Finite Sample Analysis of Two-Time-Scale Natural Actor-Critic Algorithm

Sajad Khodadadian, Thinh T. Doan, Justin Romberg, Siva Theja Maguluri

***Abstract*— Actor-critic style two-time-scale algorithms are one of the most popular methods in reinforcement learning, and have seen great empirical success. However, their performance is not completely understood theoretically. In this paper, we characterize the *global* convergence of an online natural actor-critic algorithm in the tabular setting using a single trajectory of samples. Our analysis applies to very general settings, as we only assume ergodicity of the underlying Markov decision process. In order to ensure enough exploration, we employ an ϵ-greedy sampling of the trajectory.**

For a fixed and small enough exploration parameter ϵ, we show that the two-time-scale natural actor-critic algorithm has a rate of convergence of $\tilde{\mathcal{O}}(1/\mathsf{T}^{1/4})$, where T is the number of samples, and this leads to a sample complexity of $\tilde{\mathcal{O}}(1/\delta^8)$ samples to find a policy that is within an error of δ from the *global optimum*. Moreover, by carefully decreasing the exploration parameter ϵ as the iterations proceed, we present an improved sample complexity of $\tilde{\mathcal{O}}(1/\delta^6)$ for convergence to the global optimum.

***Index Terms*— Actor-Critic, Machine Learning, Reinforcement Learning, Two-Time-Scale**

I. Introduction

In reinforcement learning (RL), an agent operating in an environment, modeled as a Markov decision process (MDP), tries to learn a policy that maximizes its long-term reward. Methods for solving this optimization problem include value function methods, such as Q-learning [1], and policy space methods, such as TRPO [2], PPO [3], and actor-critic [4].

Policy space methods explicitly search for the maximum of the value function V^π, which codifies the expected long-term reward, through iterative optimization over the policy π. Although in general V^π is a nonconvex function of π [5], global optimality can be obtained by employing either gradient descent [5], [6], mirror descent [7], [8], or natural gradient descent [5]. These methods, however, assume access to an oracle that returns the gradient of the value function for any given policy. In many practical scenarios, and in particular when the parameters of the MDP are only partially known, these gradients have to be estimated from observations or simulations.

Actor-critic (AC) techniques integrate estimation of the gradient into the policy search. In this framework, a critic estimates the value (Q-function) of a policy, usually through a temporal difference iteration. The actor then uses this estimate to form a gradient to improve the policy. AC algorithms have been observed to converge quickly relative to other methods [9], [10], and have enjoyed success in several applications including robotics [11], computer games [12], and power networks [13].

AC algorithms can be classified into batch vs. online. In the batch setting, in each iteration of the AC, the critic evaluates the policy using a set of collected data. This type of batch update, however, cannot be implemented in an online manner, and requires simulations that need to be restarted in specific states, making its implementation appropriate in artificial environments such as Atari games [3], but not in scenarios that require the agent to "learn as they go".

A truly online and two-time-scale AC variant was first proposed in [4], where at every iteration the actor and critic updates depend only on one sample observed from the environment using the current policy. Later [14], [15] presented a version of this algorithm using a natural policy gradient. These methods can be viewed as two-time-scale stochastic approximation (SA) algorithms, where the actor and the critic operate at the "slow" and "fast" time scales, respectively.

AC algorithms often use low-order approximations for the value and policy functions. While this type of function approximation can dramatically simplify the learning process, thus allowing us to apply the algorithm to complex, real-world problems, these approximations introduce non-vanishing, systematic errors, as the truly optimal policy typically does not lie in the set of functions considered. In this paper, however, we are completely focussed on recovering the globally optimal policy, and so we will operate in the "tabular setting" that considers every possible distribution over the (finite number of) actions for every possible state.

Main contributions:

- We analyze the two-time-scale natural AC algorithm in the tabular setting. Our setting can be seen as a two-time-scale linear stochastic approximation with a time-varying Markovian noise. Unlike several recent papers, we do not make an extensive set of assumptions. The *only* assumption we make is the ergodicity of the underlying Markov chain under all policies.

Sajad Khodadadian and Siva Theja Maguluri are with H. Milton Stewart School of Industrial & Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia, 30332, (email: {skhodadadian, siva.theja}@gatech.edu)

Thinh T. Doan is with Bradley Department of Electrical and Computer Department, Virginia Tech, Perry St, Blacksburg, Virginia, 24060 (email: thinhdoan@vt.edu)

Justin Romberg is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, Georgia, 30332, (email: jrom@ece.gatech.edu)

- Our analysis show the importance of the exploration in AC type algorithms. We argue that a naive execution of natural AC fails to properly explore all of the state-action pairs, a fact that we illustrate with a simple example. Therefore, we employ ϵ-greedy exploration to guarantee global convergence.
- For a fixed and small enough exploration parameter ϵ, we show that using T number of samples, two-time-scale natural AC Algorithm 1 converges to within $\tilde{\mathcal{O}}\left(\frac{1}{\epsilon T^{1/4}}+\epsilon\right)$ of the global optimum. We show that for carefully chosen ϵ, Algorithm 1 finds a policy within a δ-ball around the global optimum using $\tilde{\mathcal{O}}(1/\delta^8)$ samples.
- We show that using a time-varying ϵ improves the sample complexity of Algorithm 1 to $\tilde{\mathcal{O}}(1/\delta^6)$.

A. Related works

Stochastic approximation: This method was first introduced in [16]. Asymptotic convergence of stochastic approximation (SA) was studied in [17], [18]. Recently, there has been a flurry of work on the finite time analysis of SA for both linear [19] and nonlinear [20], [21] operators, under both i.i.d [22] and Markovian [23] noise, with batch [24] or two-time-scale [17] updates. Our setting in this paper can be categorized as a linear two-time-scale SA with the noise generated from a time-varying Markov chain.

Actor-critic: AC algorithm was first proposed in [4] as a two-time-scale stochastic approximation [25]–[28] variant of the policy gradient algorithm [29]. In this algorithm a faster time scale is used to collect samples for gradient estimation, and a slower time scale is used to perform an update to the policy. In this paper we are interested in such a two-time-scale version of the natural policy gradient [30]. While natural gradient descent is closely related to mirror descent [31], [32], in the context of Markov decision processes they are known to be identical [5], [7]. Even though the objective V^π is a nonconvex function of the policy π, convergence rate of natural policy gradient to the global optimum under the planning setting (when the exact gradients are known) is recently established in [5], [7], [33], [34]. Natural AC, which is a variant of AC with natural policy gradient in the actor, was studied in [14], [15], [35], [36].

While the asymptotic convergence of AC methods including natural AC is well-understood [4], [15], [17], [37], [38], their finite-time convergence was largely unknown until recently [24], [39]–[52]. The authors in [42]–[44], [52] provide the convergence rate of AC where the parameter of the critic is updated using a number of collected samples instead of only one single sample. Such a setting, referred to as batch AC, cannot be implemented in an online fashion since at any iteration the critic has to implement the current policy for a number of time steps to collect enough data. A similar batch approach was used in [24], [40], [46]–[49], [51] to study natural AC and in [39], [45] to study the TRPO algorithm, which is a variant of mirror descent. The authors in [24], [51] study the finite time convergence bound of off-policy natural AC algorithm under constant step size. However, due to the constant step size they do not have convergence to the global optimum. [40], [53] study the finite time convergence of a regularized variant of natural AC with batch data update. In [50] the convergence of two-time-scale AC is analyzed. However, in [50] the convergence only to the local optimal is established. In a concurrent work with us, [54] studies the converging point of single trajectory AC for linear MDPs.

In this paper, we study the original AC method [4] without considering a batch update. In other words, data is collected through a single trajectory of a time-varying Markov chain and the update is performed in a two-time-scale manner. To the best of our knowledge, the only paper in the literature that considers such a setting is [50] which studies the AC algorithm under function approximation. Although their results are remarkable, they make several assumptions on the space of approximation functions. In Section II-A we will explain why these assumptions cannot be satisfied in the tabular setting with zero approximation error. Another related work is [48] where the authors claim to have a single trajectory algorithm. However, as explained in [24, Appendix C], the proposed algorithm in [48] is not single trajectory.

ϵ-greedy: One of the differences between our algorithm and the previous work is the inclusion of ϵ-greedy to the natural AC. This greedy step ensures sufficient exploration of our algorithm, while keeping the algorithm online. ϵ-greedy [55] is commonly employed in various settings such as Q-learning [56], multi-armed bandit [57], [58] and contextual bandit [59]. In these algorithms, ϵ-greedy is usually employed in order to ensure sufficient exploration [55]. In this paper, we show that this greedy step can ensure the global convergence of the AC as well, and we characterize the rate of this convergence.

To summarize, the work in the literature over the last two decades has looked at various challenges thrown by AC algorithms under various assumptions and different simplifying models. This paper studies the greedy version of this algorithm and consequently in one place addresses several analytical challenges which include: (i) two-time-scale analysis, (ii) an online or single trajectory update, (iii) Markovian data samples, (iv) time-varying Markov chain, (v) asynchronous update in tabular setting, (vi) diminishing step sizes, and (vii) global convergence with minimal assumptions.

II. Natural Actor-Critic Methods for Reinforcement Learning

The environment of our RL problem is modeled by a Markov Decision Process (MDP) specified by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are finite sets of states and actions, $\mathcal{P}$ is the set of transition probability matrices, $\gamma \in (0,1)$ is the discount factor, and $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the reward function, where without loss of generality we assume that the rewards are in $[0,1]$. We focus on randomized stationary policies, where each policy π assigns to each state $s \in \mathcal{S}$ a probability distribution $\pi(\cdot \,|\, s)$ over $\mathcal{A}$. Each policy π on the MDP, induces a Markov chain with transition probability $P^\pi(s'|s) = \sum_a \mathcal{P}(s'|s,a)\pi(a|s)$ on the states. Assuming that this Markov chain is irreducible, it induces a stationary distribution over states, which we denote by μ^π. By definition, this distribution satisfies $(\mu^\pi)^T P^\pi = (\mu^\pi)^T$ [60].

For a fixed policy π, a sample trajectory of the states and actions is generated according to $S_{k+1} \sim \mathcal{P}(\cdot|S_k, A_k), A_{k+1} \sim \pi(\cdot|S_{k+1})$. The value function associated with π and the state s is defined as the expected discounted cumulative reward, i.e.$V^\pi(s) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k \mathcal{R}(S_k, A_k) \,|\, S_0 = s, A_k \sim \pi(\cdot|S_k)\right]$. Furthermore, given an initial state distribution P over $\mathcal{S}$, we denote the expected cumulative reward for a policy π as $V^\pi(\mathrm{P})$. The goal is to find a policy that maximizes this expected cumulative reward:

$$\pi^* \in \arg\max_{\pi} V^\pi(\mathrm{P}). \tag{1}$$

It can be shown [61] that the optimal policy π^* is independent of the initial distribution P, and hence throughout this paper we assume P as fixed and we denote $V^\pi \equiv V^\pi(\mathrm{P})$. Furthermore, we denote V^{π^*} as V^*. It can be shown that value function can be written as $V^\pi \equiv V^\pi(\mathrm{P}) = \sum_{s,a} d_{\mathrm{P}}^\pi(s)\pi(a|s)\mathcal{R}(s,a)$, where d_{P}^π, denoted as the discounted state visitation [29], is defined as $d_{\mathrm{P}}^\pi(s) = (1-\gamma)\sum_{k=0}^{\infty}\gamma^k P^\pi(S_k = s \,|\, s_0 \sim \mathrm{P})$, with $P^\pi(S_k = s \,|\, S_0 \sim \mathrm{P})$ being the probability that $S_k = s$ after executing policy π starting from the initial distribution P at $k = 0$. Throughout, we denote $d_{\mathrm{P}}^{\pi^*}$ as d^*.

Given policy π_t, the Natural Policy Gradient (NPG) algorithm [5], [30] under the softmax parametrization updates the policy in every time step according to

$$\pi_{t+1}(a|s) = \frac{\pi_t(a|s)\exp(\beta_t Q^{\pi_t}(s,a))}{\sum_{a'}\pi_t(a'|s)\exp(\beta_t Q^{\pi_t}(s,a'))}, \forall s, a, \tag{2}$$

where $Q^\pi(s,a) = \mathbb{E}[\sum_{k=0}^{\infty}\gamma^k\mathcal{R}(S_k,A_k) \,|\, S_0 = s, A_0 = a, A_k \sim \pi(\cdot|S_k)]$ is the Q-function corresponding to the policy π. Here β_t is the step size which might be time dependent.

The update rule in (2) has multiple interpretations [62]. Firstly, as explained in [7], it can be seen as the update of the mirror descent for problem in (1) using negative entropy as the divergence generating function. Secondly, the NPG update in (2) can be seen as a pre-conditioned gradient ascent with softmax parameterization, where the pseudoinverse of the Fisher information matrix [63] multiplies the gradient as a preconditioner [30]. While mirror descent and natural gradient descent are distinct but related algorithms in general [7], [31], [32], they are both identical to (2) in our setting of solving the problem in (1) using softmax policy parametrization.

In the setting above, the NPG method finds a globally optimal policy with a provable rate; [5] shows that after T iterations of the update (2) with constant step size $\beta_t = \beta$, it finds a policy whose expected cumulative reward is within $\mathcal{O}(1/T)$ of the optimal policy. The convergence bound in [5] is for the "MDP setting", where the Q-function is computed exactly for every candidate policy π_t. In the vast majority of reinforcement learning applications, however, Q^{π_t} has to be estimated from simulations or observations.

A. Two-Time-Scale Natural Actor-Critic Algorithm

In order to perform the NPG update (2) for an unknown environment, one can first estimate Q^{π_t} using a batch of samples of state-action-rewards. However, using batch of data for the update of the Q-function has practical drawbacks. In particular, sampling of the batch data requires the state of the system to be reset frequently, which is not possible in environments such as robotics. A truly online, completely data-driven technique that keeps a running estimate of the Q-functions while performing NPG updates based on these estimates is presented in Algorithm 1 with $\epsilon_t = 0$. In this algorithm, the "critic" implements an asynchronous update to the Q-function, where the only entry in the table that is changed at every iteration is the one corresponding to the observed state-action pair (S_t, A_t). After this, the "actor" uses the estimated Q-table to update the policy using a natural policy gradient update of the form in (2). The critic and the actor use different step sizes (α_t and β_t, respectively), a fact that helps maintain the algorithm's stability.

Due to the existence of two different step sizes, Algorithm 1 can be viewed as a variant of two-time-scale stochastic approximation [25]. Intuitively, the critic has to collect information about the gradient at a faster time scale than the time scale at which the actor executes the gradient update — in other types of policy gradient algorithms, this takes the form of multiple samples being generated in an inner loop. Since the AC method performs both updates from a single sample, we can achieve a similar effect by having the actor take a more conservative step.

One of the main differentiators of our work with the existing literature on the convergence of AC algorithms is the update of the policy that mixes in a small multiple of the uniform distribution. This mixing is necessary to ensure sufficient exploration of the state-action space. In this algorithm, at each iteration t, the action A_{t+1} is sampled from the policy $\hat{\pi}_t$, which is a convex combination of the policy π_t and the uniform distribution. This strategy ensures that the sampling policy $\hat{\pi}_t$ attains at least ϵ_t weight in all it's elements, even though some elements of π_t might be arbitrary small. Furthermore, with introducing this step, we can ignore the technical assumptions which is made by the previous works. In Section IV, we give an example of an MDP with 4 states and 2 actions where a naive implementation of AC without this ϵ-greedy exploration step results in a suboptimal policy.

In the existing literature, this exploration is ensured through more stringent conditions on the problem structure, which *if* satisfied, can guarantee enough *exploration* by the AC algorithm (Assumption 1 in [48] and [47], Assumption 4.2 in [49], Assumption 4.3 in [64], and Assumption 4.1 in [50]). These assumptions, however, need not necessarily be satisfied in the tabular MDP setting. In particular, all these assumptions require $\pi_t(a|s)$ to be bounded away from zero for all states and actions s, a uniformly over time. However, we know that in an MDP, there always exist a deterministic policy which is a global optimal. This means that $\pi_t(a|s)$ can very likely go to zero for some sate-action pair s, a, and the assumption can be violated.

III. Main Result: Finite Time Convergence Bounds of Greedy Natural Actor-Critic

In this section, we provide a finite-time performance guarantee for Algorithm 1. In this algorithm, we can either choose

Algorithm 1: Two-time-scale natural AC algorithm with ϵ-greedy exploration

Input: Iteration number $T > 0$, step sizes α_t, β_t, exploration parameter ϵ_t, $Q_0(s,a) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $\pi_0(a|s) = \hat{\pi}_0(a|s) = \frac{1}{|\mathcal{A}|}, \forall s, a$.

