
Order-Optimal Regret with Novel Policy Gradient Approaches in Infinite-Horizon Average Reward MDPs

Swetha Ganesh
Indian Institute of Science
Purdue University

Washim Uddin Mondal
Indian Institute of Technology,
Kanpur

Vaneet Aggarwal
Purdue University

Abstract

We present two Policy Gradient-based algorithms with general parametrization in the context of infinite-horizon average reward Markov Decision Process (MDP). The first one employs Implicit Gradient Transport for variance reduction, ensuring an expected regret of the order $\tilde{O}(T^{2/3})$. The second approach, rooted in Hessian-based techniques, ensures an expected regret of the order $\tilde{O}(\sqrt{T})$. These results significantly improve the state-of-the-art $\tilde{O}(T^{3/4})$ regret and achieve the theoretical lower bound. We also show that the average-reward function is approximately L -smooth, a result that was previously assumed in earlier works.

1 INTRODUCTION

Reinforcement Learning (RL) encompasses a set of challenges where a learner interacts iteratively with an unknown Markovian environment, aiming to maximize the total rewards earned. This framework finds application in various domains, such as networking, transportation, queueing theory, and epidemic control [Geng et al., 2020, Al-Abbasi et al., 2019, Aggarwal et al., 2022, Ling et al., 2024]. RL problems are commonly analyzed under three major frameworks: episodic, infinite-horizon with discounted rewards, and infinite-horizon with average rewards. Among these, the infinite-horizon average reward setup holds particular importance due to its ability to capture long-term goals in practical situations. To address these challenges, both model-free and model-based approaches have been proposed in the literature.

While model-based approaches have been extensively investigated for the average reward setup [Auer et al., 2008, Agrawal and Jia, 2017, Wei et al., 2021], they, unfortunately, require a large memory to store model parameters, limiting their practical applicability in large state space scenarios. Additionally, although model-free algorithms have been studied primarily in the tabular setups [Wei et al., 2020], they require all the policies to be represented in a tabular format, which creates a barrier in efficiently handling large state spaces. One way to tackle the issue of large state space is via general policy parametrization, where policies are indexed by a $d \ll SA$ dimensional parameter where S , and A denote sizes of the state and action spaces of the underlying Markov Decision Process (MDP). This approach has been explored recently for the average reward setup [Bai et al., 2024]. However, their proposed algorithm yields an expected regret of $\tilde{O}(T^{3/4})$, which is far from the theoretical lower bound of $\Omega(\sqrt{T})$. This prompts the following question:

Is it possible to improve the state-of-the-art regret performance and attain the lower bound of $\tilde{O}(\sqrt{T})$ in the average-reward setup with general policy parametrization?

In this paper, we affirmatively address the above question by presenting two algorithms. The first one is a Policy Gradient-based approach that employs implicit gradient transport for variance reduction. Notably, this algorithm does not require any second-order information and achieves an expected regret of $\tilde{O}(T^{2/3})$. The second algorithm uses a Hessian-based technique within the policy gradient framework to achieve an expected regret bound of $\tilde{O}(\sqrt{T})$. This bridges the gap between the upper and lower bounds of regret for the average reward setup in parametrized scenarios in terms of T .

1.1 Related Works

Policy Gradient Algorithms in Discounted Reward MDPs: Recent works on policy gradient-

Algorithm	Regret	Policy Parametrization
MDP-OOMD [Wei et al., 2020]	$\tilde{O}(\sqrt{T})$	Tabular
POLITEX [Abbasi-Yadkori et al., 2019]	$\tilde{O}(T^{3/4})$	Tabular
MDP-EXP2 [Wei et al., 2021]	$\tilde{O}(\sqrt{T})$	Tabular
GBPA [Murthy et al., 2023]	$\tilde{O}(T^{2/3})$	Tabular
Parametrized Policy Gradient [Bai et al., 2024]	$\tilde{O}(T^{3/4})$	General
Algorithm 1 (This Paper)	$\tilde{O}(T^{2/3})$	General
Algorithm 3 (This Paper)	$\tilde{O}(\sqrt{T})$	General
Lower bound [Auer et al., 2008]	$\Omega(\sqrt{T})$	N/A

Table 1: This table summarizes policy-based algorithms from the literature for infinite-horizon average reward MDPs. In the tabular setting, the function approximation error for the policy is zero.

based algorithms have primarily focused on discounted reward MDPs. For instance, Mondal and Aggarwal [2024] and Fatkhullin et al. [2023] provide approaches achieving a sample complexity of $\tilde{O}(\epsilon^{-2})$ for general parameterized policies. More specifically, Mondal and Aggarwal [2024] combines Accelerated Stochastic Gradient Descent with NPG to obtain the above-mentioned sample complexity bound. Whereas Fatkhullin et al. [2023] uses (N)-HARPG, a recursive variance-reduction technique that uses Hessian estimates, to obtain their sample complexity bounds. Moreover, they also propose N-PG-IGT, which combines Policy Gradient (PG) with Implicit Gradient Transport (IGT), and show that this algorithm obtains a sample complexity of $\mathcal{O}(\epsilon^{-2.5})$. Our approach is inspired by these two methods. However, our analysis deviates considerably for several reasons, including the notably different properties of the (biased) policy gradient and Hessian estimates, which emerge due to the lack of access to simulators to restart the MDP (more details in Section 1.2).

Tabular Model-Based/Model-Free Average Reward MDPs: One of the first bounds in this setting was established in Auer et al. [2008], which introduced the model-based UCRL2 algorithm, achieving a regret of $\tilde{O}(\sqrt{T})$. However, model-based approaches become computationally challenging in large state spaces.

In the realm of model-free algorithms, one approach is purely value-based, where the algorithm learns value functions and extracts the optimal policy only using the learned functions. This approach was adopted in Zhang and Xie [2023], achieving an optimal regret bound of $\mathcal{O}(\sqrt{T})$ for weakly-communicating MDPs. However, such methods often struggle with scalability in large state-action spaces, as they require maximizing the Q -function over the entire action space at each iteration. In contrast, policy gradient methods

are more suitable for large state-action spaces and are widely used in practice.

Another class of methods updates policies using a learned value function, avoiding action-space maximization at each step [Wei et al., 2020, 2022, Abbasi-Yadkori et al., 2019, Murthy et al., 2023]. However, these methods assume tabular policy parameterization. More specifically, the transferred function approximation error in (18), and consequently, ϵ_{bias} in Assumption 3, is zero. [Wei et al., 2020] introduced an online mirror descent algorithm achieving $\tilde{O}(\sqrt{T})$ regret for ergodic MDPs. They also showed that optimistic-Q learning attains $\tilde{O}(T^{2/3})$ regret in weakly communicating MDPs, though their approach relies on tabular value function approximation. This assumption was later relaxed to linear value function approximation [Abbasi-Yadkori et al., 2019, Wei et al., 2021]. In Murthy et al. [2023], policy parameterizations beyond softmax were explored, achieving $\mathcal{O}(T^{2/3})$ regret. However, their approach still requires tabular policies and value functions.

In contrast, our paper develops efficient algorithms for MDPs with general parameterized *policies*. Notably, analyzing the general parameterization case is not a straightforward extension of the tabular setting with an added function approximation error; instead, it introduces fundamental differences. One key distinction stems from the additional monotonicity property that ensures per-step improvements in the tabular case, a property that does not naturally extend to the function approximation setting [Agarwal et al., 2021].

Parametrized Policies in Average Reward Setup: It is worth stating that the policy gradient-based approaches with general parametrization are scalable and useful in multiple applications. Although most of the algorithms with general parametrization focus on the discounted reward setup, our work aims to solve the average reward MDPs. The first regret analysis within this framework was undertaken in [Bai et al., 2024]. Using a policy-gradient-based approach, they established a regret bound of $\tilde{O}(T^{3/4})$. In another recent study within a similar framework, a global convergence bound of $\tilde{O}(T^{-1/4})$ was achieved employing a Multi-level Monte Carlo approach [Patel et al., 2024]. However, these works assume that the average reward function, J , is L -smooth. In this paper, we enhance these findings by presenting two algorithms: one utilizing second-order information and one without. Our results demonstrate improvements over prior work, with the second-order information algorithm achieving the optimal regret order. Table 1 summarizes the key results and comparisons.

1.2 Technical Novelty and Contributions

Difference from Discounted Setup Results:

Infinite-horizon discounted MDPs with general policy parametrization have been extensively studied in the literature [Mondal and Aggarwal, 2024, Fatkhullin et al., 2023, Liu et al., 2020]. However, despite being more aligned toward real-world scenarios, the average-reward setting is less investigated due to several unique challenges. For example, in average-reward MDPs, crucial quantities such as the value function, advantage function, and policy gradient estimator can become unbounded. In contrast, in the discounted case, these quantities are bounded, with the bounds scaling with powers of $\frac{1}{1-\gamma}$ (where γ is the discount factor, which equals 1 in the average reward setup). Additionally, most works in the discounted setup assume access to a generative model (simulator), which allows them to restart the state distribution at will, a practice we avoid due to its impracticality. Such avoidance also makes the biases of the policy gradient and Hessian estimators non-zero, further complicating the analysis.

Technical Novelty: We note that this work improves the state of the art for average reward MDP, by proposing two algorithms and analyzing their regret with the general policy parametrization. The key technical novelties in this work are summarized as follows.

- **Showing the smoothness property of average-reward function:** It is well-known that the average reward function, J , is L -smooth in the discounted MDP [Liu et al., 2020]. This implies that the Hessian of J is bounded by L . The proof is straightforward in this context, as the Hessian can be calculated using the policy gradient estimator and its derivative, both of which are bounded. However, as stated before, this does not hold in the average reward case, and prior works have treated this as an unverifiable assumption [Bai et al., 2024]. We address this issue by proving an approximate L -smoothness result by identifying a function \bar{J} that is L -smooth and works as a close approximation to J (see Section 4.1). We note that while Kumar et al. [2025] provides a smoothness result for average reward MDPs, it considers the tabular setup which uses a substantially different analysis. They show smoothness of J with respect to the policy itself, while we show smoothness with respect to the policy parameter θ . As a result, their analysis crucially relies on the linearity of the policy, reward and transition kernel P with respect to the parameter α (see (13) and (14) of Kumar et al. [2025]).

- **Efficient Bounds on Gradient Estimator:** In prior research on policy gradient methods in the average-reward parameterized framework, it was observed that even with the exact advantage function $A^{\pi_\theta}(s, a)$, the trajectory sampling distribution introduced an error term proportional to $\frac{1}{T^{1/4}}$. This error would persist despite employing variance-reduction techniques, thereby hindering the achievement of $\mathcal{O}(\sqrt{T})$ regret. We address this issue by adopting a simple approach that estimates the advantage and the policy gradient at a slightly later point along the trajectory, effectively diminishing this error term without impairing the convergence rate and, consequently, the regret.
- **Efficient Variance Reduction Analysis in the Average Reward Setup:** In the analysis of variance-reduced methods for discounted settings, the gradient and Hessian estimators are typically unbiased. However, this does not hold in the average reward case. Note that bias accumulation can hinder convergence, so one ought to minimize it for faster convergence. Unfortunately, minimizing the bias of the gradient and Hessian estimators is a difficult endeavor. The approximate function, \bar{J} (stated above), helps us in this regard by allowing us to construct its unbiased gradient and Hessian estimators (see Section 4.1) which are close to their actual values.

Key Contributions: We provide two Policy Gradient-based approaches with general parametrization and provide regret guarantees under the assumption of an ergodic MDP. More specifically,

- We show that our first approach (Algorithm 1) has an expected regret of $\tilde{\mathcal{O}}(T^{2/3})$. This method utilizes implicit gradient transport for variance reduction without requiring importance sampling or curvature information (such as Hessian estimates). Moreover, this algorithm only samples a single trajectory per iteration.
- Our second approach (Algorithm 3) uses a Hessian-based approach to obtain an expected regret of $\tilde{\mathcal{O}}(\sqrt{T})$, which is optimal in T . Although this algorithm uses Hessian estimates, the required updates can be computed with memory and computational complexities that are similar to that of Hessian-free methods.
- Additionally, we provide the first approximate smoothness-type result for the average-reward function using our novel Hessian-based analysis.

For both Algorithms 1 and 3, the obtained regret bounds significantly improve the existing state-of-the-art $\tilde{\mathcal{O}}(T^{3/4})$ regret of Bai et al. [2024].

2 SETUP

In this paper, we explore an infinite-horizon reinforcement learning problem with an average reward criterion, modeled by a Markov Decision Process (MDP) represented as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, P, \rho)$. Here, \mathcal{S} denotes the state space with size S , \mathcal{A} is the action space with a size of A , $r : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ represents the reward function, $P : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$ defines the state transition function, where $\Delta(\mathcal{S})$ denotes the probability simplex over \mathcal{S} , and $\rho \in \Delta(\mathcal{S})$ signifies the initial distribution of states. A policy $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$ determines the distribution of the action to be taken given the current state. It gives rise to a transition function $P^\pi : \mathcal{S} \rightarrow \Delta(\mathcal{S})$ defined as $P^\pi(s, s') = \sum_{a \in \mathcal{A}} P(s'|s, a)\pi(a|s)$, for all $s, s' \in \mathcal{S}$. It can be seen that for any given policy π , the sequence of states produced by the MDP forms a Markov chain. We will assume the following throughout the paper.

Assumption 1. *The MDP \mathcal{M} is ergodic. In other words, the Markov chain induced by every policy π is irreducible and aperiodic.*

The assumption of ergodicity is frequently employed in the analysis of Markov Decision Processes (MDPs) [Pesquerel and Maillard, 2022, Gong and Wang, 2020, Bai et al., 2024].

