}{e}\right) + \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right) + \log|\mathcal{A}|$. Now, since $\tau_i = 2^{-i}\,\tau_0$, the suboptimality ϵ_i has an exponential rate in terms of the number of executed stages:

$$\leq 2^{-i}\,\tau_0\, C_1 \tag{202}$$

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

$$2^{N_{\text{stages}}} \leq \frac{\tau_0\, C_1}{\epsilon} \implies N_{\text{stages}} \leq \log_2\left(\frac{\tau_0\, C_1}{\epsilon}\right). \tag{203}$$

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

$$T_i = \frac{\log\left(\frac{\tau_{i-1}}{\tau_i} C_2\right)}{\eta_i\, \tau_i\, c_i^2}, \tag{204}$$

where $C_2 = 1 + W\left(\frac{|\mathcal{A}|-1}{e}\right) + \log|\mathcal{A}|$. According to Lemma 13 from [10], we have $c_i = \min_{\text{last}_{i-1} \leq t < \text{last}_i} \min_a \pi_{\theta_t}(a) \geq \hat{c}_i$, where $\hat{c}_i = 1/(|\mathcal{A}| \exp(\frac{1}{\tau_i}) \exp(4(||\theta_{\text{last}_{i-1}}||_\infty + \frac{1}{\tau_i})\sqrt{|\mathcal{A}|}))$. We also assumed that $\hat{c} = \inf_{i>0} \hat{c}_i$ is a constant. Since $c_i \geq \hat{c}_i \geq \hat{c}$, we have

$$\leq \frac{\log\left(\frac{\tau_{i-1}}{\tau_i} C_2\right)}{\eta_i \, \tau_i \, \hat{c}^2} \tag{205}$$

Since $\eta_i = 1/L^{\tau_i} = 1/\left(\frac{5}{2} + \tau_i \, 5\,(1 + \log|\mathcal{A}|)\right)$,

$$= \frac{\left[\frac{5}{2} + \tau_i \, 5\,(1 + \log|\mathcal{A}|)\right] \, \log\left(\frac{\tau_{i-1}}{\tau_i} C_2\right)}{\tau_i \, \hat{c}^2} \tag{206}$$

$$= \frac{\frac{5}{2} \, \log\left(\frac{\tau_{i-1}}{\tau_i} C_2\right)}{\tau_i \, \hat{c}^2} + C_3 \tag{207}$$

where $C_3 = 5\,(1 + \log|\mathcal{A}|) \, \log(2\,C_2)/\hat{c}^2$. Since $\tau_i = 2^{-i} \, \tau_0$, we have

$$= \frac{\frac{5}{2} \, \log(2\,C_2) \, 2^i}{\tau_0 \, \hat{c}^2} + C_3, \tag{208}$$

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

$$T_{\text{Total}} = \sum_{i=1}^{N_{\text{stages}}} T_i \leq \sum_{i=1}^{N_{\text{stages}}} \left[\frac{\frac{5}{2} \, \log(2\,C_2) \, 2^i}{\tau_0 \, \hat{c}^2} + C_3\right] \tag{209}$$

$$= \frac{\frac{5}{2} \, \log(2\,C_2) \, \sum_{i=1}^{N_{\text{stages}}} 2^i}{\tau_0 \, \hat{c}^2} + C_3 \, N_{\text{stages}} \tag{210}$$

$$= \frac{\frac{5}{2} \, \log(2\,C_2) \, [2^{N_{\text{stages}}+1} - 2]}{\tau_0 \, \hat{c}^2} + C_3 \, N_{\text{stages}} \tag{211}$$

$$\leq \frac{5 \, \log(2\,C_2) \, 2^{N_{\text{stages}}}}{\tau_0 \, \hat{c}^2} + C_3 \, N_{\text{stages}} \tag{212}$$

Using Eq. (203), we have

$$\leq \frac{5\,C_1 \, \log(2\,C_2)}{\epsilon \, \hat{c}^2} + C_3 \, \log_2\left(\frac{\tau_0 \, C_1}{\epsilon}\right) \tag{213}$$

$$\implies T_{\text{Total}} \in \mathcal{O}\left(\frac{1}{\epsilon}\right). \tag{214}$$

■

Appendix E. Stochastic Policy Gradient with Entropy Regularization Proofs

E.1. Proof of Theorem 5

Lemma 16 *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

$$\frac{T}{\log T} > 2 > C. \tag{215}$$

Otherwise, if $C \geq 2$,

$$2\,C\,\log C = C(\log C + \log C) \tag{216}$$

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

$$\geq C(\log C + \log(2\log C)) = C\log(2\,C\,\log C) \tag{217}$$

$$\implies \frac{2\,C\,\log C}{\log(2\,C\,\log C)} \geq C. \tag{218}$$

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

$$\frac{T}{\log T} \geq \frac{2\,C\,\log C}{\log(2\,C\,\log C)} \geq C. \tag{219}$$

■

Lemma 17 *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

$$\frac{T}{\log^2 T} > 75 > C. \tag{220}$$

Otherwise, if $C \geq 75$,

$$4\,C\,\log^2 C = C(\log C + \log C)^2 \tag{221}$$

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

$$\geq C(\log C + \log(4\log^2 C))^2 = C\log^2(4\,C\,\log^2 C) \tag{222}$$

$$\implies \frac{4\,C\,\log^2 C}{\log^2(4\,C\,\log^2 C)} \geq C. \tag{223}$$

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

$$\frac{T}{\log^2 T} \geq \frac{4\,C\,\log^2 C}{\log^2(4\,C\,\log^2 C)} \geq C. \tag{224}$$

■

Lemma 18 *Assuming* $c = \min_{t_1 \leq t < t_2} \min_a \pi_{\theta_t}(a) > 0$. *Using Update 4 with exponential step-size* $\eta_t = \eta_{t_1}\, \alpha^{t-t_1}$, *where* $\eta_{t_1} = 1/L^\tau$ *and* $\alpha = (1/T)^{1/T}$, *where* $T = t_2 - t_1 > 0$, $\hat{\epsilon}$*-suboptimality is achieved if* $T = \max(5583, 2\,Y_1\,\log Y_1, 4\,Y_2\,\log^2 Y_2)$, *where*

$$Y_1 = \frac{\log\left(\frac{2\,B_1\,\mathbb{E}[f^{*\tau} - f^\tau(\theta_{t_1})]}{\hat{\epsilon}}\right)}{2\,B_2\,\tau\,c^2}, \quad Y_2 = \frac{B_3\,b}{2\,\tau^2\,c^4\,\hat{\epsilon}}, \tag{225}$$

where $B_1 = \exp\left(\frac{8}{5}\right)$, $B_2 = \frac{1.38}{5+10(1+\log K)}$, *and* $B_3 = \frac{5}{e^2}\left(\frac{5}{2} + 5(1+\log K)\right)\exp(\frac{8}{5})$.