Draw S_0 from some initial distribution and $A_0 \sim \pi_0(\cdot|s_0)$

for t=0,1,2,...,T **do**

Sample $S_{t+1} \sim \mathcal{P}(\cdot|S_t, A_t), A_{t+1} \sim \hat{\pi}_t(\cdot|S_{t+1})$

$\alpha_t(s,a) = \alpha_t \mathbb{1}\{S_t = s, A_t = a\}, \forall s, a$

$Q_{t+1}(s,a) = Q_t(s,a) + \alpha_t(s,a)\big[\mathcal{R}(S_t, A_t) + \gamma Q_t(S_{t+1}, A_{t+1}) - Q_t(S_t, A_t)\big], \forall s, a$

$\pi_{t+1}(a|s) = \pi_t(a|s)\frac{\exp(\beta_t Q_{t+1}(s,a))}{\sum_{a'} \pi_t(a'|s)\exp(\beta_t Q_{t+1}(s,a'))}, \forall s, a$

$\hat{\pi}_{t+1} = \frac{\epsilon_t}{|\mathcal{A}|} + (1-\epsilon_t)\pi_{t+1}$

end for

Sample $\hat{T}$ from $\{0, 1, \ldots, T\}$ by distribution $P(\hat{T} = i) = \frac{\beta_i}{\sum_{j=0}^{T} \beta_j}$

Output: $\hat{\pi}_{\hat{T}}$

a constant ϵ-greedy parameter, or a time-varying one. The advantage of the constant ϵ is faster rate of convergence to a neighborhood of the optimal, and the advantage of the time-varying greedy parameter is global convergence without any necessary pre-specified error.

In order to characterize our convergence results, first we make the following assumption.

Assumption 1: For every deterministic policy π, the Markov chain induced by the transition probability P^π is ergodic.

For more explanation regarding this assumption, look at Section IV.

We now present the main result of the paper. We bound the deviation of the value of the policy returned by 1 from the value of the optimal policy.

Theorem 1: Suppose Assumption 1 holds. Consider Algorithm 1 under the following step size parameters

$$\alpha_t = \frac{\alpha}{(t+1)^\nu}, \quad \beta_t = \frac{\beta}{(t+1)^\sigma}, \quad \epsilon_t = \frac{\epsilon}{(t+1)^\xi}, \tag{3}$$

with $0 \leq \xi < \nu < \sigma < 1,\ \alpha, \epsilon \leq 1$. Then,

$$\mathbb{E}[V^* - V^{\hat{\pi}_{\hat{T}}}] \leq \mathcal{O}(T^{\sigma-1}) + \begin{cases} \tilde{\mathcal{O}}(T^{-\sigma}) & \text{if} \quad 1 > 2\sigma \\ \tilde{\mathcal{O}}(T^{\sigma-1}) & \text{o.w} \end{cases}$$
$$+ \begin{cases} \tilde{\mathcal{O}}(\epsilon T^{\sigma-1}) & \text{if} \quad \xi + \sigma > 1 \\ \tilde{\mathcal{O}}(\epsilon T^{-\xi}) & \text{o.w} \end{cases}$$
$$+ \begin{cases} \tilde{\mathcal{O}}(T^{0.5(\nu+\xi-1)}/\epsilon^{0.5}) & \text{if} \quad \nu + \xi + 1 > 2\sigma, \\ \tilde{\mathcal{O}}(T^{\sigma-1}/\epsilon^{0.5}) & \text{o.w.} \end{cases}$$
$$+ \begin{cases} \tilde{\mathcal{O}}(t^{0.5(\xi-\nu)}/\epsilon^{0.5}) & \text{if} \quad 2 + \xi > \nu + 2\sigma, \\ \tilde{\mathcal{O}}(T^{\sigma-1}/\epsilon^{0.5}) & \text{o.w.} \end{cases}$$
$$+ \begin{cases} \tilde{\mathcal{O}}(t^{-0.5}) & \text{if} \quad 1 > 2\sigma, \\ \tilde{\mathcal{O}}(T^{\sigma-1}) & \text{o.w.} \end{cases}$$
$$+ \begin{cases} \tilde{\mathcal{O}}(t^{0.5(\xi+\nu-2\sigma)}/\epsilon^{0.5}) & \text{if} \quad 2 + \xi + \nu > 4\sigma, \\ \tilde{\mathcal{O}}(T^{\sigma-1}/\epsilon^{0.5}) & \text{o.w,} \end{cases}$$
$$+ \begin{cases} \tilde{\mathcal{O}}(t^{\xi+\nu-\sigma}/\epsilon) & \text{if} \quad 2 + 2\nu + 2\xi > 4\sigma, \\ \tilde{\mathcal{O}}(T^{\sigma-1}/\epsilon) & \text{o.w.} \end{cases}$$

where $\tilde{\mathcal{O}}(\cdot)$ ignores the $\log(T)$ terms.

The proof of Theorem 1 is provided in Section V.

Note that while V^* is not random, but $V^{\hat{\pi}_T}$ is, since the policy $\hat{\pi}_T$ is a function of all the random variables drawn in Algorithm 1.

Furthermore, we state two corollaries of Theorem 1 for different choices of ξ.

Corollary 1.1: Suppose Assumption 1 holds. Consider Algorithm 1 under the parameters in (3). Suppose $\xi = 0$, $\nu = 0.5$, and $\sigma = 0.75$. We have:

$$\mathbb{E}[V^* - V^{\hat{\pi}_{\hat{T}}}] \leq \tilde{\mathcal{O}}\left(\frac{1}{\epsilon T^{1/4}} + \epsilon\right). \tag{4}$$

Hence, the algorithm requires $\tilde{\mathcal{O}}(1/\delta^4)$ number of samples to get $\epsilon + \delta/\epsilon$ close to the global optimum. Furthermore, by taking $\epsilon = \mathcal{O}(\delta)$, we get $\mathbb{E}[V^* - V^{\hat{\pi}_{\hat{T}}}] \leq \tilde{\mathcal{O}}(\delta)$ after $T = \tilde{\mathcal{O}}(1/\delta^8)$ iteration of Algorithm 1.

The sample complexity in Corollary 1.1 is relatively poor due to the $\frac{1}{\epsilon}$ term on the upper bound in (4), which is due to a constant exploration factor in AC. In the next corollary we show how to achieve a better sample complexity by gradually reducing the exploration factor ϵ_t.

Corollary 1.2: Suppose Assumption 1 holds. Consider Algorithm 1 under the parameters in (3). Suppose $\xi = 1/6$, $\nu = 0.5$, and $\sigma = 5/6$. We have:

$$\mathbb{E}[V^* - V^{\hat{\pi}_{\hat{T}}}] \leq \tilde{\mathcal{O}}(1/T^{1/6}).$$

In particular, we have $\mathbb{E}[V^* - V^{\hat{\pi}_{\hat{T}}}] \leq \delta$ after $T = \tilde{\mathcal{O}}(1/\delta^6)$ iterations of Algorithm 1.

Corollaries 1.1 and 1.2 are direct application of Theorem 1. In particular, in the case of $\xi = 0$, the term $\epsilon T^{-\xi}$ in the bound of Theorem 1 will be a constant proportional to ϵ, and the best rate of convergence is obtained by picking $\nu = 0.5$ and $\sigma = 0.75$ which gives Corollary 1.1. Also assuming $\xi > 0$, the best rate of convergence can be obtained by $\xi = 1/6, \nu = 0.5, \sigma = 5/6$.

We should emphasize that Corollaries 1.1 and 1.2 characterize the sample complexity for global convergence of Algorithm 1 with the only assumption of ergodicity of the underlying MDP. This is indeed a much weaker assumption compared to the related work.

IV. Need for Exploration and Ergodicity

In this section we explain the necessity of the ϵ-greedy step in Algorithm 1 and the ergodicity Assumption 1. In iteration t of the natural AC algorithm, the objective of the critic is to estimate the Q-function corresponding to the policy π_t. In two-time-scale natural AC, in each iteration t the algorithm estimates the Q-function by updating only a single random element $(s = S_t, a = A_t)$ of the Q_t table. In our analysis, the ϵ-greedy step ensures that in each iteration of the algorithm, each of the actions are being sampled with probability at least ϵ_t. Furthermore, Assumption 1 ensures that all the states of the MDP are visited infinitely often. In the following we show

This article has been accepted for publication in IEEE Transactions on Automatic Control. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAC.2022.3190032

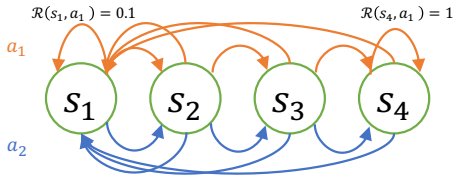

Fig. 1: A 4 states, and 2 actions MDP. Orange and blue correspond to the non zero transition probabilities of actions a_1 and a_2 respectively.

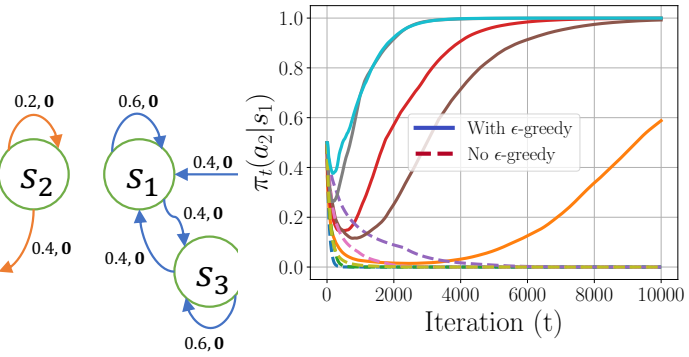

Fig. 2: $\pi_t(a_2|s_1)$ for 10 trajectories generated by the natural AC algorithm over the MDP in Fig. 1. Straight and dashed lines show the result with and without ϵ-greedy, respectively. Since here $\pi^*(a_2|s_1) = 1$, it shows that the algorithm without ϵ-greedy converges to a suboptimal policy.

why both ϵ-greedy and Assumption 1 are essential for the convergence of the natural AC algorithm.

(I) ϵ-greedy: Following the update of the policy in Algorithm 1, we have

$$\pi_t(a|s) = \frac{\exp(\sum_{l=0}^{t-1} \beta_l Q_{l+1}(s,a))}{\sum_{a'} \exp(\sum_{l=0}^{t-1} \beta_l Q_{l+1}(s,a'))}. \tag{5}$$

If for some state s, action $\tilde{a}$ satisfies $Q_k(s,\tilde{a}) \ll Q_k(s,a)$, $\forall a \neq \tilde{a}$, by (5), $\pi_t(\tilde{a}|s)$ converges to zero geometrically. Thus, with high probability $(s,\tilde{a})$ will not be explored, and we might converge to a suboptimal policy. Note that the scenario explained here can very likely happen when $\mathcal{R}(s,\tilde{a})$ is negligible with respect to $\mathcal{R}(s,a)$ for other actions a.

The following experiment illustrates the necessity of the ϵ-greedy policy update. Consider the MDP depicted in Fig. 1. This MDP has 4 states and 2 actions. All the transition probabilities depicted in the figure are positive, and the rest are zero. Furthermore, $\mathcal{R}(s_1,a_1) = 0.1$ and $\mathcal{R}(s_4,a_1) = 1$, and the rest of the rewards are zero. Suppose $\mathcal{P}(s_1|s_i,a_1) = 0.999$, $i = 1,2,3$, $\mathcal{P}(s_{i+1}|s_i,a_1) = 0.001$, $i = 1,2,3$, $\mathcal{P}(s_4|s_4,a_1) = 0.999$, $\mathcal{P}(s_1|s_4,a_1) = 0.001$, $\mathcal{P}(s_1|s_i,a_2) = 0.001$, $i = 2,3$, $\mathcal{P}(s_{i+1}|s_i,a_2) = 0.999$, $i = 2,3$, $\mathcal{P}(s_2|s_1,a_2) = 1$, $\mathcal{P}(s_4|s_4,a_2) = 0.001$, $\mathcal{P}(s_1|s_4,a_2) = 0.999$. In this MDP, the optimal policy in state s_1 is to play action a_2. Fig. 2 shows $\pi_t(a_2|s_1)$ for 10 trajectories achieved by the natural AC. The straight lines show the output of the algorithm when ϵ-greedy is employed, and the dashed lines are the output without ϵ-greedy. It is clear that almost always the trajectories of the algorithm without ϵ-greedy converge to a suboptimal policy.

(II) Ergodicity Assumption: This assumption implies that under all policies, the induced Markov chain over the states of the MDP is irreducible and aperiodic. We discuss these two assumptions separately in the next two paragraphs.

First, for an example of an MDP which does not satisfy irreducibility assumption, consider any episodic MDP, where there is a terminal state in which the episode ends [55]. This system can be modeled as an infinite horizon MDP with an absorbing state corresponding to the terminal state. It is clear that this MDP does not satisfy irreducibility assumption. Furthermore, since after a finite time, with high probability we reach to the absorbing state, it is impossible to find the optimal policy using a *single* trajectory.

Second, since here we assume finite state and action spaces, the aperiodicity assumption along with irreducibility is equivalent to the existence of a mixing time, which is common in the literature [40], [47], [48], [50], [64], [65]. We make this more precise in Lemma 7.

V. Proof of Theorem 1: Two-Time-Scale Analysis

Next we provide the proof of Theorem 1. Before expressing the proof, we state the following Proposition on the convergence of the natural AC along with its proof.

Proposition 1: Consider the two-time-scale natural AC Algorithm 1 with T iterations, and the output $\hat{\pi}_{\hat{T}}$. Suppose the step size β_t and the exploration parameter ϵ_t are non-increasing with respect to t. We have the following:

$$\mathbb{E}[V^* - V^{\hat{\pi}_{\hat{T}}}] \leq \frac{1}{\sum_{t=0}^{T} \beta_t} \Bigg\{ \frac{2\beta}{(1-\gamma)^2} + \frac{\log|\mathcal{A}|}{(1-\gamma)} + \frac{2}{1-\gamma} \sum_{t=0}^{T} \Big[\beta_t \mathbb{E}\|Q^{\hat{\pi}_t} - Q_{t+1}\| + \frac{L_1\sqrt{|\mathcal{A}|}}{(1-\gamma)} \beta_t^2 + \epsilon_t \beta_t \Big] \Bigg\},$$

where $\|\cdot\|$ is the euclidean norm and L_1 is a constant whose precise value is given in Lemma 9 in Section V-D.

A. Proof of Proposition 1

In this section we provide the proof of Proposition 1. A similar result was proved for NPG in [5], when the actor has access to the exact Q-function. However, since the actor in Algorithm 1 has only access to $Q_t(s,a)$, rather than the exact Q-function $Q^{\pi_t}(s,a)$, establishing the bound in Proposition 1 is more challenging. Note that $Q_t(s,a)$ is obtained by the critic carrying out only one step of the TD-learning using a single sample update at each time step. Consequently, the error bound in Proposition 1 involves the term $\frac{1}{T}\sum_{t=0}^{T} \beta_t \mathbb{E}\|Q^{\hat{\pi}_t} - Q_{t+1}\|$, which accounts for the time-average error in the critic's estimate of the Q-function. Proposition 1 is also different from the results in [5] is terms of the step size. In particular, while [5] only considers the case of constant step size, the result in Proposition 1 is stated for general choice of non-increasing step sizes. Furthermore, a similar type of upper bound has been established in [24], [51] for the analysis of off-policy natural AC. However, in those works the ϵ_T term is absent.

When the actor has access to the exact Q-functions, it was observed in [5] that using a constant step size results

in $\mathcal{O}(1/T)$ rate of convergence. This result can be reproduced from Proposition 1 by eliminating the $\frac{1}{T}\sum_{t=0}^{T}\beta_t\mathbb{E}\|Q^{\hat{\pi}_t}-Q_{t+1}\|$ term in the upper bound, and taking a constant β_t, and choosing $\epsilon_T=0$. However, due to the existence of ϵ_T, and the $\frac{1}{T}\sum_{t=0}^{T}\beta_t\mathbb{E}\|Q^{\hat{\pi}_t}-Q_{t+1}\|$ term, the optimal convergence rate can only be obtained by a carefully diminishing step size, which has been shown in Theorem 1.

Next we provide the proof of Proposition 1.

proof: We will use a Lyapunov drift based argument to prove the proposition using the KL-divergence [66] as a Lyapunov or potential function. This is a natural choice because it is known to be the right potential function for mirror descent [67] in optimization and it is known [5], [7] that the natural gradient ascent is equivalent to mirror descent.