We consider a parameterized class of policies Π , which consists of all policies π_θ such that $\theta \in \Theta$, where $\Theta = \mathbb{R}^d$. It is a well-known fact that if \mathcal{M} is ergodic, then for all $\theta \in \Theta$, there exists a unique stationary distribution denoted as $d^{\pi_\theta} \in \Delta(\mathcal{S})$, that is independent of the initial distribution ρ and satisfies $(P^{\pi_\theta})^\top d^{\pi_\theta} = d^{\pi_\theta}$. We define the long-term average reward for a given policy π_θ as follows:

$$J_\rho^{\pi_\theta} := \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E} \left[\sum_{t=0}^{T-1} r(s_t, a_t) \middle| s_0 \sim \rho \right] \\ = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} r(s, a) d^{\pi_\theta}(s) \pi_\theta(a|s),$$

where the expectation is computed over all state-action trajectories generated by following the action execution process $a_t \sim \pi(\cdot|s_t)$ and the state transition rule $s_{t+1} \sim P(\cdot|s_t, a_t)$ for all $t \in \{0, 1, \dots\}$. Similar to the stationary distribution, $J_\rho^{\pi_\theta}$ is also independent of ρ . We simplify the notation of J^{π_θ} by expressing it as $J(\theta)$. Our target is to maximize $J(\cdot)$ without any knowledge about the transition function P . The maximum value and the corresponding maximizing policy are denoted as J^* and π^* respectively.

We introduce a few notations before delving into our proposed algorithms to solve this optimization. The

state value function V is defined below.

$$V^{\pi_\theta}(s) = \mathbb{E}_\theta \left[\sum_{t=0}^{\infty} (r(s_t, a_t) - J(\theta)) \middle| s_0 = s \right], \quad (1)$$

where $r^{\pi_\theta}(s) := \sum_a r(s, a) \pi_\theta(a|s)$, $\forall s \in \mathcal{S}$, $\forall \theta \in \Theta$, and $\mathbb{E}_\theta[\cdot]$ is expectation over all trajectories induced by the policy π_θ . Similarly, $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$, the state-action value function $Q^{\pi_\theta}(s, a)$ can be written as shown below.

$$Q^{\pi_\theta}(s, a) = \mathbb{E}_\theta \left[\sum_{t=0}^{\infty} (r(s_t, a_t) - J(\theta)) \middle| s_0 = s, a_0 = a \right]. \quad (2)$$

We can now define the advantage function $A^{\pi_\theta} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ such that $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$, we have,

$$A^{\pi_\theta}(s, a) := Q^{\pi_\theta}(s, a) - V^{\pi_\theta}(s). \quad (3)$$

Algorithm 1 Parameterized Policy Gradient with Implicit Gradient Transport

- 1: **Input:** Initial parameters θ_0 and θ_1 , stepsizes $\{\gamma_k\}_{k \geq 1}$, $\{\eta_k\}_{k \geq 1}$, $s_0 \sim \rho(\cdot)$, epoch length H , epoch number K , $N = 7t_{\text{mix}} \log_2 T$
- 2: **for** $k \in \{1, \dots, K\}$ **do**
- 3: $\hat{\theta}_k \leftarrow \theta_k + \frac{1-\eta_k}{\eta_k}(\theta_k - \theta_{k-1})$, $\tilde{\tau}_k \leftarrow \phi$
- 4: **for** $t \in \{(k-1)H, \dots, kH-1\}$ **do**
- 5: Execute $a_t \sim \pi_{\hat{\theta}_k}(\cdot|s_t)$, receive reward $r(s_t, a_t)$ and observe s_{t+1}
- 6: $\tilde{\tau}_k \leftarrow \tilde{\tau}_k \cup \{(s_t, a_t)\}$
- 7: **end for**
- 8: **for** $t \in \{(k-1)H + N, \dots, kH-1\}$ **do**
- 9: Obtain $\hat{A}^{\pi_{\hat{\theta}_k}}(\tilde{\tau}_k, s_t, a_t)$ via Algorithm 2
- 10: **end for**
- 11: Obtain $g(\tilde{\theta}_k, \tilde{\tau}_k)$ via (7) and d_k via (13)
- 12: Update policy parameter as

$$\theta_{k+1} = \theta_k + \gamma_k \frac{d_k}{\|d_k\|} \quad (4)$$

13: **end for**

It is known that ergodicity also implies the existence of a finite mixing time. Specifically, if \mathcal{M} is ergodic, then the mixing time can be defined as follows.

Definition 1. *The mixing time of an MDP \mathcal{M} with respect to a parameter θ is given as, $t_{\text{mix}}^\theta := \min \left\{ t \geq 1 \middle| \|(P^{\pi_\theta})^t(s, \cdot) - d^{\pi_\theta}\| \leq \frac{1}{4}, \forall s \in \mathcal{S} \right\}$. We also define $t_{\text{mix}} := \sup_{\theta \in \Theta} t_{\text{mix}}^\theta$ as the overall mixing time which is finite due to ergodicity.*

The mixing time of an MDP measures how quickly the MDP approaches its stationary distribution when the same policy is executed repeatedly. Additionally, we define the hitting time below.

Definition 2. The hitting time of an MDP \mathcal{M} with respect to a policy parameter, θ is defined as, $t_{\text{hit}}^\theta := \max_{s \in \mathcal{S}} \frac{1}{d^{\pi_\theta}(s)}$. We also define $t_{\text{hit}} := \sup_{\theta \in \Theta} t_{\text{hit}}^\theta$ as the overall hitting time which is finite due to ergodicity.

For a given MDP \mathcal{M} and a time horizon T , the regret of an algorithm \mathbb{A} is defined as follows.

$$\text{Reg}_T(\mathbb{A}, \mathcal{M}) := \sum_{t=0}^{T-1} (J^* - r(s_t, a_t)), \quad (5)$$

where J^* denotes the optimal long-term average reward and $\{a_t\}_{t \geq 0}$, are selected by the algorithm, \mathbb{A} , based on the history up to time, t . The state at the next time step is obtained by following the state transition function, P . We simplify the notation of regret to Reg_T wherever there is no ambiguity. We are now ready to describe our proposed algorithms.

Algorithm 2 State and state-action value function estimation

```

1: Input: State  $s$ , action  $a$ , policy parameter  $\theta$ , and trajectory  $\tau = \{(s_t, a_t)\}_{t=t_s}^{t_e}$ 
2: Define:  $N = 7t_{\text{mix}} \log_2 T$ 
3: Initialize:  $i \leftarrow 0$ ,  $\xi \leftarrow t_s$ 
4: while  $\xi \leq t_e - N$  do
5:   if  $s_\xi = s$  then
6:      $i \leftarrow i + 1$ ,
7:      $\xi_i \leftarrow \xi$ 
8:      $y_i \leftarrow \sum_{t=\xi}^{\xi+N-1} r(s_t, a_t)$ .
9:      $\xi \leftarrow \xi + 2N$ .
10:  else
11:     $\xi \leftarrow \xi + 1$ .
12:  end if
13: end while
14: if  $i > 0$  then
15:   Obtain  $\hat{V}^{\pi_\theta}(\tau, s)$  and  $\hat{Q}^{\pi_\theta}(\tau, s, a)$  using Eq. (8) and (9)
16: else
17:    $\hat{V}^{\pi_\theta}(\tau, s) \leftarrow 0$ ,  $\hat{Q}^{\pi_\theta}(\tau, s, a) \leftarrow 0$ 
18: end if
19: return  $\hat{Q}^{\pi_\theta}(\tau, s, a)$  and  $\hat{V}^{\pi_\theta}(\tau, s)$ 

```

3 PROPOSED ALGORITHMS

This paper proposes two algorithms. The first one (Algorithm 1) is a PG method that utilizes Implicit Gradient Transport (IGT) for variance reduction. The second one (Algorithm 3) applies a Hessian-based approach for variance-reduction. Given a horizon of length T , both algorithms run $K = T/H$ number of epochs where H is the duration of each epoch. At the k th epoch, both algorithms update the current estimate of the policy parameter, θ_k , along some direc-

tion, d_k , which is computed as a function of the previous direction, d_{k-1} , and the gradient and Hessian estimates of the function $J(\cdot)$ at some appropriate parameter values. Note that the recursive definition of the update direction, d_k , is the typical characteristic of a momentum-based algorithm, which contrasts with that of a simple policy gradient algorithm. Below we describe the process to compute gradient and Hessian estimates of $J(\cdot)$ at an arbitrary parameter, θ .

Gradient Estimation: The idea of estimating $\nabla_\theta J(\theta)$ is mimicking the following result [Sutton et al., 1999].

$$\nabla_\theta J(\theta) = \mathbb{E}_{s \sim d^{\pi_\theta}, a \sim \pi_\theta(\cdot|s)} [A^{\pi_\theta}(s, a) \nabla_\theta \log \pi_\theta(a|s)]. \quad (6)$$

In particular, if $\tau = \{(s_t, a_t)\}_{t=t_s}^{t_e}$ is a π_θ -induced trajectory, and $\hat{A}^{\pi_\theta}(\tau, s, a)$ is an estimate of the advantage value $A^{\pi_\theta}(s, a)$ corresponding to τ , $\forall (s, a)$, then the estimate of $\nabla_\theta J(\theta)$ corresponding to τ is the following expression where $|\tau| = t_e - t_s + 1$ is the length of τ and $N = \mathcal{O}(t_{\text{mix}} \log T)$.

$$g(\theta, \tau) = \frac{1}{|\tau| - N} \sum_{t=t_s+N}^{t_e} \hat{A}^{\pi_\theta}(\tau, s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t), \quad (7)$$

Note that the first N elements of τ are discarded to compute the gradient estimate. This is to ensure that the remaining trajectory's initial state distribution is close to the stationary one. It also guarantees that the estimates obtained at different epochs are fairly independent of each other. Algorithm 2 gives a procedure to calculate the estimate $\hat{A}^{\pi_\theta}(\tau, s, a)$. In particular, for a π_θ -induced trajectory $\tau = \{(s_t, a_t)\}_{t=t_s}^{t_e}$, Algorithm 2 first locates its disjoint sub-trajectories of length N that start with a given state s and are at least N step apart from each other. This guarantees that the estimates from each sub-trajectory are fairly independent. If i indicates the number of such sub-trajectories, and the sum of rewards and the initial time of the j th such sub-trajectory are y_j and ξ_j , respectively, then the estimates $\hat{V}^{\pi_\theta}(\tau, s)$ and $\hat{Q}^{\pi_\theta}(\tau, s, a)$ are given as follows for $i > 0$.

$$\hat{V}^{\pi_\theta}(\tau, s) = \frac{1}{i} \sum_{j=1}^i y_j \quad (8)$$

and

$$\hat{Q}^{\pi_\theta}(\tau, s, a) = \frac{1}{i \pi_\theta(a|s)} \sum_{j=1}^i y_j \mathbf{1}(a_{\xi_j} = a), \quad (9)$$

where $\mathbf{1}(\cdot)$ is an indicator function. For $i = 0$, each estimate is taken to be zero. These expressions

help obtain the advantage estimate as $\hat{A}^{\pi_\theta}(\tau, s, a) = \hat{Q}^{\pi_\theta}(\tau, s, a) - \hat{V}^{\pi_\theta}(\tau, s)$, $\forall (s, a)$.

Hessian Estimation: Let $\tau = \{(s_t, a_t)\}_{t=t_s}^{t_e}$ be a π_θ -induced trajectory. Define the following.

$$\Phi(\theta, \tau) := \frac{1}{|\tau| - N} \sum_{t=t_s+N}^{t_e} \left\{ \Psi_t^{(1)}(\tau) \log \pi_\theta(a_t|s_t) + \frac{\Psi_t^{(2)}(\tau)}{\pi_\theta(a_t|s_t)} \right\}, \quad (10)$$

where $|\tau| = t_e - t_s + 1$ is the length of the trajectory, τ ,

$$\Psi_t^{(1)}(\tau) := -\hat{V}^{\pi_\theta}(\tau, s_t)$$

and

$$\Psi_t^{(2)}(\tau) := -\hat{Q}^{\pi_\theta}(\tau, s_t, a_t) \pi_\theta(a_t|s_t)$$

where $\hat{V}^{\pi_\theta}(\tau, \cdot)$ and $\hat{Q}^{\pi_\theta}(\tau, \cdot, \cdot)$ are defined by (8) and (9). It is easy to verify that $\Psi_t^k(\tau)$, $k \in \{1, 2\}$ depend only on (τ, t) and not on π_θ .

Let $p(\tau, \theta, \bar{\rho}) := \bar{\rho}(s_{t_s}) \prod_{h=t_s+1}^{t_e} \pi_\theta(a_h|s_h) P(s_{h+1}|s_h, a_h)$. $p(\tau, \theta, \bar{\rho})$ is the probability of occurrence of τ that initiates with a θ -independent state distribution $\bar{\rho}$. We estimate the Hessian as follows:

$$B(\theta, \tau) := \nabla_\theta \Phi(\theta, \tau) \nabla_\theta \log p(\tau, \theta, \bar{\rho})^\top + \nabla_\theta^2 \Phi(\theta, \tau). \quad (11)$$

It is to be emphasized that although $p(\tau, \theta, \bar{\rho})$ requires knowledge of P and $\bar{\rho}$ to be calculated, $\nabla_\theta \log p(\tau, \theta, \bar{\rho})$ does not. One can establish (Appendix A) that $B(\theta, \tau)$ is an unbiased estimator of $\nabla_\theta^2 \bar{J}(\theta)$ where \bar{J} is an auxiliary function defined in section 4.1 that closely approximates J . Appendix A also exhibits that $\nabla_\theta \Phi(\theta, \tau) = g(\theta, \tau)$ and derives an expression to compute $\nabla_\theta^2 \Phi(\theta, \tau)$. These results enable the computation of $B(\theta, \tau)$.

