Regret minimization of Tabular Policy Gradient*

Tal Dim† **Yishay Mansour**‡

1 Introduction

Reinforcement Learning (RL) learns to control a complex, unknown environment through interaction. RL algorithms were highly successfully applied to various domains. Policy gradient methods optimize directly a parameterized policy by computing gradients of the value function and updating the parameters. Policy gradient methods have been highly successful in many applications areas (such as playing GO [Silver et al. 2016] and robotics [Deisenroth et al. 2013]) and are among the most effective methods in Reinforcement Learning.

The policy gradient theorem [Sutton et al. 1999] gives a method of computing the gradient as a function of the observable quantities in the MDP. Using Monte-Carlo methods and Reinforce Williams (1992) or actor-critic methodology give popular implementations of the policy gradient methodology. However, those fundamental results do not derive convergence bounds or performance guarantees.

Our starting point is the work of Agarwal et al. (2020b) which studies the number of gradient steps required to reach an ϵ approximation of the optimal value function, assuming we receive exact gradients. The main contribution of our work is to avoid the assumption of exact gradients, and directly approximate the gradients from observations and in addition derive vanishing regret bounds.

More specifically, this work focuses on finite horizon Markov Decision Process (MDP) with tabular policy parameterizations, i.e., there is a parameter per state-action pair. Our policy gradient algorithms approximate the policy gradient using episodes sampled from the MDP, i.e., we do not assume direct access to exact policy gradient.

We analyze the algorithms with respect to the regret, i.e., the difference between the expected return of the optimal policy and the expected return of the policies chosen by the online algorithm (which is running a policy gradient algorithm). We stress that our focus is on understanding the performance of the widely used policy gradient methodology and not on deriving new algorithms or new performance guarantees. (See the related work section for a variety of regret minimization algorithms.)

We consider two parameterized policy classes (similar to Agarwal et al. (2020b)). The first, *direct parameterization*, simply encodes the policy using a lookup table. The second, *softmax*, encodes a weight per state-action, and selects an action using a softmax function of the weights. At a high level, our algorithms work in phases, where during each phase they sample episodes using the current policy, and at the end of the phase compute an approximation of the policy gradient and update the policy.

In the direct parameterization for every state-action pair s, a the parameter $\theta_{s,a}$ is the probability to choose action a in state s. In every phase we perform a policy gradient update of the parameters, and project the new parameters to the policy simplex. For the direct parameterization we show a regret bound of $\widetilde{O}(K^{\frac{5}{6}} H^{\frac{7}{6}} |S|^{\frac{2}{3}} |A|^{\frac{4}{3}} D)$,[1] where K is the total number of episodes, S and A are the set of states and actions, H is the horizon, and we have an MDP parameter D which we will define later.

*This project has received funding from the European Research Council (ERC) under the European Union'sHorizon 2020 research and innovation program (grant agreement No. 882396), by the Israel Science Foundation(grant number 993/17), Tel Aviv University Center for AI and Data Science (TAD), and the Yandex Initiative for Machine Learning at Tel Aviv University

†Blavatnik School of Computer Science, Tel Aviv Univercity taldim1@gmail.com

‡Blavatnik School of Computer Science, Tel Aviv Univercity and Google Research. mansour.yishay@gmail.com

[1]The notation $\widetilde{O}(\cdot)$ ignores logarithmic factors.

In the softmax parameterization we use a variant of the natural policy gradient algorithm, where for a parameter vector $\theta \in \mathbb{R}^{|S|\times|A|}$, the probability of choosing action a in state s is proportional to $\exp(\theta_{s,a})$. In the softmax setting, in order to have efficient approximations of the policy gradient, we need to make some assumption about the MDP:

- In the first setting we assume that with some probability $\lambda > 0$, the initial state of the MDP is uniformly chosen and a random action is performed. After this, the algorithm follows the current policy. The regret of the policy gradient with softmax parameterization in this setting is bounded by $\widetilde{O}(H^2|S|^{\frac{1}{3}}|A|^{\frac{1}{3}}K^{\frac{2}{3}}\lambda^{-\frac{1}{3}})$.
- In the second setting, where potentially $\lambda = 0$, we assume a reset for the initial state. Namely, the algorithm can choose an initial state and action to start from. To ensure a "meaningful" regret bound, the return of episodes not starting at state s_0 are ignored in the regret analysis. The regret of the softmax policy gradient algorithm in the this settings is bounded by $\widetilde{O}(H^{\frac{9}{4}}\sqrt{|S|\ |A|}K^{\frac{3}{4}})$.

While our results apply to tabular policy parameterizations, we believe that they shade light on the success of the policy gradient in general. Having a vanishing regret is a very stringent requirement, which highlight the soundness on the policy gradient methodology.

1.1 Related Works

Regret minimization using the optimism in face of uncertainty principle: There is a vast literature on regret minimization in RL for the tabular setting, that mostly builds on the optimism in face of uncertainty principle.

Jaksch et al. (2010) Presents the UCRL2 algorithm, using the approach of optimism in the face of uncertainty. That is, it defines a set of plausible MDPs given the observations so far, chooses an optimistic MDP with respect to the set of models, and executes it. In the regret analysis they use a notation of average reward, which means every episode instead of the random reward that the agent received, they use the expected value of the reward given the MDP and the policy. They define the diameter D of an MDP, which is the maximal path between any two states, and prove a regret bound of $\tilde{O}(DS\sqrt{AK})$ after K steps assuming average reward, and show it is near optimal by presenting a lower bound of $\Omega(\sqrt{DSAK})$ for every online learning algorithm.

Bartlett and Tewari (2009) presents an algorithm called REGAL, which is inspired by the UCRL2 algorithm. The algorithm obtains a $\tilde{O}(cS\sqrt{AK})$ regret bound in the larger class of weakly communicating MDPs, where c is the bound on the span of the bias function.

Fruit et al. (2018) presents the SCAL algorithm, with a regret bound of $\tilde{O}(c\sqrt{\Gamma SAK})$, where $\Gamma \leq S$ possible next states for every state, and c is the bound for the span of the optimal bias function (Similar to the value of c in Bartlett and Tewari (2009)).

Abbasi-Yadkori et al. (2019) presents the POLITEX algorithm, using a model-free settings and function approximation. They assume that the value function error after running a policy for τ time steps scales as $\epsilon(\tau) = \epsilon_0 + \tilde{O}(\sqrt{d/\tau})$, where ϵ_0 is the worst-case approximation error and d is the number of features in a compressed representation of the state-action space. They show a regret bound of $\tilde{O}(d^{1/2}K^{3/4} + \epsilon_0 K)$.

Regret minimization with finite horizon MDPs Osband et al. (2016) presents a randomized least-squares value iteration (RLSVI) algorithm using linearly parameterized value functions. They present a regret bound of $\tilde{O}(\sqrt{H^3SAK}$, where H is the finite horizon.

Dann et al. (2017) presents a new framework for theoretically measuring the performance of RL algorithms. To show the benefits of the new framework they present an algorithm called Uniform-PAC with a regret bound of $\tilde{O}(H^2\sqrt{SAK} + S^3A^2)$, and simultaneously achieves the optimal regret and PAC guarantees except for a factor of the horizon.

Azar et al. (2017) show that an optimistic modification to value iteration achieves a regret bound of $\tilde{O}(\sqrt{HSAK})$ assuming K is large enough ($K \geq H^3S^3A$) and $SA \geq H$. They define Bernstein-based "exploration bonuses" that use the empirical variance of the estimated values at the next states.

Jin et al. (2018) addresses Model-free RL, and show that a Q-learning algorithm with UCB exploration achieves a regret bound of $\tilde{O}(\sqrt{H^3SAK})$.

Zanette and Brunskill (2019) presents an algorithm for finite horizon discrete MDPs which not only conforms with the previously known state of the art regret bound of $\tilde{O}(\sqrt{HSAK})$, but also provably obtains much tighter guarantees if the domain has a small variance of the the quality function Q^* distribution, or a small bound in the possible achievable reward.

Efroni et al. (2019) focus on model-based RL in the finite-state finite-horizon MDP, and establish that exploring with greedy policies can achieve tight regret bound of $\tilde{O}(\sqrt{HSAK})$. Therefore, full-planning in model-based RL can be avoided, and by doing so the computational complexity decreases by a factor of S.

Cai et al. (2020) presents an Optimistic variant of the Proximal Policy Optimization (OPPO) algorithm, which follows an "optimistic version" of the policy gradient direction. They assume that the MDP is linear, i.e., that the transition dynamics are linear in a feature map, and show a regret bound of $\tilde{O}(\sqrt{d^2H^3K})$, where d is the dimension of the feature map and H is the episode horizon.

Zhang et al. (2020) presents the Monotonic Value Propagation (MVP) algorithm, which relies on a new Bernstein-type bonus. To handle long planning horizon problems, they present a regret bound of $\tilde{O}(\sqrt{SAK} + S^2A)$, which has only logarithmic dependence on the horizon H. This regret bound approaches the $\Omega(\sqrt{SAK})$ lower bound of contextual bandits up to logarithmic terms.

Sample complexity: Works on sample complexity can be traced back to [Kearns and Singh 2002, Brafman and Tennenholtz 2002, Kakade 2003]. Dann et al. (2019) presents an upper bounds for the PAC model: $\widetilde{O}(\frac{SAH^2}{\epsilon^2} + \frac{S^2AH^3}{\epsilon})$ while Dann and Brunskill (2015) gives a lower bound of $\widetilde{\Omega}(\frac{SAH^2}{\epsilon^2})$. Additional sample complexity bounds bounds appear in [Lattimore and Hutter 2012, Azar et al. 2012, Dann et al. 2017].

Policy gradient in non-tabular setting: The work of Shen et al. (2019) gives a sample complexity bound of $O(\epsilon^{-3})$ in the non-tabular case to reach a stationary point. Yang et al. (2020) presents a second order stationary point with a sample complexity of $\widetilde{O}(\epsilon^{-9/2})$, which is guaranteed to converge to a local maxima.

Policy gradient in tabular setting: In the works of Kakade and Langford (2002) they assume to have an ϵ-greedy policy chooser that chooses a next policy that maximizes the expected advantage function of the new policy w.r.t. the previous policy. They prove a convergence bound of $V(\pi^*) - V(\pi) \leq \frac{\epsilon}{(1-\gamma)^2} D_\infty$ where D_∞ is similar to our definition of D, and γ is the discounted return factor, after $O(\epsilon^{-2})$ calls to the ϵ-greedy policy chooser.

Mei et al. (2020) shows that with a softmax parameterization, the policy gradient algorithm converges to $V^*(\rho) - V^{\pi_t}(\rho) \leq \epsilon$ after $O(\frac{|S|}{\epsilon c(1-\gamma)^6} D_\infty \| \frac{1}{\mu} \|_\infty)$ assuming ρ is some starting state distribution, μ is the starting state distribution used by the RL algorithm, and c is defined as: $c = \inf_{s \in S. t \geq 1} \pi_t(a*(s)|s)$.

Cen et al. (2020) analyses the Natural Policy Gradient algorithm using entropy regularization and show that $\widetilde{O}(\frac{|S||A|}{(1-\gamma)^8\epsilon^2})$ samples are needed to find an ϵ-optimal policy.

Shani et al. (2020) presents the Uniform Trust region policy optimization (TRPO) algorithm, and shows that the algorithm finds an ϵ-optimal policy using the regularization constant λ given $\widetilde{O}(\frac{A^5(1+\lambda \log(A))S}{(1-\gamma)^4\epsilon^3\lambda})$ samples. In Trust Region methods a sum of two terms is iteratively being minimized: a linearization of the objective function and a proximity term which restricts two consecutive updates to be 'close' to each other

As mentioned before, the work of Agarwal et al. (2020b) analyzes the sample complexity of policy gradient in tabular setting for discounted return, assuming access to the true policy gradients. For the direct parameterization, they bound the number of the gradient updates by $O(D_\infty^2|S||A|(1-\gamma)^{-6}\epsilon^{-2})$, where D_∞ is similar to our definition of D, and γ is the discounted return factor. For the softmax parameterization, they bound the number of gradient updates by $O((1-\gamma)^{-2}\epsilon^{-1})$.

We stress that our aim is not to improve the best regret bounds but to understand the performance of the widely used policy gradient methods.

2 Model

Markov Decision Process (MDP) are defined by $M = (S, s_0, A, P, r)$, where S is a finite set of states and $s_0 \in S$ is the initial state, A is a finite set of actions, P is the transition probability function, where $P(s'|s,a)$ is the probability of reaching state s' when we are in state s and performing action a, and r is the expected reward, where $r(s,a) \in [0,1]$ is the expected reward of performing action a in state s.

We consider the finite horizon return, which is the sum of the first H rewards. We assume, w.l.o.g.,[2] that the state space is levelled, i.e., the state space S is partitioned to $H+1$ subsets $S_0, \ldots S_H$, where $S_0 = \{s_0\}$, $S_H = \{s_H\}$ and we can move only between adjacent levels. Formally, for any $s_i \in S_i$ and $s_j \in S_j$, where $j \neq i+1$ and for any action a we have that $P(s_j|s_i,a) = 0$. In addition, we assume w.l.o.g. that any state s is reachable from s_0.

A *policy* is a mapping of states to a distribution over actions, i.e., $\pi : S \to \Delta(A)$, where $\Delta(A)$ is the set of distributions over A. An *episode* using a policy π is a sequence $(s_0, a_0, r_0, s_1, \ldots s_{H-1}, a_{H-1}, r_{H-1}, s_H)$, where a_i is sampled using $\pi(\cdot|s_i)$ and s_{i+1} is sampled using $P(\cdot|s_i,a_i)$, and $r_i = r(s_i,a_i)$. The return of an episode is $\sum_{i=0}^{H-1} r_i$. We denote by $\Pr^\pi[\cdot]$ the probability distribution generated by sampling using policy π and by $\mathbb{E}^\pi$ the expectation w.r.t. $\Pr^\pi$.

Policy Parameterizations - We consider two different policy parameterizations:
Direct Parameterization: Policies are parameterized by $\theta \in [0,1]^{|S|\times|A|}$, and $\pi(a|s;\theta) = \theta_{s,a}$, where for any state s we have $\sum_{a\in A} \theta_{s,a} = 1$. A ρ-stochastic policy has $\theta \in [\rho,1]^{|S|\times|A|}$, i.e., for such policies π we have $\pi(a|s;\theta) \geq \rho$. The set of ρ-stochastic policies is denoted by Π_ρ.
Softmax Parameterization: Policies are parameterized by $\theta \in \mathbb{R}^{|S|\times|A|}$, where

$$\pi(a|s;\theta) = \frac{\exp(\theta_{s,a})}{\sum_{a'\in A}\exp(\theta_{s,a'})}$$

We will use interchangeably the notation π and θ to denote the policy.

We define the standard value functions for an MDP as follows. Given a policy π and a state $s \in S$ the value of $V^\pi(s)$ is the expected finite horizon return, when we start at state s and run until we reach s_H, i.e., for a state $s \in S_i$ we have

$$V^\pi(s) = \mathbb{E}^\pi\left[\sum_{j=i}^{H} r(s_j,a_j)|s_i = s\right]$$

where a_j is sampled using $\pi(\cdot|s_j)$ and s_{j+1} is sampled using $P(\cdot|s_j,a_j)$. Note that $V^\pi(s) \leq H$.

Given a policy π, and a state-action pair s,a, the value of $Q^\pi(s,a)$ is the expected finite horizon return, when we start at state s and perform action a, and follow the policy π until we reach s_H, i.e., for a state $s \in S_i$ we have

$$Q^\pi(s,a) = \mathbb{E}^\pi\left[\sum_{j=i}^{H} r(s_j,a_j)|s_i = s, a_i = a\right]$$

where a_j is sampled using $\pi(\cdot|s_j)$ and s_{j+1} is sampled using $P(\cdot|s_j,a_j)$.

The *advantage function* for a policy π and a state-action pair s,a is the difference between the $V^\pi(s)$ and $Q^\pi(s,a)$, i.e., $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$.

Let π^* be the optimal policy, i.e., $\pi^* = \arg\max_\pi V^\pi(s_0)$, and denote V^{π^*}, Q^{π^*} and A^{π^*} by V^*, Q^* and A^*, respectively.

Definition 2.1 *The regret of an algorithm, over K episodes, using policies $\pi_1, \pi_2, \ldots, \pi_K$, is:*

$$\mathit{Regret} = \sum_{l=1}^{K} V^*(s_0) - V^{\pi_l}(s_0)$$

[2] To avoid the assumption we can create an equivalent MDP, where there are H levels and each level includes all S states. Every edge will be as in the original MDP, only the edges move from one level to the next. The new MDP will be equivalent, and the number of states will increase by a multiplication of H

Given a policy π we define the vector of *occupancy measure* d^π as follows. For a state $s \in S$ we have that d^π_s is the probability that we reach state s when we start from state s_0 and generate an episode using policy π, i.e., $d^\pi_s = \Pr^\pi(s_t = s)$. Observe that for any policy π, $\sum_{s\in S} d^\pi_s = H$. For a level $i \in [H]$ the levelled occupancy measure $d^{\pi,i}_s = \Pr^\pi(s_i = s)$, which is a distribution over the states in level i, i.e., S_i.

For an MDP M and a parameter ρ, we define the parameter $D_\rho = \max_{\pi\in\Pi_\rho} \max_s \frac{d^{\pi*}_s}{d^\pi_s}$.[3]

Notations: A function $f : X \to Y$ is β-*smooth*, for $\beta \geq 0$, with respect to a norm $\|\cdot\|$ if for all $x_1, x_2 \in X$ we have $\|\nabla f(x_1) - \nabla f(x_2)\| \leq \beta\|x_1 - x_2\|$.

For an integer $n \geq 1$ let $[n] = \{0, \ldots, n-1\}$.

For a matrix $A \in \mathbb{R}^{n\times m}$, the Moore Penrose pseudo-inverse matrix is denoted by $A^\dagger$.
The indicator function $I(\cdot)$ equals 1 when the condition holds, and 0 otherwise.

[3]As discussed in Agarwal et al. (2020b), this parameter is essential to overcome some inherently hard to learn MDPs, such as combination locks. The difference with D_∞ in Agarwal et al. (2020a) is due to their discounted return, and their requirement that the initial state distribution has full support

3 Direct parameterization

At a high level the policy gradient algorithm works as follows. It has N phases, and in each phase it samples m episodes to estimate the policy gradient, and uses the policy gradient estimate to update the parameters. Overall there are $K = Nm$ episodes. In each phase t, using the m episodes, we compute an estimator $\widetilde{\nabla_\theta V^{\pi^t}(s_0)}$ and show that with high probability $\|\widetilde{\nabla_\theta V^{\pi^t}(s_0)} - \nabla_\theta V^{\pi^t}(s_0)\|_\infty \leq \epsilon$.

In order to create an unbiased estimator for the policy gradient, we would need approximately $m = O(1/q^2)$ episodes every phase, where $q = \min_{s\in S, a\in A} \pi^t(a|s)$. This bound would become infeasible as π^t becomes a near-deterministic policy, i.e., $\pi^t(a|s) \approx 0$. To overcome that, we consider ρ-stochastic policies, which lower bound the probability of an action by ρ, and hence $m = O(1/\rho^2)$. We stress that we do not assume that the optimal policy is ρ-stochastic.

In each phase t we compute an unbiased estimator for the gradient $\nabla_\theta V^{\pi^t}(s_0)$, which we denote by $\widetilde{\nabla_\theta V^{\pi^t}(s_0)}$. The estimator is composed by averaging m unbiased estimators, one for each episode we run in phase t.

After we compute the estimator $\widetilde{\nabla_\theta V^{\pi^t}(s_0)}$, we update the policy parameters to $\theta^{t+1} = prox_\rho(\theta^t + \eta\widetilde{\nabla_\theta V^{\pi^t}(s_0)})$ where $prox_\rho(\theta) \colon = \arg\min_{\theta' \in \Pi_\rho}\{\|\theta' - \theta\|_2^2\}$ is a projection operator to the class of ρ-stochastic policies.

For the final regret bound, we optimize the regret with respect to ρ and ϵ. In the following we elaborate on each component of the algorithm, and in the supplementary material we provide detailed proofs.