Proof According to Theorem 1 from [9], using exponential step-size $\eta_t = \eta_{t_1}\,\alpha^{t-t_1}$, where $\eta_{t_1} = 1/L^\tau$ and $\alpha = (1/T)^{1/T}$, where $T = t_2 - t_1$, we have

$$\mathbb{E}[f^{*\tau} - f^\tau(\theta_{t_2})] \leq X_1\,\exp\left(-X_2\,\mu\,\frac{T}{\log T}\right)\mathbb{E}[f^{*\tau} - f^\tau(\theta_{t_1})] + \frac{X_3\,b}{\mu^2\,\frac{T}{\log^2 T}}, \tag{226}$$

where

$$X_1 = \exp\left(\frac{2\,\mu}{L^\tau\,\log T}\right), \quad X_2 = \frac{0.69}{L^\tau}, \quad X_3 = \frac{5\,L^\tau\,X_1}{e^2}. \tag{227}$$

According to Theorem 31 and since $0 \leq \tau \leq 1$, we have $\mu = 2\,\tau\,\min_a \pi(\theta) \leq 2$. Furthermore, $\frac{5}{2} \leq L^\tau = \frac{5}{2} + 5\,\tau(1+\log K) \leq \frac{5}{2} + 5(1+\log K)$ and $\log T \geq 1$. Therefore,

$$X_1 \leq B_1 = \exp\left(\frac{8}{5}\right), \tag{228}$$

$$X_2 \geq B_2 = \frac{0.69}{\frac{5}{2} + 5(1+\log K)}, \tag{229}$$

$$X_3 \leq B_3 = \frac{5\left(\frac{5}{2} + 5(1+\log K)\right)\exp(\frac{8}{5})}{e^2}. \tag{230}$$

Hence, we can safely substitute variables X_1, X_2, X_3 with their corresponding constants B_1, B_2, B_3. Now, assuming that $c = \inf_{t \geq 0} \min_a \pi_{\theta_t}(a) > 0$, according to Theorem 31, we have $\mu \geq 2\,\tau\,c^2$. Therefore,

$$\mathbb{E}[f^{*\tau} - f^\tau(\theta_{t_2})] \leq B_1\,\exp\left(-2\,B_2\,\tau\,c^2\,\frac{T}{\log T}\right)\mathbb{E}[f^{*\tau} - f^\tau(\theta_{t_1})] + \frac{B_3\,b}{4\,\tau^2\,c^4\,\frac{T}{\log^2 T}}. \tag{231}$$

We show that if the inequalities $\frac{T}{\log T} \geq Y_1$ and $\frac{T}{\log^2 T} \geq Y_2$ are satisfied, where

$$Y_1 = \frac{\log\left(\frac{2\,B_1\,\mathbb{E}[f^{*\tau} - f^\tau(\theta_{t_1})]}{\hat{\epsilon}}\right)}{2\,B_2\,\tau\,c^2}, \quad Y_2 = \frac{B_3\,b}{2\,\tau^2\,c^4\,\hat{\epsilon}}, \tag{232}$$

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

$$\mathbb{E}[f^{*\tau} - f^\tau(\theta_{t_2})] \leq B_1\,\exp\left(-2\,B_2\,\tau\,c^2\,\frac{1}{2\,B_2\,\tau\,c^2}\log\left(\frac{2\,B_1\,[f^{*\tau} - f^\tau(\theta_{t_1})]}{\hat{\epsilon}}\right)\right)\mathbb{E}[f^{*\tau} - f^\tau(\theta_{t_1})] \tag{233}$$

$$+ \frac{B_3\,b}{4\,\tau^2\,c^4\,\frac{B_3\,b}{2\,\tau^2\,c^4\,\hat{\epsilon}}} = \frac{\hat{\epsilon}}{2} + \frac{\hat{\epsilon}}{2} = \hat{\epsilon}. \tag{234}$$

Now, according to Theorem 16, for $\frac{T}{\log T} \geq Y_1$ to hold, it suffices that $T \geq \max(2, 2\,Y_1\,\log Y_1)$. Furthermore, according to Theorem 17, for $\frac{T}{\log^2 T} \geq Y_2$ to hold, it suffices that $T \geq \max(5583, 4\,Y_2\,\log^2 Y_2)$. Therefore, the required number of iterations to achieve $\hat{\epsilon}$-suboptimality is $T = \max(5583, 2\,Y_1\,\log Y_1, 4\,Y_2\,\log^2 Y_2)$. ■

Theorem 5 *Assuming* $c := \inf_{t\geq 0} \min_a \pi_{\theta_t}(a) > 0$. *Using Update 4 with* $\tau = \epsilon / \left(2\left(W\left(|\mathcal{A}|^{-1}/e\right) + \log|\mathcal{A}|\right)\right)$ *and using exponential decreasing step-size* $\eta_t = \eta_0\,\alpha^t$, *where* $\eta_0 = 1/L^\tau$, *achieves* ϵ*-suboptimality after* $\tilde{\mathcal{O}}\left(1/\epsilon + b/\epsilon^3\right)$ *iterations.*

Proof We have

$$\mathbb{E}[\delta_t] = \mathbb{E}[(\pi^* - \pi_{\theta_t})^\top r] \tag{235}$$

$$= (\pi^* - \pi^*_\tau)^\top r + \mathbb{E}[(\pi^*_\tau - \pi_{\theta_t})^\top r] \tag{236}$$

Now, using Theorem 11,

$$\leq \tau W\left(\frac{|\mathcal{A}| - 1}{e}\right) + \mathbb{E}[(\pi^*_\tau - \pi_{\theta_t})^\top r] \tag{237}$$