Parameterized Policy Gradient with Implicit Gradient Transport (Algorithm 1): The idea of Implicit Gradient Transport was originally proposed in Cutkosky and Mehta [2020] in the context of Stochastic Optimization. It was later used in discounted RL Fatkhullin et al. [2023]. Algorithm 1 obtains an auxiliary parameter θ_k , at each epoch k , using the current and previous policy parameters θ_k, θ_{k-1} as follows.

$$\tilde{\theta}_k = \theta_k + \frac{1 - \eta_k}{\eta_k} (\theta_k - \theta_{k-1}) \quad (12)$$

where η_k is an appropriately chosen learning parameter. Next, a $\pi_{\tilde{\theta}_k}$ -induced trajectory, $\tilde{\tau}_k$, of length H is obtained. The update direction, d_k , is then recursively computed using the following relation.

$$d_k = (1 - \eta_k) d_{k-1} + \eta_k g(\tilde{\theta}_k, \tilde{\tau}_k) \quad (13)$$

where $g(\tilde{\theta}_k, \tilde{\tau}_k)$ is the estimate of the policy gradient given by (7). Finally, θ_k is updated following (4) where γ_k is an appropriate learning rate.

The following remarks are worth mentioning. Firstly, d_k 's are computed recursively to capture the momentum. Secondly, unlike other momentum-based algorithms, the update equation of d_k involves a gradient estimate computed at the auxiliary parameter $\tilde{\theta}_k$, rather than it being computed at θ_k . Note that $\tilde{\theta}_k$ is defined such that it can be interpreted as taking a "look-ahead" extrapolation from iterates θ_k and θ_{k-1} (further insights into this approach can be found in [Cutkosky and Mehta, 2020, Sec. 3]). Finally, the update of θ_k is proportional to the normalized vector $d_k/\|d_k\|$ rather than it being proportional to d_k itself.

Algorithm 3 Parameterized Hessian-aided Policy Gradient

- 1: **Input:** Initial parameters θ_0 and θ_1 , stepsizes $\{\gamma_k\}_{k \geq 1}$, momentum parameters $\{\eta_k\}_{k \geq 1}$, initial state $s_0 \sim \rho(\cdot)$, epoch length H , number of epochs K , $N = 7t_{\text{mix}} \log_2 T$, $d_0 = \mathbf{0}$
- 2: **for** $k \in \{1, \dots, K\}$ **do**
- 3: $q_k \sim \mathcal{U}([0, 1])$, $\hat{\theta}_k = q_k \theta_k + (1 - q_k) \theta_{k-1}$
- 4: $\tau_k \leftarrow \phi$, $\hat{\tau}_k \leftarrow \phi$
- 5: **for** $t \in \{(k-1)H, \dots, (k-1)H + \frac{H}{2} - 1\}$ **do**
- 6: Execute $a_t \sim \pi_{\theta_k}(\cdot|s_t)$, receive reward $r(s_t, a_t)$ and observe s_{t+1}
- 7: $\tau_k \leftarrow \tau_k \cup \{(s_t, a_t)\}$
- 8: **end for**
- 9: **for** $t \in \{(k-1)H + \frac{H}{2}, \dots, kH - 1\}$ **do**
- 10: Execute $a_t \sim \pi_{\hat{\theta}_k}(\cdot|s_t)$, receive reward $r(s_t, a_t)$ and observe s_{t+1}
- 11: $\hat{\tau}_k \leftarrow \hat{\tau}_k \cup \{(s_t, a_t)\}$
- 12: **end for**
- 13: **for** $t \in \{(k-1)H + N, \dots, (k-1)H + \frac{H}{2} - 1\}$ **do**
- 14: Using Algorithm 2, and τ_k , compute $\hat{A}^{\pi_{\theta_k}}(\tau_k, s_t, a_t)$
- 15: **end for**
- 16: **for** $t \in \{(k-1)H + \frac{H}{2} + N, \dots, kH - 1\}$ **do**
- 17: Using Algorithm 2, and $\hat{\tau}_k$, compute $\hat{Q}^{\pi_{\hat{\theta}_k}}(\hat{\tau}_k, s_t, a_t)$ and $\hat{V}^{\pi_{\hat{\theta}_k}}(\hat{\tau}_k, s_t)$
- 18: **end for**
- 19: Compute $g(\theta_k, \tau_k)$ as in (7), $v(\hat{\theta}_k, \hat{\tau}_k)$ as in (16) and d_k using (15)
- 20: Update policy parameter as

$$\theta_{k+1} = \theta_k + \gamma_k \frac{d_k}{\|d_k\|} \quad (14)$$

21: **end for**

Parameterized Hessian-aided Policy Gradient (Algorithm 3): This approach has been explored in

the context of Reinforcement Learning (RL) with an infinite-horizon discounted rewards setup in [Salehkaileybar et al., 2024, Fatkhullin et al., 2023]. The algorithm shares similarities with the variance reduction method proposed in [Cutkosky and Orabona, 2019] but incorporates second-order information as opposed to the difference between consecutive stochastic gradients [Tran and Cutkosky, 2022].

At the k th epoch, Algorithm 3 generates two trajectories, each of length $H/2$. The first one is π_{θ_k} induced where θ_k is the current estimate of the policy parameter while the second one is $\pi_{\hat{\theta}_k}$ induced where $\hat{\theta}_k = q_k \theta_k + (1 - q_k) \theta_{k-1}$, q_k being a uniform random number in $[0, 1]$. The trajectories are named τ_k and $\hat{\tau}_k$ respectively. The policy parameter θ_k is updated by following (14) where γ_k is a learning parameter and the direction d_k is recursively obtained as follows.

$$d_k = (1 - \eta_k)(d_{k-1} + v(\hat{\theta}_k, \hat{\tau}_k)) + \eta_k g(\theta_k, \tau_k) \quad (15)$$

where η_k is a learning rate and the second order correction $v(\hat{\theta}_k, \hat{\tau}_k)$ is computed as shown below.

$$v(\hat{\theta}_k, \hat{\tau}_k) := B(\hat{\theta}_k, \hat{\tau}_k)(\theta_k - \theta_{k-1}). \quad (16)$$

where $B(\hat{\theta}_k, \hat{\tau}_k)$ is the Hessian estimate defined by (11). The term $\hat{\theta}_k$ is defined such that it ensures $v(\hat{\theta}_k, \hat{\tau}_k)$ to be an unbiased estimate of $\nabla_{\theta} J(\theta_k) - \nabla_{\theta} J(\theta_{k-1})$. Algorithm 3 does not need to store the Hessian estimate $B(\hat{\theta}_k, \hat{\tau}_k)$ since it only uses the matrix-vector product $v(\hat{\theta}_k, \hat{\tau}_k)$.

Remark: It is to be mentioned that Algorithm 2 draws inspiration from Algorithm 2 in Bai et al. [2024]. A major distinction lies in the length of the input trajectory $|\tau| = H$. While Bai et al. [2024] utilizes $H = \tilde{\mathcal{O}}(\sqrt{T})$ -length trajectories to estimate the advantage function, we use $H = \mathcal{O}(T^{1/6} \log^2 T)$ for Algorithm 1 and $H = \mathcal{O}(\log^2 T)$ for Algorithm 3 (see Theorem 1, 2). The variance reduction techniques allow us to use a shorter epoch length, which plays a pivotal role in improving the regret bound.

4 MAIN RESULTS

In this section, we establish the global convergence of Algorithms 1 and 3. Before delving into the details, we would like to highlight a few assumptions necessary for proving these results.

Assumption 2 (Policy parametrization regularity). *For all $\theta, \theta_1, \theta_2 \in \Theta$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$, the following statements hold:*

$$(a) \|\nabla_{\theta} \log \pi_{\theta}(a|s)\| \leq G \quad (b) \|\nabla_{\theta}^2 \log \pi_{\theta_1}(a|s)\| \leq B. \quad (17)$$

Remark 1. *The Lipschitz and smoothness properties for the log-likelihood are quite common in the field of policy gradient algorithm [Agarwal et al., 2020, Zhang et al., 2021, Liu et al., 2020]. These properties were shown to hold for various examples recently including Gaussian policies with linearly parameterized means and certain neural parametrizations [Liu et al., 2020, Fatkhullin et al., 2023].*

Define the transferred function approximation error

$$\begin{aligned} L_{d^{\pi^*}, \pi^*}(\omega_{\theta}^*, \theta) \\ = \mathbb{E}_{s \sim d^{\pi^*} a \sim \pi^*(\cdot|s)} \left[\left(\nabla_{\theta} \log \pi_{\theta}(a|s) \cdot \omega_{\theta}^* - A^{\pi_{\theta}}(s, a) \right)^2 \right], \end{aligned} \quad (18)$$

where π^* is the optimal policy and ω_{θ}^* is given as

$$\begin{aligned} \omega_{\theta}^* = \arg \min_{\omega \in \mathbb{R}^d} \\ \mathbb{E}_{s \sim d^{\pi_{\theta}} a \sim \pi^*(\cdot|s)} \left[\left(\nabla_{\theta} \log \pi_{\theta}(a|s) \cdot \omega - A^{\pi_{\theta}}(s, a) \right)^2 \right]. \end{aligned} \quad (19)$$

It is worth mentioning that ω_{θ}^* defined in (19) can be alternatively written as,

$$\omega_{\theta}^* = F(\theta)^{\dagger} \mathbb{E}_{s \sim d^{\pi_{\theta}} a \sim \pi^*(\cdot|s)} [\nabla_{\theta} \log \pi_{\theta}(a|s) A^{\pi_{\theta}}(s, a)],$$

where \dagger symbolizes the Moore-Penrose pseudoinverse operation and $F(\theta)$ is the Fisher information matrix as defined below:

$$F(\theta) = \mathbb{E}_{s \sim d^{\pi_{\theta}} a \sim \pi^*(\cdot|s)} [\nabla_{\theta} \log \pi_{\theta}(a|s) (\nabla_{\theta} \log \pi_{\theta}(a|s))^{\top}]. \quad (20)$$

Assumption 3. *We assume that the error satisfies $L_{d^{\pi^*}, \pi^*}(\omega_{\theta}^*, \theta) \leq \epsilon_{\text{bias}}$ for any $\theta \in \Theta$ where ϵ_{bias} is a positive constant.*

Remark 2. *The transferred function approximation error, defined by (18) and (19), quantifies the expressivity of the policy class in consideration. It has been shown that the softmax parametrization Agarwal et al. [2021] or linear MDP Jin et al. [2020] admits $\epsilon_{\text{bias}} = 0$. When parameterized by a restricted policy class that does not contain all the policies, ϵ_{bias} turns out to be strictly positive. However, for a rich neural network parametrization, the ϵ_{bias} is small [Wang et al., 2019]. A similar assumption has been adopted in Liu et al. [2020] and Agarwal et al. [2021].*

Assumption 4 (Fisher non-degenerate policy). *There exists a constant $\mu > 0$ such that $F(\theta) - \mu I_d$ is positive semidefinite where I_d denotes an identity matrix.*

Assumption 4 requires that the eigenvalues of the Fisher information matrix can be bounded from below.

This is commonly used in obtaining global complexity bounds for PG based methods [Liu et al., 2020, Zhang et al., 2021, Bai et al., 2023, Fatkhullin et al., 2023]. Assumptions 2-4 are widely use in PG literature and are shown to hold simultaneously for parametrization including Gaussian policies with linearly parameterized means and certain neural parametrizations Liu et al. [2020], Fatkhullin et al. [2023].

Next, we provide the regret bounds of Algorithm 1 and 3 in the following theorems, with their proof outlined in Appendix B.

Theorem 1 (Regret bound for Algorithm 1). *Let $\{\theta_k\}_{k=1}^K$ be the outputs of Algorithm 1. If Assumptions 1, 2, 3 and 4 hold, $\nabla_\theta J(\theta)$ is L_h -smooth, $\gamma_k = \frac{6G}{\mu(k+2)}$ and $\eta_k = \left(\frac{2}{k+2}\right)^{4/5}$ then the following inequality holds for $K = T/H$ where $H = 63t_{\text{mix}}t_{\text{hit}}(\log_2 T)^2T^{1/6}$*

$$\begin{aligned} \mathbb{E}[\text{Reg}_T] &\leq T\sqrt{\epsilon_{\text{bias}}} + \mathcal{O}\left(\left(G^2At_{\text{mix}}^{7/5}t_{\text{hit}}^{2/5}(\log T)^{9/5}\right.\right. \\ &\quad \left.+ G^3\mu^{-2}t_{\text{mix}}^{2/5}t_{\text{hit}}^{2/5}L_h(\log T)^{4/5} + \sqrt{ABG}t_{\text{mix}}^{7/5}t_{\text{hit}}^{2/5}(\log T)^{9/5}\right. \\ &\quad \left.+ \sqrt{AG^2}t_{\text{mix}}^{12/5}t_{\text{hit}}^{7/5}(\log T)^{14/5}\right) \cdot \mu^{-1}T^{2/3}\Big). \end{aligned}$$

Note that the above result requires $\nabla_\theta J$ to be L_h smooth. Unlike the discounted case where smoothness follows from the standard assumptions on the score function, this is not the case for the average-reward setting and we assume it to hold instead. However, our Hessian-based algorithm (Algorithm 3) does not require any such assumption and obtains an order-optimal regret bound.