3.1 Approximating the Gradient

We bound the number of episodes m needed to approximate the gradient $\nabla_\theta V^\pi(s_0)$, with error ϵ in norm L_∞, where ϵ is a parameter we will later optimize. The policy gradient theorem states:

Theorem 3.1 (Sutton et al. (1999)) *For a policy π which is parameterized by the parameter θ:*

$$\nabla_\theta V^\pi(s) \;=\; \sum_{s'\in S} \Pr^\pi[s_t = s'|s_0 = s] \sum_{a\in A} \nabla_\theta \pi(a|s';\theta) Q^\pi(s', a) \tag{1}$$

For ρ-stochastic policies, given a parameter $\theta \in \Pi_\rho$ the gradient $\nabla_\theta \pi(a|s;\theta)$ is a unit vector that equals 1 at the (s, a) coordinate, and 0 at all other coordinates. The following Lemma is well known (see, e.g., Agarwal et al. (2020b)), and follows from the policy gradient theorem and the value of $\nabla_\theta \pi(a|s;\theta)$. We give the proof for completeness.

Lemma 3.2 *For a policy π and a state-action pair s, a, we have $\nabla_\theta V^\pi(s_0)_{s,a} = d_s^\pi Q^\pi(s, a)$.*

Consider m episodes, $\tau_0, \tau_1, \ldots, \tau_{m-1}$, sampled using a policy $\pi \in \Pi_\rho$. For each episode $\tau_i = (s_0^i, a_0^i, r_0^i, s_1^i, \ldots s_{H-1}^i, a_{H-1}^i, r_{H-1}^i, s_H^i)$ we define a vector $\vec{X^\pi}(\tau_i)$ which is an unbiased estimator of the policy gradient. For a state $s \in S_l$ (a state in level l) and an action $a \in A$ the entry s, a in the vector $\vec{X^\pi}(\tau)$ is defined as:

$$\vec{X^\pi}(\tau_i)_{s,a} = \frac{1}{\pi(a|s)} \sum_{j=l}^{H} r_j^i I(s_l^i = s, a_l^i = a) \tag{2}$$

Note that since the state space is leveled, no state can be reached twice in a single episode, therefore this is an unbiased estimator for the return from s.

Given the m sampled episodes, $\tau_0, \tau_1, \ldots, \tau_{m-1}$, using the policy π, we define an estimate of the policy gradient as:

$$\widetilde{\nabla_\theta V^\pi(s_0)} = \frac{1}{m}(\vec{X^\pi}(\tau_0) + \cdots + \vec{X^\pi}(\tau_{m-1})) \tag{3}$$

The next theorem gives a high probability bound on the deviation of the estimate from the true gradient. It establishes the required number of episodes, m, as a function of the error, ϵ, and the confidence, δ.

Lemma 3.3 *Fix $\epsilon > 0$, $\delta > 0$ and let $m = \frac{H^2|A||S|}{\epsilon^2\rho^2}\log(\frac{2|A||S|K}{\delta})$. With probability at least $1-\delta$, for all phases $t \in [N]$ we have, $\|\nabla_\theta \widetilde{V^{\pi^t}}(s_0) - \nabla_\theta V^{\pi^t}(s_0)\|_\infty \leq \epsilon$.*

3.2 Regret analysis

In the following section we sketch the analysis of the regret. We will assume that in each of the N phases we approximate the gradient with error at most ϵ. We will derive a general regret bound as a function of all our parameters, and then we will optimize over those parameters. The final regret bound would be as follows.

Theorem 3.4 *Let K be the total number of episodes, the number of phases be $N = K^{1/3}(H|A|^5|S|\log(\frac{2|A||S|K}{\delta}))^{-1/3}$ and the number of episodes in each phase be $m = K^{2/3}(H|A|^5|S|\log(\frac{2|A||S|K}{\delta}))^{1/3}$. For the parameters $\eta = (\frac{2}{3}|A|H^3+1)^{-1}$ and $\rho = \frac{|S|^{2/3}H^{1/2}\log(\frac{2|A||S|K}{\delta})^{1/6}}{|A|^{2/3}K^{1/6}}$ we have, with probability $1-\delta$, that the regret is bounded by:*

$$Regret \leq 9K^{\frac{5}{6}}D_\rho H^{\frac{7}{6}}|S|^{\frac{2}{3}}|A|^{\frac{4}{3}}\log^{\frac{1}{6}}\left(\frac{2|A||S|k}{\delta}\right)$$

Our proof strategy will be to show first that the difference between the value function of the optimal policy and the current policy is bounded by the gradient of the current policy in some direction (as in [Kakade and Langford (2002), Agarwal et al. (2020b)]). Note that z is an arbitrary stochastic policy, and not limited to a ρ-stochastic policy.

Lemma 3.5 *For any policy π,*

$$V^*(s_0) - V^\pi(s_0) \leq \frac{D_\rho}{H}\max_{z\in\Delta(A)^{|S|}}(z-\pi)^\top \nabla_\theta V^\pi(s_0)$$

The above lemma bounds the regret of a policy π as a function of its gradient $\nabla_\theta V^\pi(s_0)$, using an inner product with an arbitrary policy. In order to bound the inner product we first prove that the value function is β-smooth (similar to Agarwal et al. (2020b)).

Lemma 3.6 *The function $V^\pi(s_0)$ is $\beta = \frac{1}{3}|A|H^3$-smooth with respect to norm $\|\cdot\|_2$. I.e, for all π,$\pi' \in \Delta(A)^{|S|}$ we have,*

$$\|\nabla_\theta V^\pi(s_0) - \nabla_\theta V^{\pi'}(s_0)\|_2 \leq \frac{1}{3}|A|H^3\|\pi-\pi'\|_2 = \beta\|\pi-\pi'\|_2 \tag{4}$$

Next we introduce a notion of *approximated gradient mapping* which is defined as follows:

$$\widetilde{G^\eta(\pi^t)} = \frac{1}{\eta}\left[prox_\rho\left(\pi^t + \eta\nabla_\theta\widetilde{V^{\pi^t}}(s_0)\right) - \pi^t\right], \tag{5}$$

where π^t is the policy used in phase t. We note that in the special case where $\pi^t + \eta\nabla_\theta\widetilde{V^{\pi^t}(s_0)} \in \Pi_\rho$, i.e., $prox_\rho(\pi^t + \eta\nabla_\theta\widetilde{V^{\pi^t}(s_0)}) = \pi^t + \eta\nabla_\theta\widetilde{V^{\pi^t}(s_0)}$, then the approximated gradient mapping equals the approximated policy gradient, i.e., $\widetilde{G^\eta(\pi^t)} = \nabla_\theta\widetilde{V^{\pi^t}(s_0)}$.

The following lemma bounds the term $\max_{z\in\Delta(A)^{|S|}}\nabla_\theta V^\pi(s_0)^\top(z-\pi)$ as a function of the norm of the approximated gradient mapping, i.e., $\|\widetilde{G^\eta(\pi)}\|_2$ and the smoothness parameter β.

Lemma 3.7 *For a phase t, assuming $\|\nabla_\theta V^{\pi^t}(s_0) - \widetilde{\nabla_\theta V^{\pi^t}}(s_0)\|_\infty \leq \epsilon$, then*

$$\max_{z\in\Delta(A)^{|S|}}\nabla_\theta V^{\pi^{t+1}}(s_0)^\top(z-\pi^{t+1}) \leq H^2|A|^2\rho + 2\sqrt{|S|}\left[(\eta\beta+1)\|\widetilde{G^\eta(\pi^t)}\|_2 + \sqrt{|A||S|}\epsilon\right]$$

The regret is the sum of the difference between the expected return of the optimal policy and policies selected in the algorithm. Lemmas 3.5, 3.6 and 3.7 bound the difference at any phase as a function of the norm of the approximated gradient mapping. The next lemma bounds the sum of the norms of the approximated gradient mapping, deriving an upper bound on the regret.

Lemma 3.8 *Assuming for all phases t we have $\|\nabla_\theta V^{\pi^t}(s_0) - \widetilde{\nabla_\theta V^{\pi^t}}(s_0)\|_\infty \leq \epsilon$, then*

$$\sum_{t=1}^{N} \eta(1 - \frac{(1+\beta)\eta}{2})\|\widetilde{G^\eta(\pi^t)}\|_2^2 \leq H + \frac{1}{2}|A||S|N\epsilon^2 \tag{6}$$

To complete the proof of the regret, we set the parameters as follows. The smoothness is $\beta = \frac{1}{3}|A|H^3$, the learning rate is $\eta = (2\beta+1)^{-1}$, the minimal action probability is $\rho = \sqrt{\frac{|S|}{|A|^3 N}}$ and maximum error is $\epsilon = \sqrt{\frac{H}{|S||A|N}}$. The regret bound of Theorem 3.4 follows by combining Lemmas 3.5, 3.6, 3.7 and 3.8 using the above parameters.

4 Softmax parameterization

For the softmax parameterization we will need to be able to induce exploration. A similar issue was in the direct parameterization, and there we imposed on the policies to be ρ-stochastic. Here we take the approach of keeping the policy purely softmax (and not mixing it with the uniform distribution over actions) but adding additional assumptions about the starting state distribution of the MDP or the ability to manipulate the start state.

Computing the policy gradient The following two Lemmas are well known (see, e.g., Agarwal et al. (2020b)), and we give the proof for completeness. Given the softmax function we can compute the partial derivatives as follows:

Lemma 4.1 *For a policy π, and two state-action pairs s, a, s', a', we have $\frac{\partial \log \pi(a|s)}{\partial \theta_{s',a'}} = I(s = s')(I(a = a') - \pi(a'|s))$*

We can relate the value of the policy gradient, for softmax parameterized policies, using the advantage function and the occupancy measure, as follows:

Lemma 4.2 *For a softmax policy π we have $\nabla_\theta V^\pi(s_0)_{s,a} = d_s^\pi \pi(a|s;\theta) A^\pi(s,a)$.*

The above lemma shows that we only need to approximate the advantage function $A^\pi(\cdot,\cdot)$ in order to approximate the policy gradient $\nabla_\theta V^\pi(s_0)$. We do not need to approximate directly d_s^π but rather we can sample states following the policy π and they will be distributed according to d^π.

It would seem we can easily use an episode sampled from the MDP as an unbiased estimator for the policy gradient. In the works of Agarwal et al. (2020a), they present a gradient ascent algorithm in the form $\theta^{(t+1)} = \theta^{(t)} + \eta \nabla_\theta V^{\pi^t}(s_0)$, and conjecture it takes an exponential number of policy gradient steps to converge, even with exact gradients. Therefore we use a variant of the *Natural Policy Gradient Ascent* algorithm [Kakade 2001] presented by Agarwal et al. (2020a), which takes a small number of phases to converge, assuming access to exact gradients.

An Approximated Natural Policy Gradient (NPG) algorithm The Approximated Natural Policy Gradient algorithm is a variant of the *Natural Policy Gradient Ascent* algorithm presented by Agarwal et al. (2020a). The approximated NPG algorithm step is defined by:

$$\begin{aligned} F(\theta) &= \mathbb{E}_{s\sim d^\pi} \mathbb{E}_{a\sim\pi(\cdot|s)} \left[\nabla_\theta \log(\pi(a|s;\theta)) \left(\nabla_\theta \log(\pi(a|s;\theta))\right)^\top \right] \\ \theta^{(t+1)} &= \theta^{(t)} + \eta F(\theta^{(t)})^\dagger \widetilde{\nabla_\theta V^{\pi^t}(s_0)} \end{aligned} \tag{7}$$

Where $M^\dagger$ is the Moore-Penrose pseudo inverse matrix. The main insight is that an equivalent form for the algorithm update is:

$$\theta^{(t+1)} = \theta^{(t)} + H\eta \widetilde{A^{\pi^t}} \tag{8}$$

The following lemma shows that the two updates are equivalent.

Lemma 4.3 *The algorithm step (7) and (8) are equivalent.*

The proof is similar to the proof given by Agarwal et al. (2020a). We first define the loss function:

$$L^\pi(w) = \left\| \mathbb{E}_{s\sim d^\pi} \mathbb{E}_{a\sim\pi(\cdot|s)} \nabla_\theta \log(\pi(a|s) \left(w^\top \nabla_\theta \log(\pi(a|s)) - \widetilde{A^\pi}(s,a) \right) \right\|_2$$

and show that for the vector $w' = \widetilde{A^{\pi^t}}(\cdot,\cdot)$ the loss function is $L^{\pi^t}(w') = 0$. We show by a property of the Moore-Penrose pseudo inverse matrix that $w = \frac{1}{H} F(\theta^{(t)})^\dagger \widetilde{\nabla_\theta V^{\pi^t}(s_0)}$ is also a global minimizer. We then prove that the loss of both vectors is 0. This implies that the vectors w and w' have the same values up to a value that's independent of the action. Finally, by the definition of the softmax parameterization, adding such value does not affect the policy, therefore the update given in (7) is equivalent to the update given in (8).

The Policy Gradient algorithm: a variant of NPG As before, we have N phases and in each phase m episodes. The starting policy at phase 0 will be the uniform policy, where $\pi(a|s) = \frac{1}{A}$ for every state-action pair s, a. At the end of each phase t, using the m episodes, we compute $\widetilde{A^{\pi^t}}$, which is an approximation to the advantage function A^{π^t}. (We later show how to compute $\widetilde{A^{\pi^t}}$ given the m episodes.) Given the approximation $\widetilde{A^{\pi^t}}$ we update the parameters as defined in (8):

$$\theta^{(t+1)} = \theta^{(t)} + H\eta\widetilde{A^{\pi^t}}$$

Note that since any vector θ represents a valid policy, there is no need for a projection after the update, unlike the case of direct parameterization.

We now show how to compute the approximation $\widetilde{A^{\pi^t}}$ to the advantage function using episodes sampled using the MDP. We will do it in two cases, depending on the properties of the MDP.

4.1 MDP with random start state

In this section we will assume that the MDP is λ random start state. This implies that every episode of the MDP, with probability $1-\lambda$ starts at state s_0, and with probability λ starts at a random $s \in S$, performs a random action $a \in A$ and continues by following the given policy π. (We can encode the random action selection also in the policy, but it is simpler to consider it as part of the random start state.)

Since the episode can start at an arbitrary level, we assume that from state s_H any action has zero reward, which will allow us to complete an episode of length H and have the return depend only on the rewards in levels before level H.

Approximating the Advantage function: Recall that we are considering λ random start state MDP, where with probability λ we start at a random state and perform a random action. In each phase t we sample m episodes, and about $\frac{\lambda}{|S||A|}m$ of the m episodes are starting at each state-action pair s, a. The sum of the rewards observed in such an episode, which starts with (s, a), is an unbiased estimator for $Q^\pi(s, a)$. The average of those unbiased estimators sampled in a single phase is used as the approximated Q-function and denoted by $\widetilde{Q^\pi}(s, a)$.

Given the unbiased estimator $\widetilde{Q^\pi}(s, a)$ we define the functions $\widetilde{V^\pi}$, $\widetilde{A^\pi}$ to be, for a state-action pair (s, a), as follows:

$$\widetilde{V^\pi}(s) = \sum_{a' \in A} \pi(a'|s)\widetilde{Q^\pi}(s, a') \qquad \text{and} \qquad \widetilde{A^\pi}(s, a) = \widetilde{Q^\pi}(s, a) - \widetilde{V^\pi}(s) \tag{9}$$

Note that $\widetilde{V^\pi}(s)$ and $\widetilde{A^\pi}(s, a)$ are unbiased estimators of $V^\pi(s)$ and $A^\pi(s, a)$, respectively.

The following lemma gives a high probability bound on the error of our estimate:

Lemma 4.4 *Fix $m \geq \frac{4H^2|S||A|}{\epsilon^2\lambda} \log(\frac{2|A||S|K}{\delta})$. With probability at least $1-\delta$, for every phase t we have $\|\widetilde{A^{\pi^t}} - A^{\pi^t}\|_\infty \leq 2\epsilon$.*

The above lemma bounds m with a dependency proportional to $1/\epsilon^2$ and $1/\lambda$. The $1/\epsilon^2$ comes from the requirement to approximate the advantage function to accuracy 2ϵ. The $1/\lambda$ dependency comes from the requirement to sample each state-action pair enough times, and the probability is at least λ (due to the λ random start state property).

Regret Analysis: Define the notation $V^\pi(\mu)$, where μ is the distribution over the starting states and actions. If the MDP starts at state s_0 (which happens with probability $1-\lambda$), then $V^\pi(\mu) = V^\pi(s_0)$ and if the MDP starts at the state-action pair s, a, then $V^\pi(\mu) = Q^\pi(s, a)$. Since we are using each policy π^t in phase t for m episodes, the regret is $Regret = m(\sum_{t=1}^N V^*(\mu) - V^{\pi^t}(\mu))$. The following theorem bounds the regret in the softmax parameterization for a λ random start state MDP.

Theorem 4.5 *Let K be the total number of episodes, the number of phases be $N = K^{1/3}(\lambda/4|S||A|\log(\frac{2|A||S|K}{\delta}))^{1/3}$ and the number of episodes in each phase be $m = K^{2/3}(4|S||A|\log(\frac{2|A||S|K}{\delta})/\lambda)^{1/3}$. For $\eta > \log(|A|)$, with probability at least $1-\delta$, the regret of the softmax policy gradient for a λ start state MDP is bounded by:*

$$Regret \leq 6H^2 \left(\frac{4|S||A|\log(\frac{2|A||S|K}{\delta})}{\lambda} \right)^{\frac{1}{3}} K^{\frac{2}{3}}$$

Our proof strategy for the regret borrows ideas from Agarwal et al. (2020b). For a policy π and a state $s \in S$ we define the function $\widetilde{Z_\pi}(s) = \sum_{a \in A} \pi(a|s) \exp(\eta H \widetilde{A^t}(s,a))$ and show that $\pi^{t+1}(a|s) = \pi^t(a|s)\frac{\exp(\eta H \widetilde{A^t}(s,a))}{\widetilde{Z_t}(s)}$. We then show that $\sum_{t=1}^N \mathbb{E}_{s\sim\pi^*} \log(\widetilde{Z_{\pi^t}}(s)) \leq \eta H^3(H + 2N\epsilon)$ where 2ϵ is the approximation error of the A-function. Namely, for every phase $t \in [N]$, and a state-action pair s, a we have, $\|\widetilde{A^{\pi^t}} - A^{\pi^t}\|_\infty \leq 2\epsilon$. We then bound the regret as follows:

$$Regret \leq m \left(2\epsilon N + \frac{\log(|A|)}{H\eta} + \frac{1}{H^2\eta} \sum_{t=1}^{N} \mathbb{E}_{s\sim\pi^*} \log(\widetilde{Z_{\pi^t}}(s)) \right). \tag{10}$$

Setting $\epsilon = \frac{H}{N}$ and the rest of the algorithm parameters as specified in the theorem proves Theorem 4.5.

4.2 Softmax parameterization with resets

In this section we remove the assumption that the MDP is λ start state. However, in order to allow for efficient exploration, we assume that the algorithm can restart the MDP at any given state s instead of s_0 for the start of the episode. Allowing reset require modifying the definition of the regret. The issue is that some high reward states might have a small probability to be reached by any policy running in the MDP. To avoid this issue, every episode that the algorithm restarts the MDP at some state $s \neq s_0$, the algorithm will have a 0 return for that episode. This incentives the algorithm to minimize the number of resets.

The algorithm: As before the algorithm will work in N phases, and in each phase we will have m episodes. During a phase the algorithm would split the m episodes to m_1 reset episodes, where it will select a random start state $s \neq s_0$ and action a, and $m - m_1$ episodes where it will start from the initial state s_0.

In each phase t we will have a current policy π^t. In the reset episodes, each state-action pair (s, a) will be sampled approximately $m_1/(|S||A|)$ times as the initial state and action. After performing action a in state s the current policy π^t would be run for the remaining episode. In the remaining $m - m_1$ episodes the current policy π^t would be run from the initial state s_0.

Approximating the Advantage function: As discussed in section 4.1, the sum of the rewards given in the m_1 episodes sampled from restarting the MDP at (s, a), is an unbiased estimator for $Q^\pi(s, a)$. The average of those unbiased estimators sampled in a single phase is used as the approximated Q-function and denoted by $\widetilde{Q^\pi}(s, a)$. As before, given $\widetilde{Q^\pi}(s, a)$ we define

$$\widetilde{V^\pi}(s) = \sum_{a' \in A} \pi(a'|s)\widetilde{Q^\pi}(s, a') \qquad \text{and} \qquad \widetilde{A^\pi}(s, a) = \widetilde{Q^\pi}(s, a) - \widetilde{V^\pi}(s)$$

Note that $\widetilde{V^\pi}(s)$ and $\widetilde{A^\pi}(s, a)$ are unbiased estimators of $V^\pi(s)$ and $A^\pi(s, a)$, respectively.