Furthermore, using Theorem 13, we have

$$\leq \tau\left(W\left(\frac{|\mathcal{A}| - 1}{e}\right) + \log|\mathcal{A}|\right) + \mathbb{E}[f^{*\tau} - f^\tau(\theta_t)] \tag{238}$$

Using $\tau = \epsilon / \left(2\left(W\left(\frac{|\mathcal{A}|-1}{e}\right) + \log|\mathcal{A}|\right)\right)$,

$$\leq \frac{\epsilon}{2} + \mathbb{E}[f^{*\tau} - f^\tau(\theta_t)]. \tag{239}$$

Therefore, to show that $\mathbb{E}[\delta_t] \leq \epsilon$, it suffices to show that

$$\mathbb{E}[f^{*\tau} - f^\tau(\theta_t)] \leq \frac{\epsilon}{2}. \tag{240}$$

Since $c = \inf_{t\geq 0} \min_a \pi_{\theta_t}(a) > 0$, according to Theorem 18, using exponential step-size $\eta_t = \eta_0\,\alpha^t$, where $\eta_0 = 1/L^\tau$ and $\alpha = (1/T)^{1/T}$, for $\mathbb{E}[f^{*\tau} - f^\tau(\theta_T)] \leq \frac{\epsilon}{2}$ to hold, it suffices that $T = \max(5583, 2\,Y_1\,\log Y_1, 4\,Y_2\,\log^2 Y_2)$, where

$$Y_1 = \frac{\log\left(\frac{4\,B_1\,\mathbb{E}[f^{*\tau} - f^\tau(\theta_{t_1})]}{\hat{\epsilon}}\right)}{2\,B_2\,\tau\,c^2}, \quad Y_2 = \frac{B_3\,b}{\tau^2\,c^4\,\epsilon}, \tag{241}$$

where $B_1 = \exp\left(\frac{8}{5}\right)$, $B_2 = \frac{1.38}{5+10(1+\log|\mathcal{A}|)}$, and $B_3 = \frac{5}{e^2}\left(\frac{5}{2} + 5(1+\log|\mathcal{A}|)\right)\exp(\frac{8}{5})$. Again, using $\tau = \epsilon / \left(2\left(W\left(\frac{|\mathcal{A}|-1}{e}\right) + \log|\mathcal{A}|\right)\right)$, we have

$$Y_1 \geq \frac{\left(W\left(\frac{|\mathcal{A}|-1}{e}\right) + \log|\mathcal{A}|\right)\log\left(\frac{4\,B_1\,[f^{*\tau} - f^\tau(\theta_0)]}{\epsilon}\right)}{B_2\,c^2\,\epsilon}, \tag{242}$$

$$Y_2 \geq \frac{4\,B_3\,\left(W\left(\frac{|\mathcal{A}|-1}{e}\right) + \log|\mathcal{A}|\right)^2\,b}{c^4\,\epsilon^3}. \tag{243}$$

Therefore, only $T \in \tilde{\mathcal{O}}\left(\frac{1}{\epsilon} + \frac{b}{\epsilon^3}\right)$ iterations are required for achieving ϵ-suboptimality. ■

E.2. Proof of Theorem 19

Theorem 19 *Assuming that* $c = \inf_i \min_{last_{i-1} \leq t < last_i} \min_a \pi_{\theta_t}(a) > 0$ *for each stage* i. *Using Algorithm 3 with exponential step-size* $\eta_{i,t} = \eta_{i,last_{i-1}} \, \alpha_i^{t-last_{i-1}}$, *where* $\eta_{i,last_{i-1}} = 1/L^{\tau_i}$ *achieves* ϵ*-suboptimality after* $\tilde{\mathcal{O}}\left(\frac{1}{\epsilon} + \frac{b}{\epsilon^3}\right)$ *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

$$T_i^{'} = \frac{\log\left(\frac{2\,B_1\,\tau_{i-1}\left(1+W\left(\frac{K-1}{e}\right)+\log K\right)}{\tau_i}\right)}{2\,B_2\,\tau_i\,c_i^2}, \quad T_i^{''} = \frac{B_3\,b}{2\,\tau_i^3\,c_i^4}, \tag{244}$$

where $B_1 = \exp\left(\frac{8}{5}\right)$, $B_2 = \frac{1.38}{5+10(1+\log K)}$, and $B_3 = \frac{5}{e^2}\left(\frac{5}{2} + 5(1+\log K)\right)\exp(\frac{8}{5})$. 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

$$f^{*\tau_0} - f^{\tau_0}(\theta_0) \leq \max(\tau_0, f^{*\tau_0} - f^{\tau_0}(\theta_0)) = \tau_0 \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right). \tag{245}$$

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. Since $c_i = \min_{\text{last}_{i-1} \leq t < \text{last}_i} \min_a \pi_{\theta_t}(a) > 0$, according to Theorem 18, using exponential step-size $\eta_{i,t} = \eta_{i,\text{last}_{i-1}} \cdot \alpha_i^{t-\text{last}_{i-1}}$, where $\eta_{i,\text{last}_{i-1}} = 1/L^{\tau_i}$ and $\alpha_i = (1/T_i)^{1/T_i}$ at stage i, 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 $T_i = \max(5583, 2\,X_i\,\log X_i, 4\,X_i^{'}\,\log^2 X_i^{'})$, where

$$X_i = \frac{\log\left(\frac{2\,B_1\,\mathbb{E}[f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_{i-1}})]}{\tau_i \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right)}\right)}{2\,B_2\,\tau_i\,c_i^2}, \quad X_i^{'} = \frac{B_3\,b}{2\,\tau_i^3\,c_i^4 \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right)}. \tag{246}$$

Now, using Theorem 14,

$$X_i \leq \frac{\log\left(\frac{2\,B_1\left(\mathbb{E}[f^{*\tau_{i-1}} - f^{\tau_{i-1}}(\theta_{\text{last}_{i-1}})] + \tau_{i-1} W\left(\frac{|\mathcal{A}|-1}{e}\right) + \tau_{i-1}\log|\mathcal{A}|\right)}{\tau_i \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right)}\right)}{2\,B_2\,\tau_i\,c_i^2} \tag{247}$$