Theorem 2 (Regret bound for Algorithm 3). *Let $\{\theta_k\}_{k=1}^K$ be generated from Algorithm 3. If Assumptions 1, 2, 3 and 4 hold, $\gamma_k = \frac{6G}{\mu(k+2)}$ and $\eta_k = \frac{2}{k+2}$ then the following inequality holds for $K = T/H$ where $H = 63t_{\text{mix}}t_{\text{hit}}(\log_2 T)^2$*

$$\begin{aligned} \mathbb{E}[\text{Reg}_T] &\leq T\sqrt{\epsilon_{\text{bias}}} + \mathcal{O}\left(\frac{\sqrt{AG^2}t_{\text{mix}}\log T}{\mu} \cdot \sqrt{T}\right. \\ &\quad \left.+ \frac{\sqrt{AG^4}t_{\text{hit}}t_{\text{mix}}^2(\log T)^{3/2}}{\mu^2} \cdot \sqrt{T}\right. \\ &\quad \left.+ \frac{\sqrt{A}(BG + G^3)t_{\text{mix}}\log T}{\mu^2} \cdot \sqrt{T}\right). \end{aligned}$$

Theorem 1 and 2 demonstrate that the regret accrued by Algorithm 1 and 3 are $\tilde{\mathcal{O}}(T^{2/3})$ and $\tilde{\mathcal{O}}(\sqrt{T})$ respectively. Note that both of these results are improvements over the current state-of-the-art bound of $\tilde{\mathcal{O}}(T^{3/4})$ [Bai et al., 2024]. Moreover, the regret of Algorithm 3 is optimal in terms of T . Observe the presence of the term $\sqrt{\epsilon_{\text{bias}}}$ in both of these results, which

appears due to the incompleteness of parameterized policy class. The proofs of both of these theorems rely on the construction of an auxiliary function, \bar{J} , which, intuitively, is a close approximation of J . However, unlike J , \bar{J} satisfies some desirable analytical properties such as L -smoothness for some $L > 0$.

4.1 Construction and Properties of the Auxiliary Function

We construct the auxiliary function $\bar{J}(\theta)$ such that our gradient and Hessian estimators are unbiased estimates of $\nabla_\theta J$ and $\nabla_\theta^2 J$, respectively. The function \bar{J} must also be close to J . Working with this function is essential, especially for Hessian-based methods, since the expected value of the Hessian estimates can potentially be far from $\nabla_\theta^2 J$, unlike the discounted case where the bias of the Hessian estimates is known to decay exponentially in γ [Masiha et al., 2022]. Moreover, \bar{J} is also L -smooth, for some $L > 0$, which helps us prove the approximate smoothness result for J . We define $\bar{J}(\theta)$ as follows.

$$\bar{J}(\theta) = J(\theta_0) + \int_0^1 f((1-q)\theta_0 + q\theta) \cdot (\theta - \theta_0) dq, \quad (21)$$

where $f(\theta) = \mathbb{E}_{\tau \sim p(\cdot, \theta, \bar{\rho})} g(\theta, \tau)$, $g(\theta, \tau)$ is defined in (7) and $p(\tau, \theta, \bar{\rho})$, as discussed in section 3, is the probability of occurrence of a π_θ -induced trajectory, τ , that initiates with state distribution, $\bar{\rho}$ and $\theta_0 \in \Theta$ is the initial parameter. Recall that $g(\theta, \tau) = \nabla_\theta \Phi(\theta, \tau)$ where $\Phi(\theta, \tau)$ is defined in (10). It follows that (see Appendix A)

$$\nabla_\theta \bar{J}(\theta) = \mathbb{E}_{\tau \sim p(\cdot, \theta, \bar{\rho})} [\nabla_\theta \Phi(\theta, \tau)] \quad (22)$$

If we sample τ according to $p(\cdot, \theta, \bar{\rho})$, $g(\theta, \tau)$ will serve as an unbiased estimator for $\nabla_\theta \bar{J}(\theta)$. Since the functions J and \bar{J} are not identical, the above estimate generates a small bias term that diminishes with increasing T . We can derive the Hessian of \bar{J} in the following manner (see Appendix A).

$$\nabla_\theta^2 \bar{J}(\theta) = \mathbb{E}_{\tau \sim p(\tau, \theta, \bar{\rho})} [B(\theta, \tau)]. \quad (23)$$

where $B(\theta, \tau)$ is defined in (11). This establishes $B(\theta, \tau)$ to be an unbiased estimator of $\nabla_\theta^2 \bar{J}$. From the above expression, we can bound $\|\nabla_\theta^2 \bar{J}(\theta)\|$ as (see Appendix B for details)

$$\|\nabla_\theta^2 \bar{J}(\theta)\| \leq L, \quad (24)$$

where L depends on H and N . Lemma 2, 3 (Appendix B) show that the following holds for any θ .

$$\begin{aligned} \|\nabla_\theta J(\theta) - \nabla_\theta \bar{J}(\theta)\| &\leq \tilde{\mathcal{O}}\left(\frac{AGt_{\text{mix}}}{T^4}\right) \\ |J(\theta) - \bar{J}(\theta)| &\leq \tilde{\mathcal{O}}\left(\frac{AGt_{\text{mix}}}{T^4} \cdot \|\theta - \theta_0\|\right) \end{aligned} \quad (25)$$

The above relations indicate that J, \bar{J} , in some sense, are close to each other. Using (24), (25), and the Taylor expansion, we show the following approximate smoothness result of J .

Theorem 3. *If Assumptions 1 and 2 hold, then the following bound holds for all $\theta, \bar{\theta} \in \Theta$.*

$$(J^* - J(\bar{\theta})) \leq (J^* - J(\theta)) - \nabla_{\theta} J(\theta)^{\top} (\bar{\theta} - \theta) + \frac{L}{2} \|\bar{\theta} - \theta\|^2 + \Delta_L,$$

where $L = \mathcal{O}(\sqrt{AG^2 t_{\text{mix}}^2 t_{\text{hit}} (\log T)^3} + B t_{\text{mix}} \log T)$ and $\Delta_L = \mathcal{O}\left(\frac{AG t_{\text{mix}} \{\|\theta - \theta_0\| + \|\bar{\theta} - \theta_0\|\}}{T^4}\right)$.

The above result is crucial for proving Theorems 1 and 2. The proof of Theorem 3 follows by observing that, from (25), we have

$$\begin{aligned} (a) \quad & \bar{J}(\theta) - \epsilon_1 \|\theta - \theta_0\| \leq J(\theta) \leq \bar{J}(\theta) + \epsilon_1 \|\theta - \theta_0\| \\ (b) \quad & \nabla \bar{J}(\theta)^{\top} (\theta - \bar{\theta}) - \epsilon_2 \|\theta - \bar{\theta}\| \leq \nabla J(\theta)^{\top} (\theta - \bar{\theta}) \\ & \leq \nabla \bar{J}(\theta)^{\top} (\theta - \bar{\theta}) + \epsilon_2 \|\theta - \bar{\theta}\| \end{aligned}$$

where ϵ_1, ϵ_2 are $\mathcal{O}(AG t_{\text{mix}} T^{-4})$. Separately, since \bar{J} is L -smooth,

$$(J^* - \bar{J}(\bar{\theta})) \leq (J^* - \bar{J}(\theta)) - \nabla \bar{J}(\theta)^{\top} (\bar{\theta} - \theta) + (L/2) \|\bar{\theta} - \theta\|^2 \quad (26)$$

Substituting (a) and (b) in (26), we obtain

$$\begin{aligned} J^* - J(\bar{\theta}) & \leq (J^* - J(\theta)) - \nabla J(\theta)^{\top} (\bar{\theta} - \theta) \\ & \quad + (L/2) \|\bar{\theta} - \theta\|^2 + \epsilon_1 (\|\theta - \theta_0\| + \|\bar{\theta} - \theta_0\|) \\ & \quad + \epsilon_2 \|\theta - \bar{\theta}\|. \end{aligned}$$

Since $\|\bar{\theta} - \theta_0\| \leq \|\theta - \theta_0\| + \|\theta - \bar{\theta}\|$, the result follows.

5 CONCLUSION

This paper considers the problem of learning an infinite-horizon average reward MDP with general parametrization. We propose two algorithms. The first, based on the idea of implicit gradient transport, achieves a regret bound of $\tilde{\mathcal{O}}(T^{2/3})$, while the second, a Hessian-based algorithm, achieves a regret of $\tilde{\mathcal{O}}(\sqrt{T})$. Both improve upon the state of the art, with the second achieving the optimal bound. In the process, we also prove an approximate smoothness property of the average-reward function, which may be of independent interest.

Following this work, the authors of [Ganesh et al., 2024] show an $\tilde{\mathcal{O}}(T^{-1/2})$ global convergence rate without requiring knowledge of the mixing time. However, their approach introduces an additional critic error due to the linear approximation of the critic. Extending these algorithms to design parameter-free algorithms is an important future direction.

ACKNOWLEDGMENT

This work was supported in part by the Overseas Visiting Doctoral Fellowship from Anusandhan National Research Foundation (ANRF) and U.S. National Science Foundation under grant CCF-2149588.

References

- Y. Abbasi-Yadkori, P. Bartlett, K. Bhatia, N. Lazic, C. Szepesvari, and G. Weisz. POLITEX: Regret bounds for policy iteration using expert prediction. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3692–3702, 2019.
- A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 64–66. PMLR, 09–12 Jul 2020.
- A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *The Journal of Machine Learning Research*, 22(1): 4431–4506, 2021.
- M. Agarwal, Q. Bai, and V. Aggarwal. Concave utility reinforcement learning with zero-constraint violations. *Transactions on Machine Learning Research*, 2022.
- S. Agrawal and R. Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. *Advances in Neural Information Processing Systems*, 30, 2017.
- A. O. Al-Abbasi, A. Ghosh, and V. Aggarwal. Deep-pool: Distributed model-free algorithm for ride-sharing using deep reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems*, 20(12):4714–4727, 2019.
- P. Auer, T. Jaksch, and R. Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.
- Q. Bai, A. S. Bedi, and V. Aggarwal. Achieving zero constraint violation for constrained reinforcement learning via conservative natural policy gradient primal-dual algorithm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6737–6744, 2023.
- Q. Bai, W. U. Mondal, and V. Aggarwal. Regret analysis of policy gradient algorithm for infinite horizon average reward markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10980–10988, 2024.

- A. Cutkosky and H. Mehta. Momentum improves normalized SGD. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 2260–2268. PMLR, 2020.
- A. Cutkosky and F. Orabona. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.
- R. Dorfman and K. Y. Levy. Adapting to mixing time in stochastic optimization with markovian data. In *International Conference on Machine Learning*, pages 5429–5446. PMLR, 2022.
- I. Fatkhullin, A. Barakat, A. Kireeva, and N. He. Stochastic policy gradient methods: Improved sample complexity for fisher-non-degenerate policies. In *International Conference on Machine Learning*, pages 9827–9869. PMLR, 2023.
- S. Ganesh, W. U. Mondal, and V. Aggarwal. Order-optimal global convergence for average reward reinforcement learning via actor-critic approach. *arXiv preprint arXiv:2407.18878*, 2024.
- N. Geng, T. Lan, V. Aggarwal, Y. Yang, and M. Xu. A multi-agent reinforcement learning perspective on distributed traffic engineering. In *2020 IEEE 28th International Conference on Network Protocols (ICNP)*, pages 1–11. IEEE, 2020.
- H. Gong and M. Wang. A duality approach for regret minimization in average-award ergodic markov decision processes. In *Learning for Dynamics and Control*, pages 862–883. PMLR, 2020.
- C. Jin, Z. Yang, Z. Wang, and M. I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2137–2143. PMLR, 09–12 Jul 2020.
- N. Kumar, Y. Murthy, I. Shufaro, K. Y. Levy, R. Srikant, and S. Mannor. Global convergence of policy gradient in average reward MDPs. In *The Thirteenth International Conference on Learning Representations*, 2025.
- L. Ling, W. U. Mondal, and S. V. Ukkusuri. Cooperating graph neural networks with deep reinforcement learning for vaccine prioritization. *IEEE Journal of Biomedical and Health Informatics*, 28(8):4891–4902, 2024. doi: 10.1109/JBHI.2024.3392436.
- 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*, 33:7624–7636, 2020.
- S. Masiha, S. Salehkaleybar, N. He, N. Kiyavash, and P. Thiran. Stochastic second-order methods improve best-known sample complexity of SGD for gradient-dominated functions. In *Advances in Neural Information Processing Systems*, 2022.
- W. U. Mondal and V. Aggarwal. Improved sample complexity analysis of natural policy gradient algorithm with general parameterization for infinite horizon discounted reward markov decision processes. In *International Conference on Artificial Intelligence and Statistics*, 2024.
- Y. Murthy, M. Moharrami, and R. Srikant. Performance bounds for policy-based average reward reinforcement learning algorithms. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- B. Patel, W. A. Suttle, A. Koppel, V. Aggarwal, B. M. Sadler, D. Manocha, and A. Bedi. Towards global optimality for practical average reward reinforcement learning without mixing time oracles. In *Forty-first International Conference on Machine Learning*, 2024.
- F. Pesquerel and O.-A. Maillard. Imed-rl: Regret optimal learning of ergodic markov decision processes. In *NeurIPS 2022-Thirty-sixth Conference on Neural Information Processing Systems*, 2022.
- S. Salehkaleybar, M. Khorasani, N. Kiyavash, N. He, and P. Thiran. Momentum-based policy gradient with second-order information. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.
- R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.
- H. Tran and A. Cutkosky. Better sgd using second-order momentum. *Advances in Neural Information Processing Systems*, 35:3530–3541, 2022.
- L. Wang, Q. Cai, Z. Yang, and Z. Wang. Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*, 2019.
- C.-Y. Wei, M. J. Jahromi, H. Luo, H. Sharma, and R. Jain. Model-free reinforcement learning in infinite-horizon average-reward markov decision processes. In *International conference on machine learning*, pages 10170–10180. PMLR, 2020.
- C.-Y. Wei, M. J. Jahromi, H. Luo, and R. Jain. Learning infinite-horizon average-reward mdp with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3007–3015. PMLR, 2021.
- H. Wei, X. Liu, and L. Ying. A provably-efficient model-free algorithm for infinite-horizon average-reward constrained markov decision processes. In

Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 3868–3876, 2022.