At the end of the phase, the parameters θ^t would be updated to θ^{t+1}, as in the approximated NPG algorithm presented in (8). Namely,

$$\theta^{(t+1)} = \theta^{(t)} + H\eta\widetilde{A^{\pi^t}},$$

The following lemma abound the error in the approximation $\widetilde{A^{\pi^t}}$ of A^{π^t} as a function of the sample size m_1.

Lemma 4.6 *Fix $m_1 \geq \frac{2H^2|S||A|}{\epsilon^2}\log(\frac{2|A||S|K}{\delta})$. With probability at least $1-\delta$, for every phase t we have $\|\widetilde{A^{\pi^t}} - A^{\pi^t}\|_\infty \leq 2\epsilon$.*

Regret Analysis: The regret analysis is similar to that of Section 4.1 and the regret bound is summarized in the following theorem.

Theorem 4.7 *Let K be the total number of episodes, the number of phases be $N = K^{1/4}((4/3)H|S|^2|A|^2)^{-1/4}(\log(2|A||S|K/\delta))^{-1/8}$, and the number of episodes in a phase be $m = K^{3/4}((4/3)H|S|^2|A|^2)^{1/4}(\log(2|A||S|K/\delta))^{1/8}$ and $m_1 = (3K/H)^{1/2}\log(\frac{2|A||S|K}{\delta})^{1/4}$. For $\eta > \log(|A|)$, with probability $1-\delta$, the regret of the softmax policy gradient for an MDP with resets is bounded by:*

$$Regret \leq 13H^{\frac{9}{4}}\sqrt{|S||A|}K^{\frac{3}{4}}\left(\log(\frac{2|A||S|K}{\delta})\right)^{\frac{1}{4}}$$

Every phase t, the m_1 episodes which use resets for the regret analysis are assumed a return of 0. The other $m - m_1$ episodes receive a similar return as in section 4.1 (since the algorithm and the advantage function approximation is the same). Therefore the regret bound as shown in (10) can be applied here, with a slight change.

$$Regret \leq m_1 NH + (m - m_1)\left(2\epsilon N + \frac{\log(|A|)}{H\eta} + \frac{1}{H^2\eta}\sum_{t=1}^{N}\mathbb{E}_{s\sim\pi^*}\log(\widetilde{Z_{\pi^t}}(s))\right).$$

We use the bound shown in section 4.1:

$$\sum_{t=1}^{N}\mathbb{E}_{s\sim\pi^*}\log(\widetilde{Z_{\pi^t}}(s)) \leq \eta H^3(H + 2N\epsilon)$$

where 2ϵ is the approximation error of the A-function. Namely, for every phase $t \in [N]$, and a state-action pair s, a we have, $\|\widetilde{A^{\pi^t}} - A^{\pi^t}\|_\infty \leq 2\epsilon$. Optimizing the algorithm parameters as specified in the theorem proves Theorem 4.7.

4.3 Comparing the two softmax settings

The two settings presented above are two different methods to ensure sufficient exploration for the agent.

In each algorithm update the parameters are updated by $\theta^{(t+1)} = \theta^{(t)} + H\eta\widetilde{A^{\pi^t}}$, as shown in (8). The expected number of visits to every state-action pair depends on the MDP and can be very low. Therefore, in order to achieve a sufficient approximation for the advantage function, we need to add some assumptions either on the MDP itself, or on the agent. Both settings includes starting the MDP at some state s in the middle of the episode, i.e., if $s \in S_i$, the episode will be of length $H - i$.

In the first setting (Section 4.1) we make an assumption about the starting state distribution of the MDP. Specifically, we assume that start state distribution gives some minimal probability for each state, namely $\lambda/|S|$. This assumption guarantees sufficient exploration, since any state can be an initial state, and no other modification to the algorithm is needed.

In the second setting (Section 4.1) introduce an assumption about the agent ability to resent the initial state to any specific state. Since the agent controls the start state, it can improve the expected payoff, compared to starting at s_0. In order to negate this agent's advantage, in the regret analysis, any episode which the agent restarts not at s_0 will have a payoff of 0 (for the regret analysis). Note that the modified regret upper bounds the true regret.

5 Conclusion and limitations

In this paper we analyzed two known tabular policy gradient algorithms in terms of regret. As presented in Agarwal et al. (2020b) the algorithms had known convergence bounds given exact gradient, and we did not aim to improve those bounds. Our goal was to purpose methods to approximate the policy gradient using sampled episodes, and give a regret bound.

The first algorithm using the direct policy parameterization does converge according to Agarwal et al. (2020b), yet the convergence rate is worse than most algorithms they presented. As expected, the regret bound presented presented in our paper for the policy parameterization algorithm was far from optimal. Yet the direct policy parameterization method allowed us to have sufficient exploration to approximate the policy gradient. Therefore we were able to bound the regret of this algorithm with no assumptions on the MDP.

The second algorithm using the softmax policy parameterization has a constant convergence rate according to Agarwal et al. (2020b) which does not depend on the size of the state space at all. As we need to approximate the policy gradient in every state-action pair parameter such result is not possible for the regret bound of course, but the algorithm seems promising for a small regret bound. As it turns out, to perform the algorithm step one needs to explore every state a sufficient number of times, which might not be possible in every MDP. To overcome that and analyze the algorithm that achieved such good convergence bounds by Agarwal et al. (2020b) we presented two different settings, each with different assumptions on the MDP, to enable us to perform the sufficient exploration. We still did not manage to reach the state of the art regret bounds ($\sqrt{K}$).

References

Y. Abbasi-Yadkori, P. Bartlett, K. Bhatia, N. Lazic, C. Szepesvari, and G. Weisz. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pages 3692–3702. PMLR, 2019.

A. Agarwal, M. Henaff, S. Kakade, and W. Sun. Pc-pg: Policy cover directed exploration for provable policy gradient learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13399–13412. Curran Associates, Inc., 2020a.

A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In J. D. Abernethy and S. Agarwal, editors, *Conference on Learning Theory, COLT*, volume 125 of *Proceedings of Machine Learning Research*, pages 64–66. PMLR, 2020b. (See also archive version abs/1908.00261).

M. G. Azar, R. Munos, and B. Kappen. On the sample complexity of reinforcement learning with a generative model. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress, 2012.

M. G. Azar, I. Osband, and R. Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org, 2017.

P. L. Bartlett and A. Tewari. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 35–42. AUAI Press, 2009.

R. I. Brafman and M. Tennenholtz. R-MAX - A general polynomial time algorithm for near-optimal reinforcement learning. *J. Mach. Learn. Res.*, 3:213–231, 2002.

Q. Cai, Z. Yang, C. Jin, and Z. Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

S. Cen, C. Cheng, Y. Chen, Y. Wei, and Y. Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *ArXiv*, abs/2007.06558, 2020.

C. Dann and E. Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NerIPS)*, pages 2818–2826, 2015.

C. Dann, T. Lattimore, and E. Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5713–5723, 2017.

C. Dann, L. Li, W. Wei, and E. Brunskill. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pages 1507–1516. PMLR, 2019.

M. P. Deisenroth, G. Neumann, and J. Peters. A survey on policy search for robotics. *Found. Trends Robotics*, 2(1-2):1–142, 2013.

Y. Efroni, N. Merlis, M. Ghavamzadeh, and S. Mannor. Tight regret bounds for model-based reinforcement learning with greedy policies. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, pages 12203–12213, 2019.

R. Fruit, M. Pirotta, A. Lazaric, and R. Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. *arXiv preprint arXiv:1802.04020*, 2018.

T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

S. Kakade and J. Langford. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

S. M. Kakade. A natural policy gradient. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*, pages 1531–1538. MIT Press, 2001.

S. M. Kakade. On the sample complexity of reinforcement learning, 2003. PhD thesis, *University of London*.

M. J. Kearns and S. P. Singh. Near-optimal reinforcement learning in polynomial time. *Mach. Learn.*, 49(2-3):209–232, 2002.

T. Lattimore and M. Hutter. Pac bounds for discounted mdps. In *International Conference on Algorithmic Learning Theory*, pages 320–334. Springer, 2012.

J. Mei, C. Xiao, C. Szepesvari, and D. Schuurmans. On the global convergence rates of softmax policy gradient methods. In *ICML*, 2020.

I. Osband, B. Van Roy, and Z. Wen. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pages 2377–2386, 2016.

L. Shani, Y. Efroni, and S. Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. *ArXiv*, abs/1909.02769, 2020.

Z. Shen, A. Ribeiro, H. Hassani, H. Qian, and C. Mi. Hessian aided policy gradient. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 5729–5738, 2019.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. P. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of go with deep neural networks and tree search. *Nat.*, 529(7587):484–489, 2016.

R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. A. Solla, T. K. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 1057–1063. The MIT Press, 1999.

R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8:229–256, 1992.

L. Yang, Q. Zheng, and G. Pan. Sample complexity of policy gradient finding second-order stationary points. *CoRR*, abs/2012.01491, 2020.

A. Zanette and E. Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312, 2019.

Z. Zhang, X. Ji, and S. S. Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. *arXiv preprint arXiv:2009.13503*, 2020.

A Proofs for section 3

Algorithm 1: Policy Gradient with direct parameterization

Input: MDP,K, δ;
$\theta = \frac{1}{|A|}\vec{1}$; `/* The dimension of the vector is |S||A| */`
$N = K^{1/3}(H|A|^5|S|\log(\frac{2|A||S|K}{\delta}))^{-1/3}$;
$m = K^{2/3}(H|A|^5|S|\log(\frac{2|A||S|K}{\delta}))^{1/3}$;
$\eta = (\frac{2}{3}|A|H^3+1)^{-1}$;
$\rho = \frac{|S|^{2/3}H^{1/2}\log(\frac{2|A||S|K}{\delta})^{1/6}}{|A|^{2/3}K^{1/6}}$;
for $n = 1, 2, \ldots, N$ **do**
  $\widetilde{\nabla_\theta V^{\pi_\theta}}(s_0) = \vec{0}$;
  for $i = 1, 2, \ldots, m$ **do**
    Run policy π_θ on the MDP and get $(s_0^i, a_0^i, r_0^i, s_1^i, \ldots, s_{H-1}^i, a_{H-1}^i, r_0^i, s_H^i)$;
    for $s, a \in S, A$ **do**
      l =The level of s;
      $\widetilde{\nabla_\theta V^{\pi_\theta}}(s_0)_{s,a} += \frac{1}{m\theta_{s,a}}\sum_{j=l}^{H} r_j^i I(s_l^i = s, a_l^i = a)$;
    end
  end
  $\theta = prox_\rho(\theta + \eta\widetilde{\nabla_\theta V^{\pi_\theta}}(s_0))$; `/* The function` $prox_\rho$ `is the projection function, which is described in Section C.1 */`
end

Using the definition of the occupancy measure d^π and the policy gradient theorem (1), we get:

$$\nabla_\theta V^\pi(s_0) = H\mathbb{E}_{s'\sim d^\pi}\sum_{a\in A}\nabla_\theta\pi(a|s';\theta)Q^\pi(s',a) \tag{11}$$

Observe that d^π is not a distribution as $\sum_{s\in S} d_s^\pi = H$. For simplicity we use the notation $s \sim d^\pi$ instead of $s \sim \frac{1}{H}d^\pi$. Given the parameter θ, the policy π is: $\pi(a|s;\theta) = \theta_{s,a}$ in the direct parameterization. Observe the gradient of the policy w.r.t. θ

$$\frac{\partial\pi_\theta(a|s)}{\partial\theta_{s,a}} = 1 \qquad \text{and} \qquad \forall(s',a') \neq (s,a),\ \frac{\partial\pi(a|s;\theta)}{\partial\theta_{s',a'}} = 0 \tag{12}$$

Therefore $\nabla_\theta\pi(a|s;\theta)$ is the unit vector that equals 1 at the (s,a) coordinate, and 0 at all other coordinates.

Proof:[Of Lemma 3.2] Fix a policy π and a state-action pair s, a.

$$\begin{aligned}\nabla_\theta V^\pi(s_0)_{s,a} &= H\mathbb{E}_{s'\sim d^\pi}\sum_{a'\in A}\nabla_\theta\pi(a'|s';\theta)_{s,a}Q^\pi(s',a') \\ &= H\mathbb{E}_{s'\sim d^\pi}\sum_{a'\in A}Q^\pi(s',a')I(s=s',a=a') \\ &= d_s^\pi Q(s,a)\end{aligned}$$

Where the first step follows by (11) and the second step follows by (12). □

Lemma A.1 *For a policy π:*

$$\mathbb{E}_{\tau\sim\pi}[\vec{X^\pi}(\tau)] = \nabla_\theta V^\pi(s_0)$$

Where $\vec{X^\pi}(\tau)$ is the the vector defined by a trajectory as in (2)

Proof: Fix a policy π, $a \in A$, $k \in [H]$, $s \in S_k$

$$\begin{aligned}\mathbb{E}_{\tau\sim\pi}[\vec{X^{\pi}}(\tau)]_{s,a} &= d_s^{\pi}\pi(a|s)\frac{1}{\pi(a|s)}\mathbb{E}_{\tau\sim\pi}[\sum_{j=k}^{H} r(s_j,a_j)I(s_k=s,a_k=a)] \\ &= d_s^{\pi}Q(s,a) \\ &= \nabla_\theta V^{\pi}(s_0)_{s,a},\end{aligned}$$

where the second step follows directly from the definition of the Q-function and the third step follows by Lemma 3.2. □

Lemma A.2 *For $\epsilon > 0$ and $\delta_1 > 0$, a state-action pair s, a, and a policy $\pi \in \Pi_\rho$, when sampling $m_1 \geq \frac{H^2}{\epsilon^2\rho^2}\log(\frac{2}{\delta_1})$ trajectories $\tau_1, \tau_2, \ldots, \tau_{m_1}$ using π, and averaging them to get an approximation of the gradient $\widetilde{\nabla_\theta V^{\pi}}(s_0)$ as given in (3) we get:*

$$P\left(\left|\left(\widetilde{\nabla_\theta V^{\pi^t}}(s_0) - \nabla_\theta V^{\pi}(s_0)\right)_{s,a}\right| \geq \epsilon\right) \leq \delta_1 \tag{13}$$

Proof: Fix $\epsilon > 0$, $\delta_1 > 0$, a state-action pair (s,a), a policy $\pi \in \Pi_\rho$ and trajectories $\tau_1, \tau_2, \ldots, \tau_{m_1}$ sampled using π. As $\frac{1}{\pi(a|s)} \leq \frac{1}{\rho}$, the value $\vec{X^{\pi}}(\tau_i)$ which is defined in (2) is non-negative and bounded by $\frac{H}{\rho}$. As the gradient $\nabla_\theta V^{\pi}(s_0)$ is non-negative the following bound holds: $|(\vec{X^{\pi}}(\tau) - \nabla_\theta V^{\pi}(s_0))_{s,a}| \leq \frac{H}{\rho}$. Using Hoeffding concentration bound (Theorem C.15) completes the proof. □

Define $\vec{\epsilon_t} := \widetilde{\nabla_\theta V^{\pi^t}}(s_0) - \nabla_\theta V^{\pi^t}(s_0)$. The policy gradient algorithm can be written as:

$$\pi_\theta^{t+1} = prox_\rho(\pi_\theta^t + \eta\widetilde{\nabla_\theta V^{\pi^t}}(s_0)) = prox_\rho(\pi_\theta^t + \eta(\nabla_\theta V^{\pi^t}(s_0) + \vec{\epsilon_t}))$$

and with high probability we have $\|\vec{\epsilon_t}\|_\infty \leq \epsilon$

Proof: [of Lemma 3.3] Let $\delta_1 = \delta\frac{1}{|A||S|K}$, and $m = \frac{H^2|A||S|}{\epsilon^2\rho^2}\log(\frac{2|A||S|K}{\delta})$. Using Lemma A.2 for a state-action pair (s,a) and a phase $t \in [N]$ after sampling $\frac{H^2}{\epsilon^2\rho^2}\log(\frac{2}{\delta_1})$ episodes we have

$$\left|\left(\widetilde{\nabla_\theta V^{\pi^t}}(s_0) - \nabla_\theta V^{\pi^t}(s_0)\right)_{s,a}\right| \leq \epsilon$$

with probability $1 - \delta_1$.

Using the union bound, sampling m trajectories at every phase $t \in [N]$, the bound $\|\widetilde{\nabla_\theta V^{\pi^t}}(s_0) - \nabla_\theta V^{\pi^t}(s_0)\|_\infty \leq \epsilon$ will hold for all $t \in [N]$ with probability $1 - |A||S|N\delta_1 = 1 - \frac{1}{m}\delta \geq 1 - \delta$. □

In the definition of the occupancy measure d^{π} we assume the execution starts at s_0. We expand the definition of the occupancy measure to $d^{\pi,\mu}$, where μ is a starting state distribution. Since the episode can start at an arbitrary level, we assume that from state s_H any action has zero reward, which will allow us to complete an episode of length H and have the return depend only on the rewards in levels before level H. We first prove the following Lemma (similar to [Kakade and Langford (2002), Agarwal et al. (2020a)]).

Lemma A.3 *For any two policies π, π' and a starting state distribution μ,*

$$V^{\pi}(\mu) - V^{\pi'}(\mu) = \frac{1}{H}\sum_{s\in S, a\in A} d_s^{\pi,\mu}\pi(a|s)A^{\pi'}(s,a)$$

Proof: Fix two policies π, π'. Let s' be some starting state at level k (i.e., $s' \in S_k$), such that $\mu_{s'} > 0$. Define $\tau \sim (\pi, s)$ to be an episode sampled using the policy π assuming the starting state is

s.

$$
\begin{aligned}
V^{\pi}(s') - V^{\pi'}(s') &= \mathbb{E}_{\tau\sim(\pi,s')}\Big[\sum_{i=k}^{H-1} r(s_i,a_i)\Big] - V^{\pi'}(s') \\
&= \mathbb{E}_{\tau\sim(\pi,s')}\Big[\left(\sum_{i=k}^{H-1} r(s_i,a_i)\right) + \left(\sum_{i=k}^{H-1} V^{\pi'}(s_i)\right) - \left(\sum_{i=k}^{H-1} V^{\pi'}(s_i)\right)\Big] - V^{\pi'}(s') \\
&= \mathbb{E}_{\tau\sim(\pi,s')}\Big[\left(\sum_{i=k}^{H-1} r(s_i,a_i)\right) + \left(\sum_{i=k+1}^{H-1} V^{\pi'}(s_i)\right) - \left(\sum_{i=k}^{H-1} V^{\pi'}(s_i)\right)\Big] \\
&= \mathbb{E}_{\tau\sim(\pi,s')}\Big[\left(\sum_{i=k}^{H-1} r(s_i,a_i)\right) + \left(\sum_{i=k}^{H-2} V^{\pi'}(s_{i+1})\right) - \left(\sum_{i=k}^{H-1} V^{\pi'}(s_i)\right)\Big] \\
&= \mathbb{E}_{\tau\sim(\pi,s')}\Big[\left(\sum_{i=k}^{H-1} r(s_i,a_i)\right) + \left(\sum_{i=k}^{H-1} V^{\pi'}(s_{i+1})\right) - \left(\sum_{i=k}^{H-1} V^{\pi'}(s_i)\right)\Big] \\
&= \mathbb{E}_{\tau\sim(\pi,s')}\left[\sum_{i=k}^{H-1}\Big(r(s_i,a_i) + V^{\pi'}(s_{i+1}) - V^{\pi'}(s_i)\Big)\right] \\
&= \mathbb{E}_{\tau\sim(\pi,s')}\Big[\sum_{i=k}^{H-1}\Big(Q^{\pi'}(s_i,a_i) - V^{\pi'}(s_i)\Big)\Big] \\
&= \mathbb{E}_{\tau\sim(\pi,s')}\Big[\sum_{i=k}^{H-1} A^{\pi'}(s_i,a_i)\Big] \\
&= \frac{1}{H}\sum_{s\in S,a\in A} d_s^{\pi,s'}\pi(a|s)A^{\pi'}(s,a)
\end{aligned}
$$

The fifth equality follows by the fact that $V^{\pi'}(s_H) = 0$ as s_H is the end of the MDP, the seventh equality follows by the fact that s_{i+1} is sampled with the distribution $P(\cdot|s_i,a_i)$, and since for all state-action pair s,a it follows that: $Q^{\pi'}(s,a) = r(s,a) + \mathbb{E}_{s'\sim P(\cdot|s_i,a_i)}[V^{\pi'}(s')]$. The eighth equality follows directly from the definition of the $A^{\pi}(s,a)$ function.