Using the inductive hypothesis,

$$\leq \frac{\log\left(\frac{2\,B_1\left(\tau_{i-1}\max\left(1,\frac{f^{*\tau_0}-f^{\tau_0}(\theta_0)}{\tau_0}\right)+\tau_{i-1}W\left(\frac{|\mathcal{A}|-1}{e}\right)+\tau_{i-1}\log|\mathcal{A}|\right)}{\tau_i\max\left(1,\frac{f^{*\tau_0}-f^{\tau_0}(\theta_0)}{\tau_0}\right)}\right)}{2\,B_2\,\tau_i\,c_i^2} \tag{248}$$

$$\leq \frac{\log\left(\frac{2\,B_1\,\tau_{i-1}\max\left(1,\frac{f^{*\tau_0}-f^{\tau_0}(\theta_0)}{\tau_0}\right)\left(1+W\left(\frac{|\mathcal{A}|-1}{e}\right)+\log|\mathcal{A}|\right)}{\tau_i\max\left(1,\frac{f^{*\tau_0}-f^{\tau_0}(\theta_0)}{\tau_0}\right)}\right)}{2\,B_2\,\tau_i\,c_i^2} \tag{249}$$

$$= \frac{\log\left(\frac{2\,B_1\,\tau_{i-1}\left(1+W\left(\frac{|\mathcal{A}|-1}{e}\right)+\log|\mathcal{A}|\right)}{\tau_i}\right)}{2\,B_2\,\tau_i\,c_i^2} = T_i^{''}. \tag{250}$$

On the other hand, we have

$$X_i^{'} \leq \frac{B_3\,b}{2\,\tau_i^3\,c_i^4} = T_i^{'}, \tag{251}$$

which is exactly the number of iterations at stage i. Therefore, $\mathbb{E}[f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_i})] \leq \tau_i \max\left(1, \frac{f^{*\tau_0}-f^{\tau_0}(\theta_0)}{\tau_0}\right)$ holds for all $i \geq 0$. As a result, we can use Theorem 13 to find an upper bound for $\mathbb{E}[(\pi^*_{\tau_i} - \pi_{\theta_{\text{last}_i}})^\top r]$ that is proportional to τ_i:

$$\mathbb{E}[(\pi^*_{\tau_i} - \pi_{\theta_{\text{last}_i}})^\top r] \leq \mathbb{E}[f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_i})] + \tau_i \log|\mathcal{A}| \tag{252}$$

$$\leq \tau_i(\max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right) + \log|\mathcal{A}|) \tag{253}$$

Therefore, using Theorem 11, the total suboptimality at the end of each stage $\epsilon_i := \mathbb{E}[(\pi^* - \pi_{\theta_{\text{last}_i}})^\top r]$ has an upper bound that is proportional to the corresponding τ_i:

$$\epsilon_i = \mathbb{E}[(\pi^* - \pi_{\theta_{\text{last}_i}})^\top r] \tag{254}$$

$$= (\pi^* - \pi^*_{\tau_i})^\top r + \mathbb{E}[(\pi^*_{\tau_i} - \pi_{\theta_{\text{last}_i}})^\top r] \tag{255}$$

$$\leq \tau_i\,C_1 \tag{256}$$

where $C_1 = W\left(\frac{|\mathcal{A}|-1}{e}\right) + \max\left(1, \frac{f^{*\tau_0}-f^{\tau_0}(\theta_0)}{\tau_0}\right) + \log|\mathcal{A}|$. Now, since $\tau_i = 2^{-i}\,\tau_0$, the suboptimality ϵ_i has an exponential rate in terms of the number of executed stages:

$$\leq 2^{-i}\,\tau_0\,C_1 \tag{257}$$

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

$$2^{N_{\text{stages}}} \leq \frac{\tau_0\,C_1}{\epsilon} \implies N_{\text{stages}} \leq \log_2\left(\frac{\tau_0\,C_1}{\epsilon}\right). \tag{258}$$

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

$$T_i = \max\left(5583, \frac{\log\left(\frac{2\,B_1\,\tau_{i-1}\,C_2}{\tau_i}\right)}{B_2\,\tau_i\,c_i^2}\,\log\left(\frac{\log\left(\frac{2\,B_1\,\tau_{i-1}\,C_2}{\tau_i}\right)}{2\,B_2\,\tau_i\,c_i^2}\right), \frac{2\,B_3\,b}{\tau_i^3\,c_i^4}\,\log^2\left(\frac{B_3\,b}{2\,\tau_i^3\,c_i^4}\right)\right) \tag{259}$$

where $C_2 = 1 + W\left(\frac{|\mathcal{A}|-1}{e}\right) + \log|\mathcal{A}|$. Since $c = \inf_{i>0} c_i$, we have

$$\leq \max\left(5583, \frac{\log\left(\frac{2\,B_1\,\tau_{i-1}\,C_2}{\tau_i}\right)}{B_2\,\tau_i\,c^2}\,\log\left(\frac{\log\left(\frac{2\,B_1\,\tau_{i-1}\,C_2}{\tau_i}\right)}{2\,B_2\,\tau_i\,c^2}\right), \frac{2\,B_3\,b}{\tau_i^3\,c^4}\,\log^2\left(\frac{B_3\,b}{2\,\tau_i^3\,c^4}\right)\right) \tag{260}$$