- J. Zhang, C. Ni, Z. Yu, C. Szepesvari, and M. Wang. On the convergence and sample efficiency of variance-reduced policy gradient method. *Advances in Neural Information Processing Systems*, 34:2228–2240, 2021.
- Z. Zhang and Q. Xie. Sharper model-free reinforcement learning for average-reward markov decision processes. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5476–5477. PMLR, 2023.

Checklist

1. For all models and algorithms presented, check if you include:
 - (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes. See Sections 2, 3 and 4]
 - (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes. See Sections 4, B and Supplementary Material]
 - (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]
2. For any theoretical claim, check if you include:
 - (a) Statements of the full set of assumptions of all theoretical results. [Yes. See Section 4]
 - (b) Complete proofs of all theoretical results. [Yes. See Section B and Supplementary Material]
 - (c) Clear explanations of any assumptions. [Yes. See Section 4]
3. For all figures and tables that present empirical results, check if you include:
 - (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable. The primary contribution of this paper is theoretical, it does not include an empirical analysis.]
 - (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable. The primary contribution of this paper is theoretical, it does not include an empirical analysis.]
 - (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable. The primary contribution of this paper is theoretical, it does not include an empirical analysis.]
 - (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable. The primary contribution of this paper is theoretical, it does not include an empirical analysis.]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
 - (a) Citations of the creator If your work uses existing assets. [Not Applicable. The primary contribution of this paper is theoretical, it does not include an empirical analysis.]
 - (b) The license information of the assets, if applicable. [Not Applicable. The primary contribution of this paper is theoretical, it does not include an empirical analysis.]
 - (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable. The primary contribution of this paper is theoretical, it does not include an empirical analysis.]
 - (d) Information about consent from data providers/curators. [Not Applicable]
 - (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable. The primary contribution of this paper is theoretical, it does not include an empirical analysis.]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
 - (a) The full text of instructions given to participants and screenshots. [Not Applicable. The primary contribution of this paper is theoretical, it does not include an empirical analysis.]
 - (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable. The primary contribution of this paper is theoretical, it does not include an empirical analysis.]
 - (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable. The primary contribution of this paper is theoretical, it does not include an empirical analysis.]

A Details of the Hessian estimator

Observe the following.

$$\begin{aligned}
 \nabla_\theta \Phi(\theta, \tau) &= \frac{1}{|\tau| - N} \sum_{t=t_s+N}^{t_e} \left\{ \Psi_t^{(1)}(\tau) \nabla_\theta \log \pi_\theta(a_t|s_t) + \Psi_t^{(2)}(\tau) \nabla_\theta \left(\frac{1}{\pi_\theta(a_t|s_t)} \right) \right\} \\
 &= \frac{1}{|\tau| - N} \sum_{t=t_s+N}^{t_e} \left\{ \Psi_t^{(1)}(\tau) \nabla_\theta \log \pi_\theta(a_t|s_t) + \Psi_t^{(2)}(\tau) \left(\frac{-\nabla_\theta \log \pi_\theta(a_t|s_t)}{\pi_\theta(a_t|s_t)} \right) \right\} \\
 &= \frac{1}{|\tau| - N} \sum_{t=t_s+N}^{t_e} (\hat{Q}^{\pi_\theta}(\tau, s_t, a_t) - \hat{V}^{\pi_\theta}(\tau, s_t)) \nabla_\theta \log \pi_\theta(a_t|s_t) \\
 &= \frac{1}{|\tau| - N} \sum_{t=t_s+N}^{t_e} \hat{A}^{\pi_\theta}(\tau, s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t) = g(\theta, \tau)
 \end{aligned} \tag{27}$$

Moreover, note that,

$$\begin{aligned}
 \nabla_\theta^2 \Phi(\theta, \tau) &= \frac{1}{|\tau| - N} \sum_{t=t_s+N}^{t_e} \nabla_\theta \left(\Psi_t^{(1)}(\tau) \nabla_\theta \log \pi_\theta(a_t|s_t) - \Psi_t^{(2)}(\tau) \left(\frac{\nabla_\theta \log \pi_\theta(a_t|s_t)}{\pi_\theta(a_t|s_t)} \right) \right) \\
 &= \frac{1}{|\tau| - N} \sum_{t=t_s+N}^{t_e} \left\{ \Psi_t^{(1)}(\tau) \nabla_\theta^2 \log \pi_\theta(a_t|s_t) - \Psi_t^{(2)}(\tau) \nabla_\theta \left(\frac{\nabla_\theta \log \pi_\theta(a_t|s_t)}{\pi_\theta(a_t|s_t)} \right) \right\} \\
 &= \frac{1}{|\tau| - N} \sum_{t=t_s+N}^{t_e} \left\{ \Psi_t^{(1)}(\tau) \nabla_\theta^2 \log \pi_\theta(a_t|s_t) \right. \\
 &\quad \left. - \Psi_t^{(2)}(\tau) \left(\frac{\nabla_\theta^2 \log \pi_\theta(a_t|s_t)}{\pi_\theta(a_t|s_t)} - \frac{\nabla_\theta \log \pi_\theta(a_t|s_t) \nabla_\theta \log \pi_\theta(a_t|s_t)^\top}{\pi_\theta(a_t|s_t)} \right) \right\}
 \end{aligned}$$

The above expression and (27) together outline the process to compute $B(\theta, \tau)$. Recall the definition of $\bar{J}(\theta)$ as stated below.

$$\bar{J}(\theta) = J(\theta_0) + \int_0^1 f((1-q)\theta_0 + q\theta) \cdot (\theta - \theta_0) dq, \tag{28}$$

where $f(\theta) = \mathbb{E}_{\tau \sim p(\cdot, \theta, \bar{\rho})} [\nabla \Phi(\theta, \tau)] = \mathbb{E}_{\tau \sim p(\cdot, \theta, \bar{\rho})} [g(\tau, \theta)]$. It follows that

$$\nabla_\theta \bar{J}(\theta) = f(\theta) = \mathbb{E}_{\tau \sim p(\cdot, \theta, \bar{\rho})} [\nabla \Phi(\theta, \tau)] = \sum_\tau p(\tau, \theta, \bar{\rho}) \nabla_\theta \Phi(\theta, \tau) \tag{29}$$

The above expression establishes that $\nabla_\theta \Phi(\theta, \tau) = g(\theta, \tau)$ is an unbiased estimate of $\nabla_\theta \bar{J}(\theta)$. The Hessian of $\bar{J}(\theta)$ can be computed as shown below.

$$\begin{aligned}
 \nabla_\theta^2 \bar{J}(\theta) &= \nabla_\theta (\nabla_\theta \bar{J}(\theta)) = \sum_\tau \nabla_\theta (p(\tau, \theta, \bar{\rho}) \nabla_\theta \Phi(\theta, \tau)) \\
 &= \sum_\tau \nabla_\theta \Phi(\theta, \tau) \nabla_\theta p(\tau, \theta, \bar{\rho})^\top + p(\tau, \theta, \bar{\rho}) \nabla_\theta^2 \Phi(\theta, \tau) \\
 &= \sum_\tau p(\tau, \theta, \bar{\rho}) [\nabla_\theta \Phi(\theta, \tau) \nabla_\theta \log p(\tau, \theta, \bar{\rho})^\top + \nabla_\theta^2 \Phi(\theta, \tau)] \\
 &= \mathbb{E}_{\tau \sim p(\cdot, \theta, \bar{\rho})} [\nabla_\theta \Phi(\theta, \tau) \nabla_\theta \log p(\tau, \theta, \bar{\rho})^\top + \nabla_\theta^2 \Phi(\theta, \tau)] = \mathbb{E}_{\tau \sim p(\cdot, \theta, \bar{\rho})} [B(\theta, \tau)]
 \end{aligned} \tag{30}$$

The above exercise shows that $B(\theta, \tau)$ is an unbiased estimate of $\nabla_\theta^2 \bar{J}(\theta)$.

Computational complexity of Algorithms 1, 2 and 3: We note that the computational complexity of Algorithm 2, which estimates the value functions has computational complexity of order $O(NH)$. This is because this method uses H iterations and the operations at each iteration are $O(N)$.

We first consider Algorithm 2. Line 3 is $O(d)$, lines 4-7 are $O(H)$, lines 8-10 are $O(NH^2)$. Computing the gradient estimate at line 11 is $O(Hd)$ and line 12 is $O(d)$. Since all these steps are done for K iterations, the final complexity of Algorithm 1 is $O(KH^2N + KHd)$.

For Algorithm 3, the computations are similar, except we compute the Hessian estimate as well. The complexity of computing the Hessian estimator in (11) is $O(Hd^2)$. To see this, note that the expression for $\nabla_\theta^2 \Phi(\theta, \tau)$ consists of an average of H matrices of order $d \times d$. The computational complexity of computing these matrices is $O(d^2)$. It follows that the complexity to compute $\nabla_\theta^2 \Phi(\theta, \tau)$ is $O(Hd^2)$. Separately, the complexity of computing $\nabla_\theta \log p(\tau, \theta, \bar{\rho})$ is $O(Hd)$ since it is the sum of $\nabla_\theta \log \pi_\theta(a_i | s_i)$ for all $(s_i, a_i) \in \tau$. $\nabla_\theta \Phi(\theta, \tau)$ is the gradient estimator and taking the outer product with $\nabla_\theta \log p(\tau, \theta, \bar{\rho})$ is $O(d^2)$. This gives a final complexity of $O(KH^2N + KHd^2)$ for Algorithm 3.

B Proof Outline

Before proceeding, we present some useful results related to different estimators used in the paper. The first result is on the first and second-order errors of advantage estimation.

Lemma 1. *Consider the advantage estimator $\hat{A}^\pi(\tau, s, a)$ (described in section 3) corresponding to a policy π , a state-action pair (s, a) , and a π -induced trajectory τ . If the length of the trajectory is $|\tau| = H \geq 63t_{\text{mix}}t_{\text{hit}}(\log_2 T)^2$ and $N = 7t_{\text{mix}} \log T$, then the following relations hold.*

$$\begin{aligned} (a) \quad \mathbb{E} \left[\left(\hat{A}^\pi(\tau, s, a) - A^\pi(s, a) \right)^2 \right] &\leq \mathcal{O} \left(\frac{t_{\text{mix}}^3 t_{\text{hit}} (\log T)^4}{H \pi(a|s)} \right) \\ (b) \quad \left| \mathbb{E} \left[\hat{A}^\pi(\tau, s, a) - A^\pi(s, a) \right] \right| &\leq \mathcal{O} \left(\frac{t_{\text{mix}} \log T}{\sqrt{\pi(a|s)} T^4} \right) \end{aligned}$$

where the expectation is over all such π -induced τ with arbitrary initial distribution $\bar{\rho}$.

Lemma 2. *Consider the gradient and the Hessian estimators given in (7) and (11) respectively. Let the π_θ -induced trajectory $\tau = \{s_t, a_t\}_{t=t_s}^{t_e}$ used in estimation is such that $|\tau| = H \geq 63t_{\text{mix}}t_{\text{hit}}(\log_2 T)^2$. If assumptions 1 and 2 hold, the following statements are true $\forall \theta \in \Theta$.*

$$\begin{aligned} (a) \quad \mathbb{E} \|g(\theta, \tau) - \nabla_\theta J(\theta)\|^2 &\leq \sigma_g^2, \text{ where } \sigma_g^2 = \mathcal{O} \left(\frac{AG^2 t_{\text{hit}} t_{\text{mix}}^3 (\log T)^4}{H} \right) \\ (b) \quad \|\mathbb{E}[g(\theta, \tau)] - \nabla_\theta J(\theta)\| &\leq \beta_g, \text{ where } \beta_g = \mathcal{O} \left(\frac{\sqrt{AG} t_{\text{mix}} \log T}{T^4} \right) \\ (c) \quad \mathbb{E} \|B(\theta, \tau)\|^2 &\leq \mathcal{O} \left(AG^4 H^2 t_{\text{mix}}^2 (\log T)^2 + B^2 t_{\text{mix}}^2 (\log T)^2 + A(B^2 + G^4) t_{\text{mix}}^2 (\log T)^2 \right) \end{aligned}$$

where the expectations are over $\tau \sim p(\cdot, \theta, \bar{\rho})$.