Let $M(\pi)=\mathbb{E}_{s\sim d^*}[D_{KL}(\pi^*(\cdot|s)||\pi(\cdot|s))]$. Then,

$$\begin{aligned}
&M(\pi_{t+1})-M(\pi_t)\\
&\quad=\mathbb{E}_{s\sim d^*}\big[\sum_a \pi^*(a|s)\log\frac{\pi_t(a|s)}{\pi_{t+1}(a|s)}\big]\\
&\quad\overset{(a)}{=}\sum_{s,a} d^*(s)\pi^*(a|s)(\log Z_t(s)-\beta_t Q_{t+1}(s,a))\\
&\quad\overset{(b)}{=}\beta_t\sum_{s,a} d^*(s)\pi^*(a|s)(Q^{\hat{\pi}_t}(s,a)-V^{\hat{\pi}_t}(s))\\
&\qquad+\beta_t\sum_{s,a} d^*(s)\pi^*(a|s)\big(Q_{t+1}(s,a)-Q^{\hat{\pi}_t}(s,a)\\
&\qquad-Q_{t+1}(s,a)+V^{\hat{\pi}_t}(s)\big)\\
&\qquad+\sum_{s,a} d^*(s)\pi^*(a|s)(\log Z_t(s)-\beta_t Q_{t+1}(s,a))\\
&\quad\overset{(c)}{=}(1-\gamma)\beta_t\left[V^{\hat{\pi}_t}-V^*\right]\\
&\qquad+\beta_t\sum_{s,a} d^*(s)\pi^*(a|s)\left[Q^{\hat{\pi}_t}(s,a)-Q_{t+1}\right]\\
&\qquad+\sum_{s,a} d^*(s)\left[\log Z_t(s)-\beta_t V^{\hat{\pi}_t}\right],
\end{aligned}$$

where, (a) is by the update rule of the policy π_t in Algorithm 1 with $Z_t(s)=\sum_a\pi_t(a|s)\exp(\beta_t Q_{t+1}(s,a))$, (b) is by adding and subtracting terms, and (c) is by Performance Difference Lemma [68]. Rearranging the terms, we get:

$$\begin{aligned}
&V^*-V^{\hat{\pi}_t}\\
&=\frac{1}{(1-\gamma)\beta_t}\left[M(\pi_t)-M(\pi_{t+1})\right]\\
&\quad+\frac{1}{1-\gamma}\sum_{s,a} d^*(s)\pi^*(a|s)\left[Q^{\hat{\pi}_t}(s,a)-Q_{t+1}(s,a)\right] \quad (6)\\
&\quad+\frac{1}{1-\gamma}\sum_{s,a} d^*(s)\left[\frac{1}{\beta_t}\log Z_t(s)-V^{\hat{\pi}_t}(s)\right]. \quad (7)
\end{aligned}$$

We bound the terms in (6) and (7) separately.

Using the Cauchy-Schwarz inequality in (6), we have:

$$\begin{aligned}
&\frac{1}{1-\gamma}\sum_{s,a} d^*(s)\pi^*(a|s)\left[Q^{\hat{\pi}_t}(s,a)-Q_{t+1}(s,a)\right]\\
&\quad\le\frac{1}{1-\gamma}\|Q^{\hat{\pi}_t}-Q_{t+1}\|.
\end{aligned}$$

In order to bound the term (7), we use the following lemma.

Lemma 1: Consider Algorithm 1 with $Z_t(s)=\sum_a\pi_t(a|s)\exp(\beta_t Q_{t+1}(s,a))$. For any $t\ge 0$ we have the following inequality:

$$\begin{aligned}
&\sum_s d^*(s)\left[\frac{1}{\beta_t}\log Z_t(s)-V^{\hat{\pi}_t}(s)\right]\le V^{\pi_{t+1}}(d^*)-V^{\pi_t}(d^*)\\
&+\|Q^{\hat{\pi}_t}-Q_{t+1}\|+\frac{2L_1\sqrt{|\mathcal{A}|}}{(1-\gamma)^2}\beta_t+\frac{\epsilon_{t-1}}{1-\gamma},
\end{aligned}$$

where $\epsilon_{-1}:=0$.

Employing Lemma 1 in (7), multiplying by β_t and summing from 0 to T, we get:

$$\begin{aligned}
&\sum_{t=0}^{T}\beta_t(V^*-V^{\hat{\pi}_t})\\
&\le\sum_{t=0}^{T}\frac{1}{(1-\gamma)}\left[M(\pi_t)-M(\pi_{t+1})\right]\\
&\quad+\frac{2\beta_t}{1-\gamma}\|Q^{\hat{\pi}_t}-Q_{t+1}\|+\frac{\beta_t}{1-\gamma}\left[V^{\pi_{t+1}}(d^*)-V^{\pi_t}(d^*)\right]\\
&\quad+\frac{2L_1\sqrt{|\mathcal{A}|}}{(1-\gamma)^2}\beta_t^2+\frac{\epsilon_{t-1}\beta_t}{1-\gamma}\\
&=\frac{1}{(1-\gamma)}\sum_{t=0}^{T}\Big\{M(\pi_t)-M(\pi_{t+1}) \quad (8)\\
&\quad+\beta_t\left[V^{\pi_{t+1}}(d^*)-V^{\pi_t}(d^*)\right]\Big\} \quad (9)\\
&\quad+\sum_{t=0}^{T}\Big\{\frac{2\beta_t}{1-\gamma}\|Q^{\hat{\pi}_t}-Q_{t+1}\|+\frac{2L_1\sqrt{|\mathcal{A}|}}{(1-\gamma)^2}\beta_t^2 \quad (10)\\
&\quad+\frac{\epsilon_{t-1}\beta_t}{1-\gamma}\Big\}. \quad (11)
\end{aligned}$$

We evaluate (8)+(9) and (10)+(11) separately. First, we have:

$$\begin{aligned}
&(8)+(9)=\frac{1}{(1-\gamma)}\sum_{t=0}^{T}\left[\beta_t\left[V^{\pi_{t+1}}(d^*)-V^{\pi_t}(d^*)\right]\right]\\
&\quad+\frac{1}{(1-\gamma)}\left[M(\pi_0)-M(\pi_{T+1})\right]\\
&\overset{(a)}{\le}\frac{1}{(1-\gamma)}\sum_{t=0}^{T}\left[\beta_t\left[V^{\pi_{t+1}}(d^*)-V^{\pi_t}(d^*)\right]\right]\\
&\quad+\frac{\log|\mathcal{A}|}{(1-\gamma)}\\
&=\frac{1}{(1-\gamma)}\sum_{t=0}^{T}(\beta_t-\beta_{t+1})V^{\pi_{t+1}}(d^*)\\
&\quad-\frac{\beta_0}{(1-\gamma)}V^{\pi_0}(d^*)+\frac{\beta_{T+1}}{(1-\gamma)}V^{\pi_{T+1}}(d^*)+\frac{\log|\mathcal{A}|}{(1-\gamma)}\\
&\le\frac{1}{(1-\gamma)}\sum_{t=0}^{T}(\beta_t-\beta_{t+1})V^{\pi_{t+1}}(d^*)\\
&\quad+\frac{\beta_{T+1}}{(1-\gamma)}V^{\pi_{T+1}}(d^*)+\frac{\log|\mathcal{A}|}{(1-\gamma)}\\
&\overset{(b)}{\le}\frac{1}{(1-\gamma)^2}\sum_{t=0}^{T}(\beta_t-\beta_{t+1})
\end{aligned}$$

$$+\frac{\beta_{T+1}}{(1-\gamma)^2}+\frac{\log|\mathcal{A}|}{(1-\gamma)}$$
$$=\frac{1}{(1-\gamma)^2}(\beta_0-\beta_{T+1})+\frac{\beta_{T+1}}{(1-\gamma)^2}+\frac{\log|\mathcal{A}|}{(1-\gamma)}$$
$$\leq\frac{2\beta}{(1-\gamma)^2}+\frac{\log|\mathcal{A}|}{(1-\gamma)}, \tag{12}$$

where (a) is due to $0\leq D_{KL}(P(X)||Unif(X))\leq\log|\mathcal{X}|$, where $P(X)$ is any distributieseseon over $\mathcal{X}$ and $Unif(X)$ is uniform distribution over $\mathcal{X}$, and $|\mathcal{X}|$ is the cardinality of the random variable X [66], and (b) is due to $V^{\pi}\in[0,\frac{1}{1-\gamma}]$, shown in Lemma 5, and β_t being non-increasing with respect to t.

Furthermore, we have

$$(10)+(11)$$
$$=\frac{2}{1-\gamma}\sum_{t=0}^{T}\left[\beta_t\|Q^{\hat{\pi}_t}-Q_{t+1}\|+\frac{L_1\sqrt{|\mathcal{A}|}}{(1-\gamma)}\beta_t^2+0.5\epsilon_{t-1}\beta_t\right]$$
$$\leq\frac{2}{1-\gamma}\sum_{t=0}^{T}\left[\beta_t\|Q^{\hat{\pi}_t}-Q_{t+1}\|+\frac{L_1\sqrt{|\mathcal{A}|}}{(1-\gamma)}\beta_t^2+\epsilon_t\beta_t\right]. \tag{13}$$

Dividing both sides of (12) and (13) with $\sum_{t=0}^{T}\beta_t$, taking expectation, and noting that $\frac{\sum_{t=0}^{T}\beta_t\mathbb{E}(V^*-V^{\hat{\pi}_t})}{\sum_{t=0}^{T}\beta_t}=\mathbb{E}[V^*-V^{\hat{\pi}_{\hat{T}}}]$, we get the proposition.■

According to Proposition 1, to establish a bound for the performance metric $\mathbb{E}[V^*-V^{\hat{\pi}_{\hat{T}}}]$, we need to characterize a bound for $\mathbb{E}\|Q^{\hat{\pi}_t}-Q_{t+1}\|$ for all $0\leq t\leq T$. Next we provide the proof of Theorem 1 which essentially corresponds to the characterization of this bound.

B. Proof of Theorem 1

First, we introduce some notations and lemmas which will be used within the proof.

$$O_t=(S_t,A_t,S_{t+1},A_{t+1})$$
$$r(O_t)=[0;0;\ldots;0;\mathcal{R}(S_t,A_t);0;\ldots;0]\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$$
$$A(O)\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$$
$$A(O)_{i,j}\equiv A(s,a,s',a')_{i,j}$$
$$=\begin{cases}\gamma-1 & i=j=(s,a)=(s',a')\\ -1 & (s,a)\neq(s',a'),i=j=(s,a)\\ \gamma & (s,a)\neq(s',a'),i=(s,a),j=(s',a')\\ 0 & \text{otherwise}\end{cases}$$
$$\theta_t=Q_t-Q^{\hat{\pi}_{t-1}} \tag{14}$$
$$\bar{A}^{\pi}=\mathbb{E}_{s\sim\mu^{\pi}(\cdot),a\sim\pi(\cdot|s),s'\sim\mathcal{P}(\cdot|s,a),a'\sim\pi(\cdot|s')}[A(s,a,s',a')] \tag{15}$$
$$\Gamma(\pi,\theta,O)=\theta^{\top}(r(O)+A(O)Q^{\pi})+\theta^{\top}(A(O)-\bar{A}^{\pi})\theta$$

Note that with the above notation, the update of the Q-function in Algorithm 1 in the vector form can be written as:

$$Q_{t+1}=Q_t+\alpha_t(r(O_t)+A(O_t)Q_t),$$

which by adding and subtracting terms, can be equivalently written as:

$$\theta_{t+1}+(Q^{\hat{\pi}_t}-Q^{\hat{\pi}_{t-1}})=\theta_t+\alpha_t(r(O_t)+A(O_t)Q^{\hat{\pi}_{t-1}}+A(O_t)\theta_t).$$

Lemmas 2 and 3 characterize an upper bound on the one step drift of $Q^{\hat{\pi}_t}$ and θ_t.

Lemma 2: One step drift of the Q-function with respect to the sampling policy $\hat{\pi}_t$ satisfies the following:

$$\|Q^{\hat{\pi}_{t+1}}-Q^{\hat{\pi}_t}\|\leq L_2(L_3\frac{\epsilon_{t-1}}{t}+L_1\beta_t),$$

where the constants L_1, L_2, and L_3 are defined in Lemmas 8 and 9.

Lemma 3: The one step drift of the vector θ_t can be bounded as

$$\|\theta_{t+1}-\theta_t\|^2\leq 2\alpha_t^2\Delta_Q^2+4{L_2}^2{L_3}^2\frac{\epsilon_{t-2}^2}{(t-1)^2}+4{L_2}^2{L_1}^2\beta_{t-1}^2,$$

where Δ_Q is defined in Lemma 16 in the Appendix.

The following lemma is directly used to create a negative drift, which is essential for the convergence proof of Theorem 1.

Lemma 4: Consider the policy $\hat{\pi}_{t-1}$ in the $t-1$'th iteration of Algorithm 1, and the vector θ_t and the matrix $\bar{A}^{\hat{\pi}_{t-1}}$ defined in (14) and (15), respectively. We have:

$$\theta_t^{\top}\bar{A}^{\hat{\pi}_{t-1}}\theta_t\leq-(1-\gamma)\frac{\epsilon_{t-2}}{|\mathcal{A}|}\mu\|\theta_t\|^2,$$

where $\mu>0$ is some absolute constant. Later in Lemma 10 we explain the intuition behind the constant μ.

The following Lemma provides some absolute bounds on the value and Q-function.

Lemma 5: Let $Q_{\max}=\frac{1}{1-\gamma}$. Then we have

1) $0\leq V^{\pi}\leq Q_{\max}$
2) $0\leq Q^{\pi}(s,a)\leq Q_{\max}$
3) $\|Q^{\pi}\|\leq\sqrt{|\mathcal{S}||\mathcal{A}|}Q_{\max}$
4) $\|Q_t\|\leq\sqrt{|\mathcal{S}||\mathcal{A}|}Q_{\max}$.

A major part of the proof of Theorem 1 is to establish a bound on $\mathbb{E}[\Gamma(\hat{\pi}_{k-1},\theta_k,O_k)]$. In the following, we provide such a bound in Lemma 6. The proof of this lemma is provided in Section V-C.

Lemma 6: For any $\tau<t$, we have:

$$\mathbb{E}\big[\Gamma(\hat{\pi}_{t-1},\theta_t,O_t)\big]\leq C_b m\rho^{\tau}+K_2\Delta_Q\tau\alpha_{t-\tau}$$
$$+(C_uL_3+K_1L_3+K_2L_2L_3)\frac{(\tau+1)^2\epsilon_{t-\tau-2}}{t-\tau-1}$$
$$+(C_uL_1+K_1L_1+K_2L_1L_2)(\tau+1)^2\beta_{t-\tau-1}.$$

We further define

$$\tau_t:=\min\{r>0|M_\rho\rho^r\leq\beta_t,r\text{ integral}\}, \tag{16}$$

where $M_\rho=(-\sigma/\ln(\rho))^{\sigma}/(\rho^{1+\sigma/\ln(\rho)})$. It is easy to see that $1\leq\tau_t\leq t$ for all t, and $\tau_t=\mathcal{O}(\log(t))=\tilde{\mathcal{O}}(1)$.

In order to establish the convergence result in Theorem 1, we use the bound in Proposition 1. By the definition of the step sizes, it is clear that $\sum_{t=0}^{T}\beta_t=\Theta(T^{1-\sigma})$. Hence, by Proposition 1 and assumptions on the step sizes, it is straightforward to show that

$$\mathbb{E}[V^*-V^{\hat{\pi}_{\hat{T}}}]$$
$$\leq\mathcal{O}\left(\frac{1}{T^{1-\sigma}}\right)+\begin{cases}\tilde{\mathcal{O}}\left(\frac{1}{T^{\sigma}}\right) & \text{if}\quad 1>2\sigma\\ \tilde{\mathcal{O}}(\frac{1}{T^{1-\sigma}}) & \text{o.w}\end{cases}$$
$$+\begin{cases}\tilde{\mathcal{O}}\left(\frac{\epsilon}{T^{1-\sigma}}\right) & \text{if}\quad \xi+\sigma>1\\ \tilde{\mathcal{O}}(\frac{\epsilon}{T^{\xi}}) & \text{o.w}\end{cases}$$

$$+\frac{1}{T^{1-\sigma}}\mathcal{O}\left\{\sum_{t=0}^{T}\beta_t\mathbb{E}\|Q^{\hat{\pi}_t}-Q_{t+1}\|\right\}. \tag{17}$$

Next, we aim at bounding the term $\sum_{t=0}^{T}\beta_t\mathbb{E}\|Q^{\hat{\pi}_t}-Q_{t+1}\|$. We have:

$$\sum_{t=0}^{T}\beta_t\mathbb{E}\|Q^{\hat{\pi}_t}-Q_{t+1}\| = \sum_{t=0}^{T}\beta_t^{1/2\sigma}\beta_t^{1-1/2\sigma}\mathbb{E}\|Q^{\hat{\pi}_t}-Q_{t+1}\|$$
$$\leq \sqrt{\sum_{t=0}^{T}\beta_t^{1/\sigma}}\times\sqrt{\sum_{t=0}^{T}\beta_t^{\frac{2\sigma-1}{\sigma}}\mathbb{E}\|Q^{\hat{\pi}_t}-Q_{t+1}\|^2}, \tag{18}$$

where the inequality is due to Cauchy–Schwarz inequality. Furthermore, we have $\sum_{t=0}^{T}\beta_t^{1/\sigma}=\mathcal{O}(\log(T))=\tilde{\mathcal{O}}(1)$. Hence, we only left to bound the last term in (18) which we will do in the rest of the proof.