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

$$= \max\left(5583, \frac{\log(4\,B_1\,C_2)\,2^i}{B_2\,\tau_0\,c^2}\,\log\left(\frac{\log(4\,B_1\,C_2)\,2^i}{2\,B_2\,\tau_0\,c^2}\right), \frac{2\,B_3\,b\,8^i}{\tau_0^3\,c^4}\,\log^2\left(\frac{B_3\,b\,8^i}{2\,\tau_0^3\,c^4}\right)\right) \tag{261}$$

$$\leq \max\left(5583, \frac{\log(4\,B_1\,C_2)\,2^i}{B_2\,\tau_0\,c^2}\,\log\left(\frac{\log(4\,B_1\,C_2)\,2^{N_{\text{stages}}}}{2\,B_2\,\tau_0\,c^2}\right), \frac{2\,B_3\,b\,8^i}{\tau_0^3\,c^4}\,\log^2\left(\frac{B_3\,b\,8^{N_{\text{stages}}}}{2\,\tau_0^3\,c^4}\right)\right) \tag{262}$$

$$= \max\left(5583, \frac{\log(4\,B_1\,C_2)\,2^i}{B_2\,\tau_0\,c^2}\,Y_1, \frac{2\,B_3\,b\,8^i}{\tau_0^3\,c^4}\,Y_2\right) \tag{263}$$

where $Y_1 = \log\left(\frac{\log(4\,B_1\,C_2)\,2^{N_{\text{stages}}}}{2\,B_2\,\tau_0\,c^2}\right)$ and $Y_2 = \log^2\left(\frac{B_3\,b\,8^{N_{\text{stages}}}}{2\,\tau_0^3\,c^4}\right)$ Consequently, we can calculate the total number of required iterations T_{Total} in terms of ϵ:

$$T_{\text{Total}} = \sum_{i=1}^{N_{\text{stages}}} T_i \tag{264}$$

$$\leq \sum_{i=1}^{N_{\text{stages}}} \max\left(5583, \frac{\log(4\,B_1\,C_2)\,2^i}{B_2\,\tau_0\,c^2}\,Y_1, \frac{2\,B_3\,b\,8^i}{\tau_0^3\,c^4}\,Y_2\right) \tag{265}$$

$$\leq \max\left(5583\,N_{\text{stages}}, \frac{\log(4\,B_1\,C_2)\,\sum_{i=1}^{N_{\text{stages}}} 2^i}{B_2\,\tau_0\,c^2}\,Y_1, \frac{2\,B_3\,b\,\sum_{i=1}^{N_{\text{stages}}} 8^i}{\tau_0^3\,c^4}\,Y_2\right) \tag{266}$$

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

$$\leq \max\left(5583\,N_{\text{stages}}, \frac{\log(4\,B_1\,C_2)\,\left[2^{N_{\text{stages}}+1}-2\right]}{B_2\,\tau_0\,c^2}\,Y_1, \frac{2\,B_3\,b\,\left[\frac{8^{N_{\text{stages}}+1}-1}{7}-1\right]}{\tau_0^3\,c^4}\,Y_2\right) \tag{267}$$

$$\leq \max\left(5583\,N_{\text{stages}}, \frac{\log(4\,B_1\,C_2)\,2^{N_{\text{stages}}+1}}{B_2\,\tau_0\,c^2}\,Y_1, \frac{2\,B_3\,b\,\frac{1}{7}\,8^{N_{\text{stages}}+1}}{\tau_0^3\,c^4}\,Y_2\right) \tag{268}$$

$$\leq \max\left(5583\,N_{\text{stages}}, \frac{2\,\log(4\,B_1\,C_2)\,2^{N_{\text{stages}}}}{B_2\,\tau_0\,c^2}\,Y_1, \frac{16\,B_3\,b\,8^{N_{\text{stages}}}}{7\,\tau_0^3\,c^4}\,Y_2\right) \tag{269}$$

Using Eq. (258), we have

$$\leq \max\left(5583\,\log_2\left(\frac{\tau_0\,C_1}{\epsilon}\right), \frac{2\,\log(4\,B_1\,C_2)\,C_1\,\log\left(\frac{\log(4\,B_1\,C_2)\,C_1}{2\,B_2\,c^2\,\epsilon}\right)}{B_2\,c^2\,\epsilon}, \frac{16\,B_3\,C_1^3\,\log^2\left(\frac{B_3\,C_1^3\,b}{2\,c^4\,\epsilon^3}\right)\,b}{7\,c^4\,\epsilon^3}\right) \tag{270}$$

$$\implies T_{\text{Total}} \in \tilde{\mathcal{O}}\left(\frac{1}{\epsilon}+\frac{b}{\epsilon^3}\right). \tag{271}$$

■

In Theorem 15, we assumed that the $c_i = \min_{\text{last}_{i-1}\leq t<\text{last}_i} \min_a \pi_{\theta_t}(a)$ constants, which come from the non-uniform Łojasiewicz constant $C(\theta) = \sqrt{2\tau}\min_a \pi_\theta(a)$ in Theorem 31, have a lower bound. In practice, we observe that c_i quickly decreases as the algorithm progresses, thus Algorithm 1 does not seem to achieve the proposed $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ rate of convergence. Our experiments suggest that we can tighten the non-uniform Łojasiewicz condition by substituting the $\min_a \pi_\theta(a)$ term with $\sqrt{f^{*\tau}-f^\tau(\theta)}$. We also know that Algorithm 1 maintains $f^{*\tau_i}-f^{\tau_i}(\theta_t)\geq\tau_i \quad \forall \text{last}_{i-1}\leq t<\text{last}_i$ at every stage i. Therefore, we can estimate the number of iterations T_i at stage i by substituting c_i with $\sqrt{\tau_i}$, which results in Algorithm 2. According to the experiments, Algorithm 2 seems to achieve an $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ rate of convergence to the global optimal policy. Similarly in Algorithm 3, we substitute c_i by $\sqrt{\tau_i}$ to calculate the number of iterations T_i at stage i.

Appendix F. Algorithms

F.1. Multi-Stage Entropy Regularized Policy Gradient

Algorithm 1: Multi-Stage Entropy Regularized Policy Gradient

Output: Policy $\pi_{\theta_t} = \text{softmax}(\theta_t)$
Initialize parameters $\theta_0 \in \mathbb{R}^{\mathcal{A}}, \tau_0 \in \mathbb{R}$;
$t \leftarrow 0$;
$\text{last}_0 \leftarrow t$;
while $i \geq 1$ **do**
  $\tau_i \leftarrow \tau_{i-1}/2$;
  $\eta_i \leftarrow 1/L^{\tau_i}$;
  $c_i \leftarrow 1$;
  do
    $c_i \leftarrow \min\left(c_i, \min_a \pi_{\theta_t}(a)\right)$; /* $c_i = \min_{\texttt{last}_{i-1}\leq t<\texttt{last}_i} \min_a \pi_{\theta_t}(a)$ */
    $\theta_{t+1} \leftarrow \theta_t + \eta_i \nabla f^{\tau_i}(\theta_t)$;
    $t \leftarrow t+1$;
    $\text{last}_i \leftarrow t$;
  while $last_i - last_{i-1} < \frac{1}{\eta_i\,\tau_i\,c_i^2}\log\left(\frac{\tau_{i-1}}{\tau_i}\left(1+W\left(\frac{K-1}{e}\right)+\log K\right)\right)$;
end