As the above is correct for all starting states s', observe:

$$
\begin{aligned}
V^{\pi}(\mu) - V^{\pi'}(\mu) &= \sum_{s'\in S}\mu_{s'}\Big(V^{\pi}(s') - V^{\pi'}(s')\Big) \\
&= \sum_{s'\in S}\mu_{s'}\left(\frac{1}{H}\sum_{s\in S,a\in A} d_s^{\pi,s'}\pi(a|s)A^{\pi'}(s,a)\right) \\
&= \frac{1}{H}\sum_{s\in S,a\in A}\pi(a|s)A^{\pi'}(s,a)\sum_{s'\in S}\mu_{s'}d_s^{\pi,s'} \\
&= \frac{1}{H}\sum_{s\in S,a\in A} d_s^{\pi,\mu}\pi(a|s)A^{\pi'}(s,a)
\end{aligned}
$$

Where the last transition holds since) $\sum_{s'\in S}\mu_{s'}d_s^{\pi,s'}$ is exactly the definition of $d_s^{\pi,\mu}$. □

Lemma A.4 *For a policy π and a state $s\in S$,*

$$\sum_{a\in A}\pi(a|s)A^{\pi}(s,a) = 0$$

Proof: Fix a policy π and a state $s \in S$.

$$\begin{aligned}
\sum_{a\in A} \pi(a|s)A^{\pi}(s,a) &= \sum_{a\in A} \pi(a|s)(Q^{\pi}(s,a) - V^{\pi}(s)) \\
&= \left(\sum_{a\in A} \pi(a|s)Q^{\pi}(s,a)\right) - \left(\sum_{a\in A} \pi(a|s)V^{\pi}(s)\right) \\
&= \mathbb{E}_{a\sim\pi(\cdot|s)}\Big[Q^{\pi}(s,a)\Big] - V^{\pi}(s)\sum_{a\in A}\pi(a|s) \\
&= V^{\pi}(s) - V^{\pi}(s) \\
&= 0
\end{aligned}$$

Where the third transition holds since $V^{\pi}(s)$ does not depend on a, and the fourth transition holds since $\sum_{a\in A}\pi(a|s) = 1$. □

Proof:[Of Lemma 3.5] Fix a policy $\pi \in \Pi_{\rho}$.

$$\begin{aligned}
V^{*}(s_0) - V^{\pi}(s_0) &= \frac{1}{H}\sum_{s\in S, a\in A} d_s^{\pi^*}\pi^{*}(a|s)A^{\pi}(s,a) \\
&\leq \frac{D_{\rho}}{H}\sum_{s\in S, a\in A} d_s^{\pi}\pi^{*}(a|s)A^{\pi}(s,a) \\
&\leq \frac{D_{\rho}}{H}\max_{\pi'\in\Delta(A)^{|S|}}\sum_{s\in S, a\in A} d_s^{\pi}\pi'(a|s)A^{\pi}(s,a) \\
&= \frac{D_{\rho}}{H}\max_{\pi'\in\Delta(A)^{|S|}}\sum_{s\in S} d_s^{\pi}\sum_{a\in A}\pi'(a|s)A^{\pi}(s,a) \\
&= \frac{D_{\rho}}{H}\max_{\pi'\in\Delta(A)^{|S|}}\sum_{s\in S} d_s^{\pi}\sum_{a\in A}(\pi'(a|s) - \pi(a|s))A^{\pi}(s,a) \\
&= \frac{D_{\rho}}{H}\max_{\pi'\in\Delta(A)^{|S|}}\sum_{s\in S} d_s^{\pi}\sum_{a\in A}(\pi'(a|s) - \pi(a|s))Q^{\pi}(s,a) \\
&\quad - \frac{D_{\rho}}{H}\max_{\pi'\in\Delta(A)^{|S|}}\sum_{s\in S} d_s^{\pi}\sum_{a\in A}(\pi'(a|s) - \pi(a|s))V^{\pi}(s) \\
&= \frac{D_{\rho}}{H}\max_{\pi'\in\Delta(A)^{|S|}}\sum_{s\in S} d_s^{\pi}\sum_{a\in A}(\pi'(a|s) - \pi(a|s))Q^{\pi}(s,a) \\
&= \frac{D_{\rho}}{H}\max_{\pi'\in\Delta(A)^{|S|}}\sum_{s\in S, a\in A} d_s^{\pi}(\pi'(a|s) - \pi(a|s))Q^{\pi}(s,a) \\
&= \frac{D_{\rho}}{H}\max_{\pi'\in\Delta(A)^{|S|}}(\pi' - \pi)^{\top}\nabla_{\theta}V^{\pi}(s_0),
\end{aligned}$$

where the first step follows by Lemma A.3, the fifth step follows by Lemma A.4, since the lemma shows that for a state $s \in S$, $\sum_{a\in A}\pi(a|s)A^{\pi}(s,a) = 0$, the seventh step follows by $\sum_{a\in A}(\pi'(a|s) - \pi(a|s)) = 0$ and since $V^{\pi}(s)$ does not depend on a, and the last step follows by Lemma 3.2, since the lemma shows that for a state-action pair s, a, $\nabla_{\theta}V^{\pi}(s_0)_{s,a} = d_s^{\pi}Q^{\pi}(s,a)$. □

The proof of Lemma 3.6 is in section C.4.

Define $\vec{\epsilon_t} = \widetilde{\nabla_{\theta}V^{\pi^t}(s_0)} - \nabla_{\theta}V^{\pi^t}(s_0)$, and note that $\|\epsilon_t\|_{\infty} \leq \epsilon$ with high probability.

Lemma A.5 *The set* $\Pi_{\rho} = \{\theta \in [0,1]^{S\times A} : \theta_{s,a} \geq \rho, \sum_{a\in A}(\theta_{s,a}) = 1\}$ *is convex.*

Proof: Let there be $x, y \in \Pi_{\rho}$ and $\lambda \in [0,1]$. Define $z = \lambda x + (1-\lambda)y$. For $s \in S, a \in A$ we have $z_{s,a} = \lambda x_{s,a} + (1-\lambda)y_{s,a}$. As $x, y \in \Pi_{\rho}$ we can see that $z_{s,a} \geq \lambda\rho + (1-\lambda)\rho = \rho$, $z_{s,a} \leq \lambda + (1-\lambda) = 1$

and $\sum_{a' \in A} z_{s,a'} = \sum_{a' \in A} \lambda x_{s,a'} + (1-\lambda) y_{s,a'} = \lambda \sum_{a' \in A} x_{s,a'} + (1-\lambda) \sum_{a' \in A} y_{s,a'} = \lambda + (1-\lambda) = 1$, therefore we can infer that $z \in \Pi_\rho$, which completes the proof that Π_ρ is convex. □

Proof:[Of Lemma 3.7] Fix a phase t and the gradient step size η. Let π^t be the policy at phase t and let the policy at phase $t+1$ be: $\pi^{t+1} = prox_\rho(\pi^t + \eta \nabla_\theta \widetilde{V^{\pi^t}}(s_0))$.

Since $\pi^{t+1} = prox_\rho(\pi^t + \eta \nabla_\theta \widetilde{V^{\pi^t}}(s_0))$ and Lemma A.5, Lemma C.5 gives us:

$$\forall z \in \Pi_\rho, (\pi^t + \eta(\nabla_\theta V^{\pi^t}(s_0) + \vec{\epsilon_t}) - \pi^{t+1})^\top (z - \pi^{t+1}) \leq 0$$

Define $u = \nabla_\theta V^{\pi^{t+1}}(s_0) - \nabla_\theta V^{\pi^t}(s_0) - \vec{\epsilon_t} - \frac{1}{\eta}(\pi^t - \pi^{t+1})$. Reorganizing that we get:

$$\forall z \in \Pi_\rho, (\eta \nabla_\theta V^{\pi^{t+1}}(s_0) - \eta u)^\top (z - \pi^{t+1}) \leq 0$$

Which is equivalent to:

$$\forall z \in \Pi_\rho, \nabla_\theta V^{\pi^{t+1}}(s_0)^\top (z - \pi^{t+1}) \leq u^\top (z - \pi^{t+1})$$

We upper bound $u^\top (z - \pi^{t+1})$. As $z, \pi^{t+1} \in \Delta(A)^{|S|}$ we have that $\|z - \pi^{t+1}\|_2 \leq 2\sqrt{|S|}$, which means that $u^\top (z - \pi^{t+1}) \leq 2\sqrt{|S|}\|u\|_2$. That implies that:

$$\forall z \in \Pi_\rho, \nabla_\theta V^{\pi^{t+1}}(s_0)^\top (z - \pi^{t+1}) \leq 2\sqrt{|S|}\|u\|_2$$

We upper bound $\|u\|_2$:

$$\begin{aligned}
\|u\|_2 &= \|\nabla_\theta V^{\pi^{t+1}}(s_0) - \nabla_\theta V^{\pi^t}(s_0) - \vec{\epsilon_t} - \frac{1}{\eta}(\pi^t - \pi^{t+1})\|_2 \\
&\leq \beta \|\pi^{t+1} - \pi^t\|_2 + \sqrt{|A||S|}\epsilon + \frac{1}{\eta}\|\pi^{t+1} - \pi^t\|_2 \\
&= (\eta\beta + 1)\|\widetilde{G^\eta(\pi^t)}\|_2 + \sqrt{|A||S|}\epsilon,
\end{aligned}$$

where the second step follows by the fact that $V^\pi(s_0)$ is β-Smooth with respect to norm $\|\cdot\|_2$ (as shown in Lemma 3.6), and the last step follows by the definition of the gradient mapping (5).

This gives us:

$$\forall z \in \Pi_\rho, \nabla_\theta V^{\pi^{t+1}}(s_0)^\top (z - \pi^{t+1}) \leq 2\sqrt{|S|}\left[(\eta\beta + 1)\|\widetilde{G^\eta(\pi^t)}\|_2 + \sqrt{|A||S|}\epsilon\right] \tag{14}$$

Note that for any $z' \in \Delta(A)^{|S|}$ there exists z^+ such that $(z' + z^+) \in \Pi_\rho$ and $\|z^+\|_\infty \leq |A|\rho$. Lemma 3.2 gives us that for a state-action pair s, a, $\nabla_\theta V^\pi(s_0)_{s,a} = d_s^\pi Q^\pi(s,a)$. We see that:

$$\begin{aligned}
\|\nabla_\theta V^\pi(s_0)\|_1 &= \sum_{s \in S, a \in A} |\nabla_\theta V^\pi(s_0)_{s,a}| \\
&= \sum_{s \in S} d_s^\pi \sum_{a \in A} Q^\pi(s,a) \\
&\leq \sum_{s \in S} d_s^\pi |A| H \\
&= |A| H^2.
\end{aligned}$$

Observe that for all $z' \in \Delta(A)^{|S|}$:

$$\begin{aligned}
\nabla_\theta V^{\pi^{t+1}}(s_0)^\top (z' - \pi^{t+1}) &= \nabla_\theta V^{\pi^{t+1}}(s_0)^\top ((-z^+) + (z' + z^+ - \pi^{t+1})) \\
&\leq \|\nabla_\theta V^{\pi^{t+1}}(s_0)\|_1 \|z^+\|_\infty + \nabla_\theta V^{\pi^{t+1}}(s_0)^\top (z' + z^+ - \pi^{t+1}) \\
&\leq H^2 |A|^2 \rho + \nabla_\theta V^{\pi^{t+1}}(s_0)^\top (z' + z^+ - \pi^{t+1}) \\
&\leq H^2 |A|^2 \rho + 2\sqrt{|S|}\left[(\eta\beta + 1)\|\widetilde{G^\eta(\pi^t)}\|_2 + \sqrt{|A||S|}\epsilon\right],
\end{aligned}$$

where the third transition follows by $\|\nabla_\theta V^\pi(s_0)\|_1 \leq |A|H^2$ and $\|z^+\|_\infty \leq |A|\rho$, and the last transition follows by (14), as $(z' + z^+) \in \Pi_\rho$. □

Proof:[Of Lemma 3.8] Fix a phase t, the gradient step size η, the policy π^{t+1} which is chosen as: $\pi^{t+1} = prox_\rho(\pi^t + \eta\widetilde{\nabla_\theta V^{\pi^t}(s_0)})$. Since $V^\pi(s_0)$ is β-smooth with respect to norm $\|\cdot\|_2$ (as shown in Lemma 3.6) and Lemma A.5, Lemma C.6 gives:

$$V^{\pi^{t+1}}(s_0) \geq V^{\pi^t}(s_0) + \nabla_\theta V^{\pi^t}(s_0)^\top \cdot (\pi^{t+1} - \pi^t) - \frac{\beta}{2}\|\pi^{t+1} - \pi^t\|_2^2. \tag{15}$$

Since $\pi^{t+1} = prox_\rho(\pi^t + \eta\nabla_\theta V^{\pi^t}(s_0) + \eta\vec{\epsilon_t})$ and $\pi^t \in \Pi_\rho$ and Lemma A.5, Lemma C.5 gives:

$$(\pi^t + \eta\nabla_\theta V^{\pi^t}(s_0) + \eta\vec{\epsilon_t} - \pi^{t+1})^\top \cdot (\pi^t - \pi^{t+1}) \leq 0$$

After some rearranging the above is equivalent to:

$$\nabla_\theta V^{\pi^t}(s_0)^\top \cdot (\pi^{t+1} - \pi^t) \geq \vec{\epsilon_t}^\top \cdot (\pi^t - \pi^{t+1}) + \frac{1}{\eta}\|\pi^{t+1} - \pi^t\|_2^2 \tag{16}$$

Combining (16) and (15) gives us:

$$V^{\pi^{t+1}}(s_0) - V^{\pi^t}(s_0) \geq \vec{\epsilon_t}^\top \cdot (\pi^t - \pi^{t+1}) + \frac{1}{\eta}\|\pi^{t+1} - \pi^t\|_2^2 - \frac{\beta}{2}\|\pi^{t+1} - \pi^t\|_2^2$$

As the above is correct for all $t \in [N]$, we sum over the the episodes, and have:

$$V^{\pi^N}(s_0) - V^{\pi^0}(s_0) \geq \sum_{t=1}^N \vec{\epsilon_t}^\top \cdot (\pi^t - \pi^{t+1}) + \sum_{t=1}^N \frac{1}{\eta}(1 - \frac{\eta\beta}{2})\|\pi^{t+1} - \pi^t\|_2^2$$

Using the gradient mapping definition (5) and the fact that $V^{\pi^N}(s_0) - V^{\pi^0}(s_0) \leq H$, we have:

$$\sum_{t=1}^N \eta(1 - \frac{\beta\eta}{2})\|\widetilde{G^\eta(\pi^t)}\|_2^2 + \sum_{t=1}^N \vec{\epsilon_t}^\top \cdot (\pi^t - \pi^{t+1}) \leq H$$

As for any two vectors $(x+y)^\top(x+y) = x^\top x + 2x^\top y + y^\top y$, observe:

$$\begin{aligned}
\sum_{t=1}^N \vec{\epsilon_t}^\top \cdot (\pi^t - \pi^{t+1}) &= \frac{1}{2}\sum_{t=1}^N (\|\vec{\epsilon_t} + \pi^t - \pi^{t+1}\|_2^2 - \|\vec{\epsilon_t}\|_2^2 - \|\pi^t - \pi^{t+1}\|_2^2) \\
&\geq \frac{1}{2}\sum_{t=1}^N (-\|\vec{\epsilon_t}\|_2^2 - \|\pi^t - \pi^{t+1}\|_2^2) \\
&\geq -\frac{1}{2}\sum_{t=1}^N \eta^2\|\widetilde{G^\eta(\pi^t)}\|_2^2 - \frac{1}{2}|A||S|N\epsilon^2
\end{aligned}$$

Where the second transition holds since a norm is non-negative. This gives us:

$$\begin{aligned}
H &\geq \sum_{t=1}^N \eta(1 - \frac{\beta\eta}{2})\|\widetilde{G^\eta(\pi^t)}\|_2^2 + \sum_{t=1}^N \vec{\epsilon_t}^\top \cdot (\pi^t - \pi^{t+1}) \\
&\geq \sum_{t=1}^N \eta(1 - \frac{(\beta+1)\eta}{2})\|\widetilde{G^\eta(\pi^t)}\|_2^2 - \frac{1}{2}|A||S|N\epsilon^2.
\end{aligned}$$

Rearranging the above completes the proof. □

Proof:[of Theorem 3.4] We show a bound on the regret with probability $1-\delta$:

$$\begin{aligned}
Regret &= m\sum_{t=1}^{N} V^*(s_0) - V^{\pi^t}(s_0) \\
&\leq \frac{m}{H}\sum_{t=1}^{N} D_\rho \max_{\pi'\in\Delta(A)^{|S|}} (\pi' - \pi^t)^\top \nabla_\theta V^{\pi^t}(s_0) \\
&\leq \frac{m}{H} D_\rho \sum_{t=1}^{N-1} \left(H^2|A|^2\rho + 2\sqrt{|S|}\left[(\eta\beta+1)\|\widetilde{G^\eta(\pi^t)}\|_2 + \sqrt{|A||S|}\epsilon\right]\right) \\
&= \frac{m}{H} D_\rho \Bigg(H^2|A|^2\rho N + 2N\epsilon\sqrt{|A|}|S| + \\
&\quad + 2\sqrt{|S|}\sqrt{\frac{1}{1-\frac{1}{2}(\beta+1)\eta}}\frac{1+\eta\beta}{\sqrt{\eta}}\sum_{t=1}^{N}\sqrt{\eta\Big(1-\frac{1}{2}(\beta+1)\eta\Big)}\|\widetilde{G^\eta(\pi)}\|_2\Bigg) \\
&\leq \frac{m}{H} D_\rho \Bigg(H^2|A|^2\rho N + 2N\epsilon\sqrt{|A|}|S| \\
&\quad + 2\sqrt{|S|}\sqrt{\frac{1}{1-\frac{1}{2}(\beta+1)\eta}}\frac{1+\eta\beta}{\sqrt{\eta}}\sqrt{N}\sqrt{\sum_{t=1}^{N}\eta\Big(1-\frac{1}{2}(\beta+1)\eta\Big)\|\widetilde{G^\eta(\pi^t)}\|_2^2}\Bigg) \\
&\leq \frac{m}{H} D_\rho \Bigg(H^2|A|^2\rho N + 2N\epsilon\sqrt{|A|}|S| + \\
&\quad + 2\sqrt{|S|}\sqrt{\frac{1}{1-\frac{1}{2}(\beta+1)\eta}}\frac{1+\eta\beta}{\sqrt{\eta}}\sqrt{N}\sqrt{H + \frac{1}{2}|A||S|N\epsilon^2}\Bigg),
\end{aligned}$$

where the second step follows by Lemma 3.5, the third step follows by Lemma 3.7, the fifth step using Cauchy-Schwarz's inequality and the sixth step follow by Lemma 3.8.

The value of η affects the value: $\sqrt{\frac{1}{1-\frac{1}{2}(\beta+1)\eta}}\frac{1+\eta\beta}{\sqrt{\eta}}$. We minimize that value by choosing $\eta = \frac{1}{2\beta+1}$. since $\beta = \frac{1}{3}|A|H^3$ we get: $\sqrt{\frac{1}{1-\frac{1}{2}(\beta+1)\eta}}\frac{1+\eta\beta}{\sqrt{\eta}} = \sqrt{2|A|H^3+2} \leq 2\sqrt{|A|}H^{\frac{3}{2}}$.

The regret is therefore upper bounded by:

$$\begin{aligned}
Regret &\leq \frac{m}{H} D_\rho \left(H^2|A|^2\rho N + 2N\epsilon\sqrt{|A|}|S| + 4\sqrt{|S||A|}H^2\sqrt{N} + \frac{4}{\sqrt{2}}|S||A|H^{\frac{3}{2}}N\epsilon\right) \\
&\leq \frac{m}{H} D_\rho \left(H^2|A|^2\rho N + 4|S||A|H^{\frac{3}{2}}N\epsilon + 4\sqrt{|S||A|}H^2\sqrt{N}\right) \\
&= KD_\rho \left(H|A|^2\rho + 4|S||A|\sqrt{H}\epsilon + 4\sqrt{|S||A|}H\frac{1}{\sqrt{N}}\right)
\end{aligned}$$

Where the third transition holds since $K = mN$. It's clear that to minimize the Regret bound, we choose $\epsilon = \sqrt{\frac{H}{|S||A|N}}$ and $\rho = \sqrt{\frac{|S|}{|A|^3N}}$.