Notice that Lemma 2(b) bounds the bias of the gradient estimate. For defining $\bar{J}(\cdot)$ with $\nabla_\theta \bar{J}(\theta) = \mathbb{E}[g(\theta, \tau)]$, we choose $|\tau| = H = 63t_{\text{mix}}t_{\text{hit}}(\log_2 T)^2$. Then, we can interpret Lemma 2(b) as an upper bound on $\|\nabla_\theta \bar{J}(\theta) - \nabla_\theta J(\theta)\|$. Lemma 2(c) implies that \bar{J} is L -smooth, where $L^2 = \mathcal{O} \left(AB^2 t_{\text{mix}}^2 (\log T)^2 + AG^4 t_{\text{mix}}^2 t_{\text{hit}}^2 (\log T)^4 \right)$. To see this, note that,

$$\|\nabla_\theta^2 \bar{J}(\theta)\| = \|\mathbb{E}[B(\theta, \tau)]\| \leq (\mathbb{E} \|B(\theta, \tau)\|^2)^{1/2}. \quad (31)$$

Substituting $H = 63t_{\text{mix}}t_{\text{hit}}(\log_2 T)^2$ in Lemma 2(c), we obtain $\|\nabla_\theta^2 \bar{J}(\theta)\| \leq L$. Observe that we do not attempt to bound the bias of the Hessian estimate, unlike that for the gradient estimate. This is because, unlike the discounted reward MDP, the expectation of the Hessian estimate might be far from the true Hessian. We instead use the fact that it is an unbiased estimate of $\nabla_\theta^2 \bar{J}(\theta)$, where $\nabla_\theta \bar{J}(\theta)$ is close to $\nabla_\theta J(\theta)$. Using Lemma 2, we can bound J and \bar{J} as follows:

Lemma 3. *Let Assumptions 1 and 2 hold. Let \bar{J} be defined by (21) and N and $|\tau|$ are the same as given in Lemma 2. Then the following statement holds for all $\theta \in \Theta$:*

$$|J(\theta) - \bar{J}(\theta)| \leq \mathcal{O} \left(\frac{AG t_{\text{mix}} \log T}{T^4} \|\theta - \theta_0\| \right) \quad (32)$$

Proof. The proof of the above result follows from the observation that

$$\begin{aligned}
 |\bar{J}(\theta) - J(\theta)| &= \left| \int_0^1 [f((1-q)\theta_0 + q\theta) - \nabla_\theta J((1-q)\theta_0 + q\theta)] \cdot (\theta - \theta_0) dq \right| \\
 &\leq \int_0^1 \|f((1-q)\theta_0 + q\theta) - \nabla_\theta J((1-q)\theta_0 + q\theta)\| \|\theta - \theta_0\| dq \\
 &\stackrel{(a)}{\leq} \mathcal{O} \left(\frac{AGt_{\text{mix}} \log T}{T^4} \|\theta - \theta_0\| \right)
 \end{aligned} \tag{33}$$

where $f(\theta) = \mathbb{E}_{\tau \sim p(\cdot, \theta, \rho)}[g(\theta, \tau)]$ and (a) follows from Lemma 2(b). \square

Lemma 4. Consider Algorithms 1 and 3 with $N = 7t_{\text{mix}} \log_2 T$ and let Assumptions 1 and 2 hold. Then the following statement holds for Algorithm 1 when number of iterations $K = T/H$:

$$\mathbb{E}[\text{Reg}_T] \leq \mathcal{O} \left(H \sum_{k=1}^K (J^* - \mathbb{E}[J(\tilde{\theta}_k)]) + G^2 \mu^{-1} t_{\text{mix}}^2 (\log^2 T) K^{4/5} \right). \tag{34}$$

Similarly, the following holds for Algorithm 3 when number of iterations $K = T/H$:

$$\mathbb{E}[\text{Reg}_T] \leq \mathcal{O} \left(H \sum_{k=1}^K (J^* - \mathbb{E}[J(\theta_k)]) + G^2 \mu^{-1} t_{\text{mix}}^2 (\log^3 T) \right). \tag{35}$$

From Lemma 4, it can be seen that regret bounds can be obtained by suitably bounding $\sum_{k=1}^K (J^* - \mathbb{E}[J(\theta_k)])$. Towards this, we have the following general lemma similar in spirit to [Lemma 5, Fatkhullin et al. [2023]] for bounding $J^* - \mathbb{E}[J(\theta_k)]$ for all $k \geq 1$.

Lemma 5. Let Assumption 2, 3 and 4 hold and $\{\theta_k\}_{k \geq 1}$ be a sequence generated by the following update rule $\forall k \in \{0, 1, \dots, K-1\}$.

$$\theta_{k+1} = \theta_k + \gamma_k \frac{d_k}{\|d_k\|},$$

where $\gamma_k = \frac{6G}{\mu(k+2)}$, $\{d_k\}_{k \geq 1}$ is any sequence in \mathbb{R}^d and $\theta_0 \in \mathbb{R}^d$ ($\theta_{k+1} = \theta_k$ if $d_k = 0$) is arbitrary. The following statement holds for every integer $K \leq T$.

$$J^* - \mathbb{E}[J(\theta_K)] \leq \frac{J^* - J(\theta_0)}{(K+1)^2} + \frac{\sum_{k=1}^K \nu_k (k+2)^2}{(K+1)^2},$$

where $\nu_k := \frac{\mu\gamma_k}{3G} \sqrt{\epsilon_{\text{bias}}} + \frac{8\gamma_k}{3} \mathbb{E} \|d_k - \nabla_\theta J(\theta_k)\| + \frac{L\gamma_k^2}{2} + \mathcal{O} \left(\frac{AGt_{\text{mix}} \log T}{T^3} \right)$.

The proof of Lemma 5 can be found in Appendix F. Lemma 5 implies that we can obtain a bound for $J^* - \mathbb{E}[J(\theta_K)]$ by bounding $\mathbb{E} \|d_k - \nabla_\theta J(\theta_k)\|$. In the following lemmas, we establish bounds for this quantity for both algorithms.

Lemma 6. Consider Algorithm 1 and let all the assumptions stated in Theorem 4 hold. Then for all $K \geq 1$, we have

$$\mathbb{E} \|d_K - \nabla_\theta J(\theta_K)\| \leq \mathcal{O} \left(\frac{Gt_{\text{mix}} \log T}{K^{2/5}} + \frac{G^2 L_h}{\mu^2 K^{2/5}} \right). \tag{36}$$

Using Lemma 6, we obtain the following last-iterate bound for Algorithm 1, for all $K \in \{1, \dots, T/H\}$. This is stronger than typical results in PG literature which only provide bounds of form $\frac{1}{K} \sum_{k=1}^K (J^* - \mathbb{E}[J(\theta_k)])$.

Theorem 4 (Last-iterate bound for Algorithm 1). Let $\{\theta_k\}_{k=1}^K$ be defined as in Algorithm 1. If Assumptions 1, 2, 3 and 4 hold, $\nabla J(\theta)$ is L_h -smooth, $\gamma_k = \frac{6G}{\mu(k+2)}$ and $\eta_k = \left(\frac{2}{k+2} \right)^{4/5}$ then the following inequality holds for all $K \in \{1, \dots, T/H\}$ and $H = 63t_{\text{mix}}t_{\text{hit}}(\log_2 T)^2$

$$J^* - \mathbb{E}[J(\theta_K)] \leq \sqrt{\epsilon_{\text{bias}}} + \mathcal{O} \left(\frac{AG^2 t_{\text{mix}} \log T}{\mu K^{2/5}} + \frac{G^3 L_h}{\mu^3 K^{2/5}} \right). \tag{37}$$

Similarly, for Algorithm 3, we have

Lemma 7. *Consider Algorithm 3 and let all the assumptions stated in Theorem 5 hold. Then for all $K \geq 1$, we have*

$$\mathbb{E} \|d_K - \nabla_{\theta} J(\theta_K)\| \leq \mathcal{O} \left(\frac{\sigma_g}{\sqrt{K}} + \frac{GM}{\sqrt{K}\mu} \right). \quad (38)$$

With this, we obtain the following bound

Theorem 5 (Last-iterate bound for Algorithm 3). *Let $\{\theta_k\}_{k=1}^K$ be defined as in Algorithm 1. If Assumptions 1, 2, 3 and 4 hold, $\gamma_k = \frac{6G}{\mu(k+2)}$ and $\eta_k = \frac{2}{k+2}$ then the following inequality holds for all $K \in \{1, \dots, T/H\}$ and $H = 63t_{\text{mix}}t_{\text{hit}}(\log_2 T)^2$*

$$\begin{aligned} & J^* - \mathbb{E}[J(\theta_K)] \\ & \leq \sqrt{\epsilon_{\text{bias}}} + \mathcal{O} \left(\frac{\sqrt{A}G^2t_{\text{mix}}\log T}{\mu\sqrt{K}} + \frac{\sqrt{A}G^4t_{\text{hit}}^2t_{\text{mix}}^2(\log T)^{3/2}}{\mu^2\sqrt{K}} + \frac{\sqrt{A}(BG + G^3)t_{\text{mix}}\log T}{\mu^2\sqrt{K}} \right). \end{aligned}$$

The proof of Lemma 6 is given in Appendix G, while the proof of Lemma 7 is given in Appendix H. Their proofs roughly proceed by recursively bounding $\mathbb{E} \|d_K - \nabla_{\theta} J(\theta_K)\|^2$. The challenges here include ensuring the accumulation of the bias is not too large and taking care of correlations in the cross-product terms involving previous estimates.

C Proof of Lemma 1

Let i denote the number of disjoint sub-trajectories of τ of length N that start with the given state s and are at least N distance apart. The advantage function estimate can be written as:

$$\hat{A}^{\pi}(\tau, s, a) = \begin{cases} \frac{1}{\pi(a|s)} \left[\frac{1}{i} \sum_{j=1}^i y_j 1(a_{\xi_j} = a) \right] - \frac{1}{i} \sum_{j=1}^i y_j & \text{if } i > 0 \\ 0 & \text{if } i = 0, \end{cases} \quad (39)$$

where ξ_j, y_j , as stated before, are the starting time and the sum of rewards in the j th sub-trajectory respectively. Using the definition of the Q function, one can show that,

$$\mathbb{E} \left[y_j \middle| s_{\xi_j} = s, a_{\xi_j} = a \right] = Q^{\pi}(s, a) + NJ^{\pi} - E_T^{\pi}(s, a), \quad (40)$$

where $E_T^{\pi}(s, a) := \sum_{s'} P(s'|s, a) \left[\sum_{j=N}^{\infty} (P^{\pi})^j(s', \cdot) - d^{\pi} \right]^{\top} r^{\pi}$. Observe that,

$$\begin{aligned} & \mathbb{E} \left[\left(\frac{1}{\pi(a|s)} y_j 1(a_{\xi_j} = a) - y_j \right) \middle| s_{\xi_j} = s \right] \\ & = \mathbb{E} \left[y_j \middle| s_{\xi_j} = s, a_{\xi_j} = a \right] - \sum_{a'} \pi(a'|s) \mathbb{E} \left[y_j \middle| s_{\xi_j} = s, a_{\xi_j} = a' \right] \\ & = Q^{\pi}(s, a) + NJ^{\pi} - E_T^{\pi}(s, a) - \sum_{a'} \pi(a'|s) [Q^{\pi}(s, a') + NJ^{\pi} - E_T^{\pi}(s, a')] \\ & = Q^{\pi}(s, a) - V^{\pi}(s) - \left[E_T(s, a) - \sum_{a'} \pi(a'|s) E_T^{\pi}(s, a') \right] \\ & = A^{\pi}(s, a) - \Delta_T^{\pi}(s, a) \end{aligned} \quad (41)$$

where $\Delta_T^{\pi}(s, a) := E_T(s, a) - \sum_{a'} \pi(a'|s) E_T^{\pi}(s, a')$. It follows from Lemma 16 that $|E_T^{\pi}(s, a)| \leq \frac{1}{T^6}$ which implies $|\Delta_T^{\pi}(s, a)| \leq \frac{2}{T^6}$. Therefore, one has the following bound.

$$\left| \mathbb{E} \left[\left(\frac{1}{\pi(a|s)} y_j 1(a_{\xi_j} = a) - y_j \right) \middle| s_{\xi_j} = s \right] - A^{\pi}(s, a) \right| \leq \frac{2}{T^6}. \quad (42)$$

Since i and the reward variables $\{y_j\}_{j=1}^i$ are correlated, we cannot directly obtain our desired result from the above bound. We use the methodology given in Wei et al. [2020] to address this challenge. In summary, we first prove bounds under an imaginary MDP where the state distribution ‘refreshes’ to the stationary distribution d^π after N time steps after the completion of a sub-trajectory. Under this MDP, i becomes decoupled from $\{y_j\}_{j=1}^i$. The resulting bounds from this framework can then be translated into the real MDP given that N is sufficiently large since this effectively makes the imaginary MDP ‘close’ to the real MDP.

More specifically, for a sub-trajectory beginning at ξ_j and ending at $\xi_j + N$, the system ‘rests’ for N additional steps before ‘refreshing’ with the state distribution d^π at $\xi_j + 2N$. The wait time between the ‘refreshing’ after the $(j-1)$ th sub-trajectory and the onset of the j th sub-trajectory is denoted as $w_j = \xi_j - (\xi_{j-1} + 2N)$ for all $j > 1$. Additionally, w_1 represents the time between the initiation of the trajectory τ and the commencement of its first sub-trajectory.

It is pertinent to notice the following: (a) w_1 relies solely on the initial state of τ , and the π -induced transition function, P^π , (b) For $j > 1$, w_j is solely contingent on the stationary distribution, d^π , and the induced transition function, P^π , (c) i is solely dependent on $\{w_1, w_2, \dots\}$, as other segments of τ maintain a fixed length of $2N$, (d) the sequence $\{w_1, w_2, \dots\}$ (and i consequently) remains independent of $\{y_1, y_2, \dots\}$.