Algorithm 2: Practical Multi-Stage Entropy Regularized Policy Gradient

Output: Policy $\pi_{\theta_t} = \mathrm{softmax}(\theta_t)$
Initialize parameters $\theta_0 \in \mathbb{R}^{|\mathcal{A}|}, \tau_0 \in \mathbb{R}$;
$t \leftarrow 0$;
$\text{last}_0 \leftarrow t$;
while $i \geq 1$ **do**
  $\tau_i \leftarrow \tau_{i-1}/2$;
  $\eta_i \leftarrow 1/L^{\tau_i}$;
  $c_i \leftarrow \sqrt{\tau_i}$; `/* c_i is substituted by sqrt(tau_i) */`
  $T_i \leftarrow \frac{1}{\eta_i \tau_i^2} \log\left(\frac{\tau_{i-1}}{\tau_i}\left(1 + W\left(\frac{|\mathcal{A}|-1}{e}\right) + \log|\mathcal{A}|\right)\right)$;
  while $t - \mathit{last}_{i-1} < T_i$ **do**
    $\theta_{t+1} \leftarrow \theta_t + \eta_i \nabla f^{\tau_i}(\theta_t)$;
    $t \leftarrow t + 1$;
  end
  $\text{last}_i \leftarrow t$;
end

F.2. Stochastic Multi-Stage Entropy Regularized Policy Gradient

Algorithm 3: Stochastic Multi-Stage Entropy Regularized Policy Gradient

Output: Policy $\pi_{\theta_t} = \mathrm{softmax}(\theta_t)$
Initialize parameters $\theta_0 \in \mathbb{R}^{|\mathcal{A}|}, \tau_0 \in \mathbb{R}$;
$t \leftarrow 0$;
$\text{last}_0 \leftarrow t$;
while $i \geq 1$ **do**
  $\tau_i \leftarrow \tau_{i-1}/2$;
  $c_i \leftarrow \sqrt{\tau_i}$; /* c_i is substituted by $\sqrt{\tau_i}$ */
  $B_1 \leftarrow \exp\left(\frac{8}{5}\right)$;
  $B_2 \leftarrow \frac{1.38}{5+10(1+\log|\mathcal{A}|)}$;
  $B_3 \leftarrow \frac{5}{e^2}\left(\frac{5}{2}+5(1+\log|\mathcal{A}|)\right)\exp(\frac{8}{5})$;
  $b \leftarrow 8(1+(\tau_i \log|\mathcal{A}|)^2)$;
  $T_i' \leftarrow \frac{1}{2\,B_2\,\tau_i\,c_i^2}\log\left(\frac{2\,B_1\,\tau_{i-1}}{\tau_i}\left(1+W\left(\frac{|\mathcal{A}|-1}{e}\right)+\log|\mathcal{A}|\right)\right)$;
  $T_i'' \leftarrow \frac{B_3\,b}{2\,\tau_i^3\,c_i^4}$;
  $T_i \leftarrow \max(5583, 2\,T_i'\,\log T_i', 4\,T_i''\,\log^2 T_i'')$;
  $\alpha_i \leftarrow (1/T_i)^{1/T_i}$;
  $\eta_{i,t} \leftarrow 1/L^{\tau_i}$;
  while $t - \mathit{last}_{i-1} < T_i$ **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
  $\text{last}_i \leftarrow t$;
end

Appendix G. Experiments

We evaluate the presented step-size schedules in single-state MDP environments in both the deterministic and stochastic setting. In the deterministic setting, we compare PG with line-search (`PGLS`) to PG with a fixed step-size (`PG`), `GNPG` [11], and PG with entropy regularization (`PG-E`) [10]. In the stochastic setting, we compare SPG with exponential step-size (`SPG-ESS`) to prior work that uses the full gradient (`J-SPG`) [11] and the reward gap (`B-SPG`) [14] when setting the step-size. Additionally, we compare the presented multi-stage PG with entropy regularization Algorithm 2 in deterministic setting (`PG-E-MS`) and Algorithm 3 in stochastic setting (`SPG-E-MS`). We find that `SPG-ESS` and `SPG-E-MS` are competitive with `J-SPG` even though they do not access to the full gradient. `B-SPG` preforms poorly since a small reward gap leads to a small step-size.

For each experiment, we plot the average (across 50 independent runs) suboptimality gap against the number of iterations. In each run, a deterministic reward vectors is randomly generated in the range of $[0,1]^{|\mathcal{A}|}$ with $|\mathcal{A}| = 10$ and, with a maximum reward gap $\bar{\Delta} := \max_{a^* \neq a} r(a^*) - r(a) = 0.5$ and vary the minimum reward gap $\underline{\Delta} := \min_{a^* \neq a} r(a^*) - r(a)$ with $\underline{\Delta} \in \{0.2, 0.1, 0.05\}$. We use uniform initialization such that $\pi_{\theta_0}(a) = \frac{1}{|\mathcal{A}|}$ for all $a \in |\mathcal{A}|$. For each entropy regularized method,

we initially set $\tau = 1$. Note that due to entropy-regularization, PG−E cannot converge to the globally optimal policy since τ is fixed. For SPG−ESS, we used $\beta = 7.5 \times 10^6$ when the minimum reward gap is 0.05 or 0.1 and $\beta = 1$ otherwise. β was determined via grid-search.