Using $\|\theta_t\|^2$ as the Lyapunov function, we have:

$$\begin{aligned}
&\|\theta_{t+1}\|^2-\|\theta_t\|^2 = 2\theta_t^\top(\theta_{t+1}-\theta_t-\alpha_t\bar{A}^{\hat{\pi}_{t-1}}\theta_t)\\
&+\|\theta_{t+1}-\theta_t\|^2+2\alpha_t\theta_t^\top\bar{A}^{\hat{\pi}_{t-1}}\theta_t\\
&\overset{(a)}{=}2\alpha_t\Gamma(\hat{\pi}_{t-1},\theta_t,O_t)+2\theta_t^\top(Q^{\hat{\pi}_{t-1}}-Q^{\hat{\pi}_t})+\|\theta_{t+1}-\theta_t\|^2\\
&+2\alpha_t\theta_t^\top\bar{A}^{\hat{\pi}_{t-1}}\theta_t\\
&\overset{(b)}{\leq}2\alpha_t\Gamma(\hat{\pi}_{t-1},\theta_t,O_t)+2\|\theta_t\|.\overbrace{\|Q^{\hat{\pi}_{t-1}}-Q^{\hat{\pi}_t}\|}^{T_1}+\overbrace{\|\theta_{t+1}-\theta_t\|^2}^{T_2}\\
&+2\alpha_t\overbrace{\theta_t^\top\bar{A}^{\hat{\pi}_{t-1}}\theta_t}^{T_3},
\end{aligned}\tag{19}$$

where (a) is by definition of Γ, and (b) is due to the Cauchy–Schwarz inequality. We bound each of the terms T_1, T_2, T_3 using Lemmas 2, 3, and 4, respectively. We have

$$\begin{aligned}
&\|\theta_{t+1}\|^2-\|\theta_t\|^2\\
\leq&2\alpha_t\Gamma(\hat{\pi}_{t-1},\theta_t,O_t)+2L_2\left(L_3\frac{\epsilon_{t-2}}{t-1}+L_1\beta_{t-1}\right)\|\theta_t\|\\
&+2\alpha_t^2\Delta_Q^2+4{L_2}^2{L_3}^2\frac{\epsilon_{t-2}^2}{(t-1)^2}\\
&+4{L_2}^2{L_1}^2\beta_{t-1}^2-\frac{2(1-\gamma)\mu}{|\mathcal{A}|}\alpha_t\epsilon_{t-2}\|\theta_t\|^2.
\end{aligned}\tag{20}$$

Define $\lambda_t=\beta_t^{\frac{2\sigma-1}{\sigma}}$. Multiplying both sides of (20) with λ_t and denoting $y_t=\lambda_t\|\theta_t\|^2$, we have $y_t\leq e_t(\|\theta_t\|^2-\|\theta_{t+1}\|^2)+u_t+h_t\sqrt{y_t}$, where $e_t=\frac{\lambda_t|\mathcal{A}|}{2(1-\gamma)\mu\alpha_t\epsilon_{t-2}}$ and $u_t=\frac{|\mathcal{A}|\lambda_t}{2(1-\gamma)\mu\alpha_t\epsilon_{t-2}}(2\alpha_t\Gamma(\hat{\pi}_{t-1},\theta_t,O_t)+2\alpha_t^2\Delta_Q^2+4{L_2}^2{L_3}^2\frac{\epsilon_{t-2}^2}{(t-1)^2}+4{L_2}^2{L_1}^2\beta_{t-1}^2)$, and $h_t=\frac{|\mathcal{A}|\sqrt{\lambda_t}}{2(1-\gamma)\mu\alpha_t\epsilon_{t-2}}2L_2\left(L_3\frac{\epsilon_{t-2}}{t-1}+L_1\beta_{t-1}\right)$. Summing from τ_t+2 to t, we have

$$\sum_{k=\tau_t+2}^{t}y_k\leq\underbrace{\sum_{k=\tau_t+2}^{t}e_k(\|\theta_k\|^2-\|\theta_{k+1}\|^2)}_{T_1}+\underbrace{\sum_{k=\tau_t+2}^{t}u_k}_{T_2}+\underbrace{\sum_{k=\tau_t+2}^{t}h_k\sqrt{y_k}}_{T_3}. \tag{21}$$

We bound the summation of each of the terms T_1, T_2, and T_3 separately. First we have

$$T_1=e_{\tau_t+1}\|\theta_{\tau_t+2}\|^2-e_t\|\theta_{t+1}\|^2+\sum_{k=\tau_t+2}^{t}(e_k-e_{k-1})\|\theta_k\|^2.$$

We have $e_k\sim\lambda_k/(\alpha_k\epsilon_k)\sim\beta_k^{\frac{2\sigma-1}{\sigma}}/\alpha_k\epsilon_k\sim k^{\nu+\xi+1-2\sigma}/\epsilon$. For the case $\nu+\xi+1-2\sigma>0$, e_k is increasing. Hence, we have

$$\begin{aligned}
T_1&\overset{(a)}{\leq}4|\mathcal{S}||\mathcal{A}|Q_{\max}^2\left[e_{\tau_t+1}+\sum_{k=\tau_t+2}^{t}(e_k-e_{k-1})\right]\\
&\overset{(b)}{\leq}4|\mathcal{S}||\mathcal{A}|Q_{\max}^2\left[e_{\tau_t+1}+e_t\right]\leq\tilde{\mathcal{O}}(t^{\nu+\xi+1-2\sigma}/\epsilon),
\end{aligned}$$

where (a) is due to Lemma 5, and (b) is due to $e_t\geq 0$. Furthermore, if $\nu+\xi+1-2\sigma<0$, we have e_k decreasing, and hence $T_1\leq\tilde{\mathcal{O}}(e_{\tau_t+1})=\tilde{\mathcal{O}}(1/\epsilon)$. It is also easy to show that for $\nu+\xi+1-2\sigma=0$, we have $T_1\leq\tilde{\mathcal{O}}(1/\epsilon)$. Hence, in total we have

$$T_1\leq\begin{cases}\tilde{\mathcal{O}}(t^{\nu+\xi+1-2\sigma}/\epsilon) & \text{if } \nu+\xi+1>2\sigma,\\ \tilde{\mathcal{O}}(1/\epsilon) & \text{o.w.}\end{cases} \tag{22}$$

Furthermore, for the term T_2 we have

$$\begin{aligned}
\mathbb{E}T_2=&\mathcal{O}(\sum_{k=\tau_t+2}^{t}\frac{\lambda_k}{\epsilon_k}\mathbb{E}\Gamma(\hat{\pi}_{k-1},\theta_k,O_k)+\frac{\lambda_k\alpha_k}{\epsilon_k}+\frac{\lambda_k\epsilon_k}{k^2\alpha_k}+\frac{\lambda_k\beta_k^2}{\alpha_k\epsilon_k})\\
\overset{(a)}{\leq}&\tilde{\mathcal{O}}(\sum_{k=\tau_t+2}^{t}\frac{\lambda_k}{\epsilon_k}(\beta_k+\alpha_k+\frac{\epsilon_k}{k})+\frac{\lambda_k\alpha_k}{\epsilon_k}+\frac{\lambda_k\epsilon_k}{k^2\alpha_k}+\frac{\lambda_k\beta_k^2}{\alpha_k\epsilon_k})\\
\overset{(b)}{\leq}&\tilde{\mathcal{O}}(\sum_{k=\tau_t+2}^{t}k^{1-2\sigma}(k^{\xi-\nu}/\epsilon+k^{-1}+k^{\xi+\nu-2\sigma}/\epsilon))\\
\leq&\begin{cases}\tilde{\mathcal{O}}(t^{2+\xi-2\sigma-\nu}/\epsilon) & \text{if } 2+\xi>\nu+2\sigma,\\ \tilde{\mathcal{O}}(1/\epsilon) & \text{o.w.}\end{cases}\\
&+\begin{cases}\tilde{\mathcal{O}}(t^{1-2\sigma}) & \text{if } 1>2\sigma,\\ \tilde{\mathcal{O}}(1) & \text{o.w.}\end{cases}\\
&+\begin{cases}\tilde{\mathcal{O}}(t^{2+\xi+\nu-4\sigma}/\epsilon) & \text{if } 2+\xi+\nu>4\sigma,\\ \tilde{\mathcal{O}}(1/\epsilon) & \text{o.w,}\end{cases}
\end{aligned}\tag{23}$$

where in (a) we use Lemma 6 with $\tau=\tau_k$, and in (b) we use the assumptions on the step sizes.

Finally, for the term T_3 we have

$$\begin{aligned}
\mathbb{E}T_3&\overset{(a)}{\leq}\sqrt{\sum_{k=\tau_t+2}^{t}h_k^2}\times\mathbb{E}\left[\sqrt{\sum_{k=\tau_t+2}^{t}y_k}\right]\\
&\overset{(b)}{\leq}\sqrt{\sum_{k=\tau_t+2}^{t}h_k^2}\times\sqrt{\sum_{k=\tau_t+2}^{t}\mathbb{E}y_k}
\end{aligned}$$

where (a) is by Cauchy–Schwarz inequality and (b) is by concavity of square root and Jensen's inequality. Denoting $G(t)=\sum_{k=\tau_t+2}^{t}h_k^2$, we have

$$G(t)\leq\tilde{\mathcal{O}}\left(\sum_{k=\tau_t+2}^{t}\frac{\lambda_k}{\alpha_k^2k^2}+\frac{\lambda_k\beta_k^2}{\alpha_k^2\epsilon_k^2}\right)$$

$$\leq \tilde{\mathcal{O}}\left(\sum_{k=\tau_t+2}^{t} k^{-1}+k^{1-4\sigma+2\nu+2\xi}/\epsilon^2\right)$$
$$=\tilde{\mathcal{O}}(1)+\begin{cases}\tilde{\mathcal{O}}(t^{2-4\sigma+2\nu+2\xi}/\epsilon^2) & \text{if } 2+2\nu+2\xi>4\sigma,\\ \tilde{\mathcal{O}}(1/\epsilon^2) & \text{o.w.}\end{cases} \tag{24}$$

Denote $H(t)=\sum_{k=\tau_t+2}^{t}\mathbb{E}y_k$. Taking expectation on both sides of (21), we have

$$H(t)\leq \mathbb{E}T_1+\mathbb{E}T_2+\sqrt{G(t)}.\sqrt{H(t)}$$
$$\implies (\sqrt{H(t)}-\frac{1}{2}\sqrt{G(t)})^2\leq \mathbb{E}T_1+\mathbb{E}T_2+1/4G(t)$$
$$\implies \sqrt{H(t)}-\frac{1}{2}\sqrt{G(t)}\leq\sqrt{\mathbb{E}T_1+\mathbb{E}T_2+1/4G(t)}$$
$$\implies H(t)\leq 2\mathbb{E}T_1+2\mathbb{E}T_2+G(t) \tag{25}$$

Combining (22), (23), (24), and (25), we have

$$H(t)\leq\begin{cases}\tilde{\mathcal{O}}(t^{\nu+\xi+1-2\sigma}/\epsilon) & \text{if } \nu+\xi+1>2\sigma,\\ \tilde{\mathcal{O}}(1/\epsilon) & \text{o.w.}\end{cases}$$
$$+\begin{cases}\tilde{\mathcal{O}}(t^{2+\xi-2\sigma-\nu}/\epsilon) & \text{if } 2+\xi>\nu+2\sigma,\\ \tilde{\mathcal{O}}(1/\epsilon) & \text{o.w.}\end{cases}$$
$$+\begin{cases}\tilde{\mathcal{O}}(t^{1-2\sigma}) & \text{if } 1>2\sigma,\\ \tilde{\mathcal{O}}(1) & \text{o.w.}\end{cases}$$
$$+\begin{cases}\tilde{\mathcal{O}}(t^{2+\xi+\nu-4\sigma}/\epsilon) & \text{if } 2+\xi+\nu>4\sigma,\\ \tilde{\mathcal{O}}(1/\epsilon) & \text{o.w,}\end{cases}$$
$$+\begin{cases}\tilde{\mathcal{O}}(t^{2-4\sigma+2\nu+2\xi}/\epsilon^2) & \text{if } 2+2\nu+2\xi>4\sigma,\\ \tilde{\mathcal{O}}(1/\epsilon^2) & \text{o.w.}\end{cases}$$
$$+\tilde{\mathcal{O}}(1). \tag{26}$$

Combining (26), (17) and (18), we get the result.■

1) Proof of Corollary 1.1: In the case of constant exploration parameter, we have $\xi=0$, and the optimal step size can be achieved by $\sigma=3/4$ and $\nu=1/2$. In this case, we get $\mathbb{E}[V^*-V^{\hat{\pi}_T}]\leq\tilde{\mathcal{O}}\left(\frac{T^{-1/4}}{\epsilon}+\epsilon\right)$. Hence, to get to a solution policy within $\delta/\epsilon+\epsilon$ of the global optimum, we need $\tilde{\mathcal{O}}(1/\delta^4)$ number of samples. Furthermore, to get δ-close to the global optimum, we should have $\tilde{\mathcal{O}}(\frac{T^{-1/4}}{\epsilon})\leq\delta/2$ and $\tilde{\mathcal{O}}(\epsilon)\leq\delta/2$, which means we have $\tilde{\mathcal{O}}(T^{-1/8})\leq\delta$. Hence, to get δ-close to the global optimum, we need $\tilde{\mathcal{O}}(1/\delta^8)$ number of samples. ■

2) Proof of Corollary 1.2: For $\xi>0$ we get $\mathbb{E}[V^*-V^{\hat{\pi}_T}]\leq\tilde{\mathcal{O}}(T^{-1/6})$ convergence to the global optimum which can be achieved by $\xi=1/6$, $\nu=1/2$, and $\sigma=5/6$. Hence, in this case to get δ-close to the global optimum, we need $\tilde{\mathcal{O}}(1/\delta^6)$ number of samples. This proves Corollaries 1.1 and 1.2. ■

C. Proof sketch of Lemma 6

In Algorithm 1, the actions $\{A_t\}_{t\geq1}$ are sampled from a time-varying policy $\hat{\pi}_{t-1}$. Hence the tuple (S_t,A_t) follows a time-varying Markov chain as follows

$$S_{t-\tau}\xrightarrow{\hat{\pi}_{t-\tau-1}}A_{t-\tau}\xrightarrow{\mathcal{P}}S_{t-\tau+1}\xrightarrow{\hat{\pi}_{t-\tau}}A_{t-\tau+1}$$
$$\dots\xrightarrow{\mathcal{P}}S_t\xrightarrow{\hat{\pi}_{t-1}}A_t\xrightarrow{\mathcal{P}}S_{t+1}\xrightarrow{\hat{\pi}_t}A_{t+1}.$$

Since the sampling policy is changing over time, the convergence analysis of this Markov chain is difficult.

In order to analyze this time-varying Markov chain, at each time step t, we construct the following auxiliary Markov chain (This idea was first employed in [69]):

$$S_{t-\tau}\xrightarrow{\hat{\pi}_{t-\tau-1}}A_{t-\tau}\xrightarrow{\mathcal{P}}\tilde{S}_{t-\tau+1}\xrightarrow{\hat{\pi}_{t-\tau-1}}\tilde{A}_{t-\tau+1}$$
$$\dots\xrightarrow{\mathcal{P}}\tilde{S}_t\xrightarrow{\hat{\pi}_{t-\tau-1}}\tilde{A}_t\xrightarrow{\mathcal{P}}\tilde{S}_{t+1}\xrightarrow{\hat{\pi}_{t-\tau-1}}\tilde{A}_{t+1}.$$

Due to the geometric mixing of the Markov chain, which is stated formally in Lemma 7, by choosing τ large enough, the distribution of $(\tilde{S}_{t+1},\tilde{A}_{t+1})$ is "sufficiently close" to the stationary distribution $\mu^{\hat{\pi}_{t-\tau-1}}\otimes\hat{\pi}_{t-\tau-1}$.

Lemma 7: Suppose Assumption 1 holds for an MDP. Then there exist $m>0$ and $\rho\in(0,1)$, such that

$$d_{TV}(\mu^\pi(\cdot),P^\pi(S_\tau=\cdot|S_1=s))\leq m\rho^\tau,\forall s\in\mathcal{S},\forall\pi, \tag{27}$$

where $d_{TV}(\cdot,\cdot)$ denotes the total variation distance between two distributions. Furthermore, aperiodicity and the existence of m and ρ in inequality (27) are equivalent, i.e., if there exist a policy such that the underlying Markov chain is periodic, then (27) does not hold.