We denote expectation taken in this system as \mathbb{E}' and probability of events similarly as \Pr' . From Lemma 14, it follows that $|A^\pi(s)| \leq \mathcal{O}(t_{\text{mix}})$. Now, define the following:

$$\Delta_j := \frac{y_j 1(a_{\xi_j} = a)}{\pi(a|s)} - y_j - A^\pi(s, a) + \Delta_T^\pi(s, a) \quad (43)$$

C.1 Proof of Lemma 1(a)

Note that $|y_j| \leq N$ and as a result $\mathbb{E}'[|\Delta_j|^2 | \{w_j\}] \leq \mathcal{O}(N^2/\pi(a|s))$. With this, we have

$$\begin{aligned} & \mathbb{E}' \left[\left(\hat{A}^\pi(s, a) - A^\pi(s, a) \right)^2 \right] \\ &= \mathbb{E}' \left[\left(\hat{A}^\pi(s, a) - A^\pi(s, a) \right)^2 \middle| i > 0 \right] \times \Pr'(i > 0) + (A^\pi(s, a))^2 \times \Pr'(i = 0) \\ &\leq 2\mathbb{E}'_{\{w_i\}} \left[\mathbb{E}' \left[\left(\frac{1}{i} \sum_{j=1}^i \Delta_j \right)^2 \middle| \{w_j\} \right] \middle| w_1 \leq |\tau| - N \right] \times \Pr'(w_1 \leq |\tau| - N) + 2(\Delta_T^\pi(s, a))^2 \\ &\quad + (A^\pi(s, a))^2 \times \Pr'(i = 0) \\ &\stackrel{(a)}{\leq} 2\mathbb{E}'_{\{w_j\}} \left[\frac{1}{i^2} \sum_{j=1}^i \mathbb{E}' [\Delta_j^2 | \{w_j\}] \middle| w_1 \leq |\tau| - N \right] \times \Pr'(w_1 \leq |\tau| - N) + \frac{8}{T^{12}} \\ &\quad + (A^\pi(s, a))^2 \times \Pr'(i = 0) \\ &\stackrel{(b)}{\leq} 2\mathbb{E}' \left[\frac{1}{i} \middle| w_1 \leq |\tau| - N \right] \mathcal{O} \left(\frac{N^2}{\pi(a|s)} \right) + \frac{8}{T^{12}} + \mathcal{O}(t_{\text{mix}}^2) \times \Pr'(i = 0), \end{aligned} \quad (44)$$

where (a) utilizes the bound $|\Delta_T^\pi(s, a)| \leq \frac{2}{T^6}$ derived earlier, and the fact that $\{\Delta_j\}$ are zero mean random variables that are conditionally independent given $\{w_j\}$. Inequality (b) is a consequence of Lemma 14. Notice that $i = 0$ is equivalent to $w_1 > |\tau| - N$. Using Lemma 17, we get:

$$\begin{aligned} \Pr'(w_1 > |\tau| - N) &\leq \left(1 - \frac{3d^\pi(s)}{4} \right)^{\frac{|\tau| - N}{N}} \leq \left(1 - \frac{3d^\pi(s)}{4} \right)^{9t_{\text{hit}} \log_2 T - 1} \\ &\stackrel{(a)}{\leq} \left(1 - \frac{3}{4t_{\text{hit}}} \right)^{8t_{\text{hit}} \log T} \leq \frac{1}{T^6} \end{aligned} \quad (45)$$

where (a) is derived from the definition of t_{hit} and the inequality that $9t_{\text{hit}} \log_2 T - 1 \geq 8t_{\text{hit}} \log T$. Towards bounding $\mathbb{E}' \left[\frac{1}{i} \middle| w_1 \leq |\tau| - N \right]$, define

$$i_0 := \frac{|\tau| - N}{2N + \frac{4N \log_2 T}{d^\pi(s)}} \quad (46)$$

Note that $i < i_0$ is identical to the event that at least one w_j is greater than $4N \log_2 T / d^\pi(s)$. From Lemma 17, we thus obtain:

$$\Pr'(i < i_0) \leq \left(1 - \frac{3d^\pi(s)}{4} \right)^{\frac{4 \log_2 T}{d^\pi(s)}} \leq \frac{1}{T^3} \quad (47)$$

Using this, we obtain

$$\begin{aligned} \mathbb{E}' \left[\frac{1}{i} \middle| i > 0 \right] &= \frac{\sum_{l=1}^{\infty} \frac{1}{l} \Pr'(i = l)}{\Pr'(i > 0)} \leq \frac{1 \times \Pr'(i \leq i_0) + \frac{1}{i_0} \Pr'(i > i_0)}{\Pr'(i > 0)} \\ &\leq \frac{1}{T^3} + \frac{2N + \frac{4N \log T}{d^\pi(s)}}{|\tau| - N} \leq \mathcal{O} \left(\frac{N \log T}{|\tau| d^\pi(s)} \right) \end{aligned} \quad (48)$$

Plugging in (45) and (48) into (44), we get

$$\mathbb{E}' \left[\left(\hat{A}^\pi(s, a) - A^\pi(s, a) \right)^2 \right] = \mathcal{O} \left(\frac{N^3 \log T}{H d^\pi(s) \pi(a|s)} \right) \quad (49)$$

We are now left with translating this result to the real MDP. Let $(\hat{A}^\pi(s, a) - A^\pi(s, a))^2 = f(X)$ where $X = (i, \xi_1, \mathcal{T}_1, \dots, \xi_i, \mathcal{T}_i)$, and $\mathcal{T}_j = (a_{\xi_j}, s_{\xi_j+1}, a_{\xi_j+1}, \dots, s_{\xi_j+N}, a_{\xi_j+N})$. We have,

$$\frac{\mathbb{E}[f(X)]}{\mathbb{E}'[f(X)]} = \frac{\sum_X f(X) \Pr(X)}{\sum_X f(X) \Pr'(X)} \leq \max_X \frac{\Pr(X)}{\Pr'(X)} \quad (50)$$

The last inequality uses the non-negativity of $f(\cdot)$. Observe that, for a fixed X , we have,

$$\begin{aligned} \Pr(X) &= \Pr(\xi_1) \times \Pr(\mathcal{T}_1 | \xi_1) \times \Pr(\xi_2 | \xi_1, \mathcal{T}_1) \times \Pr(\mathcal{T}_2 | \xi_2) \times \dots \\ &\quad \times \Pr(\xi_i | \xi_{i-1}, \mathcal{T}_{i-1}) \times \Pr(\mathcal{T}_i | \xi_i) \times \Pr(s_t \neq s, \forall t \in [\xi_i + 2N, t_e - N] | \xi_i, \mathcal{T}_i), \\ \Pr'(X) &= \Pr(\xi_1) \times \Pr(\mathcal{T}_1 | \xi_1) \times \Pr'(\xi_2 | \xi_1, \mathcal{T}_1) \times \Pr'(\mathcal{T}_2 | \xi_2) \times \dots \\ &\quad \times \Pr'(\xi_i | \xi_{i-1}, \mathcal{T}_{i-1}) \times \Pr'(\mathcal{T}_i | \xi_i) \times \Pr(s_t \neq s, \forall t \in [\xi_i + 2N, t_e - N] | \xi_i, \mathcal{T}_i), \end{aligned}$$

Thus, the difference between $\Pr(X)$ and $\Pr'(X)$ arises because $\Pr(\xi_{j+1} | \xi_j, \mathcal{T}_j) \neq \Pr'(\xi_{j+1} | \xi_j, \mathcal{T}_j)$, $\forall j \in \{1, \dots, i-1\}$. Note that the ratio of these two terms can be bounded as follows.

$$\begin{aligned} &\frac{\Pr(\xi_{j+1} | \xi_j, \mathcal{T}_j)}{\Pr'(\xi_{j+1} | \xi_j, \mathcal{T}_j)} \\ &= \frac{\sum_{s' \neq s} \Pr(s_{\xi_j+2N} = s' | \xi_j, \mathcal{T}_j) \times \Pr(s_t \neq s, \forall t \in [\xi_j + 2N, \xi_{j+1} - 1], s_{\xi_{j+1}} = s | s_{\xi_j+2N} = s')}{\sum_{s' \neq s} \Pr'(s_{\xi_j+2N} = s' | \xi_j, \mathcal{T}_j) \times \Pr(s_t \neq s, \forall t \in [\xi_j + 2N, \xi_{j+1} - 1], s_{\xi_{j+1}} = s | s_{\xi_j+2N} = s')} \\ &\leq \max_{s'} \frac{\Pr(s_{\xi_j+2N} = s' | \xi_j, \mathcal{T}_j)}{\Pr'(s_{\xi_j+2N} = s' | \xi_j, \mathcal{T}_j)} \\ &= \max_{s'} 1 + \frac{\Pr(s_{\xi_j+2N} = s' | \xi_j, \mathcal{T}_j) - d^\pi(s')}{d^\pi(s')} \stackrel{(a)}{\leq} \max_{s'} 1 + \frac{1}{T^6 d^\pi(s')} \leq 1 + \frac{t_{\text{hit}}}{T^6} \leq 1 + \frac{1}{T^5} \end{aligned} \quad (51)$$

where (a) is a consequence of Lemma 16. We have,

$$\frac{\Pr(X)}{\Pr'(X)} \leq \left(1 + \frac{1}{T^5}\right)^i \leq e^{\frac{i}{T^5}} \stackrel{(a)}{\leq} e^{\frac{1}{T^4}} \stackrel{(b)}{\leq} \left(1 + \frac{2}{T^4}\right) \quad (52)$$

where (a) uses the fact that $i \leq T$ while (b) is a consequence of the inequality that $\exp(x) \leq 1 + 2x$ for sufficiently small $x > 0$. Combining (50) and (52), we get,

$$\begin{aligned} \mathbb{E} \left[\left(\hat{A}^\pi(s, a) - A^\pi(s, a) \right)^2 \right] &\leq \mathcal{O} \left(1 + \frac{1}{T^4} \right) \mathbb{E}' \left[\left(\hat{A}^\pi(s, a) - A^\pi(s, a) \right)^2 \right] \\ &\stackrel{(a)}{\leq} \mathcal{O} \left(\frac{N^3 \log T}{H d^\pi(s) \pi(a|s)} \right) \end{aligned} \quad (53)$$

where (a) follows from (49).

C.2 Proof of Lemma 1(b)

Observe the following relations.

$$\begin{aligned} &\left| \mathbb{E}' \left[\hat{A}^\pi(s, a) - A^\pi(s, a) \right] \right| \\ &= \left| \mathbb{E}' \left[\hat{A}^\pi(s, a) - A^\pi(s, a) \middle| i > 0 \right] \times \Pr'(i > 0) + A^\pi(s, a) \times \Pr'(i = 0) \right| \\ &= \left| \mathbb{E}' \left[\frac{1}{i} \sum_{j=1}^i \Delta_j - \Delta_T^\pi(s, a) \middle| i > 0 \right] \times \Pr'(i > 0) + A^\pi(s, a) \times \Pr'(i = 0) \right| \\ &\stackrel{(a)}{=} \left| \mathbb{E}' \left[-\Delta_T^\pi(s, a) \middle| i > 0 \right] \times \Pr'(i > 0) + A^\pi(s, a) \times \Pr'(i = 0) \right| \\ &\leq \mathbb{E}' \left[\left| \Delta_T^\pi(s, a) \right| \middle| i > 0 \right] + |A^\pi(s, a)| \times \Pr'(i = 0) \stackrel{(b)}{=} \mathcal{O}(t_{\text{mix}} T^{-6}), \end{aligned} \quad (54)$$

where (a) follows from the fact that $\{\Delta_j\}$ are zero mean random variables and (b) utilizes Lemma 14 and the bound $|\Delta_T^\pi(s, a)| \leq \frac{2}{T^6}$ that was derived earlier. Define $X = (i, \xi_1, \mathcal{T}_1, \dots, \xi_i, \mathcal{T}_i)$ and $f(X) = \hat{A}^\pi(s, a) - A^\pi(s, a)$ where $\mathcal{T}_j = (a_{\xi_j}, a_{\xi_j}, s_{\xi_j+1}, a_{\xi_j+1}, \dots, s_{\xi_j+N}, a_{\xi_j+N})$. For any $f(\cdot)$ (not necessarily non-negative), observe that

$$\begin{aligned} |\mathbb{E}[f(X)] - \mathbb{E}'[f(X)]| &= \left| \sum_X f(X) \Pr(X) - \sum_X f(X) \Pr'(X) \right| \\ &\leq \sum_X |f(X)| |\Pr(X) - \Pr'(X)|. \end{aligned}$$

We now focus on bounding $|\Pr(X) - \Pr'(X)|$. We have the result stated below from (52).

$$\Pr(X) \leq \left(1 + \frac{2}{T^4}\right) \Pr'(X) \quad (55)$$

Furthermore, one can derive the following lower bound using Lemma 16.

$$\begin{aligned} &\frac{\Pr(\xi_{j+1} | \xi_j, \mathcal{T}_j)}{\Pr'(\xi_{j+1} | \xi_j, \mathcal{T}_j)} \\ &= \frac{\sum_{s' \neq s} \Pr(s_{\xi_j+2N} = s' | \xi_j, \mathcal{T}_j) \times \Pr(s_t \neq s, \forall t \in [\xi_j + 2N, \xi_{j+1} - 1], s_{\xi_{j+1}} = s | s_{\xi_j+2N} = s')}{\sum_{s' \neq s} \Pr'(s_{\xi_j+2N} = s' | \xi_j, \mathcal{T}_j) \times \Pr(s_t \neq s, \forall t \in [\xi_j + 2N, \xi_{j+1} - 1], s_{\xi_{j+1}} = s | s_{\xi_j+2N} = s')} \\ &\geq \min_{s'} \frac{\Pr(s_{\xi_j+2N} = s' | \xi_j, \mathcal{T}_j)}{\Pr'(s_{\xi_j+2N} = s' | \xi_j, \mathcal{T}_j)} \\ &= \min_{s'} 1 + \frac{\Pr(s_{\xi_j+2N} = s' | \xi_j, \mathcal{T}_j) - d^\pi(s')}{d^\pi(s')} \geq \min_{s'} 1 - \frac{1}{T^6 d^\pi(s')} \geq 1 - \frac{t_{\text{hit}}}{T^6} \geq 1 - \frac{1}{T^5} \end{aligned} \quad (56)$$

By utilising a similar argument used in deriving (52), we arrive at the following:

$$\frac{\Pr(X)}{\Pr'(X)} \geq \left(1 - \frac{1}{T^5}\right)^i \stackrel{(a)}{\geq} e^{-\frac{2i}{T^5}} \geq e^{-\frac{2}{T^4}} \stackrel{(b)}{\geq} 1 - \frac{4}{T^4}. \quad (57)$$

where (a) and (b) are consequences of the inequality that $1 - x \geq \exp(-2x) \geq 1 - 4x$ for sufficiently small $x > 0$. Therefore, we arrive at the following.

$$\left(1 - \frac{4}{T^4}\right) \Pr'(X) \leq \Pr(X) \leq \left(1 + \frac{2}{T^4}\right) \Pr'(X)$$

which is equivalent to the following relation.

$$-\frac{4}{T^4} \cdot \Pr'(X) \leq \Pr(X) - \Pr'(X) \leq \frac{2}{T^4} \cdot \Pr'(X)$$

It follows that

$$\begin{aligned} \sum_X |f(X)| |\Pr(X) - \Pr'(X)| &\leq \frac{4}{T^4} \sum_X |f(X)| \Pr'(X) \\ &= \frac{4}{T^4} \mathbb{E}'[|f(X)|] \\ &= \frac{4}{T^4} \mathbb{E}'[|\hat{A}^\pi(s, a) - A^\pi(s, a)|] \\ &\leq \frac{4}{T^4} \mathbb{E}'[(\hat{A}^\pi(s, a) - A^\pi(s, a))^2]^{1/2} \stackrel{(a)}{\leq} \mathcal{O}\left(\frac{t_{\text{mix}} \log T}{\sqrt{\pi(a|s)} T^4}\right) \end{aligned} \quad (58)$$

where (a) results from (49). We finally have

$$\begin{aligned} \left| \mathbb{E} [\hat{A}^\pi(s, a) - A^\pi(s, a)] \right| &= \left| \mathbb{E} [f(X)] \right| \leq \left| \mathbb{E}' [f(X)] \right| + \sum_X |f(X)| |\Pr(X) - \Pr'(X)| \\ &\leq \mathcal{O}\left(\frac{t_{\text{mix}} \log T}{\sqrt{\pi(a|s)} T^4}\right) \end{aligned}$$

This concludes the proof.