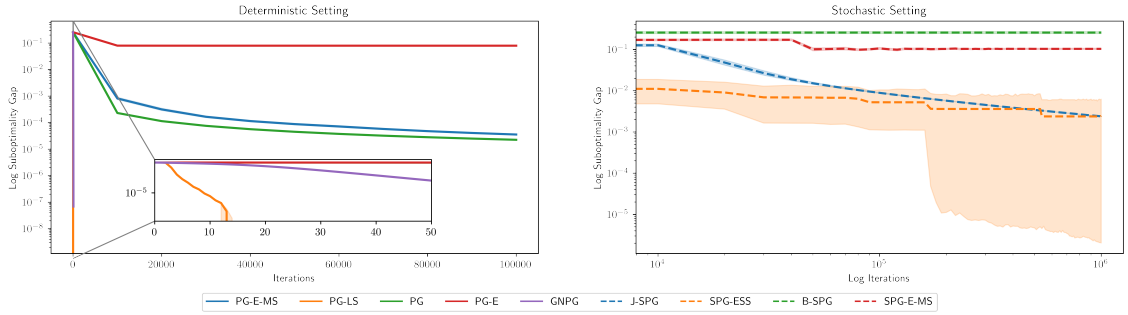

Figure 1: Minimum reward gap: 0.05

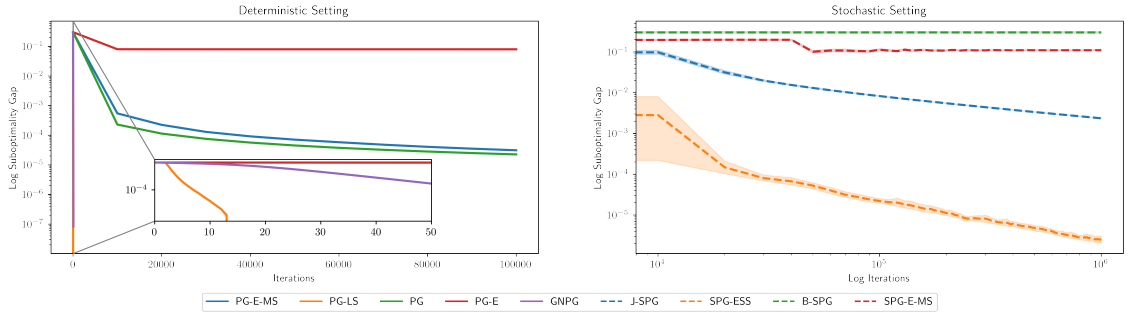

Figure 2: Minimum reward gap: 0.1

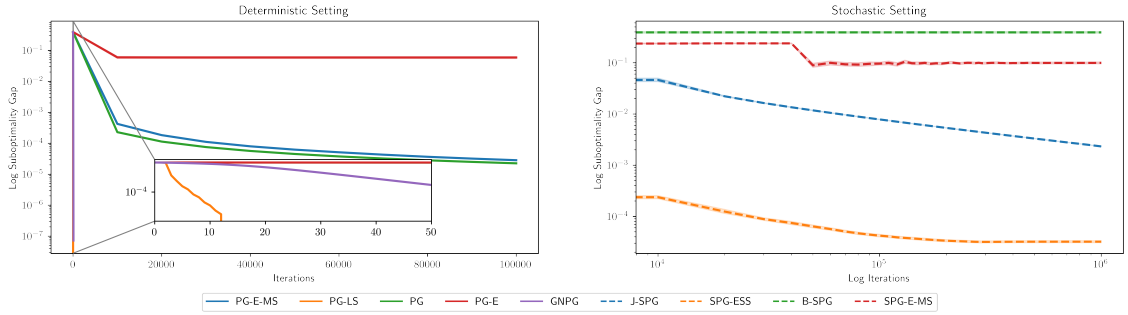

Figure 3: Minimum reward gap: 0.2

Appendix H. Extra Lemmas

For completeness, we append external lemmas here.

Lemma 20 (Ascent lemma for smooth function (Lemma 18 in [10])) *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a L-smooth function, $\theta \in \mathbb{R}^d$ and $\theta' = \theta + \frac{1}{L}\nabla f(\theta)$. We have,*

$$f(\theta) - f(\theta') \leq -\frac{1}{2L}\left\|\nabla f(\theta)\right\|_2^2 \tag{272}$$

H.1. Policy Gradients

Lemma 21 (Uniform Smoothness (Lemma 2 in [10])) $\forall r \in [0,1]^K$ $\theta \mapsto \pi_\theta^\top r$ *is* $5/2$*-smooth.*

Lemma 22 (Non-uniform Smoothness (Lemma 2 in [12])) *Denote $\theta_\zeta := \theta + \zeta(\theta' - \theta)$ with some $\zeta \in [0,1]$. For any $r \in [0,1]^K$, $\theta \mapsto \pi_\theta^\top r$ satisfies $\beta(\theta_\zeta)$ non-uniform smoothness with $\beta(\theta_\zeta) = 3\left\|\frac{d\pi_{\theta_\zeta}^\top r}{d\theta_\zeta}\right\|_2$*

Lemma 23 (Lemma 3 in [12]) *Let $\theta' = \theta + \eta \frac{\frac{d\pi_\theta^\top r}{d\theta}}{\left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2}$ Denote $\theta_\zeta := \theta + \zeta(\theta' - \theta)$ with some $\zeta \in [0,1]$. We have for all $\eta \in (0, \frac{1}{3})$,*

$$\left\|\frac{d\pi_{\theta_\zeta}^\top r}{d\theta_\zeta}\right\|_2 \leq \frac{1}{1-3\eta}\left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2 \tag{273}$$

H.2. Stochastic Policy Gradients

[On-policy IS (Definition 1 from [11])] At iteration t, sample one action $a \sim \pi_{\theta_t}(\,)$. The IS reward estimator $\hat{r}_t$ is constructed as $\hat{r}_t(a) = \frac{\mathbb{1}\{a_t = a\}}{\pi_{\theta_t}(a)} r(a)$ for all $a \in [K]$.