Define $\tilde{O}_t=(\tilde{S}_t,\tilde{A}_t,\tilde{S}_{t+1},\tilde{A}_{t+1})$. We have:

$$\Gamma(\hat{\pi}_{t-1},\theta_t,O_t)=\Gamma(\hat{\pi}_{t-1},\theta_t,O_t)-\Gamma(\hat{\pi}_{t-\tau-1},\theta_t,O_t) \tag{28}$$
$$+\Gamma(\hat{\pi}_{t-\tau-1},\theta_t,O_t)-\Gamma(\hat{\pi}_{t-\tau-1},\theta_{t-\tau},O_t) \tag{29}$$
$$+\Gamma(\hat{\pi}_{t-\tau-1},\theta_{t-\tau},O_t)-\Gamma(\hat{\pi}_{t-\tau-1},\theta_{t-\tau},\tilde{O}_t) \tag{30}$$
$$+\Gamma(\hat{\pi}_{t-\tau-1},\theta_{t-\tau},\tilde{O}_t). \tag{31}$$

We bound each of the terms above separately. Due to the Lipschitzness of Γ with respect to it's first and second arguments, the terms (28) and (29) can be bounded as follows:

$$\Gamma(\hat{\pi}_{t-1},\theta_t,O_t)-\Gamma(\hat{\pi}_{t-\tau-1},\theta_t,O_t)\leq\mathcal{O}\left(\|\hat{\pi}_{t-1}-\hat{\pi}_{t-\tau-1}\|\right)$$
$$\leq\mathcal{O}\left(\sum_{i=t-\tau}^{t-1}\|\hat{\pi}_i-\hat{\pi}_{i-1}\|\right)\leq\mathcal{O}\left(\tau\frac{\epsilon_t}{t}+\tau\beta_t\right).$$

$$\Gamma(\hat{\pi}_{t-\tau-1},\theta_t,O_t)-\Gamma(\hat{\pi}_{t-\tau-1},\theta_{t-\tau},O_t)\leq\mathcal{O}\left(\|\theta_t-\theta_{t-\tau}\|\right)$$
$$\leq\mathcal{O}\left(\sum_{i=t-\tau+1}^{t}\|\theta_i-\theta_{i-1}\|\right)\leq\mathcal{O}\left(\tau\alpha_t+\tau\frac{\epsilon_t}{t}+\tau\beta_t\right).$$

In order to bound the remaining two terms (30) and (31), we first apply conditional expectation on both sides. Bounding (30) is slightly technical and is presented in Lemma 13. The main idea is as follows. Since the policy $\hat{\pi}_t$ does not change very fast over time, the conditional expectation of $\Gamma(\hat{\pi}_{t-\tau-1},\theta_{t-\tau},O_t)$ and $\Gamma(\hat{\pi}_{t-\tau-1},\theta_{t-\tau},\tilde{O}_t)$ are close. Denoting $\bar{\mathcal{F}}_{t-\tau}=\{S_{t-\tau},\hat{\pi}_{t-\tau-1},\theta_{t-\tau}\}$, we have

$$\mathbb{E}[\Gamma(\hat{\pi}_{t-\tau-1},\theta_{t-\tau},O_t)-\Gamma(\hat{\pi}_{t-\tau-1},\theta_{t-\tau},\tilde{O}_t)|\bar{\mathcal{F}}_{t-\tau}]$$
$$\leq\tilde{\mathcal{O}}\left(\sum_{i=t-\tau}^{t}\|\hat{\pi}_i-\hat{\pi}_{t-\tau-1}\|\right)\leq\mathcal{O}\left(\tau\frac{\epsilon_t}{t}+\tau\beta_t\right).$$

Finally, denoting $O'_t=(S'_t,A'_t,S'_{t+1},A'_{t+1})$, where $S'_t\sim\mu^{\hat{\pi}_{t-\tau-1}}$, $A'_t\sim\hat{\pi}_{t-\tau-1}(\cdot|S'_t)$, $S'_{t+1}\sim\mathcal{P}(\cdot|S'_t,A'_t)$, and $A'_{t+1}\sim\hat{\pi}_{t-\tau-1}(\cdot|S'_{t+1})$, we have $\mathbb{E}\left[\Gamma(\hat{\pi}_{t-\tau-1},\theta_{t-\tau},O'_t)\middle|S_{t-\tau},\hat{\pi}_{t-\tau-1}\right]=0$ due to the

Bellman equation. According to Lemma 7, the distribution of the auxiliary chain $\tilde{O}_t = (\tilde{S}_t, \tilde{A}_t, \tilde{S}_{t+1}, \tilde{A}_{t+1})$ converges geometrically fast to the distribution of $O'_t = (S'_t, A'_t, S'_{t+1}, A'_{t+1})$. Hence, we have

$$\mathbb{E}\left[\Gamma(\hat{\pi}_{t-\tau-1}, \theta_{t-\tau}, \tilde{O}_t)\middle| S_{t-\tau}, \hat{\pi}_{t-\tau-1}\right] \leq \mathcal{O}(\rho^\tau).$$

Putting all the above bounds together, we get the result. □

D. Explanation of the main lemmas

Lemma 2 which provides a bound on one step drift of the Q-function with respect to the sampling policy $\hat{\pi}_t$ can be derived from Lemmas 8 and 9 below.

Lemma 8: For every pair of policies π_1 and π_2, we have:

$$\|Q^{\pi_1} - Q^{\pi_2}\| \leq L_2 \|\pi_1 - \pi_2\|,$$

where $L_2 = \frac{\gamma|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$.

Lemma 9: The policy $\hat{\pi}_t$, satisfies the following:

$$\|\hat{\pi}_{t+1} - \hat{\pi}_t\| \leq L_1 \beta_t + L_3 \frac{\epsilon_{t-1}}{t}, \ \forall t \geq 1,$$

where $L_1 = Q_{\max}\sqrt{|\mathcal{A}||\mathcal{S}|}$ and $L_3 = \xi\sqrt{|\mathcal{S}|}(\frac{1}{\sqrt{|\mathcal{A}|}} + 1)$.

Lemma 8 characterizes the Lipschitzness of the Q^π function with respect to the policy π, and Lemma 9 provides an upper bound on the drift of the sampling policy $\hat{\pi}_t$.

Finally, Lemma 10 below provides an intuition regarding the constant μ in Lemma 4.

Lemma 10: Suppose Assumption 1 holds. There exist a constant $\mu > 0$ such that for all the policies π, the stationary distribution μ^π satisfies

$$\mu^\pi(s) \geq \mu, \forall s \in \mathcal{S}.$$

Lemma 10 is a direct consequence of the ergodicity of the underlying MDP. In particular, the ergodicity Assumption 1 ensures that for all the policies π, under the stationary distribution μ^π, all the states are being visited with rate at least μ. As explained in Section IV this is indeed essential for the convergence of AC algorithm.

REFERENCES

[1] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[2] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International conference on machine learning*, 2015, pp. 1889–1897.

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

[4] V. R. Konda and J. N. Tsitsiklis, "Actor-critic algorithms," in *Advances in neural information processing systems*, 2000, pp. 1008–1014.

[5] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, "On the theory of policy gradient methods: Optimality, approximation, and distribution shift," *Preprint arXiv:1908.00261*, 2019.

[6] J. Mei, C. Xiao, C. Szepesvari, and D. Schuurmans, "On the global convergence rates of softmax policy gradient methods," *preprint arXiv:2005.06392*, 2020.

[7] M. Geist, B. Scherrer, and O. Pietquin, "A theory of regularized markov decision processes," *preprint arXiv:1901.11275*, 2019.

[8] L. Shani, Y. Efroni, and S. Mannor, "Adaptive trust region policy optimization: Global convergence and faster rates for regularized MDPs," *arXiv preprint arXiv:1909.02769*, 2019.

[9] Z. Wang, V. Bapst, N. Heess, V. Mnih, R. Munos, K. Kavukcuoglu, and N. de Freitas, "Sample efficient actor-critic with experience replay," *preprint arXiv:1611.01224*, 2016.

[10] D. Bahdanau, P. Brakel, K. Xu, A. Goyal, R. Lowe, J. Pineau, A. Courville, and Y. Bengio, "An actor-critic algorithm for sequence prediction," *preprint arXiv:1607.07086*, 2016.

[11] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel *et al.*, "Soft actor-critic algorithms and applications," *preprint arXiv:1812.05905*, 2018.

[12] L. Espeholt, H. Soyer, R. Munos, K. Simonyan, V. Mnih, T. Ward, Y. Doron, V. Firoiu, T. Harley, I. Dunning *et al.*, "Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures," *arXiv preprint arXiv:1802.01561*, 2018.

[13] G. Gajjar, S. Khaparde, P. Nagaraju, and S. Soman, "Application of actor-critic learning algorithm for optimal bidding problem of a genco," *IEEE Transactions on Power Systems*, vol. 18, no. 1, pp. 11–18, 2003.

[14] J. Peters and S. Schaal, "Natural actor-critic," *Neurocomputing*, vol. 71, no. 7-9, pp. 1180–1190, 2008.

[15] S. Bhatnagar, R. S. Sutton, M. Ghavamzadeh, and M. Lee, "Natural actor–critic algorithms," *Automatica*, vol. 45, no. 11, pp. 2471–2482, 2009.

[16] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

[17] V. S. Borkar, *Stochastic approximation: a dynamical systems viewpoint*. Springer, 2009, vol. 48.

[18] A. Benveniste, M. Métivier, and P. Priouret, *Adaptive algorithms and stochastic approximations*. Springer Science & Business Media, 2012, vol. 22.

[19] W. Mou, C. J. Li, M. J. Wainwright, P. L. Bartlett, and M. I. Jordan, "On linear stochastic approximation: Fine-grained polyak-ruppert and non-asymptotic concentration," in *Conference on Learning Theory*. PMLR, 2020, pp. 2947–2997.

[20] Z. Chen, S. T. Maguluri, S. Shakkottai, and K. Shanmugam, "Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[21] T. T. Doan, "Finite-time convergence rates of nonlinear two-time-scale stochastic approximation under markovian noise," *arXiv preprint arXiv:2104.01627*, 2021.

[22] B. Liu, J. Liu, M. Ghavamzadeh, S. Mahadevan, and M. Petrik, "Finite-sample analysis of proximal gradient td algorithms." in *UAI*. Citeseer, 2015, pp. 504–513.

[23] M. Kaledin, E. Moulines, A. Naumov, V. Tadic, and H.-T. Wai, "Finite time analysis of linear two-timescale stochastic approximation with markovian noise," in *Conference on Learning Theory*. PMLR, 2020, pp. 2144–2203.

[24] S. Khodadadian, Z. Chen, and S. T. Maguluri, "Finite-sample analysis of off-policy natural actor-critic algorithm," *arXiv preprint arXiv:2102.09318*, 2021.

[25] V. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

[26] V. S. Borkar and S. Pattathil, "Concentration bounds for two time scale stochastic approximation," in *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2018, pp. 504–511.

[27] M. Kaledin, E. Moulines, A. Naumov, V. Tadic, and H.-T. Wai, "Finite time analysis of linear two-timescale stochastic approximation with markovian noise," *preprint arXiv:2002.01268*, 2020.

[28] T. T. Doan, "Finite-time analysis and restarting scheme for linear two-time-scale stochastic approximation," *preprint arXiv:1912.10583*, 2019.

[29] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Advances in neural information processing systems*, 2000, pp. 1057–1063.

[30] S. M. Kakade, "A natural policy gradient," in *Advances in neural information processing systems*, 2002, pp. 1531–1538.

[31] G. Raskutti and S. Mukherjee, "The information geometry of mirror descent," *IEEE Transactions on Information Theory*, vol. 61, no. 3, pp. 1451–1457, 2015.

[32] S. Gunasekar, B. Woodworth, and N. Srebro, "Mirrorless mirror descent: A more natural discretization of riemannian gradient flow," *arXiv preprint arXiv:2004.01025*, 2020.

[33] S. Cen, C. Cheng, Y. Chen, Y. Wei, and Y. Chi, "Fast global convergence of natural policy gradient methods with entropy regularization," *arXiv preprint arXiv:2007.06558*, 2020.

[34] S. Cayci, N. He, and R. Srikant, "Linear convergence of entropy-regularized natural policy gradient with linear function approximation," *arXiv preprint arXiv:2106.04096*, 2021.

[35] T. Morimura, E. Uchibe, J. Yoshimoto, and K. Doya, "A generalized natural actor-critic algorithm," in *Advances in neural information processing systems*, 2009, pp. 1312–1320.

[36] P. S. Thomas, W. C. Dabney, S. Giguere, and S. Mahadevan, "Projected natural actor-critic," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 2337–2345. [Online]. Available: http://papers.nips.cc/paper/5184-projected-natural-actor-critic.pdf

[37] R. J. Williams and L. C. Baird, "A mathematical analysis of actor-critic architectures for learning optimal controls through incremental dynamic programming," in *Proceedings of the Sixth Yale Workshop on Adaptive and Learning Systems*. Citeseer, 1990, pp. 96–101.

[38] S. Zhang, B. Liu, H. Yao, and S. Whiteson, "Provably convergent two-timescale off-policy actor-critic with function approximation," in *International Conference on Machine Learning*. PMLR, 2020, pp. 11 204–11 213.

[39] L. Shani, Y. Efroni, and S. Mannor, "Adaptive Trust Region Policy Optimization: Global Convergence and Faster Rates for Regularized MDPs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 5668–5675.

[40] G. Lan, "Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes," *arXiv preprint arXiv:2102.00135*, 2021.

[41] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Basar, "Finite-sample analysis for decentralized batch multi-agent reinforcement learning with networked agents," *IEEE Transactions on Automatic Control*, 2021.

[42] K. Zhang, A. Koppel, H. Zhu, and T. Başar, "Convergence and iteration complexity of policy gradient method for infinite-horizon reinforcement learning," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 7415–7422.

[43] S. Qiu, Z. Yang, J. Ye, and Z. Wang, "On the finite-time convergence of actor-critic algorithm," in *Optimization Foundations for Reinforcement Learning Workshop at Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[44] H. Kumar, A. Koppel, and A. Ribeiro, "On the sample complexity of actor-critic method for reinforcement learning with function approximation," *preprint arXiv:1910.08412*, 2019.

[45] B. Liu, Q. Cai, Z. Yang, and Z. Wang, "Neural proximal/trust region policy optimization attains globally optimal policy," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[46] L. Wang, Q. Cai, Z. Yang, and Z. Wang, "Neural policy gradient methods: Global optimality and rates of convergence," *preprint arXiv:1909.01150*, 2019.

[47] T. Xu, Z. Wang, and Y. Liang, "Non-asymptotic convergence analysis of two time-scale (natural) actor-critic algorithms," *arXiv preprint arXiv:2005.03557*, 2020.

[48] ——, "Improving sample complexity bounds for actor-critic algorithms," *preprint arXiv:2004.12956*, 2020.

[49] Y. Liu, K. Zhang, T. Basar, and W. Yin, "An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[50] Y. Wu, W. Zhang, P. Xu, and Q. Gu, "A finite time analysis of two time-scale actor critic methods," *preprint arXiv:2005.01350*, 2020.

[51] Z. Chen, S. Khodadadian, and S. T. Maguluri, "Finite-sample analysis of off-policy natural actor-critic with linear function approximation," *arXiv preprint arXiv:2105.12540*, 2021.

[52] T. Xu, Z. Yang, Z. Wang, and Y. Liang, "Doubly robust off-policy actor-critic: Convergence and optimality," *Preprint arXiv:2102.11866*, 2021.

[53] W. Zhan, S. Cen, B. Huang, Y. Chen, J. D. Lee, and Y. Chi, "Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence," *arXiv preprint arXiv:2105.11066*, 2021.

[54] Y. Hu, Z. Ji, and M. Telgarsky, "Actor-critic is implicitly biased towards high entropy optimal policies," *arXiv preprint arXiv:2110.11280*, 2021.

[55] P. R. Montague, "Reinforcement learning: An introduction, by sutton, rs and barto, ag," *Trends in cognitive sciences*, vol. 3, no. 9, p. 360, 1999.

[56] M. Wunder, M. L. Littman, and M. Babes, "Classes of multiagent q-learning dynamics with epsilon-greedy exploration," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. Citeseer, 2010, pp. 1167–1174.

[57] V. Kuleshov and D. Precup, "Algorithms for multi-armed bandit problems," *arXiv preprint arXiv:1402.6028*, 2014.

[58] M. Tokic, "Adaptive ε-greedy exploration in reinforcement learning based on value differences," in *Annual Conference on Artificial Intelligence*. Springer, 2010, pp. 203–210.

[59] D. Bouneffouf, A. Bouzeghoub, and A. L. Gançarski, "A contextual-bandit algorithm for mobile context-aware recommender system," in *International conference on neural information processing*. Springer, 2012, pp. 324–331.

[60] B. Hajek, *Random processes for engineers*. Cambridge university press, 2015.

[61] M. L. Puterman, "Markov decision processes," *Handbooks in operations research and management science*, vol. 2, pp. 331–434, 1990.

[62] S. Khodadadian, P. R. Jhunjhunwala, S. M. Varma, and S. T. Maguluri, "On the linear convergence of natural policy gradient algorithm," *arXiv preprint arXiv:2105.01424*, 2021.

[63] J. J. Rissanen, "Fisher information and stochastic complexity," *IEEE transactions on information theory*, vol. 42, no. 1, pp. 40–47, 1996.

[64] Z. Fu, Z. Yang, and Z. Wang, "Single-timescale actor-critic provably finds globally optimal policy," *arXiv preprint arXiv:2008.00483*, 2020.

[65] D. A. Levin and Y. Peres, *Markov chains and mixing times*. American Mathematical Soc., 2017, vol. 107.

[66] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[67] N. Bansal and A. Gupta, "Potential-function proofs for gradient methods," *Theory of Computing*, vol. 15, no. 1, pp. 1–32, 2019.

[68] S. Kakade and J. Langford, "Approximately optimal approximate reinforcement learning," in *ICML*, vol. 2, 2002, pp. 267–274.

[69] J. Bhandari, D. Russo, and R. Singal, "A finite time analysis of temporal difference learning with linear function approximation," in *Conference on learning theory*. PMLR, 2018, pp. 1691–1692.

[70] A. Davis, "Markov chains as random input automata," *The American Mathematical Monthly*, vol. 68, no. 3, pp. 264–267, 1961.

[71] A. Beck, *First-order methods in optimization*. SIAM, 2017.

Sajad Khodadadian is a PhD student in the Milton Stewart School of Industrial and Systems Engineering at Georgia Tech. He received his bachelor degree at Sharif University of Technology, in Electrical Engineering and Physics. His research interests include Reinforcement Learning Theory and Algorithms, Non-convex Optimization, Stochastic Approximation, and Causality.