For simplicity, we define a new parameter ϵ_1 and find the optimal ratio between m and N to get the average regret below ϵ_1:

$$\frac{1}{K}Regret \leq D_\rho 9H\sqrt{|S||A|}\frac{1}{\sqrt{N}} \leq \epsilon_1, \tag{17}$$

which occurs when:

$$N \geq 81H^2|A||S|D_\rho^2\frac{1}{\epsilon_1^2}$$

Namely, we perform $N = 81H^2|A||S|D_\rho^2 \frac{1}{\epsilon_1^2}$ gradient steps. In each gradient step $m = \frac{H^2|A||S|}{\epsilon^2\rho^2}\log(\frac{2|A||S|K}{\delta})$ samples are sampled. Since ϵ is chosen as $\epsilon = \sqrt{\frac{H}{|A||S|N}}$, and ρ is chosen as $\rho = \sqrt{\frac{|S|}{|A|^3N}}$ the total number of episodes taken every policy gradient step is $m = N^2H|A|^5|S|\log(\frac{2|A||S|K}{\delta})$. Observe that the above is equivalent to $K = N^3H|A|^5|S|\log(\frac{2|A||S|K}{\delta})$, which means that:

$$N = K^{\frac{1}{3}}(\frac{1}{H|A|^5|S|\log(\frac{2|A||S|K}{\delta})})^{\frac{1}{3}}$$

and since $m = \frac{K}{N}$ the value of m is:

$$m = K^{\frac{2}{3}}(H|A|^5|S|\log(\frac{2|A||S|K}{\delta}))^{\frac{1}{3}}$$

The final regret is bounded by:

$$\begin{aligned} Regret &\leq D_\rho 9H\sqrt{|S||A|}\frac{K}{\sqrt{N}} \\ &\leq 9K^{\frac{5}{6}}D_\rho H^{\frac{7}{6}}|S|^{\frac{2}{3}}|A|^{\frac{4}{3}}\log^{\frac{1}{6}}(\frac{2|A||S|k}{\delta}) \end{aligned}$$

□

For completeness we show that D_ρ is well defined.

Lemma A.6 *For a state $s \in S$ and a policy $\pi \in \Pi_\rho$ in the direct parameterization, the occupancy measure is positive. Namely, $d_s^\pi > 0$.*

Proof: Fix a state $s \in S$, and let j be the level where the state s is. Namely, $s \in S_j$. From the assumption that all states are reachable from s_0, there exists s_{j-1} and an action a_{j-1} such that $P(s|s_{j-1}, a_{j-1}) > 0$. Doing the same for every $i \in [j]$ we can generate an episode from s_0 to s: $s_0, a_0, r_0, s_1, ..., s_{j-1}, a_{j-1}, r_{j-1}, s_j$, where $s = s_j$, such that for every $i \in [j-1]$, $P(s_i|s_{i-1}, a_{i-1}) > 0$. It is clear that:

$$d_s^\pi = Pr^\pi(s_j = s) \geq \Pi_{i=0}^{j-1}\pi(a_i|s_i)P(s_{i+1}|s_i, a_i) \geq \rho^j\Pi_{i=0}^{j-1}P(s_{i+1}|s_i, a_i) > 0$$

Where the third step holds due to $\pi \in \Pi_\rho$ and the fourth step follows by $P(s_i|s_{i-1}, a_{i-1}) > 0$ for all $i \in [1, j]$ as stated above. □

B Proofs for section 4

Algorithm 2: Policy Gradient with softmax parameterization - with random start state

Input: MDP,K, δ, λ;
$\theta = \frac{1}{|A|}\vec{1}$; `/* The dimension of the vector is` $|S||A|$ `*/`
$N = K^{1/3}(\lambda/4|S||A|\log(\frac{2|A||S|K}{\delta}))^{1/3}$;
$m = K^{2/3}(4|S||A|\log(\frac{2|A||S|K}{\delta})/\lambda)^{1/3}$;
$\eta = \log(|A|)$;
for $t = 1, 2, \ldots, N$ **do**
  $\widetilde{Q^t} = \vec{0}$;
  $\widetilde{V^t} = \vec{0}$;
  $\widetilde{A^t} = \vec{0}$;
  for $i = 1, 2, \ldots, m$ **do**
    Run policy π_θ on the MDP and get $(s_0^i, a_0^i, r_0^i, s_1^i, \ldots, s_{H-1}^i, a_{H-1}^i, r_0^i, s_H^i)$;
    for $s, a \in S, A$ **do**
      l =The level of s;
      $\widetilde{Q^t} += \frac{1}{m}\sum_{j=l}^{H} r_j^i I(s_l^i = s, a_l^i = a)$;
    end
  end
  for $s \in S$ **do**
    for $a \in A$ **do**
      $\widetilde{V^t}_s += \pi^\theta(a|s)\widetilde{Q^\theta}_{s,a}$
    end
    for $a \in A$ **do**
      $\widetilde{A^t}_{s,a} = \widetilde{Q^t}_{s,a} - \widetilde{V^t}_s$
    end
  end
  $\theta = \theta + \eta\widetilde{A^t}$
end

Proof:[Of Lemma 4.1] Fix a policy π, and two state-action pairs s, a, s', a'.
If $s \neq s'$, $\pi(a|s)$ does not depend on $\theta_{s',a'}$, so $\frac{\partial \log \pi(a|s)}{\partial \theta_{s',a'}} = 0$.
If $s' = s$, and $a' \neq a$,

$$\frac{\partial \log \pi(a|s)}{\partial \theta_{s',a'}} = -\pi(a'|s)$$

If $s' = s$, and $a' = a$,

$$\frac{\partial \log \pi(a|s)}{\partial \theta_{s',a}} = 1 - \pi(a|s)$$

□

Lemma B.1 *For a policy π and a state-action pair s, a,*

$$\mathbb{E}_{a' \sim \pi(\cdot|s;\theta)}\left[\nabla_\theta \log(\pi(a'|s;\theta)\right]_{s,a} = 0$$

Algorithm 3: Policy Gradient with softmax parameterization - with resets

Input: MDP,K, δ;
$\theta = \frac{1}{|A|}\vec{1}$; `/* The dimension of the vector is |S||A| */`
$N = K^{1/4}((4/3)H|S|^2|A|^2)^{-1/4}(\log(2|A||S|K/\delta))^{-1/8}$;
$m = K^{3/4}((4/3)H|S|^2|A|^2)^{1/4}(\log(2|A||S|K/\delta))^{1/8}$;
$m_1 = (3K/H)^{1/2}\log(\frac{2|A||S|K}{\delta})^{1/4}$;
$\eta = \log(|A|)$;
for $t = 1, 2, \ldots, N$ **do**
  $\widetilde{Q^t} = \vec{0}$;
  $\widetilde{V^t} = \vec{0}$;
  $\widetilde{A^t} = \vec{0}$;
  for $i = 1, 2, \ldots, m$ **do**
    if $i \leq m_1$ **then**
      Run policy π_θ on the MDP starting from a random state and choosing a random first action, and get $(s_0^i, a_0^i, r_0^i, s_1^i, \ldots, s_{H-1}^i, a_{H-1}^i, r_0^i, s_H^i)$;
      for $s, a \in S, A$ **do**
        l =The level of s;
        $\widetilde{Q^t} += \frac{1}{m_1}\sum_{j=l}^{H} r_j^i I(s_l^i = s, a_l^i = a)$;
      end
    else
      Run policy π_θ on the MDP starting from s_0
    end
  end
  for $s \in S$ **do**
    for $a \in A$ **do**
      $\widetilde{V^t}_s += \pi^\theta(a|s)\widetilde{Q^\theta}_{s,a}$
    end
    for $a \in A$ **do**
      $\widetilde{A^t}_{s,a} = \widetilde{Q^t}_{s,a} - \widetilde{V^t}_s$
    end
  end
  $\theta = \theta + \eta\widetilde{A^t}$
end

Proof: Fix a policy π and a state-action pair s, a,

$$\begin{aligned}
\mathbb{E}_{a'\sim\pi(\cdot|s;\theta)}\left[\nabla_\theta \log(\pi(a'|s;\theta))\right]_{s,a} &= \sum_{a'\in A}\pi(a'|s;\theta)(I(a=a') - \pi(a|s;\theta)) \\
&= \pi(a|s;\theta) - \sum_{a'\in A}\pi(a'|s;\theta)\pi(a|s;\theta) \\
&= \pi(a|s;\theta) - \pi(a|s;\theta)\sum_{a'\in A}\pi(a'|s;\theta) \\
&= \pi(a|s;\theta) - \pi(a|s;\theta) \\
&= 0
\end{aligned}$$

□

Lemma B.2 *For a policy π_θ,*

$$\nabla_\theta V^\pi(s_0) = H\mathbb{E}_{s\sim d^\pi}\mathbb{E}_{a\sim\pi(\cdot|s;\theta)}\left[\nabla_\theta \log(\pi(a|s;\theta)A^\pi(s,a)\right]$$

Proof: The policy gradient theorem (1) states that:

$$\nabla_\theta V^\pi(s_0) = H\mathbb{E}_{s\sim d^\pi}\mathbb{E}_{a\sim\pi(\cdot|s;\theta)}\left[\nabla_\theta \log(\pi(a|s;\theta))Q^\pi(s,a)\right] \tag{18}$$

We have that:

$$\begin{aligned}
\nabla_\theta V^\pi(s_0) &= H\mathbb{E}_{s\sim d^\pi}\mathbb{E}_{a\sim\pi(\cdot|s;\theta)}\left[\nabla_\theta \log(\pi(a|s;\theta)Q^\pi(s,a)\right] \\
&= H\mathbb{E}_{s\sim d^\pi}\mathbb{E}_{a\sim\pi(\cdot|s;\theta)}\left[\nabla_\theta \log(\pi(a|s;\theta)Q^\pi(s,a) - \nabla_\theta \log(\pi(a|s;\theta)V^\pi(s))\right] \\
&= H\mathbb{E}_{s\sim d^\pi}\mathbb{E}_{a\sim\pi(\cdot|s;\theta)}\left[\nabla_\theta \log(\pi(a|s;\theta)(Q^\pi(s,a) - V^\pi(s))\right] \\
&= H\mathbb{E}_{s\sim d^\pi}\mathbb{E}_{a\sim\pi(\cdot|s;\theta)}\left[\nabla_\theta \log(\pi(a|s;\theta)A^\pi(s,a))\right],
\end{aligned}$$

where the second transition follows by Lemma B.1 since $\mathbb{E}_{a\sim\pi(\cdot|s;\theta)}\left[\nabla_\theta \log(\pi(a|s;\theta)\right]$ is the 0 vector. □

Proof:[Of Lemma 4.2] Fix a policy π and a state-action pair s, a.

$$\begin{aligned}
\nabla_\theta V^\pi(s_0)_{s,a} &= H\mathbb{E}_{s'\sim d^\pi}\mathbb{E}_{a'\sim\pi(\cdot|s;\theta)}\left[\nabla_\theta \log(\pi(a'|s';\theta)_{s,a}A^\pi(s',a'))\right] \\
&= H\mathbb{E}_{s'\sim d^\pi}\sum_{a'\in A}\pi(a'|s';\theta)\nabla_\theta \log(\pi(a'|s';\theta)_{s,a}A^\pi(s',a')) \\
&= d_s^\pi\sum_{a'\in A}\pi(a'|s;\theta)\nabla_\theta \log(\pi(a'|s;\theta))_{s,a}A^\pi(s,a')) \\
&= d_s^\pi\sum_{a'\in A}\pi(a'|s;\theta)(I(a=a') - \pi(a|s))A^\pi(s,a') \\
&= d_s^\pi\pi(a|s;\theta)A^\pi(s,a) - d_s^\pi\sum_{a'\in A}\pi(a'|s;\theta)\pi(a|s)A^\pi(s,a') \\
&= d_s^\pi\pi(a|s;\theta)A^\pi(s,a) - d_s^\pi\pi(a|s;\theta)\sum_{a'\in A}\pi(a'|s;\theta)A^\pi(s,a') \\
&= d_s^\pi\pi(a|s;\theta)A^\pi(s,a),
\end{aligned}$$

where the first transition follows by Lemma B.2, the fourth transition follows by Lemma 4.1 and the last transition follows by Lemma A.4 which does not depend on the parameterization. □

Lemma B.3 *For a policy π and a state $s \in S$,*

$$\sum_{a\in A}\pi(a|s)\widetilde{A^\pi}(s,a) = 0$$

Proof: Fix a policy π and a state $s \in S$.

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

where the first transition follows by the definition of $\widetilde{A^\pi}$ as in (9), the third step holds since $\widetilde{V^\pi}(s)$ does not depend on a, the fourth step follows by the fact that $\pi(\cdot|s)$ is a distribution, i.e., $\sum_{a\in A}\pi(a|s) = 1$, and the fifth step follows by the definition of $\widetilde{V^\pi}(s)$ as in (9). □

Definition B.4 *Define the approximated policy gradient as:*

$$\widetilde{\nabla_\theta V^\pi}(s_0) = \sum_{s,a} d_s^\pi \pi(a|s;\theta) \nabla_\theta \log(\pi(a|s;\theta)) \widetilde{A^\pi}(s,a) \tag{19}$$

The same as in Lemma 4.2, we get that for every state-action pair s, a and policy π:

$$\widetilde{\nabla_\theta V^\pi}(s_0)_{s,a} = d_s^\pi \pi(a|s;\theta) \widetilde{A^\pi}(s,a) \tag{20}$$

Proof:[Of Lemma 4.3] Consider the loss function:

$$L^\theta(w) = \left\| \mathbb{E}_{s\sim d^\pi} \mathbb{E}_{a\sim\pi(\cdot|s;\theta)} \nabla_\theta \log(\pi(a|s;\theta) \left(w^\top \nabla_\theta \log(\pi(a|s;\theta)) - \widetilde{A^\pi}(s,a) \right) \right\|_2$$

Let $w' = \widetilde{A^\pi}(\cdot,\cdot)$. Observe that for a state-action pair s, a:

$$\begin{aligned} \nabla_\theta \log(\pi(a|s))^\top w' - \widetilde{A^\pi}(s,a) &= \sum_{a'\in A} (I(a=a') - \pi(a'|s)) \widetilde{A^\pi}(s,a') - \widetilde{A^\pi}(s,a) \\ &= \widetilde{A^\pi}(s,a) - \sum_{a'\in A} \pi(a'|s) \widetilde{A^\pi}(s,a') - \widetilde{A^\pi}(s,a) \\ &= 0, \end{aligned} \tag{21}$$

where the last transition follows by Lemma B.3. Therefore $L^\theta(w') = 0$. Reorganizing the loss function, we get that:

$$L^\theta(w) = \left\| F(\theta) w - \mathbb{E}_{s\sim d^\pi} \mathbb{E}_{a\sim\pi(\cdot|s;\theta)} \nabla_\theta \log(\pi(a|s;\theta)) \widetilde{A^\pi}(s,a) \right\|_2$$

Since $L^\theta(w') = 0$, we know that w' is a global minimizer of L^θ, and that $L^\theta(w) = 0$. Consider the vector w_θ^* which satisfies the following:

$$F(\theta) w_\theta^* = \mathbb{E}_{s\sim d^\pi} \mathbb{E}_{a\sim\pi(\cdot|s;\theta)} \nabla_\theta \log(\pi(a|s;\theta)) \widetilde{A^\pi}(s,a).$$

Recall that the Moore Penrose inverse of a matrix A satisfies $A^\dagger b = \arg\min_{x:Ax=b} \|x\|_2$ (see Section C.4). Therefore

$$w_\theta^* = F(\theta)^\dagger \mathbb{E}_{s\sim d^\pi} \mathbb{E}_{a\sim\pi(\cdot|s;\theta)} \left[\nabla_\theta \log(\pi(a|s;\theta)) \widetilde{A^\pi}(s,a) \right] \tag{22}$$

is a global minimizer of $L^\theta(w)$. By the definition of the approximated gradient (19), We get that:

$$w_\theta^* = \frac{1}{H} F(\theta)^\dagger \widetilde{\nabla_\theta V^\pi}(s_0).$$

Note that the loss function is a norm of a vector, therefore as $L^\theta(w_\theta^*) = 0$, every coordinate in the vector is 0. Consider the s, a coordinate:

$$\begin{aligned} 0 &= \frac{1}{H} d_s^\pi \sum_{a'\in A} \pi(a'|s)(I(a'=a) - \pi(a|s)) \left(w^\top \nabla_\theta \log(\pi(a'|s;\theta)) - \widetilde{A^\pi}(s,a') \right) \\ &= \frac{1}{H} d_s^\pi \pi(a|s) \left[\left(w^\top \nabla_\theta \log(\pi(a|s;\theta)) - \widetilde{A^\pi}(s,a) \right) - \sum_{a'\in A} \pi(a'|s) \left(w^\top \nabla_\theta \log(\pi(a'|s;\theta)) - \widetilde{A^\pi}(s,a') \right) \right] \end{aligned}$$

Define the function $B(s,a,w) = w^\top \nabla_\theta \log(\pi(a|s;\theta)) - \widetilde{A^\pi}(s,a)$. The above implies that $B(s,a,w) = \sum_{a'\in A} \pi(a'|s) B(s,a',w)$. Therefore, B is independent on a, i.e., we can view B as a function of s

and w: $B(s,w)$. Define the vector v as $v = w^*_\theta - w'_\theta$. For a state-action pair s,a, consider $B(s,w^*_\theta)$:

$$\begin{aligned}
B(s,w^*_\theta) &= B(s,a,w^*_\theta) \\
&= \nabla_\theta \log(\pi(a|s;\theta))^\top w^*_\theta - \widetilde{A^\pi}(s,a) \\
&= \nabla_\theta \log(\pi(a|s;\theta))^\top (w'_\theta + v) - \widetilde{A^\pi}(s,a) \\
&= \nabla_\theta \log(\pi(a|s;\theta))^\top v \\
&= v_{s,a} - \sum_{a' \in A} \pi(a'|s) v_{s,a'},
\end{aligned}$$

where the fourth transition follows by (21). Therefore,

$$v_{s,a} = \sum_{a' \in A} \pi(a'|s) v_{s,a'} + B(s,w^*_\theta)$$

This implies that $v_{s,a}$ is independent of a, therefore we can view v as a vector that depends on the state alone: v_s, yet it is still a $|S||A|$ dimensional vector. Observe that for all state-action pair s,a the algorithm step (7) is equivalent to (8) up to a value that does not depend on the action a:

$$\begin{aligned}
\theta^{(t+1)}_{s,a} &= \theta^{(t)}_{s,a} + \eta \left[F(\theta^{(t)})^\dagger \widetilde{\nabla_\theta V^{\pi^t}}(s_0) \right]_{s,a} \\
&= \theta^{(t)}_{s,a} + \eta H w^*_{\theta^{(t)},s,a} \\
&= \theta^{(t)}_{s,a} + \eta H w'_{\theta^{(t)},s,a} + \eta H v_s \\
&= \theta^{(t)}_{s,a} + \eta H \widetilde{A^{\pi^t}}(s,a) + \eta H v_s
\end{aligned}$$

Where the second step follows by (22), the third step holds since $v = w^*_\theta - w'_\theta$, and the forth step follows by the definition of w'_θ. Observe the effect of v_s on the policy:

$$\begin{aligned}
\pi^{(t+1)}(a|s) &= \frac{exp(\theta^{(t)}_{s,a} + \eta H \widetilde{A^{\pi^t}}(s,a) + \eta H v_s)}{\sum_{a' \in A} exp(\theta^{(t)}_{s,a'} + \eta H \widetilde{A^{\pi^t}}(s,a') + \eta H v_s)} \\
&= \frac{exp(\theta^{(t)}_{s,a} + \eta H \widetilde{A^{\pi^t}}(s,a))}{\sum_{a' \in A} exp(\theta^{(t)}_{s,a'} + \eta H \widetilde{A^{\pi^t}}(s,a'))}.
\end{aligned}$$

This implies that the value of v is irrelevant for the algorithm, therefore the algorithm step (7) is equivalent to algorithm step (8). □

Proof:[Of Lemma 4.4] Let $m = \frac{H^2|S||A|}{\epsilon^2 \lambda} \log(\frac{2|S||A|K}{\delta})$ be the number of episodes sampled every phase. At phase t of the algorithm, for every state-action pair s,a, let $m_{s,a}$ be the number of episodes where the MDP restarted at state s and action a. Note that $m_{s,a}$ is a random number, and $\mathbb{E}[m_{s,a}] = \frac{\lambda}{|S||A|} m = H^2 \frac{1}{\epsilon^2} \log(\frac{2|S||A|K}{\delta})$. We will use both Chernoff and Hoeffding concentration bound (Theorem C.15,C.16). We use Chernoff concentration bound to get a lower bound (with some probability) on $m_{s,a}$, and Hoeffding concentration bound on the approximation of the Q-function assuming the bound on $m_{s,a}$ holds. Let $\tau_1, \tau_2, \ldots, \tau_{m_{s,a}}$ be the episodes sampled by starting at state s, taking action a and following the policy π^t, and let $r_1, r_2, \ldots, r_{m_{s,a}}$ be the corresponding sum of the rewards. Define

$$\widetilde{Q^{(t)}}(s,a) = \frac{1}{m_{s,a}} \sum_{i=1}^{m_{s,a}} r_i$$

By the Chernoff concentration bound, for some state-action pair s,a, to assure that

$$m_{s,a} > \frac{H^2}{2\epsilon^2} \log(\frac{2|S||A|K}{\delta}) \tag{23}$$

with probability $1 - \frac{1}{2|S||A|K}\delta$, we would need that $\frac{1}{2|S||A|K}\delta = e^{-\left(\frac{1}{2}\right)^2 \lambda m \frac{1}{|S||A|}\frac{1}{2}}$ which is equivalent to:

$$m \geq 8|S||A|\frac{1}{\lambda}\log(\frac{2|S||A|K}{\delta}), \tag{24}$$

Note that the requirement of the Lemma ($m \geq \frac{H^2|S||A|}{\epsilon^2\lambda}\log(\frac{2|A||S|K}{\delta})$ fulfills condition (24) assuming that $H^2\frac{1}{\epsilon^2} \geq 8$.