Lemma 24 (Equation (115 - 117) from [11] / Proof of Theorem 2 in [12]) *Denote $\theta_\zeta := \theta + \zeta(\theta' - \theta)$ with some $\zeta \in [0,1]$. By Theorem 22 we have*

$$\left|(\pi_{\theta_{t+1}} - \pi_{\theta_t})^\top r - \left\langle \frac{d\pi_{\theta_t}^\top}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle\right| \leq \frac{3}{2}\left\|\frac{d\pi_{\theta_{\zeta_t}}^\top r}{d\theta_{\zeta_t}}\right\|_2 \left\|\theta_{t+1} - \theta_t\right\|_2^2 \tag{274}$$

By Lemma (23) with $\eta \in (0, \frac{1}{3})$

$$\leq \frac{3}{2}\frac{1}{1-3\eta}\left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2 \left\|\theta_{t+1} - \theta_t\right\|_2^2 \tag{275}$$

Lemma 25 (Lemma 5 from [11]) *Let $\hat{r}$ be the IS estimator using on-policy sampling $a \sim \pi_{\theta_t}(\,)$. Then stochastic softmax PG estimator is:*
Unbiased: $\mathbb{E}_{a\sim\pi_\theta}\left[\nabla\widetilde{f}(\theta)\right] = \nabla f(\theta)$
Bounded Variance: $\mathbb{E}_{a\sim\pi_\theta}\left\|\nabla\widetilde{f}(\theta)\right\|_2^2 \leq 2 \Rightarrow \sigma^2 := \mathbb{E}_{a\sim\pi_\theta}\left[\nabla\widetilde{f}(\theta) - \nabla f(f)\right] = \mathbb{E}_{a\sim\pi_\theta}\left\|\nabla\widetilde{f}(\theta)\right\|_2^2 - \mathbb{E}_{a\sim\pi_\theta}\|\nabla f(\theta)\|_2^2 \leq 2$

Note: Following Lemma 4.3 in [15], we can also show that the variance converges to 0 as the policy becomes deterministic
Proof

$$\mathbb{E}_t\left[\nabla\widetilde{f}(\theta_t)\right] = \sum_{a\in[K]}\left[\left\|\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}\right\|_2^2 \| a_t = a\right] \tag{276}$$

$$\leq 2R_{\max}^2 \sum_{a\in[K]} \pi_{\theta_t}(a)(1-\pi_{\theta_t}(a))^2 \tag{277}$$

Let $k_t := \arg\max_{a\in[K]} \pi_{\theta_t}(a)$

$$\tag{278}$$

$$= 2R_{\max}^2\left[\pi_{\theta_t}(k_t)(1-\pi_{\theta_t}(k_t))^2 + \sum_{a\neq k_t}\pi_{\theta_t}(a)(1-\pi_{\theta_t}(a))^2\right] \tag{279}$$

Since $\pi_{\theta_t}(a) \in (0,1)$

$$\leq 4R_{\max}^2(1-\pi_{\theta_t}(k_t))^2 \tag{280}$$

■

Lemma 26 *$\theta \to \pi_\theta^\top(r - \tau\log\pi_\theta)$ is $\frac{5}{2} + \tau\, 5\,(1+\log K)$-smooth.*

Proof Starting with the definition of L-smooth

$$\left|\pi_{\theta'}^\top(r-\tau\log\pi_{\theta'}) - \pi_\theta^\top(r-\tau\log\pi_\theta) - \left\langle\frac{d\pi_\theta^\top(r-\tau\log\pi_\theta)}{d\theta}, \theta'-\theta\right\rangle\right|$$

$$= \left|(\pi_{\theta'}-\pi_\theta)^\top r + \tau(-\pi_{\theta'}^\top\log\pi_{\theta'} - (-\pi_\theta^\top\log\pi_\theta) - \left\langle\frac{d\pi_\theta^\top r}{d\theta}, \theta'-\theta\right\rangle - \tau\left\langle\frac{d-\pi_\theta^\top\log\pi_\theta}{d\theta}, \theta'-\theta\right\rangle\right| \tag{281}$$

let $h(\theta) = -\pi_\theta^\top\log\pi_\theta$

$$\leq \left|(\pi_{\theta'}-\pi_\theta)^\top r - \left\langle\frac{d\pi_\theta^\top r}{d\theta}, \theta'-\theta\right\rangle\right| + \tau\left|h(\theta') - h(\theta) - \left\langle\frac{\partial h(\theta_t)}{\partial\theta}, \theta'-\theta\right\rangle\right| \tag{282}$$

By Theorem 21 and Theorem 28

$$\leq \frac{5/2 + \tau\, 5\,(1+\log K)}{2}\|\theta'-\theta\|^2 \tag{283}$$

■

Lemma 27 (Lemma 4.3 in [14]) *Using Update 2, we have for all $t \geq 1$,*

$$\mathbb{E}_t\left[\left\|\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}\right\|_2^2\right] \leq \frac{8\, R_{\max}^3\, K^{3/2}}{\Delta^2}\left\|\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}\right\|_2 \tag{284}$$

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

Lemma 28 (Lemma 14 from [10]) *$\theta \to -\pi_\theta^\top \log \pi_\theta$ is $5\,(1 + \log K)$-smooth*

Lemma 29 (Non-uniform Łojasiewicz (Lemma 3 in [10])) *Assume r has one unique optimal action. Let $\pi^* = \max_{\pi\in\Pi} \pi^\top r$ Then*

$$\left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2 \geq C(\theta)\,(\pi^* - \pi_\theta)^\top r \tag{285}$$

with

$$C(\theta) := \pi_\theta(a^*) \tag{286}$$

Lemma 30 (Non-uniform Łojasiewicz (Lemma 8 in [10])) *We have,*

$$\left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta}\right\|_2 \geq C(\theta)\,(V^*(\rho) - V^{\pi_\theta})(\rho)) \tag{287}$$

with

$$C(\theta) := \frac{\min_s \pi_\theta(a^*(s)|s)}{\sqrt{|\mathcal{S}|}\left\|d_\rho^{\pi^*}/d_\mu^{\pi_\theta}\right\|_\infty} \tag{288}$$

Lemma 31 (Proposition 5 in [10]) *In the single-state MDP setting the non-uniform Łojasiewicz condition is*

$$\left\|\frac{d\{\pi_\theta^\top (r - \tau\log\pi_\theta)\}}{d\theta}\right\|_2 \geq C(\theta)\left(\mathbb{E}_{a\sim\pi_\tau^*}[r(a) - \tau\log\pi_\tau^*] - \mathbb{E}_{a\sim\pi_\theta}[r(a) - \tau\log\pi_\theta]\right)^{\frac{1}{2}} \tag{289}$$

with

$$C(\theta) := \sqrt{2\tau}\,\min_a \pi_\theta(a) \tag{290}$$