Thinh T. Doan is an Assistant Professor in the Department of Electrical and Computer Engineering at Virginia Tech. He obtained his Ph.D. degree at the University of Illinois, Urbana-Champaign, his master degree at the University of Oklahoma, and his bachelor degree at Hanoi University of Science and Technology, Vietnam, all in Electrical Engineering. His research interests span on the intersection of control theory, optimization, machine learning, reinforcement learning, and applied probability theory.

Justin Romberg is the Schlumberger Professor in the School of Electrical and Computer Engineering at the Georgia Institute of Technology, where he has been on the faculty since 2006. Dr. Romberg received the B.S.E.E. (1997), M.S. (1999) and Ph.D. (2004) degrees from Rice University in Houston, Texas; in 2010, he was named a Rice University Outstanding Young Engineering Alumnus. From Fall 2003 until Fall 2006, he was a Postdoctoral Scholar in Applied and Computational Mathematics at the California Institute of Technology. His current research interest lie at the intersection of statistical signal processing, machine learning, optimization, and applied probability

Siva Theja Maguluri is Fouts Family Early Career Professor and Assistant Professor in the School of Industrial and Systems Engineering at Georgia Tech. He obtained his Ph.D. and MS in ECE as well as MS in Applied Math from UIUC, and B.Tech in Electrical Engineering from IIT Madras. His research interests span the areas of Networks, Control, Optimization, Algorithms, Applied Probability and Reinforcement Learning. He is a recipient of the biennial "Best Publication in Applied Probability" award in 2017, "CTL/BP Junior Faculty Teaching Excellence Award" in 2020 and "Student Recognition of Excellence in Teaching: Class of 1934 CIOS Award" in 2020.

APPENDIX

The supplementary material is organized as follows: in Section I the details of the Proof of Theorem 1 is presented, and in Section II details of the proof of Proposition 1 is provided.

APPENDIX I
DETAILS OF THE PROOF OF THEOREM 1

A. Proof of Useful Lemmas

Proof of Lemma 3:

$$\begin{aligned}\|\theta_{t+1}-\theta_t\|^2 &= \|Q_{t+1}-Q_t+Q^{\hat{\pi}_{t-1}}-Q^{\hat{\pi}_t}\|^2\\ &\le 2\|Q_{t+1}-Q_t\|^2+2\|Q^{\hat{\pi}_{t-1}}-Q^{\hat{\pi}_t}\|^2\\ &\overset{(a)}{\le} 2\alpha_t^2\Delta_Q^2+2{L_2}^2\|\hat{\pi}_{t-1}-\hat{\pi}_t\|^2\\ &\overset{(b)}{\le} 2\alpha_t^2\Delta_Q^2+2{L_2}^2\left(L_3\frac{\epsilon_{t-2}}{t-1}+L_1\beta_{t-1}\right)^2\\ &\le 2\alpha_t^2\Delta_Q^2+4{L_2}^2{L_3}^2\frac{\epsilon_{t-2}^2}{(t-1)^2}+4{L_2}^2{L_1}^2\beta_{t-1}^2,\end{aligned}$$

where (a) is due to Lemmas 8 and 16, and (b) is due to Lemma 9. □

Proof of Lemma 4: We prove this lemma for a slightly more general case. Assume a finite state Markov chain $\{X_k\}_{k=0,1,\dots}$ with state space $\mathcal{X}=\{x_1,x_2,\dots,x_{|\mathcal{X}|}\}$ and stationary distribution ν. Define $M:=diag(\nu)$ a diagonal matrix with diagonal entries equal to the elements of ν. Clearly, $M=M^\top$. Further denote P as the transition matrix of the Markov chain. Define $V=\gamma P-I$, where I is the identity matrix. Assuming $X_k\sim\nu$, for any function $F(\cdot):\mathcal{X}\to\mathbb{R}$, we have:

$$\mathbb{E}\left[F\left(X_k\right)^2\right]=\mathbb{E}\left[F\left(X_{k+1}\right)^2\right].$$

By Cauchy–Schwarz inequality, we have:

$$\begin{aligned}\mathbb{E}\left[F\left(X_k\right)F\left(X_{k+1}\right)\right] &\le \sqrt{\mathbb{E}\left[F^2(X_k)\right]}.\sqrt{\mathbb{E}\left[F^2(X_{k+1})\right]}\\ &=\mathbb{E}\left[F^2(X_k)\right].\end{aligned}\tag{32}$$

Denoting $F=[F(x_1);F(x_2);\dots;F(x_{|\mathcal{X}|})]$ as a $|\mathcal{X}|$ dimentional vector, we have:

$$\mathbb{E}[F^2(X_k)]=\sum_{x\in\mathcal{X}}\nu(x)F^2(x)=F^\top MF,\tag{33}$$

$$\begin{aligned}\mathbb{E}[F(X_k)F(X_{k+1})] &= \sum_{x,y\in\mathcal{X}}\nu(x)P(y|x)F(x)F(y)\\ &=F^\top MPF=F^\top P^\top MF,\end{aligned}\tag{34}$$

where the last equality is due to $\mathbb{E}[F(X_k)F(X_{k+1})]$ being a scalar. Combining (32), (33), and (34), we have:

$$\begin{aligned}&F^\top MPF\le F^\top MF,\quad\forall F\\ \implies\ &MP\le M\\ \implies\ &M(\gamma P-I)\le-(1-\gamma)M.\end{aligned}\tag{35}$$

Next, in the case of MDP, for a fixed policy π, we define $M^\pi\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$ and $P^\pi\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$ matrices as follows:

$$\begin{aligned}M^\pi_{(s,a),(s',a')} &= \begin{cases}\mu^\pi(s)\pi(a|s) & (s,a)=(s',a'),\\ 0 & o.w\end{cases}\\ P^\pi_{(s,a),(s',a')} &= \mathcal{P}(s'|s,a)\pi(a'|s').\end{aligned}$$

It is easy to see that:

$$\begin{aligned}&\bar{A}^\pi_{(s,a),(s',a')}\\ &=\begin{cases}\mu^\pi(s)\pi(a|s)\left(\gamma\mathcal{P}\left(s'|s,a\right)\pi\left(a'|s'\right)-1\right) & s=s',a=a',\\ \gamma\mu^\pi(s)\pi(a|s)\mathcal{P}\left(s'|s,a\right)\pi\left(a'|s'\right) & s\ne s' \text{ or } a\ne a'\end{cases}\\ &\implies \bar{A}^\pi=M^\pi(\gamma P^\pi-I)\le-(1-\gamma)M^\pi,\end{aligned}$$

where the last inequality follows from (35). As a result, we have:

$$\begin{aligned}\theta_t^\top\bar{A}^{\hat{\pi}_{t-1}}\theta_t &\le -(1-\gamma)\sum_{s,a}\mu^\pi(a)\hat{\pi}_{t-1}(a|s)\theta_{t,s,a}^2\\ &\le -(1-\gamma)\frac{\epsilon_{t-2}}{|\mathcal{A}|}\mu||\theta_t||^2,\end{aligned}$$

where the last inequality follows from $\hat{\pi}_{t-1}(a|s)\ge\frac{\epsilon_{t-2}}{|\mathcal{A}|}$ and Lemma 10. □

Proof of Lemma 5:

1) By the assumption on the reward function $\mathcal{R}(s,a)\ge0$, we have $V^\pi(s)=\mathbb{E}\left[\sum_{k=0}^\infty\gamma^k\mathcal{R}(S_k,A_k)\,|\,S_0=s\right]\ge0$. Furthermore, due to $\mathcal{R}(s,a)\le1$, we have $V^\pi(s)=\mathbb{E}\left[\sum_{k=0}^\infty\gamma^k\mathcal{R}(S_k,A_k)\,|\,S_0=s\right]\le\mathbb{E}\left[\sum_{k=0}^\infty\gamma^k\,|\,S_0=s\right]=\frac{1}{1-\gamma}$ for all $s\in\mathcal{S}$.
2) Similarly, we have $Q^\pi(s,a)\in[0,\frac{1}{1-\gamma}]$ for all s,a.
3) $\|Q^\pi\|=\sqrt{\sum_{s,a}Q^{\pi^2}(s,a)}\le\frac{\sqrt{|\mathcal{S}||\mathcal{A}|}}{1-\gamma}$
4) In order to prove this, first we show $\|Q_t\|_\infty\le\frac{1}{1-\gamma}$ for all $t\ge0$. We construct this bound by induction. Due to the initialization, the inequality holds for $t=0$. Assuming the inequality holds for t, we prove it holds for $t+1$. For all s,a, we have:

$$\begin{aligned}&|Q_{t+1}(s,a)|=\Big|(1-\alpha_t(s,a))Q_t(s,a)\\ &\quad+\alpha_t(s,a)(\mathcal{R}(s,a)+\gamma Q_t(s_{t+1},a_{t+1}))\Big|\\ &\le(1-\alpha_t(s,a))|Q_t(s,a)|\\ &\quad+\alpha_t(s,a)|\mathcal{R}(s,a)+\gamma Q_t(S_{t+1},A_{t+1})|\\ &\le(1-\alpha_t(s,a))Q_{\max}+\alpha_t(s,a)(1+\gamma Q_{\max})\\ &=(1-\alpha_t(s,a))\frac{1}{1-\gamma}+\alpha_t(s,a)(1+\gamma\frac{1}{1-\gamma})=\frac{1}{1-\gamma}.\end{aligned}$$

The bound for $\|Q_t\|$ follows directly. □

Proof of Lemma 6: Given time indices t and $\tau < t$, we consider the following auxiliary chain of state-actions:

$$S_{t-\tau} \xrightarrow{\hat{\pi}_{t-\tau-1}} A_{t-\tau} \xrightarrow{\mathcal{P}} \tilde{S}_{t-\tau+1} \xrightarrow{\hat{\pi}_{t-\tau-1}} \tilde{A}_{t-\tau+1} \dots \xrightarrow{\mathcal{P}} \tilde{S}_t \xrightarrow{\hat{\pi}_{t-\tau-1}} \tilde{A}_t \xrightarrow{\mathcal{P}} \tilde{S}_{t+1} \xrightarrow{\hat{\pi}_{t-\tau-1}} \tilde{A}_{t+1}.$$

Note that the original chain is as follows:

$$S_{t-\tau} \xrightarrow{\hat{\pi}_{t-\tau-1}} A_{t-\tau} \xrightarrow{\mathcal{P}} S_{t-\tau+1} \xrightarrow{\hat{\pi}_{t-\tau}} A_{t-\tau+1} \dots \xrightarrow{\mathcal{P}} S_t \xrightarrow{\hat{\pi}_{t-1}} A_t \xrightarrow{\mathcal{P}} S_{t+1} \xrightarrow{\hat{\pi}_t} A_{t+1}.$$

Further, we define $\tilde{O}_t = (\tilde{S}_t, \tilde{A}_t, \tilde{S}_{t+1}, \tilde{A}_{t+1})$. We have:

$$\begin{aligned}
\Gamma(\hat{\pi}_{t-1},\theta_t,O_t) &= \Gamma(\hat{\pi}_{t-1},\theta_t,O_t) - \Gamma(\hat{\pi}_{t-\tau-1},\theta_t,O_t) && (36)\\
&+ \Gamma(\hat{\pi}_{t-\tau-1},\theta_t,O_t) - \Gamma(\hat{\pi}_{t-\tau-1},\theta_{t-\tau},O_t) && (37)\\
&+ \Gamma(\hat{\pi}_{t-\tau-1},\theta_{t-\tau},O_t) - \Gamma(\hat{\pi}_{t-\tau-1},\theta_{t-\tau},\tilde{O}_t) && (38)\\
&+ \Gamma(\hat{\pi}_{t-\tau-1},\theta_{t-\tau},\tilde{O}_t). && (39)
\end{aligned}$$

We bound each of the terms above separately. Firstly:

$$\begin{aligned}
&\Gamma(\hat{\pi}_{t-1},\theta_t,O_t) - \Gamma(\hat{\pi}_{t-\tau-1},\theta_t,O_t) \overset{(a)}{\leq} K_1\|\hat{\pi}_{t-1} - \hat{\pi}_{t-\tau-1}\|\\
&\leq K_1 \sum_{i=t-\tau}^{t-1} \|\hat{\pi}_i - \hat{\pi}_{i-1}\| \overset{(b)}{\leq} K_1 \sum_{i=t-\tau}^{t-1} \left[L_3\frac{\epsilon_{i-2}}{i-1} + L_1\beta_{i-1}\right]\\
&\leq K_1\tau\left[L_3\frac{\epsilon_{t-\tau-2}}{t-\tau-1} + L_1\beta_{t-\tau-1}\right],
\end{aligned}$$

where (a) is due to Lemma 11, and (b) is due to Lemma 9. Second, we have:

$$\begin{aligned}
&\Gamma(\hat{\pi}_{t-\tau-1},\theta_t,O_t) - \Gamma(\hat{\pi}_{t-\tau-1},\theta_{t-\tau},O_t)\\
&\overset{(a)}{\leq} K_2\|\theta_t - \theta_{t-\tau}\|\\
&\leq K_2 \sum_{i=t-\tau+1}^{t} \|\theta_i - \theta_{i-1}\|\\
&\overset{(b)}{\leq} K_2 \sum_{i=t-\tau+1}^{t} \Delta_Q\alpha_{i-1} + L_2L_3\frac{\epsilon_{i-3}}{i-2} + L_1L_2\beta_{i-2}\\
&\leq K_2\tau\left[\Delta_Q\alpha_{t-\tau} + L_2L_3\frac{\epsilon_{t-\tau-2}}{t-\tau-1} + L_1L_2\beta_{t-\tau-1}\right],
\end{aligned}$$

where (a) due to Lemma 12, and (b) is due to Lemma 16. Third, denoting $\mathcal{F}_{t-\tau} := \{S_{t-\tau}, \hat{\pi}_{t-\tau-1}, \theta_{t-\tau}\}$, we have:

$$\begin{aligned}
&\mathbb{E}\left[\Gamma(\hat{\pi}_{t-\tau-1},\theta_{t-\tau},O_t) - \Gamma(\hat{\pi}_{t-\tau-1},\theta_{t-\tau},\tilde{O}_t)\middle|\mathcal{F}_{t-\tau}\right]\\
&\overset{(a)}{\leq} C_u\mathbb{E}\left[\sum_{i=t-\tau}^{t} \|\hat{\pi}_i - \hat{\pi}_{t-\tau-1}\|\middle|\mathcal{F}_{t-\tau}\right]\\
&\overset{(b)}{\leq} C_u(\tau+1)^2\left[L_3\frac{\epsilon_{t-\tau-2}}{t-\tau-1} + L_1\beta_{t-\tau-1}\right],
\end{aligned}$$

where (a) is due to Lemma 13 and (b) is due to Lemma 15. Finally, by Lemma 14 we have:

$$\mathbb{E}\left[\Gamma(\hat{\pi}_{t-\tau-1},\theta_{t-\tau},\tilde{O}_t)\middle|\mathcal{F}_{t-\tau}\right] \leq C_bm\rho^\tau.$$

Combining the bounds above, and noticing $\tau \geq 1$, we get the result. □

Proof of Lemma 7: Suppose π be an arbitrary stochastic policy. π can be written as a $|\mathcal{S}|$ by $|\mathcal{A}|$ stochastic matrix, which has non-negative elements, and each row sums up to one. Hence, by [70, Theorem 1], π can be written as a convex combination of at most $N = |\mathcal{S}|(|\mathcal{A}|-1)+1$ deterministic policies $\{\pi_i\}_{i=1}^{|\mathcal{S}|(|\mathcal{A}|-1)+1}$. In other words, there exist coefficients $\{\alpha_i\}_{i=1}^N$, such that $\alpha_i \geq 0$ and $\sum_i \alpha_i = 1$, and $\pi = \sum_i \alpha_i\pi_i$. By definition of P^π, we have $P^\pi = \sum_i \alpha_i P^{\pi_i}$.

Due to ergodicity Assumption 1, for every policy π_i, there exist a finite integer r_i, such that $(P^{\pi_i})^{\bar{r}_i}$ is a positive matrix with minimum element $e_i > 0$ for all $\bar{r}_i \geq r_i$. Since we have a finite number of r_i and e_i's, we have $r = \max_i r_i$ is a finite integer, and $e = \min_i e_i > 0$. Furthermore, we have

$$\begin{aligned}
[(P^\pi)^r]_{m,n} &= \left[\left(\sum_i \alpha_i P^{\pi_i}\right)^r\right]_{m,n} \overset{(a)}{\geq} \sum_i \alpha_i^r \left[(P^{\pi_i})^r\right]_{m,n}\\
&\geq Ne\frac{1}{N}\sum_i \alpha_i^r \overset{(b)}{\geq} Ne\left(\frac{1}{N}\sum_i \alpha_i\right)^r = Ne\left(\frac{1}{N}\right)^r > 0,
\end{aligned}$$

where (a) is due to non-negativity of matrices $\alpha_i P^{\pi_i}$ and (b) is by Jensen's inequality. Hence, by [65, Theorem 4.9], we can show the existence of $\rho \in (0,1)$ and $m > 0$.