We presented a lower bound for the random number $m_{s,a}$ for every state-action pair s, a (23), we can now show a bound for $|\widetilde{Q^\pi}(s,a) - Q^\pi(s,a)|$ given a lower bound on the number of episodes $m_{s,a}$. For every i, $\mathbb{E}[r_i] = Q^{\pi^t}(s,a)$, and $0 \leq r_i \leq H$, therefore we use Hoeffding concentration bound again to show that for $\epsilon > 0$, a state action pair s, a and a policy π, with probability at least $1 - \frac{1}{2|S||A|K}\delta$, we have that:

$$|\widetilde{Q^\pi}(s,a) - Q^\pi(s,a)| \leq \epsilon \tag{25}$$

using the lower bound of $m_{s,a}$ (23). Using (23) and (25) and the union bound, with probability $1-\delta$ for every phase t and state-action pair s, a, $m_{s,a} \geq H^2\frac{1}{2\epsilon^2}\log(\frac{2|S||A|K}{\delta})$ and $\|\widetilde{Q^\pi} - Q^\pi\|_\infty \leq \epsilon$.

can now bound the advantage function using the bound on the Q-function:

$$\begin{aligned}
|A^\pi(s,a) - \widetilde{A^\pi}(s,a)| &= |Q^\pi(s,a) - V^\pi(s) - \widetilde{Q^\pi}(s,a) + \widetilde{V^\pi}(s)| \\
&\leq |Q^\pi(s,a) - \widetilde{Q^\pi}(s,a)| + |V^\pi(s,a) - \widetilde{V^\pi}(s)| \\
&\leq \epsilon + |\sum_{a' \in A}\pi(a'|s)(Q^\pi(s,a') - \widetilde{Q^\pi}(s,a')| \\
&\leq \epsilon + \sum_{a' \in A}\pi(a'|s)\epsilon \\
&= 2\epsilon
\end{aligned}$$

Where the first transition follows by the definition of $A^\pi(s,a)$ and $\widetilde{A^\pi}(s,a)$, and the third transition follows by (25). Therefore, for every phase t,

$$\|A^\pi - \widetilde{A^\pi})\|_\infty \leq 2\epsilon \tag{26}$$

□

Define for a phase t and a state-action pair s, a:

$$\epsilon_{t,s,a} = A^{\pi^t}(s,a) - \widetilde{A^{\pi^t}}(s,a) \tag{27}$$

By Lemma 4.4, with high probability, we have

$$|\epsilon_{t,s,a}| \leq 2\epsilon. \tag{28}$$

Definition B.5 *For a policy π and a state $s \in S$, define the function $Z_\pi(s)$:*

$$\widetilde{Z_\pi}(s) = \sum_{a \in A}\pi(a|s)exp(\eta H\widetilde{A^t}(s,a))$$

To simplify the notation - we write Z_t instead of Z_{π^t} during phase t.

Lemma B.6 *For a phase t, a state-action pair s, a, and the policy gradient update (8)*

$$\pi^{t+1}(a|s) = \pi^t(a|s)\frac{exp(\eta H\widetilde{A^t}(s,a))}{\widetilde{Z_t}(s)}$$

Proof: Fix a phase t, and a state-action pair s, a:

$$\begin{aligned}
\pi_{t+1}(a|s) &= \frac{exp(\theta_{s,a}^{t+1})}{\sum_{a' \in A} exp(\theta_{s,a}^{t+1})} \\
&= \frac{exp(\theta_{s,a}^{t} + \eta H \widetilde{A^t}(s,a))}{\sum_{a' \in A} exp(\theta_{s,a'}^{t} + \eta H \widetilde{A^t}(s,a'))} \\
&= \frac{exp(\theta_{s,a}^{t}) exp(\eta H \widetilde{A^t}(s,a))}{\sum_{a' \in A} \left[\pi_t(a|s) exp(\eta H \widetilde{A^t}(s,a')) \sum_{a'' \in A} exp(\theta_{s,a''}^{t}) \right]} \\
&= \frac{exp(\theta_{s,a}^{t}) exp(\eta H \widetilde{A^t}(s,a))}{\sum_{a'' \in A} exp(\theta_{s,a''}^{t}) \sum_{a' \in A} \left[\pi_t(a|s) exp(\eta H \widetilde{A^t}(s,a')) \right]} \\
&= \pi(a|s) \frac{exp(\eta H \widetilde{A^t}(s,a))}{\sum_{a' \in A} \left[\pi_t(a|s) exp(\eta H \widetilde{A^t}(s,a')) \right]} \\
&= \pi(a|s) \frac{exp(\eta H \widetilde{A^t}(s,a))}{\widetilde{Z_t}(s)}
\end{aligned}$$

□

Lemma B.7 *For a policy π and a state s,*

$$\log(\widetilde{Z_\pi}(s)) \geq 0$$

Proof: Fix a policy π and a state s,

$$\begin{aligned}
\log(\widetilde{Z_\pi}(s)) &= \log \left(\sum_{a \in A} \pi(a|s) exp(\eta H \widetilde{A^t}(s,a)) \right) \\
&\geq \sum_{a \in A} \pi(a|s) \log \left(exp(\eta H \widetilde{A^t}(s,a)) \right) \\
&= \sum_{a \in A} \pi(a|s) \eta H \widetilde{A^t}(s,a) \\
&= 0,
\end{aligned}$$

where the second transition follows by Jensen's inequality, ($\log()$ is a concave function and $\sum_{a \in A} \pi(a|s) = 1$), and the last transition follows by Lemma B.3. □

Lemma B.8

$$\frac{1}{\eta H^2} \sum_{t=1}^{T} \sum_{s \in S} d_s^* \log(\widetilde{Z_t}(s)) \leq H^2 + 2HT\epsilon \tag{29}$$

Proof: From Lemma A.3 we have that for every distribution μ:

$$V^{\pi}(\mu) - V^{\pi'}(\mu) = \frac{1}{H} \sum_{s \in S, a \in A} d_s^{\pi,\mu} \pi(a|s) A^{\pi'}(s,a)$$

Recall the notation $d^{\pi,\mu}$ which equals the steady state distribution of the states assuming the starting state is distributed according to μ, and the episodes are run according to the policy π. Consider the distribution $d^* = d^{\pi^*}$,

$$V^{\pi^{t+1}}(d^*) - V^{\pi^t}(d^*) = \frac{1}{H} \sum_{s \in S} d_s^{\pi^{t+1}, d^*} \sum_{a \in A} \pi^{t+1}(a|s) A^{\pi^t}(s,a)$$

$$
\begin{aligned}
&= \frac{1}{H}\sum_{s\in S} d_s^{\pi^{t+1},d^*}\sum_{a\in A}\pi^{t+1}(a|s)\epsilon_{t,s,a} + \frac{1}{H}\sum_{s\in S} d_s^{\pi^{t+1},d^*}\sum_{a\in A}\pi^{t+1}(a|s)\widetilde{A^{\pi^t}}(s,a)\\
&\geq -2\epsilon + \frac{1}{H^2\eta}\sum_{s\in S} d_s^{\pi^{t+1},d^*}\sum_{a\in A}\pi^{t+1}(a|s)\log\left(\frac{\pi^{t+1}(a|s)\widetilde{Z_t}(s)}{\pi^t(a|s)}\right)\\
&= -2\epsilon + \frac{1}{H^2\eta}\sum_{s\in S} d_s^{\pi^{t+1},d^*}\sum_{a\in A} KL(\pi^{t+1}(\cdot|s),\pi^t(\cdot|s))\\
&\qquad + \frac{1}{H^2\eta}\sum_{s\in S} d_s^{\pi^{t+1},d^*}\sum_{a\in A}\pi^{t+1}(a|s)\log(\widetilde{Z_t}(s))\\
&\geq -2\epsilon + \frac{1}{H^2\eta}\sum_{s\in S} d_s^{\pi^{t+1},d^*}\sum_{a\in A}\pi^{t+1}(a|s)\log(\widetilde{Z_t}(s))\\
&\geq -2\epsilon + \frac{1}{H^2\eta}\sum_{s\in S} d_s^{\pi^{t+1},d^*}\log(\widetilde{Z_t}(s)),
\end{aligned}
$$

where the first transition follows by Lemma A.3, the second transition follows by Lemma B.6 and (28), the third transition follows by the definition of KL (46), the fourth transition holds since $KL(\cdot,\cdot)$ is non-negative by Lemma C.13, and the fifth transition holds since $\log(\widetilde{Z_t}(s))$ does not depend on a, and $\sum_{a\in A}\pi^{t+1}(a|s)=1$. When we sum over the phases, we get:

$$
V^{\pi^N}(d^*) - V^{\pi^0}(d^*) \geq -2N\epsilon + \frac{1}{\eta H^2}\sum_{t=1}^{N}\sum_{s\in S} d_s^{\pi^{t+1},d^*}\log(\widetilde{Z_t}(s))
$$

Observe that $d_s^{\pi^{t+1},d^*} \geq \frac{1}{H}d^*$, therefor:

$$
V^{\pi^N}(d^*) - V^{\pi^0}(d^*) \geq -2N\epsilon + \frac{1}{\eta H^3}\sum_{t=1}^{N}\sum_{s\in S} d_s^*\log(\widetilde{Z_t}(s))
$$

The fact that $V^{\pi^N}(d^*) - V^{\pi^0}(d^*) \leq H$ finishes the proof. □

Lemma B.9 *Assuming the algorithm is run for N phases, and in each phase the advantage function is approximated with the approximation error: 2ϵ, i.e., $\|A^{\pi^t} - \widetilde{A^{\pi^t}}\|_\infty \leq 2\epsilon$, then:*

$$
\sum_{t=1}^{N} V^*(\mu) - V^t(\mu) \leq 3H(H+N\epsilon).
$$

Proof: Fix a phase t,

$$
\begin{aligned}
V^*(\mu) - V^t(\mu) &= \frac{1}{H}\sum_{s\in S} d_s^{\pi^*}\sum_{a\in A}\pi^*(a|s)A^{\pi^t}(s,a)\\
&\leq 2\epsilon + \frac{1}{H^2\eta}\sum_{s\in S} d_s^{\pi^*}\sum_{a\in A}\pi^*(a|s)\log\left(\frac{\pi^{t+1}(a|s)\widetilde{Z_t}(s)}{\pi^t(a|s)}\right)\\
&= 2\epsilon + \frac{1}{H^2\eta}\sum_{s\in S} d_s^{\pi^*}\sum_{a\in A}\pi^*(a|s)\log\left(\frac{\pi^*(a|s)\pi^{t+1}(a|s)\widetilde{Z_t}(s)}{\pi^t(a|s)\pi^*(a|s)}\right)\\
&= 2\epsilon + \frac{1}{H^2\eta}\sum_{s\in S} d_s^{\pi^*}\left(KL(\pi^*(\cdot|s),\pi^t(\cdot|s)) - KL(\pi^*(\cdot|s),\pi^{t+1}(\cdot|s)) + \sum_{a\in A}\pi^*(a|s)\log\widetilde{Z_t}(s)\right)\\
&= 2\epsilon + \frac{1}{H^2\eta}\sum_{s\in S} d_s^{\pi^*}\left(KL(\pi^*(\cdot|s),\pi^t(\cdot|s)) - KL(\pi^*(\cdot|s),\pi^{t+1}(\cdot|s)) + \log\widetilde{Z_t}(s)\right),
\end{aligned}
$$

where the first transition holds since $|A^{\pi^t}(s,a) - \widetilde{A^{\pi^t}}(s,a)| \leq 2\epsilon$. When we sum over the phases, we have:

$$\begin{aligned}
\sum_{t=1}^{N} V^*(\mu) - V^t(\mu) &\leq 2\epsilon N + \frac{1}{H^2\eta}\sum_{t=1}^{N}\sum_{s\in S} d_s^{\pi^*}\left(KL(\pi^*(\cdot|s),\pi^t(\cdot|s)) - KL(\pi^*(\cdot|s),\pi^{t+1}(\cdot|s)) + \log\widetilde{Z_t}(s)\right) \\
&= 2\epsilon N + \frac{1}{H^2\eta}\sum_{s\in S} d_s^{\pi^*}\left(KL(\pi^*(\cdot|s),\pi^1(\cdot|s)) - KL(\pi^*(\cdot|s),\pi^{N+1}(\cdot|s)) + \sum_{t=1}^{N}\log\widetilde{Z_t}(s)\right) \\
&\leq 2\epsilon N + \frac{1}{H^2\eta}\sum_{s\in S} d_s^{\pi^*}\left(KL(\pi^*(\cdot|s),\pi^1(\cdot|s)) + \sum_{t=1}^{N}\log\widetilde{Z_t}(s)\right) \\
&\leq 2\epsilon N + \frac{1}{H^2\eta}\sum_{s\in S} d_s^{\pi^*}\left(\log(|A|) + \sum_{t=1}^{N}\log\widetilde{Z_t}(s)\right) \\
&= 2\epsilon N + \frac{\log(|A|)}{H\eta} + \frac{1}{H^2\eta}\sum_{s\in S} d_s^{\pi^*}\sum_{t=1}^{N}\log\widetilde{Z_t}(s) \\
&\leq 2\epsilon N + \frac{\log(|A|)}{H\eta} + H^2 + 2HN\epsilon \\
&\leq H^2 + 2(H+1)N\epsilon + \frac{\log(|A|)}{H\eta},
\end{aligned}$$

where the third transition follows by Lemma C.13, the fourth transition follows by Lemma C.14 assuming the starting policy is the uniform policy as states in the algorithm description, the sixth transition follows by Lemma B.8. Assuming the step size is large enough: $\eta \geq \frac{1}{H}\log(|A|)$, we have that $\frac{\log(|A|)}{H\eta} \leq H^2$ which completes the proof. □

Note that the above lemma applies to both settings of the softmax parameterization.

Proof:[Of Theorem 4.5] Using Lemma B.9, observe the regret:

$$Regret = m\sum_{t=1}^{N} V^*(\mu) - V^t(\mu) \leq 3mH(H + N\epsilon)$$

To minimize the regret, we optimize ϵ to $\epsilon = \frac{H}{N}$. From Lemma 4.4:

$$m = \frac{4H^2|S||A|}{\epsilon^2\lambda}\log(\frac{2|A||S|K}{\delta}) = \frac{4N^2|S||A|}{\lambda}\log(\frac{2|A||S|K}{\delta})$$

Multiplying both sides by N gives:

$$K = mN = \frac{4N^3|S||A|}{\lambda}\log(\frac{2|A||S|K}{\delta}).$$

This implies

$$N = K^{\frac{1}{3}}\left(\frac{\lambda}{4|S||A|\log(\frac{2|A||S|K}{\delta})}\right)^{\frac{1}{3}},$$

and

$$m = \frac{K}{N} = K^{\frac{2}{3}}\left(\frac{4|S||A|\log(\frac{2|A||S|K}{\delta})}{\lambda}\right)^{\frac{1}{3}}$$

□

Proof:[Of Lemma 4.6] Let $m_1 = \frac{2H^2|S||A|}{\epsilon^2}\log(\frac{2|A||S|K}{\delta})$ be the number of episodes sampled by restarting the algorithm at some state $s \neq s_0$ every phase. At phase t of the algorithm, for every

state-action pair s, a, let $\tau_1, \tau_2, \ldots, \tau_{\frac{m_1}{|S||A|}}$ be the trajectories sampled by starting at state s, taking action a and then following the policy π^t. Let $r_1, r_2, \ldots, r_{\frac{m_1}{|S||A|}}$ be their corresponding sum of the rewards. Define

$$\widetilde{Q^{(t)}}(s,a) = \frac{|S||A|}{m_1} \sum_{i=1}^{\frac{m_1}{|S||A|}} (r_i)$$

Observe that for every i, $\mathbb{E}[r_i] = Q^{(t)}(s,a)$, and $0 \leq r_i \leq H$, therefore we use Hoeffding concentration bound (Theorem C.15) to show that for $\epsilon > 0$, a state action pair s, a and a policy π, with probability at least $1 - \frac{1}{|S||A|K}\delta$, we get the following approximation of the gradient

$$|\widetilde{Q^{\pi}}(s,a) - Q^{\pi}(s,a)| \leq \epsilon, \tag{30}$$

Using the union bound we get that with probability $1 - \delta$, for every phase t the following bound hold: $\|\widetilde{Q^{\pi}} - Q^{\pi^t}\|_\infty \leq \epsilon$.

We define the approximated value function $\widetilde{V^{\pi}}$ and the approximated advantage function $\widetilde{A^{\pi}}$ using the approximated Q-function just as before. For a state-action pair s, a:

$$\widetilde{V^{\pi}}(s) = \sum_{a' \in A} \pi(a'|s)\widetilde{Q^{\pi}}(s,a') \qquad \text{and} \qquad \widetilde{A^{\pi}}(s,a) = \widetilde{Q^{\pi}}(s,a) - \widetilde{V^{\pi}}(s)$$

And similar to (26) we get that

$$\|\widetilde{A^{\pi}} - A^{\pi}\|_\infty \leq 2\epsilon.$$

□

Proof:[Of Theorem 4.7] As in Lemma B.9:

$$\sum_{t=1}^{N} V^*(s_0) - V^t(s_0) \leq 3H(H + N\epsilon)$$

Observe the regret:

$$Regret = \sum_{t=1}^{N} m_1 H + m_2(V^*(s_0) - V^t(s_0)) \leq H m_1 N + 3(m - m_1)H(H + N\epsilon)$$

To minimize the regret we take $\epsilon = \frac{H}{N}$. As m_1 is defined using ϵ in Lemma 4.6, we see that:

$$\begin{aligned} Regret &\leq H^2 m_1 \frac{1}{\epsilon} + 6(m - m_1)H^2 \\ &= 2H^3 \sqrt{2|S||A| \log(\frac{2|A||S|K}{\delta})} m_1^{\frac{3}{2}} + 6(m - m_1)H^2 \\ &\leq 2H^3 \sqrt{2|S||A| \log(\frac{2|A||S|K}{\delta})} m_1^{\frac{3}{2}} + 6mH^2, \end{aligned}$$

where the second transition follows by Lemma 4.6 as $\epsilon = \sqrt{\frac{2H^2|S||A|}{m_1} \log(\frac{2|A||S|K}{\delta})}$. We optimize m_1 to be:

$$m_1 = \left(\frac{4.5}{H^2|S||A| \log(\frac{2|A||S|K}{\delta})} \right)^{\frac{1}{3}} m^{\frac{2}{3}} \tag{31}$$

Note the value of m as a function of N. Since we optimized ϵ to be $\epsilon = \frac{H}{N}$ and m_1 in (31), we get that:

$$\begin{aligned} m &= m_1^{\frac{3}{2}} \sqrt{\frac{2}{9} H^2 |S||A| \log(\frac{2|A||S|K}{\delta})} \\ &= \left(\frac{2H^2|S||A|}{\epsilon^2}\right)^{\frac{3}{2}} \sqrt{\frac{2}{9} H^2 |S||A| \log(\frac{2|A||S|K}{\delta})} \\ &= \left(\frac{N}{H}\right)^3 \frac{4}{3} H^4 |S|^2 |A|^2 \sqrt{\log(\frac{2|A||S|K}{\delta})} \\ &= \frac{4}{3} H |S|^2 |A|^2 \sqrt{\log(\frac{2|A||S|K}{\delta})} N^3 \end{aligned}$$

Multiplying both sides by N gives us:

$$K = Nm = \frac{4}{3} H |S|^2 |A|^2 \sqrt{\log(\frac{2|A||S|K}{\delta})} N^4$$

This implies

$$N = K^{\frac{1}{4}} \left(\frac{3}{4H|S|^2|A|^2\sqrt{\log(\frac{2|A||S|K}{\delta})}}\right)^{\frac{1}{4}},$$

and

$$m = \frac{K}{N} = K^{\frac{3}{4}} \left(\frac{4}{3} H |S|^2 |A|^2 \sqrt{\log(\frac{2|A||S|K}{\delta})}\right)^{\frac{1}{4}}$$

Which yields the final regret bound:

$$\begin{aligned} Regret &\leq 12mH^2 \\ &\leq 12K^{\frac{3}{4}} \left(\frac{4}{3} H^9 |S|^2 |A|^2 \sqrt{\log(\frac{2|A||S|K}{\delta})}\right)^{\frac{1}{4}} \end{aligned}$$

□

C Additional Proofs

C.1 The projection algorithm w.r.t. the euclidean norm

We show how to implement the projection operation for the set $\Pi_\rho = \{z \in [0,1]^d \colon \sum_{i=0}^{d-1} z_i = 1, \forall i \in [d]\ z_i \geq \rho\}$. We assume that $\rho \leq \frac{1}{d}$, otherwise Π_ρ would be the empty set. Given a vector $y \in \mathbb{R}^d$, we would like to compute its projection to Π_ρ, i.e., $\theta^* = \arg\min_{z \in \Pi_\rho} \|z - y\|_2$. For simplicity and w.l.o.g., we assume that $y_0 \leq y_1, \ldots \leq y_{d-1}$.

For each $m \in [d]$ define x^m as follows, $x^m_i = \rho$ for $i \leq m-1$ and $x^m_i = y_i - \lambda_m$ for $i \geq m$, where $\lambda_m = \frac{1}{d-m}\Big(\sum_{i \geq m}(y_i) + m\rho - 1\Big)$. Let

$$\theta = \arg\min_{m : x^m \in \Pi_\rho} \|x^m - y\|_2. \tag{32}$$

We show that there exists at least one value m for which $x^m \in \Pi_\rho$ and that $\theta = \theta^* = \arg\min_{z \in \Pi_\rho} \|z - y\|_2$

Lemma C.1 *Let $\theta^* = \arg\min_{z \in \Pi_\rho} \|z - y\|_2^2$, we have that:*

$$\forall i,\ \theta^*_i \leq \theta^*_{i+1}$$

Proof: Assume for contradiction that there exists $k < j$ such that $\theta^*_k > \theta^*_j$. Define a vector z which equals the average of θ^*_k and θ^*_j in the k and j coordinates, and equals θ^* in every other coordinate, i.e., $z_k = z_j = \frac{1}{2}(\theta^*_k + \theta^*_j)$ and $z_i = \theta^*_i$ for $i \notin \{j,k\}$. Clearly, $z \in \Pi_\rho$. It follows that that:

$$\begin{aligned}
\sum_{i=0}^{d-1}(\theta^*_i - y_i)^2 - \sum_{i=0}^{d-1}(z_i - y_i)^2 &= (\theta^*_k - y_k)^2 + (\theta^*_j - y_j)^2 - (\frac{\theta^*_k + \theta^*_j}{2} - y_k)^2 - (\frac{\theta^*_k + \theta^*_j}{2} - y_j)^2 \\
&= \theta^{*2}_k - 2\theta^*_k y_k + \theta^{*2}_j - 2\theta^*_j y_j - \frac{1}{2}\theta^{2*}_k - \frac{1}{2}\theta^{2*}_j - \theta^*_j\theta^*_k + (y_k + y_j)(\theta^*_k + \theta^*_j) \\
&= \frac{1}{2}\theta^{*2}_k - \theta^*_k\theta^*_j + \frac{1}{2}\theta^{*2}_j - \theta^*_k y_k - \theta^*_j y_j + \theta^*_k y_j + \theta^*_j y_k \\
&= \frac{1}{2}(\theta^*_k - \theta^*_j)^2 + (\theta^*_k - \theta^*_j)(y_j - y_k) \\
&\geq \frac{1}{2}(\theta^*_k - \theta^*_j)^2 \\
&> 0
\end{aligned}$$

where we use the fact that $y_j \geq y_k$ and $\theta^*_k > \theta^*_j$. This contradicts to the optimality of θ^*. □

Lemma C.2 *Let $0 \leq m \leq d-1$ be the number of values in θ^* which are ρ. Then $x^m \in \Pi_\rho$*

Proof: First we show that $\sum_{i=0}^{d-1} x^m_i = 1$:

$$\sum_{i=0}^{d-1} x^m_i = m\rho + \sum_{i \geq m} y_i - (d-m)\lambda_m = m\rho + \sum_{i \geq m} y_i - \sum_{i \geq m} y_i - m\rho + 1 = 1$$

Assume by contradiction that there exists j such that $x^m_j < \rho$. Clearly $j \geq m$, since $x^m_i = \rho$ for $i \leq m-1$. If $m+1 = d$ then $x^m = (\rho, \ldots, \rho, 1 - \rho(d-1)) \in \Pi_\rho$, and $1 - \rho(d-1) \geq \rho$ because of the assumption $\rho \leq \frac{1}{d}$ as stated above. We can now assume that $m+1 < d$. For $i \geq m$, we have $x^m_i \leq x^m_{i+1}$ since $x^m_i = y_i - \lambda_m \leq y_{i+1} - \lambda_m = x^m_{i+1}$. Since we assume that $j \geq m$ and $x^m_j < \rho$, we have that $y_m - \lambda_m \leq y_j + \lambda_m = x^m_j < \rho$. Consider the expression $(|A| - m)(\rho + \lambda_m - y_m)$ which is

strictly positive. We have,

$$\begin{aligned}
0 < (d-m)(\rho + \lambda_m - y_m) &= (d-m)\rho + (d-m)\lambda_m - (d-m)y_m \\
&= (d-m)\rho + \sum_{i \geq m} y_i + m\rho - 1 - (d-m)y_m \\
&\leq d\rho - 1 + y_{d-1} - y_m + \sum_{i=m}^{d-2} y_i - y_m \\
&\leq d\rho - 1 + y_d - y_m
\end{aligned}$$

Where the second transition follows by the definition of $\lambda_m = \frac{1}{d-m}(\sum_{i \geq m} y_i + m\rho - 1)$. This implies that $y_{d-1} > 1 - d\rho + y_m$.

By Lemma C.1, since the values are non-decreasing in θ^* we have that $\theta^*_1 = \cdots = \theta^*_m = \rho$.

We define z' which equals θ^* except that we reduce the m coordinate to ρ and increase the $d-1$ coordinate accordingly, i.e., look at z', where $z'_m = \rho$, $z'_{d-1} = \theta^*_{d-1} + \theta^*_m - \rho$, and $z'_i = \theta^*_i$ for $i \notin \{m, d-1\}$. Clearly $\sum_{i=0}^{d-1} z'_i = 1$ and $z'_i \geq \rho$ so $z' \in \Pi_\rho$.

Define $\alpha = \theta^*_m - \rho > 0$ and look at:

$$\begin{aligned}
\sum_{i=0}^{d-1}(\theta^*_i - y_i)^2 - \sum_{i=0}^{d-1}(z'_i - y_i)^2 &= (\rho + \alpha - y_m)^2 + (\theta^*_{d-1} - y_{d-1})^2 - (\rho - y_m)^2 - (\theta^*_{d-1} + \alpha - y_{d-1})^2 \\
&= 2\rho\alpha - 2\alpha y_m - 2\alpha\theta^*_{d-1} + 2\alpha y_{d-1} \\
&= 2\alpha(\rho - y_m - \theta^*_{d-1} + y_{d-1}) \\
&> 2\alpha(\rho - y_m - \theta^*_{d-1} + 1 - d\rho + y_m) \\
&= 2\alpha(\rho - \theta^*_{d-1} + 1 - d\rho) \\
&= 2\alpha(1 - (d-1)\rho - \theta^*_{d-1}) \\
&= 2\alpha(\sum_i \theta^*_i - (d-1)\rho - \theta^*_d) \\
&= 2\alpha(\sum_{i<d-1} \theta^*_i - (d-1)\rho) \geq 0
\end{aligned}$$

which implies that $\|\theta^* - y\|_2 > \|z' - y\|_2$, contradicting the optimality of θ^*. □

Theorem C.3

$$\theta^* = \operatorname*{arg\,min}_{m : x^m \in \Pi_\rho} \|x^m - y\|_2$$

Proof: First, by Lemma C.2, we have that there exists m such that $x^m \in \Pi_\rho$ where m is the number of coordiantes of value ρ in θ^*. We will show that $\theta^* = x^m$.

Second, by Lemma C.1 we have that the coordinates of θ^* are non-decreasing, i.e., $\theta^*_i \leq \theta^*_{i+1}$

Assume for contradiction that $\theta^* \neq x^m$. Since $x^m, \theta^* \in \Pi_\rho$ and $x^m \neq \theta^*$ there exists $k \neq j$ such that $\theta^*_k = x^m_k + a$ and $\theta^*_j = x^m_j + b$ where $ab < 0$ and $|a| < |b|$. Consider z which is equal to θ^* in all the coordinates except k where we subtract a and coordinate j where we add a. We first show that $z \in \Pi_\rho$. It's clear that $z_k = \theta^*_k - a = x^m_k$ and $z_j = \theta^*_j + a = x^m_j + a + b$. It is clear that $\sum_{i=1}^{d} z_i = 1$. As $z_k = x^m_k$ it is clear that $z_k \in [\rho, 1]$. If b is negative then $\theta^*_j = z_j - a < z_j < z_j - (b + a) = x^m_j$ then it is clear that $z_j \in [\rho, 1]$ (since both x^m_j, θ^*_j are in the valid range), and if b is positive then $x^m_j < x^m_j + a + b = z_j < z_j - a = \theta^*_j$ then it is clear that $z_j \in [\rho, 1]$. This gives us that $z \in \Pi_\rho$.

We now consider the difference in the norms

$$\begin{aligned}\sum_{i=0}^{d-1}(\theta_i^*-y_i)^2-\sum_{i=0}^{d-1}(z_i-y_i)^2 &= (\theta_j^*-y_j)^2+(\theta_k^*-y_k)^2-(\theta_j^*+a-y_j)^2-(\theta_k^*-a-y_k)^2\\ &= (y_j-\lambda_m+b-y_j)^2+(y_k-\lambda_m+a-y_k)^2\\ &\quad -(y_j-\lambda_m+b+a-y_j)^2-(y_k-\lambda_m+a-a-y_k)^2\\ &= (b-\lambda_m)^2+(a-\lambda_m)^2-(a+b-\lambda_m)^2-\lambda_m^2\\ &= -2ab>0\end{aligned}$$

Which is a contradiction to the optimality of θ^*. □
This gives us that among $x^0, x^1, \ldots, x^{d-1}$ is the optimal solution. Which means that in order to find the projection for the vector y, we need to calculate $x^0, x^1, \ldots, x^{d-1}$, remove the vectors who are not in Π_ρ, and take the one that minimizes $\|x^i - y\|_2$.

C.2 Definitions

Definition C.4 *For a matrix $A \in \mathbb{R}^{n\times m}$, the Moore Penrose inverse matrix is $A^\dagger$, which follows the following properties:*

$$1.\ AA^\dagger A = A \qquad 2.\ A^\dagger AA^\dagger = A^\dagger \tag{33}$$

$$3.\ (A^\dagger A)^\top = A^\dagger A \qquad 4.\ (AA^\dagger)^\top = AA^\dagger \tag{34}$$

$$5.\ A^\dagger b = \underset{x,\ Ax=b}{\arg\min}\|x\|_2 \tag{35}$$

Where 3, 4 only holds when the values of A are real.

C.3 General vector proofs

Lemma C.5 *for $x \in \mathbb{R}^n$ and $y = \arg\min_{z\in C}\|z-x\|_2^2$ assuming C is convex, we have:*

$$\max_{z\in C}(x-y)^\top(z-y)\le 0$$

Proof: Let there be $z \in C$, $t \in (0,1]$ and define $z_t := tz + (1-t)y$. As C is convex it is clear that $z_t \in C$. We observe:

$$\begin{aligned}\|x-y\|_2^2-\|x-z_t\|_2^2 &= \|x-y\|_2^2-\|(x-y)+(y-z_t)\|_2^2\\ &= \|x-y\|_2^2-\left[\|x-y\|_2^2+\|y-z_t\|_2^2+2(x-y)^\top(y-z_t)\right]\\ &= -\|y-z_t\|_2^2+2(x-y)^\top(z_t-y)\\ &= 2t(x-y)^\top(z-y)-t^2\|y-z\|_2^2\end{aligned}$$

Where the last equality follows by: $z_t - y = tz + (1-t)y - y = tz - ty = t(z-y)$. because $y = \arg\min_{z\in C}\|z-x\|_2^2$, we get: $\|x-y\|_2^2 - \|x-z_t\|_2^2 \le 0$ so:

$$2t(x-y)^\top(z-y)\le t^2\|y-z\|_2^2$$

Dividing by $2t$ gives:

$$(x-y)^\top(z-y)\le\frac{t}{2}\|y-z\|_2^2$$

As this is true for every $t \in (0,1]$, we get the the wanted inequality. □

Lemma C.6 *Let $f:\mathbb{R}^n\to\mathbb{R}$ be a β-smooth function with respect to norm $\|\cdot\|_2$, where $dom(f)$ is a convex set. then for every $x, y \in dom(f)$:*

$$f(y)\ge f(x)+\nabla f(x)^\top(y-x)-\frac{\beta}{2}\|y-x\|_2^2 \tag{36}$$

Proof: Fix $x, y \in dom(f)$. By the fundamental theorem of calculus,

$$f(y) - f(x) = \int_0^1 \nabla f(x + t(y-x))^\top (y-x) dt$$

Therefore:

$$f(y) - f(x) = \nabla f(x)^\top (y-x) + \int_0^1 (\nabla f(x + t(y-x)) - \nabla f(x))^\top (y-x) dt$$

Since $\nabla f(x)^\top (y-x)$ does not depend on t. Taking the absolute value of the above:

$$\begin{aligned}
|f(y) - f(x) - \nabla f(x)^\top (y-x)| =& |\int_0^1 (\nabla f(x + t(y-x)) - \nabla f(x))^\top (y-x) dt| \\
&\leq \int_0^1 |(\nabla f(x + t(y-x)) - \nabla f(x))^\top (y-x)| dt \\
&\leq \int_0^1 \|(\nabla f(x + t(y-x)) - \nabla f(x)\|_2 \|y-x\|_2 dt \\
&\leq \int_0^1 \beta \|x + t(y-x) - x\|_2 \|y-x\|_2 dt \\
&= \int_0^1 t\beta \|y-x\|_2^2 dt = \frac{\beta}{2} \|y-x\|_2^2
\end{aligned}$$

Where the third step holds due to the fact that for vectors x_1, x_2 we have $|{x_1}^\top x_2| \leq \|x_1\|_2 \|x_2\|_2$. Since $f(x) + \nabla f(x)^\top (y-x) - f(y) \leq |f(y) - f(x) - \nabla f(x)^\top (y-x)|$ and the above we get:

$$f(x) + \nabla f(x)^\top (y-x) - f(y) \leq \frac{\beta}{2} \|y-x\|_2^2$$

Reorganizing that finishes the proof. □

Lemma C.7 *For a matrix $M \in \mathbb{R}^{n \times m}$, a vector $x \in \mathbb{R}^n$, and a number $L > 0$, assume that $\sum_{i=1}^n \sum_{j=1}^m (M_{i,j})^2 \leq L$, then we get:*

$$\|Mx\|_2^2 \leq L \|x\|_2^2$$

Proof: Observe:

$$\begin{aligned}
\|Mx\|_2^2 &= \sum_{i=1}^m (M_i x)^2 \\
&\leq \sum_{i=1}^m \|M_i\|_2^2 \|x\|_2^2 \\
&= \|x\|_2^2 \sum_{i=1}^m \|M_i\|_2^2 \\
&= \|x\|_2^2 \sum_{i=1}^m \sum_{j=1}^n (M_{i,j})^2 \\
&\leq L \|x\|_2^2
\end{aligned}$$

Where M_i is the i-th column in the matrix M. The second step follows by Cauchy-Schwarz's inequality. The last step follows by the assumption $\sum_{i=1}^n \sum_{j=1}^m (M_{i,j})^2 \leq L$. □

C.4 Smoothness proofs

Lemma C.8 *For a policy $\pi \in \Delta(A)^{|S|}$ and a vector $u \in \mathbb{R}^{|S||A|}$ such that for all states $s \in S$, $\|u_{\cdot,s}\|_2 = 1$, we get:*

$$\left|u^\top \nabla^2 V^\pi(s_0) u\right| \leq \frac{1}{3} H^3 |A|$$

Where ∇^2 is the Hessian matrix, where for two state-action pairs: s, a and s', a', $\nabla^2 V^\pi(s_0)_{(s,a),(s',a')} = \frac{\partial \partial V^\pi(s_0)}{\partial \theta_{s,a} \partial \theta_{s',a'}}$.