Furthermore, if the underlying Markov chain under a policy π is periodic with period d, then we have $\lim_{t\to\infty} P(S_{dt} = i|S_0 = i) > 0$ while $\lim_{t\to\infty} P(S_{dt+1} = i|S_0 = i) = 0$, and hence (27) does not hold. □

Proof of Lemma 8: By the policy gradient theorem [5], we know that for any distribution μ, we have $\frac{\partial V^\pi(\mu)}{\partial \pi(a|s)} = \frac{1}{1-\gamma} d_\mu^\pi(s) Q^\pi(s,a)$. As a result:

$$\left\|\frac{\partial V^\pi(\mu)}{\partial \pi}\right\| \leq \frac{1}{1-\gamma}\sqrt{\sum_{s,a} {Q^\pi}^2(s,a)} \leq \frac{\sqrt{|\mathcal{S}||\mathcal{A}|}}{(1-\gamma)^2}.$$

Furthermore, we have:

$$\frac{\partial Q^\pi(s,a)}{\partial \pi} = \gamma\sum_{s'} \mathcal{P}(s'|s,a)\frac{\partial V^\pi(s')}{\partial \pi},$$

which implies

$$\begin{aligned}
&\left\|\frac{\partial Q^\pi(s,a)}{\partial \pi}\right\| \leq \gamma\sum_{s'} \mathcal{P}(s'|s,a)\left\|\frac{\partial V^\pi(s')}{\partial \pi}\right\| \leq \gamma\frac{\sqrt{|\mathcal{S}||\mathcal{A}|}}{(1-\gamma)^2}\\
&\implies |Q^{\pi_1}(s,a) - Q^{\pi_2}(s,a)| \leq \gamma\frac{\sqrt{|\mathcal{S}||\mathcal{A}|}}{(1-\gamma)^2}\|\pi_1 - \pi_2\|.
\end{aligned}$$

Using this, we have:

$$\begin{aligned}
\|Q^{\pi_1} - Q^{\pi_2}\| &\leq \sqrt{\sum_{s,a} \frac{\gamma^2|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4}\|\pi_1 - \pi_2\|^2}\\
&= \frac{\gamma|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}\|\pi_1 - \pi_2\| = L_2\|\pi_1 - \pi_2\|,
\end{aligned}$$

where $L_2 := \frac{\gamma|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$. □

Proof of Lemma 9: Policy π_t can be parameterized by the vector $\theta^t \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ as $\pi_t(a|s) = \frac{\exp(\theta^t_{s,a})}{\sum_{a'}\exp(\theta^t_{s,a'})}$. It is

straightforward to see that the multiplicative weight update of the policy in Algorithm 1 is equivalent to [51, Lemma 3.1]

$$\theta^{t+1} = \theta^t + \beta_t Q_{t+1}.$$

We have

$$\begin{aligned}\|\pi_{t+1} - \pi_t\|^2 &= \sum_s \|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|^2 \\ &\overset{(a)}{\leq} \sum_s \|\beta_t Q_{t+1}(s,\cdot)\|^2 \leq \beta_t^2 |\mathcal{A}||\mathcal{S}| Q_{\max}^2,\end{aligned} \tag{40}$$

where in (a) we use 1-Lipschitzness of the softmax function [71].

As a result, we have:

$$\begin{aligned}&\|\hat{\pi}_{t+1} - \hat{\pi}_t\| \\ &= \|(\epsilon_t - \epsilon_{t-1})(\frac{1}{\mathcal{A}} - \pi_{t+1}) + (1-\epsilon_{t-1})(\pi_{t+1} - \pi_t)\| \\ &\overset{(a)}{\leq} |\epsilon_t - \epsilon_{t-1}|\sqrt{|\mathcal{S}|}(\frac{1}{\sqrt{|\mathcal{A}|}} + 1) + \|\pi_{t+1} - \pi_t\| \\ &\overset{(b)}{\leq} \frac{\xi \epsilon_{t-1}}{t}\sqrt{|\mathcal{S}|}(\frac{1}{\sqrt{|\mathcal{A}|}} + 1) + \beta_t Q_{\max}\sqrt{|\mathcal{A}||\mathcal{S}|} \\ &= L_3 \frac{\epsilon_{t-1}}{t} + L_1 \beta_t,\end{aligned}$$

where (a) is due to triangle inequality and (b) is due to the assumption on ϵ_t and (40). Here $L_3 = \xi\sqrt{|\mathcal{S}|}(\frac{1}{\sqrt{|\mathcal{A}|}} + 1)$ and $L_1 = Q_{\max}\sqrt{|\mathcal{A}||\mathcal{S}|}$.□

Proof of Lemma 10: Ergodicity assumption 1 implies that the underlying Markov chain induced by all the policies is irreducible. The proof follows from [65, Proposition 1.14]. □

B. Auxiliary Lemmas

Lemma 11: For any π_1, π_2, θ, and $O = (S, A, S', A')$,

$$|\Gamma(\pi_1, \theta, O) - \Gamma(\pi_2, \theta, O)| \leq K_1 \|\pi_1 - \pi_2\|,$$

where $K_1 = 2Q_{\max}\sqrt{2|\mathcal{S}||\mathcal{A}|}L_2 + 8Q_{\max}^2|\mathcal{S}|^2|\mathcal{A}|^3\left(\lceil \log_\rho m^{-1}\rceil + \frac{1}{1-\rho} + 2\right)$.

Lemma 12: For any π, Q_1, Q_2, and $O = (S, A, S', A')$,

$$|\Gamma(\pi, \theta_1, O) - \Gamma(\pi, \theta_2, O)| \leq K_2 \|\theta_1 - \theta_2\|,$$

where $K_2 = 1 + 9\sqrt{2|\mathcal{S}||\mathcal{A}|}Q_{\max}$.

Lemma 13: Consider original tuples $O_t = (S_t, A_t, S_{t+1}, A_{t+1})$ and the auxiliary tuples $\tilde{O}_t = (\tilde{S}_t, \tilde{A}_t, \tilde{S}_{t+1}, \tilde{A}_{t+1})$. Denote $\mathcal{F}_{t-\tau} := \{S_{t-\tau}, \hat{\pi}_{t-\tau-1}, \theta_{t-\tau}\}$. For any time indices $t > \tau > 1$, we have

$$\begin{aligned}&\mathbb{E}\left[\Gamma(\hat{\pi}_{t-\tau-1}, \theta_{t-\tau}, O_t) - \Gamma(\hat{\pi}_{t-\tau-1}, \theta_{t-\tau}, \tilde{O}_t) \mid \mathcal{F}_{t-\tau}\right] \\ &\qquad \leq C_u \mathbb{E}\left[\sum_{i=t-\tau}^{t} \|\hat{\pi}_i - \hat{\pi}_{t-\tau-1}\| \mid \mathcal{F}_{t-\tau}\right],\end{aligned}$$

where $C_u = 4Q_{\max}|\mathcal{S}|^{1.5}|\mathcal{A}|^{1.5}(1 + 3Q_{\max}|\mathcal{S}||\mathcal{A}|)$.

Lemma 14: Consider the auxiliary tuple $\tilde{O}_t = (\tilde{S}_t, \tilde{A}_t, \tilde{S}_{t+1}, \tilde{A}_{t+1})$. Denote $\mathcal{F}_{t-\tau} = \{S_{t-\tau}, \hat{\pi}_{t-\tau-1}, \theta_{t-\tau}\}$. For any time indices $t > \tau > 1$, we have:

$$\mathbb{E}\left[\Gamma(\hat{\pi}_{t-\tau-1}, \theta_{t-\tau}, \tilde{O}_t)\middle|\mathcal{F}_{t-\tau}\right] \leq C_b m \rho^\tau,$$

where $C_b = 4Q_{\max}|\mathcal{S}||\mathcal{A}|(1 + 3|\mathcal{S}||\mathcal{A}|Q_{\max})$.

Lemma 15: For any time indices $t > \tau > 1$, the policies generated by Algorithm 1 satisfy the following:

$$\sum_{i=t-\tau}^{t} \|\hat{\pi}_i - \hat{\pi}_{t-\tau-1}\| \leq (\tau+1)^2\left[L_3\frac{\epsilon_{t-\tau-2}}{t-\tau-1} + L_1\beta_{t-\tau-1}\right].$$

Lemma 16: We have the following bounds:

1) $\|A(O)\| \leq \sqrt{1+\gamma^2} \leq \sqrt{2}$,
2) $\|r(O)\| \leq 1$,
3) $\|\bar{A}^\pi\| \leq \sqrt{2}$,
4) $\|\mathbb{E}[r(O_1) - r(O_2)]\|_1 \leq 2|\mathcal{S}||\mathcal{A}| d_{TV}(O_1, O_2)$
5) $\|\mathbb{E}[A(O_1) - A(O_2)]\|_1 \leq 2|\mathcal{S}||\mathcal{A}| d_{TV}(O_1, O_2)$,
6) $\|Q_{t+1} - Q_t\| \leq \alpha_t \Delta_Q := \alpha_t(2Q_{\max} + 1)$,
7) $\|\theta_t - \theta_{t-1}\| \leq \Delta_Q \alpha_{t-1} + L_2 L_3 \frac{\epsilon_{t-3}}{t-2} + L_1 L_2 \beta_{t-2}$,

where $Q_{\max} = \frac{1}{1-\gamma}$, and the constants L_1, L_2, L_3 are defined in Lemmas 8 and 9.

Lemma 17: Consider $O_t = (S_t, A_t, S_{t+1}, A_{t+1})$ and $\tilde{O}_t = (\tilde{S}_t, \tilde{A}_t, \tilde{S}_{t+1}, \tilde{A}_{t+1})$. Denote $\mathcal{F}_{t-\tau} := \{S_{t-\tau}, \hat{\pi}_{t-\tau-1}, \theta_{t-\tau}\}$. We have:

$$\begin{aligned}&d_{TV}\left(P(O_t \in \cdot|\mathcal{F}_{t-\tau})||P(\tilde{O}_t \in \cdot|\mathcal{F}_{t-\tau})\right) \\ &\qquad \leq \sqrt{|\mathcal{A}||\mathcal{S}|}\mathbb{E}\left[\sum_{i=t-\tau}^{t} \|\hat{\pi}_i - \hat{\pi}_{t-\tau-1}\| \mid \mathcal{F}_{t-\tau}\right]\end{aligned}$$

Lemma 18 (Lemma A.1 in [50]): Denote $M = \lceil \log_\rho m^{-1}\rceil + \frac{1}{1-\rho}$. For any π_1 and π_2 policies, we have the following inequality:

$$\begin{aligned}&d_{TV}\left(\mu^{\pi_1} \otimes \pi_1 \otimes P \otimes \pi_1, \mu^{\pi_2} \otimes \pi_2 \otimes P \otimes \pi_2\right) \\ &\qquad \leq |\mathcal{A}|\,(M+2)\,\|\pi_1 - \pi_2\|\end{aligned}$$

C. Proofs of the auxiliary Lemmas

Proof of Lemma 11:

$$\begin{aligned}&\Gamma(\pi_1, \theta, O) - \Gamma(\pi_2, \theta, O) \\ &= \theta^\top A(O)(Q^{\pi_1} - Q^{\pi_2}) - \theta^\top(\bar{A}^{\pi_1} - \bar{A}^{\pi_2})\theta \\ &\overset{(a)}{\leq} \|\theta\|.\|A(O)\|.\|Q^{\pi_1} - Q^{\pi_2}\| + \|\theta\|^2.\|\bar{A}^{\pi_1} - \bar{A}^{\pi_2}\| \\ &\overset{(b)}{\leq} 2Q_{\max}\sqrt{2|\mathcal{S}||\mathcal{A}|}\|Q^{\pi_1} - Q^{\pi_2}\| + 4Q_{\max}^2|\mathcal{S}||\mathcal{A}|.\|\bar{A}^{\pi_1} - \bar{A}^{\pi_2}\| \\ &\overset{(c)}{\leq} 2Q_{\max}\sqrt{2|\mathcal{S}||\mathcal{A}|}L_2\|\pi_1 - \pi_2\| + 8Q_{\max}^2|\mathcal{S}|^2|\mathcal{A}|^2 \\ &\quad \times d_{TV}\left(\mu^{\pi_1} \otimes \pi_1 \otimes \mathcal{P} \otimes \pi_1, \mu^{\pi_2} \otimes \pi_2 \otimes \mathcal{P} \otimes \pi_2\right) \\ &\overset{(d)}{\leq} 2Q_{\max}\sqrt{2|\mathcal{S}||\mathcal{A}|}L_2\|\pi_1 - \pi_2\| \\ &\quad + 8Q_{\max}^2|\mathcal{S}|^2|\mathcal{A}|^3\left(\lceil \log_\rho m^{-1}\rceil + \frac{1}{1-\rho} + 2\right)\|\pi_1 - \pi_2\| \\ &= K_1\|\pi_1 - \pi_2\|\end{aligned}$$

where (a) is due to Cauchy–Schwarz inequality, (b) is due to Lemmas 5 and 16, (c) is due to Lemmas 8 and 16, (d) is due to Lemma 18.□

Proof of Lemma 12:

$$\begin{aligned}&|\Gamma(\pi, \theta_1, O) - \Gamma(\pi, \theta_2, O)| \\ &\overset{(a)}{\leq} (\|r(O)\| + \|A(O)\|.\|Q^\pi\|)\|\theta_1 - \theta_2\| \\ &\quad + \|A(O) - \bar{A}^\pi\|.\|\theta_1 - \theta_2\|(\|\theta_1\| + \|\theta_2\|)\end{aligned}$$

$$\overset{(b)}{\leq}(1+9\sqrt{2|\mathcal{S}||\mathcal{A}|}Q_{\max})\|\theta_1-\theta_2\|$$

where (a) follows from Cauchy–Schwarz and triangle inequality, and (b) is due to Lemmas 5 and 16. □

Proof of Lemma 13: We have:

$$\begin{aligned}
&\mathbb{E}\big[\Gamma(\hat{\pi}_{t-\tau-1},\theta_{t-\tau},O_t)-\Gamma(\hat{\pi}_{t-\tau-1},\theta_{t-\tau},\tilde{O}_t)|\mathcal{F}_{t-\tau}\big]\\
=&\theta_{t-\tau}^\top\mathbb{E}\left[r\left(O_t\right)-r(\tilde{O}_t)+\Big(A\left(O_t\right)-A(\tilde{O}_t)\Big)Q^{\hat{\pi}_{t-\tau-1}}|\mathcal{F}_{t-\tau}\right]\\
&+\theta_{t-\tau}^\top\mathbb{E}\left[A(O_t)-A(\tilde{O}_t)|\mathcal{F}_{t-\tau}\right]\theta_{t-\tau}\\
\overset{(a)}{\leq}&\|\theta_{t-\tau}\|_\infty\left\|\mathbb{E}\left[r\left(O_t\right)-r(\tilde{O}_t)|\mathcal{F}_{t-\tau}\right]\right\|_1\\
&+\|\theta_{t-\tau}\|_\infty\left\|\mathbb{E}\left[\Big(A\left(O_t\right)-A(\tilde{O}_t)\Big)|\mathcal{F}_{t-\tau}\right]Q^{\hat{\pi}_{t-\tau-1}}\right\|_1\\
&+\|\theta_{t-\tau}\|_\infty\left\|\mathbb{E}\left[A\left(O_t\right)-A(\tilde{O}_t)|\mathcal{F}_{t-\tau}\right]\theta_{t-\tau}\right\|_1\\
\overset{(b)}{\leq}&\|\theta_{t-\tau}\|_\infty\left\|\mathbb{E}\left[r\left(O_t\right)-r(\tilde{O}_t)|\mathcal{F}_{t-\tau}\right]\right\|_1\\
&+\|\theta_{t-\tau}\|_\infty\left\|\mathbb{E}\left[\Big(A\left(O_t\right)-A(\tilde{O}_t)\Big)|\mathcal{F}_{t-\tau}\right]\right\|_1\left\|Q^{\hat{\pi}_{t-\tau-1}}\right\|_1\\
&+\|\theta_{t-\tau}\|_\infty.\left\|\mathbb{E}\left[A\left(O_t\right)-A(\tilde{O}_t)|\mathcal{F}_{t-\tau}\right]\right\|_1.\|\theta_{t-\tau}\|_1\\
\overset{(c)}{\leq}&2Q_{\max}\times2|\mathcal{S}||\mathcal{A}|d_{TV}\left(O_t,\tilde{O}_t|\mathcal{F}_{t-\tau}\right)\\
&+2Q_{\max}\times2|\mathcal{S}||\mathcal{A}|d_{TV}\left(O_t,\tilde{O}_t|\mathcal{F}_{t-\tau}\right)\times Q_{\max}|\mathcal{S}||\mathcal{A}|\\
&+2Q_{\max}\times2|\mathcal{S}||\mathcal{A}|d_{TV}\left(O_t,\tilde{O}_t|\mathcal{F}_{t-\tau}\right)\times2Q_{\max}|\mathcal{S}||\mathcal{A}|\\
=&4Q_{\max}|\mathcal{S}||\mathcal{A}|(1+3Q_{\max}|\mathcal{S}||\mathcal{A}|)d_{TV}\left(O_t,\tilde{O}_t|\mathcal{F}_{t-\tau}\right)
\end{aligned}$$

where (a) is due to the Hölder's inequality, (b) is due to definition of matrix induced norm, (c) is due to Lemma 16. Using Lemma 17, we get the result. □

Proof of Lemma 14: Consider the tuple $O'^t=(S'_t,A'_t,S'_{t+1},A'_{t+1})$, where $S'_t\sim\mu^{\hat{\pi}_{t-\tau-1}}$, $A'_t\sim\hat{\pi}_{t-\tau-1}(\cdot|S'_t)$, $S'_{t+1}\sim\mathcal{P}(\cdot|S'_t,A'_t)$, and $A'_{t+1}\sim\hat{\pi}_{t-\tau-1}(\cdot|S'_{t+1})$. We have

$$\begin{aligned}
&\mathbb{E}\big[\Gamma(\hat{\pi}_{t-\tau-1},\theta_{t-\tau},O'^t)\big|\mathcal{F}_{t-\tau}\big]\\
&=\theta_{t-\tau}^\top\mathbb{E}\left[r(O'_t)+A(O'_t)Q^{\hat{\pi}_{t-\tau-1}}|\mathcal{F}_{t-\tau}\right]\\
&\quad+\theta_{t-\tau}^\top\mathbb{E}\left[A(O'_t)|\mathcal{F}_{t-\tau}\right]\theta_{t-\tau}-\theta_{t-\tau}^\top\bar{A}^{\hat{\pi}_{t-\tau-1}}\theta_{t-\tau}=0,
\end{aligned}$$

where the last equality is due to the Bellman equation and the definition of $\bar{A}^{\pi_{t-\tau-1}}$. As a result, we have:

$$\begin{aligned}
&\mathbb{E}\big[\Gamma(\hat{\pi}_{t-\tau-1},\theta_{t-\tau},\tilde{O}_t)\big|\mathcal{F}_{t-\tau}\big]\\
&=\mathbb{E}\left[\Gamma(\hat{\pi}_{t-\tau-1},\theta_{t-\tau},\tilde{O}_t)-\Gamma(\hat{\pi}_{t-\tau-1},\theta_{t-\tau},O'^t)\big|\mathcal{F}_{t-\tau}\right]\\
&\overset{(a)}{\leq}\|\theta_{t-\tau}\|_\infty\left\|\mathbb{E}\left[r(\tilde{O}_t)-r\left(O'_t\right)|\mathcal{F}_{t-\tau}\right]\right\|_1\\
&+\|\theta_{t-\tau}\|_\infty\left\|\mathbb{E}\left[A(\tilde{O}_t)-A(O'_t)\big|\mathcal{F}_{t-\tau}\right]\right\|_1\left\|Q^{\hat{\pi}_{t-\tau-1}}\right\|_1\\
&+\|\theta_{t-\tau}\|_\infty.\left\|\mathbb{E}\left[A(\tilde{O}_t)-A(O'_t)\big|\mathcal{F}_{t-\tau}\right]\right\|_1.\|\theta_{t-\tau}\|_1\\
&\overset{(b)}{\leq}2Q_{\max}\times2|\mathcal{S}||\mathcal{A}|d_{TV}\left(\tilde{O}_t,O'_t|\mathcal{F}_{t-\tau}\right)\\
&+2Q_{\max}\times2|\mathcal{S}||\mathcal{A}|d_{TV}(\tilde{O}_t,O'_t|\mathcal{F}_{t-\tau})\times3|\mathcal{S}||\mathcal{A}|Q_{\max}\\
&\overset{(c)}{=}C_b\sum_{s,a,s',a'}|P(\tilde{S}_t=s|\mathcal{F}_{t-\tau})\hat{\pi}_{t-\tau-1}(a|s)\mathcal{P}(s'|s,a)\hat{\pi}_{t-\tau-1}(a'|s')\\
&\quad-P(S'_t=s|\mathcal{F}_{t-\tau})\hat{\pi}_{t-\tau-1}(a|s)\mathcal{P}(s'|s,a)\hat{\pi}_{t-\tau-1}(a'|s')|\\
&=C_b\sum_{s,a,s',a'}\hat{\pi}_{t-\tau-1}(a|s)\mathcal{P}(s'|s,a)\hat{\pi}_{t-\tau-1}(a'|s')\\
&\times|P(\tilde{S}_t=s|\mathcal{F}_{t-\tau})-P(S'_t=s|\mathcal{F}_{t-\tau})|\\
&=C_b\sum_s|P(\tilde{S}_t=s|\mathcal{F}_{t-\tau})-P(S'_t=s|\mathcal{F}_{t-\tau})|\\
&\overset{(d)}{\leq}C_bm\rho^\tau,
\end{aligned}$$

where (a) follows from Hölder's inequality and the definition of the matrix norm, (b) follows from Lemma 5, in (c) we defined $C_b=4Q_{\max}|\mathcal{S}||\mathcal{A}|(1+3|\mathcal{S}||\mathcal{A}|Q_{\max})$, and (d) is due to the Lemma 7. □

Proof of Lemma 15:

$$\begin{aligned}
&\sum_{i=t-\tau}^{t}\|\hat{\pi}_i-\hat{\pi}_{t-\tau-1}\|=\sum_{i=t-\tau}^{t}\left\|\sum_{j=t-\tau}^{i}\hat{\pi}_j-\hat{\pi}_{j-1}\right\|\\
\leq&\sum_{i=t-\tau}^{t}\sum_{j=t-\tau}^{i}\|\hat{\pi}_j-\hat{\pi}_{j-1}\|\overset{(a)}{\leq}\sum_{i=t-\tau}^{t}\sum_{j=t-\tau}^{i}\left[L_3\frac{\epsilon_{j-2}}{j-1}+L_1\beta_{j-1}\right]\\
\leq&(\tau+1)^2\left[L_3\frac{\epsilon_{t-\tau-2}}{t-\tau-1}+L_1\beta_{t-\tau-1}\right]
\end{aligned}$$

where (a) follows from Lemma 9. □

Proof of Lemma 16:

1) The proof follows directly by Frobenius norm upper bound on the two norm of a matrix.
2) Follows directly from assumption $\mathcal{R}(s,a)\leq1\ \forall s,a$.
3) $\|\bar{A}^\pi\|=\|\mathbb{E}_\pi A(O)\|\leq\mathbb{E}_\pi\|A(O)\|\leq\sqrt{2}$.
4) $\|\mathbb{E}[r\left(O_1\right)-r(O_2)]\|_1=\sum_{s,a}\left|\mathbb{E}\left(r(O_1)-r(O_2)\right)_{s,a}\right|$ $\leq2|\mathcal{S}||\mathcal{A}|d_{TV}(O_1,O_2)$
   where the inequality is due to $|r(O)_{s,a}|\leq1$.
5) $\|\mathbb{E}[A\left(O_1\right)-A(O_2)]\|_1$
   $\overset{(a)}{=}\max_{s',a'}\sum_{s,a}\left|\mathbb{E}\left(A(O_1)-A(O_2)\right)_{s,a,s',a'}\right|$
   $\overset{(b)}{\leq}\max_{s',a'}2|\mathcal{S}||\mathcal{A}|d_{TV}(O_1,O_2)=2|\mathcal{S}||\mathcal{A}|d_{TV}(O_1,O_2)$,
   where (a) is due to the definition of matrix norm, and (b) is due to $|A(O)_{s,a,s',a'}|\leq1$.
6) $\|Q_{t+1}-Q_t\|\leq\sqrt{\sum_{s,a}\alpha_t^2(s,a)(2Q_{\max}+1)^2}=\alpha_t(2Q_{\max}+1)$
7) $\|\theta_t-\theta_{t-1}\|\leq\|Q_t-Q_{t-1}\|+\left\|Q^{\hat{\pi}_{t-1}}-Q^{\hat{\pi}_{t-2}}\right\|$
   $\leq\Delta_Q\alpha_{t-1}+L_2\left(L_3\frac{\epsilon_{t-3}}{t-2}+L_1\beta_{t-2}\right)$
   where the last inequality follows from the previous part, and Lemmas 8 and 9.□

Proof of Lemma 17:

$$\begin{aligned}
&d_{TV}(P(O_t\in\cdot\mid\mathcal{F}_{t-\tau})||P(\tilde{O}_t\in\cdot\mid\mathcal{F}_{t-\tau}))\\
=&\sum_{s,a,s',a'}\Big|P(\overbrace{S_t=s,A_t=a}^{\mathcal{H}_t},S_{t+1}=s',A_{t+1}=a'|\mathcal{F}_{t-\tau})\\
&-P(\tilde{S}_t=s,\tilde{A}_t=a,\tilde{S}_{t+1}=s',\tilde{A}_{t+1}=a'|\mathcal{F}_{t-\tau})\Big|\\
=&\sum_{s,a,s',a'}\big|\mathbb{E}[\hat{\pi}_t(a'|s')|\mathcal{F}_{t-\tau},\mathcal{H}_t]\mathcal{P}(s'|s,a)P(S_t=s,A_t=a|\mathcal{F}_{t-\tau})\\
&-\hat{\pi}_{t-\tau-1}(a'|s')\mathcal{P}(s'|s,a)P(\tilde{S}_t=s,\tilde{A}_t=a|\mathcal{F}_{t-\tau})\big|
\end{aligned}$$

$$\leq \sum_{s,a,s',a'} \mathcal{P}(s'|s,a)P(S_t=s,A_t=a|\mathcal{F}_{t-\tau}) \times \left|\mathbb{E}[\hat{\pi}_t(a'|s')|\mathcal{F}_{t-\tau},\mathcal{H}_t]-\hat{\pi}_{t-\tau-1}(a'|s')\right| \quad (I_1)$$

$$+\sum_{s,a}\left|P(S_t=s,A_t=a|\mathcal{F}_{t-\tau})-P(\tilde{S}_t=s,\tilde{A}_t=a|\mathcal{F}_{t-\tau})\right|. \quad (I_2)$$

We bound I_1 and I_2 separately:

$$\begin{aligned}
I_1 &\leq \sum_{s,a,s',a'} \mathcal{P}(s'|s,a)P(S_t=s,A_t=a|\mathcal{F}_{t-\tau}) \\
&\times \mathbb{E}[|\hat{\pi}_t(a'|s')-\hat{\pi}_{t-\tau-1}(a'|s')|\big|\mathcal{F}_{t-\tau},\mathcal{H}_t] \\
&\leq \sum_{s,a,s',a'} P(S_t=s,A_t=a|\mathcal{F}_{t-\tau}) \\
&\times \mathbb{E}[|\hat{\pi}_t(a'|s')-\hat{\pi}_{t-\tau-1}(a'|s')|\big|\mathcal{F}_{t-\tau},\mathcal{H}_t] \\
&= \sum_{s',a'} \mathbb{E}[|\hat{\pi}_t(a'|s')-\hat{\pi}_{t-\tau-1}(a'|s')|\big|\mathcal{F}_{t-\tau}] \\
&\leq \sqrt{|\mathcal{A}||\mathcal{S}|}\mathbb{E}[\|\hat{\pi}_t-\hat{\pi}_{t-\tau-1}\|\big|\mathcal{F}_{t-\tau}],
\end{aligned}$$

$$\begin{aligned}
I_2 =& \sum_{s,a}\Big|\sum_{s'',a''}P(S_t=s,A_t=a,S_{t-1}=s'',A_{t-1}=a''|\mathcal{F}_{t-\tau}) \\
&- P(\tilde{S}_t=s,\tilde{A}_t=a,\tilde{S}_{t-1}=s'',\tilde{A}_{t-1}=a''|\mathcal{F}_{t-\tau})\Big| \\
\leq& \sum_{s,a,s'',a''}\Big|P(S_t=s,A_t=a,S_{t-1}=s'',A_{t-1}=a''|\mathcal{F}_{t-\tau}) \\
&- P(\tilde{S}_t=s,\tilde{A}_t=a,\tilde{S}_{t-1}=s'',\tilde{A}_{t-1}=a''|\mathcal{F}_{t-\tau})\Big| \\
=& d_{TV}(P(O_{t-1}\in\cdot|\mathcal{F}_{t-\tau})||P(\tilde{O}_{t-1}\in\cdot|\mathcal{F}_{t-\tau})).
\end{aligned}$$

Combining the above bounds, we get:

$$\begin{aligned}
&d_{TV}(P(O_t\in\cdot|\mathcal{F}_{t-\tau})||P(\tilde{O}_t\in\cdot|\mathcal{F}_{t-\tau})) \\
\leq&\sqrt{|\mathcal{A}||\mathcal{S}|}\mathbb{E}\left[\|\hat{\pi}_t-\hat{\pi}_{t-\tau-1}\|\Big|\mathcal{F}_{t-\tau}\right] \\
&+ d_{TV}(P(O_{t-1}\in\cdot|\mathcal{F}_{t-\tau})||P(\tilde{O}_{t-1}\in\cdot|\mathcal{F}_{t-\tau})).
\end{aligned}$$

Following this induction, and noting that $P(S_{t-\tau}=s,A_{t-\tau}=a|\mathcal{F}_{t-\tau}) = P(\tilde{S}_{t-\tau}=s,\tilde{A}_{t-\tau}=a|\mathcal{F}_{t-\tau})$ (due to the definition of $\tilde{S}$ and $\tilde{A}$), we get the result. □

Proof of Lemma 18: The proof follows directly from Lemma A.1 in [50]. □

APPENDIX II
DETAILS OF THE PROOF OF PROPOSITION 1

A. Useful lemmas

Proof of Lemma 1: We have:

$$\begin{aligned}
\log Z_t(s) &= \log\sum_{a'}\pi_t(a|s)\exp(\beta_t Q_{t+1}(s,a)) \\
&\geq \sum_a \pi_t(a|s)\beta_t Q_{t+1}(s,a), \quad (41)
\end{aligned}$$

where the inequality is due to the concavity of $\log(\cdot)$ function and Jensen's inequality. Furthermore, we have:

$$\begin{aligned}
&V^{\pi_{t+1}}(\mu)-V^{\pi_t}(\mu) \\
\overset{(a)}{=}&\frac{1}{1-\gamma}\sum_{s,a}d_\mu^{\pi_{t+1}}(s)\pi_{t+1}(a|s)\big[Q_{t+1}(s,a)+Q^{\pi_t}(s,a) \\
&-Q_{t+1}(s,a)-V^{\pi_t}(s)\big] \\
\overset{(b)}{=}&\frac{1}{1-\gamma}\sum_{s,a}d_\mu^{\pi_{t+1}}(s)\pi_{t+1}(a|s)\left[\frac{1}{\beta_t}\log\frac{\pi_{t+1}(a|s)}{\pi_t(a|s)}\right. \\
&\left.+\frac{1}{\beta_t}\log Z_t(s)+Q^{\pi_t}(s,a)-Q_{t+1}(s,a)-V^{\pi_t}(s)\right] \\
\overset{(c)}{\geq}&\frac{1}{1-\gamma}\sum_{s,a}d_\mu^{\pi_{t+1}}(s)\pi_{t+1}(a|s)\left[\frac{1}{\beta_t}\log Z_t(s)+Q^{\pi_t}(s,a)\right. \\
&\left.-Q_{t+1}(s,a)-V^{\pi_t}(s)\right] \\
=&\frac{1}{1-\gamma}\left[\sum_{s,a}d_\mu^{\pi_{t+1}}(s)\pi_t(a|s)\left[\frac{1}{\beta_t}\log Z_t(s)-Q_{t+1}(s,a)\right]\right. \\
&+\sum_{s,a}d_\mu^{\pi_{t+1}}(s)(\pi_{t+1}(a|s)-\pi_t(a|s))\times \\
&\left.[Q^{\pi_t}(s,a)-Q_{t+1}(s,a)]\right] \\
\overset{(d)}{\geq}&\sum_{s,a}\mu(s)\pi_t(a|s)\left[\frac{1}{\beta_t}\log Z_t(s)-Q_{t+1}(s,a)\right] \\
&-\frac{2Q_{\max}L_1\sqrt{|\mathcal{A}|}}{1-\gamma}\beta_t \\
=&\sum_s\mu(s)\left[\frac{1}{\beta_t}\log Z_t(s)-V^{\hat{\pi}_t}(s)\right] \\
&+\sum_{s,a}\mu(s)\hat{\pi}_t(a|s)\left[Q^{\hat{\pi}_t}(s,a)-Q_{t+1}(s,a)\right] \\
&+\sum_{s,a}\mu(s)(\hat{\pi}_t(a|s)-\pi_t(a|s))Q_{t+1}(s,a) \\
&-\frac{2Q_{\max}L_1\sqrt{|\mathcal{A}|}}{1-\gamma}\beta_t,
\end{aligned}$$

where (a) is due to Performance Difference Lemma [68], (b) is by the update rule in Algorithm 1, (c) is by positivity of the KL-divergence [66], and (d) is by the definition of $d^{\pi_{t+1}}$ and (41). Taking $\mu=d^*$, we have:

$$\begin{aligned}
&\sum_s d^*(s)\left[\frac{1}{\beta_t}\log Z_t(s)-V^{\hat{\pi}_t}(s)\right]\leq V^{\pi_{t+1}}(d^*)-V^{\pi_t}(d^*) \\
&+\|Q^{\hat{\pi}_t}-Q_{t+1}\|+\frac{2L_1\sqrt{|\mathcal{A}|}}{(1-\gamma)^2}\beta_t+\frac{\epsilon_{t-1}}{1-\gamma}
\end{aligned}$$

which gets the result.□