Proof: Fix $\theta \in \Delta(A)^{|S|}$ and let π be the policy which is parameterized by θ. Let $u \in [0,1]^{|S||A|}$ be such that for every state $s \in S$, $\|u_{\cdot,s}\|_2 = 1$. For a scalar α let $\pi_\alpha = \pi + \alpha u$. For a state $s \in S$:

$$\sum_{a \in A} \left| \frac{d\pi_\alpha(a|s)}{d\alpha}\Big|_{\alpha=0} \right| = \sum_{a \in A} |u_{s,a}| \leq \sqrt{|A|}$$

and for all actions $a \in A$:

$$\frac{d^2 \pi_\alpha(a|s)}{(d\alpha)^2}\Big|_{\alpha=0} = 0$$

Define the state-action transition matrix under π as follows:

$$[\widetilde{P}(\alpha)]_{(s,a)->(s',a')} = \pi_\alpha(a'|s') P(s'|s,a).$$

It's clear to see that for all $n \geq 1$ the vector $\widetilde{P}(0)^n \vec{1}$ consists of only non-negative values. This implies that for an arbitrary vector x, a state-action pair s, a and a number $n \geq 1$,

$$\left| \left[\widetilde{P}(\alpha)^n \Big|_{\alpha=0} x \right]_{s,a} \right| \leq \|x\|_\infty \left| \left[\widetilde{P}(\alpha)^n \vec{1} \right]_{s,a} \right|, \tag{37}$$

For a state-action pair s, a,

$$\begin{aligned}
\widetilde{P}(\alpha)\Big|_{\alpha=0} \vec{1}_{s,a} &= \sum_{s' \in S, a' \in A} \pi(a'|s') P(s'|s,a) \\
&= \sum_{s' \in S} P(s'|s,a) \sum_{a' \in A} \pi(a'|s') \\
&= \sum_{s' \in S} P(s'|s,a) \\
&= 1
\end{aligned}$$

Combining that with (37) gives us that for a vector x, a state-action pair s, a and a number $n \geq 1$:

$$\left| \left[\widetilde{P}(\alpha)^n \Big|_{\alpha=0} x \right]_{s,a} \right| \leq \|x\|_\infty$$

Which can also be written as:

$$\left\| \widetilde{P}(\alpha)^n \Big|_{\alpha=0} x \right\|_\infty \leq \|x\|_\infty \tag{38}$$

Consider the first derivative of $\widetilde{P}(\alpha)$ at $\alpha = 0$. For an arbitrary vector x, and a state-action pair s, a:

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

We bound the absolute value of the above:

$$\begin{aligned}
\left|\left[\left.\frac{d\widetilde{P}(\alpha)}{d\alpha}\right|_{\alpha=0} x\right]_{s,a}\right| &= \left|\sum_{s'\in S,a'\in A} \left.\frac{d\pi_\alpha(a'|s')}{d\alpha}\right|_{\alpha=0} P(s'|s,a)x_{a',s'}\right| \\
&\leq \sum_{s'\in S,a'\in A} \left|\left.\frac{d\pi_\alpha(a'|s')}{d\alpha}\right|_{\alpha=0}\right| P(s'|s,a)|x_{a',s'}| \\
&\leq \sum_{s'\in S} P(s'|s,a)\|x\|_\infty \sum_{a\in A}\left|\frac{d\pi_\alpha(a'|s')}{d\alpha}\right| \\
&\leq \sum_{s'\in S} P(s'|s,a)\|x\|_\infty \sqrt{|A|} \\
&= \sqrt{|A|}\|x\|_\infty \sum_{s'\in S} P(s'|s,a) \\
&= \sqrt{|A|}\|x\|_\infty
\end{aligned}$$

Which can also be written as:

$$\left\|\left.\frac{d\widetilde{P}(\alpha)}{d\alpha}\right|_{\alpha=0} x\right\|_\infty \leq \sqrt{|A|}\|x\|_\infty \tag{39}$$

When differentiating $\widetilde{P}(\alpha)$ twice w.r.t. α at $\alpha = 0$, we get for an arbitrary vector x:

$$\left|\left[\left.\frac{d^2\widetilde{P}(\alpha)}{(d\alpha)^2}\right|_{\alpha=0} x\right]_{s,a}\right| = \left|\sum_{s'\in S,a'\in A} \left.\frac{d^2\pi_\alpha(a'|s')}{(d\alpha)^2}\right|_{\alpha=0} P(s'|s,a)x_{a',s'}\right| \tag{40}$$

$$\leq \sum_{s'\in S,a'\in A} \left|\left.\frac{d^2\pi_\alpha(a'|s')}{(d\alpha)^2}\right|_{\alpha=0}\right| P(s'|s,a)|x_{a',s'}| \tag{41}$$

$$\leq \sum_{s'\in S,a'\in A} \left|\left.\frac{d^2\pi_\alpha(a'|s')}{(d\alpha)^2}\right|_{\alpha=0}\right| P(s'|s,a)\|x\|_\infty \tag{42}$$

$$= \sum_{s'\in S} P(s'|s,a)\|x\|_\infty \sum_{a'\in A}\left|\left.\frac{d^2\pi_\alpha(a'|s')}{(d\alpha)^2}\right|_{\alpha=0}\right| \tag{43}$$

$$= 0 \tag{44}$$

For some action $a \in A$, let $Q^\alpha(s_0, a)$ be the corresponding Q-function for the policy π_α at state s_0 and action a. Observe that $Q^\alpha(s_0, a)$ can be written as:

$$Q^\alpha(s_0,a) = e_{s_0,a}^\top \sum_{n=0}^{H-1} \widetilde{P}(\alpha)^n r$$

Where r is the reward vector. Observe the absolute value of the first derivative of $Q^\alpha(s_0, a)$ w.r.t.

α at $\alpha = 0$:

$$\begin{aligned}
\left|\frac{dQ^\alpha(s_0,a)}{d\alpha}\Big|_{\alpha=0}\right| &= \left|e_{s_0,a}^\top \sum_{n=0}^{H-1}\sum_{k=0}^{n-1} \widetilde{P}(0)^k \frac{d\widetilde{P}(\alpha)}{d\alpha}\Big|_{\alpha=0} \widetilde{P}(0)^{n-1-k} r\right| \\
&\le \sum_{n=0}^{H-1}\sum_{k=0}^{n-1} \left|e_{s_0,a}^\top \widetilde{P}(0)^k \frac{d\widetilde{P}(\alpha)}{d\alpha}\Big|_{\alpha=0} \widetilde{P}(0)^{n-1-k} r\right| \\
&\le \sum_{n=0}^{H-1}\sum_{k=0}^{n-1} \sqrt{|A|}\|r\|_\infty \\
&\le \sum_{n=0}^{H-1}\sum_{k=0}^{n-1} \sqrt{|A|} \\
&= \frac{1}{2}H(H-1)\sqrt{|A|}
\end{aligned}$$

Where the third step follows by (39) and (38).

observe the absolute value of the second derivative of $Q^\alpha(s_0,a)$ w.r.t. α at $\alpha = 0$:

$$\begin{aligned}
\left|\frac{d^2Q^\alpha(s_0,a)}{(d\alpha)^2}\Big|_{\alpha=0}\right| &= \left|e_{s_0,a}^\top \sum_{n=0}^{H-1}\sum_{k=0}^{n-1} \left(\sum_{l=0}^{k-1} \widetilde{P}(0)^l \frac{d\widetilde{P}(\alpha)}{d\alpha}\Big|_{\alpha=0} \widetilde{P}(0)^{k-1-l} \frac{d\widetilde{P}(\alpha)}{d\alpha}\Big|_{\alpha=0} \widetilde{P}(0)^{n-1-k} r\right)\right. \\
&\quad + \left(\widetilde{P}(0)^k \frac{d^2\widetilde{P}(\alpha)}{(d\alpha)^2}\Big|_{\alpha=0} \widetilde{P}(0)^{n-1-k} r\right) \\
&\quad \left. + \left(\sum_{l=0}^{n-1-k-1} \widetilde{P}(0)^k \frac{d\widetilde{P}(\alpha)}{d\alpha}\Big|_{\alpha=0} \widetilde{P}(0)^l \frac{d\widetilde{P}(\alpha)}{d\alpha}\Big|_{\alpha=0} \widetilde{P}(0)^{n-1-k-1-l} r\right)\right| \\
&\le \sum_{n=0}^{H-1}\sum_{k=0}^{n-1} \left(\sum_{l=0}^{k-1} \left|e_{s_0,a}^\top \widetilde{P}(0)^l \frac{d\widetilde{P}(\alpha)}{d\alpha}\Big|_{\alpha=0} \widetilde{P}(0)^{k-1-l} \frac{d\widetilde{P}(\alpha)}{d\alpha}\Big|_{\alpha=0} \widetilde{P}(0)^{n-1-k} r\right|\right) \\
&\quad + \left(\left|e_{s_0,a}^\top \widetilde{P}(0)^k \frac{d^2\widetilde{P}(\alpha)}{(d\alpha)^2}\Big|_{\alpha=0} \widetilde{P}(0)^{n-1-k} r\right|\right) \\
&\quad + \left(\sum_{l=0}^{n-1-k-1} \left|e_{s_0,a}^\top \widetilde{P}(0)^k \frac{d\widetilde{P}(\alpha)}{d\alpha}\Big|_{\alpha=0} \widetilde{P}(0)^l \frac{d\widetilde{P}(\alpha)}{d\alpha}\Big|_{\alpha=0} \widetilde{P}(0)^{n-1-k-1-l} r\right|\right) \\
&\le \sum_{n=0}^{H-1}\sum_{k=0}^{n-1} \left(\sum_{l=0}^{k-1} |A|\|r\|_\infty\right) + \left(\sum_{l=0}^{n-1-k-1} |A|\|r\|_\infty\right) \\
&\le \frac{1}{3}|A|H(H-1)(H-2)
\end{aligned}$$

Where the third step follows by (39), (38) and (40).

Consider the identity:

$$\widetilde{V}(\alpha) = \sum_{a\in A} \pi_\alpha(a|s_0) Q^\alpha(s_0,a)$$

By differentiating $\widetilde{V}(\alpha)$ twice w.r.t. α we get:

$$\frac{d^2\widetilde{V}(\alpha)}{(d\alpha)^2} = \sum_{a\in A} \frac{d^2\pi_\alpha(a|s_0)}{(d\alpha)^2} Q^\alpha(s_0,a) + 2\sum_{a\in A} \frac{d\pi_\alpha(a|s_0)}{d\alpha}\frac{dQ^\alpha(s_0,a)}{d\alpha} + \sum_{a\in A} \pi_\alpha(a|s_0)\frac{d^2Q^\alpha(s_0,a)}{(d\alpha)^2}$$

Hence,

$$\begin{aligned}
\left|\frac{d^2\widetilde{V}(\alpha)}{(d\alpha)^2}\Big|_{\alpha=0}\right| &\leq \sum_{a\in A}\left|\frac{d^2\pi_\alpha(a|s_0)}{(d\alpha)^2}\Big|_{\alpha=0}\right| |Q^\pi(s_0,a)| \\
&\quad + 2\sum_{a\in A}\left|\frac{d\pi_\alpha(a|s_0)}{d\alpha}\Big|_{\alpha=0}\right|\left|\frac{dQ^\alpha(s_0,a)}{d\alpha}\Big|_{\alpha=0}\right| \\
&\quad + \sum_{a\in A}|\pi(a|s_0)|\left|\frac{d^2Q^\alpha(s_0,a)}{(d\alpha)^2}\Big|_{\alpha=0}\right| \\
&\leq 0 + 2\frac{1}{2}H(H-1)\sqrt{|A|}\sum_{a\in A}\left|\frac{d\pi_\alpha(a|s_0)}{d\alpha}\Big|_{\alpha=0}\right| + \frac{1}{3}|A|H(H-1)(H-2)\sum_{a\in A}|\pi(a|s_0)| \\
&= H(H-1)|A| + \frac{1}{3}|A|H(H-1)(H-2) \\
&= \frac{1}{3}H(H-1)(H+1)|A| \\
&\leq \frac{1}{3}H^3|A|
\end{aligned}$$

Observe the first derivative of $\widetilde{V}(\alpha)$ w.r.t. α:

$$\begin{aligned}
\frac{d\widetilde{V}(\alpha)}{d\alpha} &= \sum_{s\in S,a\in A}\frac{\partial V^{\pi_\alpha}}{\partial\theta_{s,a}}\frac{\partial\pi_\alpha}{\partial\theta_{s,a}} \\
&= \sum_{s\in S,a\in A}\frac{\partial V^{\pi_\alpha}(s_0)}{\partial\theta_{s,a}}u_{s,a}
\end{aligned}$$

Observe The second derivative of $\widetilde{V}(\alpha)$ w.r.t. α:

$$\begin{aligned}
\frac{d^2\widetilde{V}(\alpha)}{(d\alpha)} &= \frac{d}{d\alpha}\frac{d\widetilde{V}(\alpha)}{d\alpha} \\
&= \frac{d}{d\alpha}\sum_{s\in S,a\in A}\frac{\partial V^{\pi_\alpha}(s_0)}{\partial\theta_{s,a}}u_{s,a} \\
&= \sum_{s'\in S,a'\in A}\sum_{s\in S,a\in A}\frac{\partial\partial V^{\pi_\alpha}}{\partial\theta_{s,a}\partial\theta_{s',a'}}u_{s,a}u_{s',a'} \\
&= u^\top\nabla^2V^{\pi_\alpha}u
\end{aligned}$$

Combining that and the fact that $\left|\frac{d^2\widetilde{V}(\alpha)}{(d\alpha)^2}\Big|_{\alpha=0}\right| \leq \frac{1}{3}H^3|A|$ completes the proof.

□

Lemma C.9 *Let $A\in\mathbb{R}^{n\times n}$ be a symmetric matrix, and $L>0$, such that for all eigenvalues λ of A, $|\lambda|\leq L$. Then,*

$$\max_{\|x\|_2=1}\|Ax\|_2 \leq L$$

Proof: Fix a symmetric matrix $A\in\mathbb{R}^{n\times n}$. As A is symmetric, it has $|S||A|$ eigenvectors $v_1,v_2,\ldots,v_{|S||A|}$, which form an orthonormal basis. Namely, for all $i\neq j$, $v_i^\top v_j=0$, and for all vector x there exists $\alpha_1,\ldots,\alpha_{|S||A|}$ such that $x=\sum_{i=1}^{|S||A|}\alpha_iv_i$. Let $\lambda_1,\lambda_2,\ldots,\lambda_{|S||A|}$ be their corresponding eigenvalue, and assume that for all i, $|\lambda_i|\leq L$. Let $x^*=\arg\max_{\|x\|_2=1}\|Ax\|_2$, and let

$\alpha_1, \alpha_2, \ldots, \alpha_{|S||A|}$ be it's representation with the eigenvectors. Namely, $x^* = \sum_{i=1}^{|S||A|} \alpha_i v_i$. See that

$$\begin{aligned}
\max_{\|x\|_2=1} \|Ax\|_2 &= \|Ax^*\|_2 \\
&= \|A \sum_{i=1}^{|S||A|} \alpha_i v_i\|_2 \\
&= \| \sum_{i=1}^{|S||A|} \alpha_i A v_i\|_2 \\
&= \| \sum_{i=1}^{|S||A|} \alpha_i \lambda_i v_i\|_2 \\
&\leq \max_{i=1}^{|S||A|} |\lambda_i| \| \sum_{i=1}^{|S||A|} \alpha_i v_i\|_2 \\
&\leq L\|x^*\|_2 \\
&= L
\end{aligned}$$

Where the first transition follows by the definition of x^*, the fourth transition holds since v_i is an eigenvector, the sixth transition follows by the assumption that the largest absolute value of the eigenvalues is bounded by L, and the seventh transition follows since x^* was chosen with a constraint of $\|x^*\|_2 = 1$. □

Lemma C.10 *For a policy π,*

$$\max_{\|x\|_2=1} |\nabla_\theta V^\pi(s_0)x| \leq \frac{1}{3}|A|H^3$$

Proof: Fix a policy $\pi \in \Delta(A)^{|S|}$. Let λ be the eigenvalue of the hessian matrix $\nabla^2_\theta V^\pi(s_0)$ whose absolute value is the largest, and let v be it's corresponding eigenvector. Let $u = \frac{1}{\|v\|_2} v$,

$$\begin{aligned}
|u^\top \nabla_\theta V^\pi(s_0) u| &= \frac{1}{\|v\|_2^2} |v^\top \nabla_\theta V^\pi(s_0) v| \\
&= \frac{1}{\|v\|_2^2} |v^\top \lambda v| \\
&= |\lambda| \frac{1}{\|v\|_2^2} |v^\top v| \\
&= |\lambda| \frac{1}{\|v\|_2^2} \|v\|_2^2 \\
&= |\lambda|
\end{aligned}$$

Using Lemma C.8, we get that

$$|\lambda| \leq \frac{1}{3}|A|H^3$$

Using Lemma C.9, we get that:

$$\max_{\|x\|_2=1} |\nabla_\theta V^\pi(s_0)x| \leq \frac{1}{3}|A|H^3$$

□

Lemma C.11 *The function $V^\pi(s_0)$ is $\frac{1}{3}|A|H^3$-smooth with respect to norm $\|\cdot\|_2$. i.e, For all π,$\pi' \in \Delta(A)^{|S|}$*

$$\|\nabla_\theta V^\pi(s_0) - \nabla_\theta V^{\pi'}(s_0)\|_2 \leq \frac{1}{3}|A|H^3 \|\pi - \pi'\|_2 \tag{45}$$

Proof: Fix policies π, π'. For a parameter $t \in [0,1]$, define the function:

$$g(t) = \nabla_\theta V^{\pi + t(\pi' - \pi)}(s_0)$$

It's clear that

$$g'(t) = \nabla^2_\theta V^{\pi + t(\pi' - \pi)}(s_0)(\pi' - \pi)$$

Observe that $\|g(1) - g(0)\|_2 = \left\|\nabla_\theta V^{\pi'}(s_0) - \nabla_\theta V^{\pi}(s_0)\right\|_2$. Using the fundamental theorem of calculus, we get that:

$$\begin{aligned}
\left\|\nabla_\theta V^{\pi'}(s_0) - \nabla_\theta V^{\pi}(s_0)\right\|_2 &= \|g(1) - g(0)\|_2 \\
&= \| \int_0^1 g'(t) dt\|_2 \\
&= \| \int_0^1 \nabla^2_\theta V^{\pi + t(\pi' - \pi)}(s_0)(\pi' - \pi) dt\|_2 \\
&\leq \int_0^1 \|\nabla^2_\theta V^{\pi + t(\pi' - \pi)}(s_0)(\pi' - \pi)\|_2 dt \\
&= \int_0^1 \left\|\nabla^2_\theta V^{\pi + t(\pi' - \pi)}(s_0) \frac{\|(\pi' - \pi)\|_2}{\|(\pi' - \pi)\|_2}(\pi' - \pi)\right\|_2 dt \\
&= \int_0^1 \left\|\nabla^2_\theta V^{\pi + t(\pi' - \pi)}(s_0) \frac{1}{\|(\pi' - \pi)\|_2}(\pi' - \pi)\right\|_2 \|(\pi' - \pi)\|_2 \, dt \\
&\leq \int_0^1 \max_{\|x\|_2 = 1} \left\|\nabla^2_\theta V^{\pi + t(\pi' - \pi)}(s_0) x\right\|_2 \|(\pi' - \pi)\|_2 \, dt \\
&\leq \int_0^1 \frac{1}{3}|A| H^3 \|(\pi' - \pi)\|_2 dt \\
&= \frac{1}{3}|A| H^3 \|(\pi' - \pi)\|_2 \int_0^1 1 dt \\
&= \frac{1}{3}|A| H^3 \|(\pi' - \pi)\|_2
\end{aligned}$$

Where the eighth step follows by Lemma C.10. □

C.5 Kullback–Leibler divergence

Definition C.12 *For two non-zero distributions $x, y \in \mathbb{R}^n$, where for all i, $x_i > 0$, $y_i > 0$ and $\sum_{i=1}^n x_i = 1$, $\sum_{i=1}^n y_i = 1$, define the Kullback–Leibler divergence $KL(x, y)$ as:*

$$KL(x, y) = \sum_{i=1}^n x_i \log(\frac{x_i}{y_i}) \tag{46}$$

Lemma C.13 *For all non-zero distributions $x, y \in \mathbb{R}^n$*

$$KL(x, y) \geq 0$$

Proof: Fix all non-zero distributions x, y

$$\begin{aligned}
KL(x,y) &= \sum_{i=1}^{n} x_i \log(\frac{x_i}{y_i}) \\
&= -\sum_{i=1}^{n} x_i \log(\frac{x_i}{y_i}) \\
&\geq -\sum_{i=1}^{n} x_i \left(\frac{y_i}{x_i} - 1\right) \\
&= -\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} x_i \\
&= -1 + 1 \\
&= 0
\end{aligned}$$

Where the third transition holds since for all c, $\log(c) \leq c - 1$, and the fourth transition holds since x, y are distributions. □

Lemma C.14 *For a non-zero distribution $x \in \mathbb{R}^n$, let $y = \frac{1}{n}\vec{1}$, then:*

$$KL(x,y) \leq \log(n)$$

Proof: Fix a non-zero distribution $x \in \mathbb{R}^n$, let $y = \frac{1}{n}\vec{1}$, then:

$$\begin{aligned}
KL(x,y) &= \sum_{i=1}^{n} x_i \log(\frac{x_i}{y_i}) \\
&= \sum_{i=1}^{n} x_i \log(nx_i) \\
&= \sum_{i=1}^{n} x_i \log(nx_i) \\
&= \log(n) \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} x_i \log(x_i) \\
&= \log(n) + \sum_{i=1}^{n} x_i \log(x_i) \\
&\leq \log(n)
\end{aligned}$$

□

C.6 Concentration bounds

Theorem C.15 *Hoeffding Theorem - Let $X_0, X_1, \ldots, X_{d-1}$ be d independent random variables such that for every $i \in [d]$, $a \leq X_i \leq b$. Let $\bar{X} = \frac{1}{d}\sum_{i=0}^{d-1} X_i$. Then with probability $1 - \delta$ we get that $|\mathbb{E}[\bar{X}] - \bar{X}| \leq \epsilon$ assuming*

$$d \geq \frac{(b-a)^2}{2\epsilon^2} \log(\frac{2}{\delta})$$

Theorem C.16 *Generalized Chernoff bound - Let $X_0, X_1 \ldots, X_{d-1}$ be independent random variables with $X_i \in \{0, 1\}$ and $Pr[X_i = 1] = p$, for $i = 0, \ldots, d-1$. Set $X := \sum_{i=1}^{n-1} X_i$ and $\mu = pd$. Then, for any $\delta \in (0, 1)$, we have*

$$Pr[X \leq (1-\delta)\mu] \leq e^{-\delta^2\mu/2}$$