D Proof of Lemma 2

D.1 Proof of Lemma 2(a)

Assume the trajectory being used for estimation is $\tau = \{(s_t, a_t)\}_{t=t_s}^{t_e}$. Observe that

$$\begin{aligned} \mathbb{E} \|g(\theta, \tau) - \nabla_\theta J(\theta)\|^2 &= \mathbb{E} \|g(\theta, \tau) - \bar{g}(\theta, \tau) + \bar{g}(\theta, \tau) - \nabla_\theta J(\theta)\|^2 \\ &\leq 2 \mathbb{E} \|g(\theta, \tau) - \bar{g}(\theta, \tau)\|^2 + 2 \mathbb{E} \|\bar{g}(\theta, \tau) - \nabla_\theta J(\theta)\|^2 \end{aligned} \quad (59)$$

where $\bar{g}(\theta, \tau) = \frac{1}{|\tau| - N} \sum_{t=t_s+N}^{t_e} A^{\pi_\theta}(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t)$. We apply the following lemma to bound the second term in the RHS of (59):

Lemma 8. [Dorfman and Levy, 2022, Lemma A.6] Fix a trajectory $\tau = \{(s_t, a_t)\}_{t \in \mathbb{N}}$ generated by following π_θ starting from some initial state $s_0 \sim \bar{\rho}$. Let $\nabla L(\theta)$ be a gradient that we wish to estimate over τ , and $l(\theta, \cdot, \cdot)$ is a function such that $\mathbb{E}_{s \sim d^{\pi_\theta}, a \sim \pi_\theta(s)} l(\theta, s, a) = \nabla L(\theta)$. Let $\|l(\theta, s, a)\|, \|\nabla L(\theta)\| \leq G_L, \forall \theta \in \Theta, \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$. Define $l^Q \triangleq \frac{1}{Q} \sum_{i=1}^Q l(\theta, s_i, a_i)$. If $P = 2t_{\text{mix}} \log T$, then the following holds as long as $Q \leq T$,

$$\mathbb{E} [\|l^Q - \nabla L(\theta)\|^2] \leq \mathcal{O}\left(G_L^2 \log(PQ) \frac{P}{Q}\right). \quad (60)$$

Applying Lemma 8, and 14, we get

$$\mathbb{E} \left[\|\bar{g}(\theta, \tau) - \nabla_{\theta} J(\theta)\|^2 \right] \leq \mathcal{O} \left(G^2 t_{\text{mix}}^2 \log T \right) \times \mathcal{O} \left(\frac{t_{\text{mix}} \log T}{|\tau| - N} \right) = \mathcal{O} \left(\frac{G^2 t_{\text{mix}}^3 (\log T)^2}{H} \right) \quad (61)$$

The first term can be bounded using Assumption 2 as follows.

$$\begin{aligned} & \mathbb{E} \|g(\theta, \tau) - \bar{g}(\theta, \tau)\|^2 \\ &= \frac{1}{(|\tau| - N)^2} \mathbb{E} \left\| \sum_{t=t_s+N}^{t_e} \hat{A}^{\pi_{\theta}}(s_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) - \sum_{t=t_s+N}^{t_e} A^{\pi_{\theta}}(s_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right\|^2 \\ &\leq \left(\frac{G^2}{|\tau| - N} \right) \mathbb{E} \left[\sum_{t=t_s+N}^{t_e} \left[\hat{A}^{\pi_{\theta}}(\tau, s_t, a_t) - A^{\pi_{\theta}}(s_t, a_t) \right]^2 \right] \\ &= \left(\frac{G^2}{|\tau| - N} \right) \mathbb{E} \left[\sum_{t=t_s+N}^{t_e} \sum_{a_t \in \mathcal{A}} \pi_{\theta}(a_t | s_t) \mathbb{E}[(\hat{A}^{\pi_{\theta}}(\tau, s_t, a_t) - A^{\pi_{\theta}}(s_t, a_t))^2 | s_t, a_t] \right] \\ &= \mathcal{O} \left(\frac{AG^2 t_{\text{hit}} N^3 \log T}{H} \right) \end{aligned} \quad (62)$$

where the last equality follows from Lemma 1.

D.2 Proof of Lemma 2(b)

We begin by observing that, for a π_{θ} -induced trajectory $\tau = \{(s_t, a_t)\}_{t=t_s}^{t_e}$, the following holds.

$$\begin{aligned} & \|\mathbb{E}[g(\theta, \tau)] - \nabla_{\theta} J(\theta)\| \\ &= \left\| \mathbb{E} \left[\frac{1}{|\tau| - N} \sum_{t=t_s+N}^{t_e} \hat{A}^{\pi_{\theta}}(\tau, s_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right] - \nabla_{\theta} J(\theta) \right\| \\ &\leq \frac{1}{|\tau| - N} \left\| \underbrace{\mathbb{E} \left[\sum_{t=t_s+N}^{t_e} \hat{A}^{\pi_{\theta}}(\tau, s_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) - \sum_{t=t_s+N}^{t_e} A^{\pi_{\theta}}(s_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right]}_{T_1} \right\| \\ &\quad + \left\| \underbrace{\mathbb{E} \left[\frac{1}{|\tau| - N} \sum_{t=t_s+N}^{t_e} A^{\pi_{\theta}}(s_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right] - \nabla_{\theta} J(\theta)}_{T_2} \right\| \end{aligned} \quad (63)$$

Note that the term T_1 can be bounded as follows.

$$\begin{aligned} T_1 &\leq \mathbb{E} \left[\sum_{t=t_s+N}^{t_e} \left| \mathbb{E} \left[\hat{A}^{\pi_{\theta}}(\tau, s_t, a_t) - A^{\pi_{\theta}}(s_t, a_t) | s_t, a_t \right] \right| \left\| \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right\| \right] \\ &\stackrel{(a)}{\leq} G \mathbb{E} \left[\sum_{t=t_s+N}^{t_e} \left| \mathbb{E}[\hat{A}^{\pi_{\theta}}(\tau, s_t, a_t) - A^{\pi_{\theta}}(s_t, a_t) | s_t, a_t] \right| \right] \\ &\leq G \mathbb{E} \left[\sum_{t=t_s+N}^{t_e} \sum_{a_t \in \mathcal{A}} \pi_{\theta}(a_t | s_t) \left| \mathbb{E}[\hat{A}^{\pi_{\theta}}(s_t, a_t) - A^{\pi_{\theta}}(s_t, a_t) | s_t, a_t] \right| \right] \\ &\stackrel{(b)}{\leq} \mathcal{O} \left(\frac{G t_{\text{mix}} \log T}{T^4} \right) \mathbb{E} \left[\sum_{t=t_s+N}^{t_e} \sum_{a_t \in \mathcal{A}} \sqrt{\pi_{\theta}(a_t | s_t)} \right] \leq \mathcal{O} \left(\frac{\sqrt{A} G t_{\text{mix}} \log T}{T^4} \right) (|\tau| - N) \end{aligned} \quad (64)$$

where (a), (b) follow from Assumption 2 and Lemma 1 respectively. To bound T_2 , define

$$f(s) := \frac{1}{|\tau| - N} \mathbb{E} \left[\sum_{t=t_s+N}^{t_e} A^{\pi_{\theta}}(s_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \middle| s_{t_s+N} = s \right], \quad \forall s \in \mathcal{S} \quad (65)$$

Using the stationarity property of d^{π_θ} and the policy gradient expression (6), we deduce

$$\nabla_\theta J(\theta) = \sum_{s \in \mathcal{S}} d^{\pi_\theta}(s) f(s) \quad (66)$$

If $\tau = \{(s_t, a_t)\}_{t=t_s}^{t_e}$ is a π_θ -induced trajectory with $s_{t_s} \sim \bar{\rho}$, then $s_{t_s+N} \sim (P^{\pi_\theta})^N \bar{\rho}$. Hence,

$$\begin{aligned} T_2 &= \left\| \sum_{s \in \mathcal{S}} [(P^{\pi_\theta})^N \bar{\rho} - d^{\pi_\theta}(s)] f(s) \right\| \\ &\leq \sum_{s \in \mathcal{S}} \left\| \left\{ \sum_{s' \in \mathcal{S}} (P^{\pi_\theta})^N(s', s) \bar{\rho}(s') \right\} - \left\{ d^{\pi_\theta}(s) \sum_{s' \in \mathcal{S}} \bar{\rho}(s') \right\} \right\| \|f(s)\| \\ &\leq \sum_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} |(P^{\pi_\theta})^N(s', s) - d^{\pi_\theta}(s)| \bar{\rho}(s') \|f(s)\| \\ &\leq \sum_{s' \in \mathcal{S}} \bar{\rho}(s') \|(P^{\pi_\theta})^N(s', \cdot) - d^{\pi_\theta}\|_1 \max_{s \in \mathcal{S}} \|f(s)\| \\ &\stackrel{(a)}{\leq} \sum_{s' \in \mathcal{S}} \bar{\rho}(s') \cdot 2 \cdot 2^{-\frac{N}{t_{\text{mix}}}} \mathcal{O}(G t_{\text{mix}}) = \mathcal{O}\left(\frac{G t_{\text{mix}}}{T^7}\right) \end{aligned} \quad (67)$$

where (s) uses Lemma 14, 15 and Assumption 2. Combining the bounds of T_1 and T_2 , we conclude the result.

D.3 Proof of Lemma 2(c)

Using the expression of $B(\theta, \tau)$ given in (11), we can derive the following.

$$\begin{aligned} \mathbb{E} \|B(\theta, \tau)\|^2 &= \mathbb{E} \|\nabla_\theta \Phi(\theta, \tau) \nabla_\theta \log p(\tau, \theta, \bar{\rho})^\top + \nabla_\theta^2 \Phi(\theta, \tau)\|^2 \\ &\leq 2 \mathbb{E} \|\nabla_\theta \Phi(\theta, \tau)\|^2 \|\nabla_\theta \log p(\tau, \theta, \bar{\rho})\|^2 + 2 \mathbb{E} \|\nabla_\theta^2 \Phi(\theta, \tau)\|^2 \end{aligned} \quad (68)$$

Observe the inequality stated below.

$$\|\nabla_\theta \log p(\tau, \theta, \bar{\rho})\|^2 = \left\| \sum_{t=t_s}^{t_e} \nabla_\theta \log \pi_\theta(a_t | s_t) \right\|^2 \leq |\tau| \sum_{t=t_s}^{t_e} \|\nabla_\theta \log \pi_\theta(a_t | s_t)\|^2 \leq G^2 |\tau|^2 \quad (69)$$

where the last inequality uses Assumption 2. It follows that

$$\mathbb{E} \|B(\theta, \tau)\|^2 \leq 2G^2 |\tau|^2 \mathbb{E} \|\nabla \Phi(\theta, \tau)\|^2 + 2 \mathbb{E} \|\nabla^2 \Phi(\theta, \tau)\|^2 \quad (70)$$

Using (27), one can write

$$\begin{aligned} \mathbb{E} \|\nabla \Phi(\theta, \tau)\|^2 &= \mathbb{E} \left\| \frac{1}{|\tau| - N} \sum_{t=t_s+N}^{t_e} \hat{A}^{\pi_\theta}(\tau, s_t, a_t) \nabla_\theta \log \pi_\theta(a_t | s_t) \right\|^2 \\ &\leq \frac{1}{|\tau| - N} \sum_{t=t_s+N}^{t_e} \mathbb{E} \left\| \hat{A}^{\pi_\theta}(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t | s_t) \right\|^2 \\ &\stackrel{(a)}{\leq} \frac{G^2}{|\tau| - N} \sum_{t=t_s+N}^{t_e} \mathbb{E} |\hat{A}^{\pi_\theta}(s_t, a_t)|^2 \\ &\stackrel{(b)}{\leq} \frac{G^2}{|\tau| - N} \sum_{t=t_s+N}^{t_e} \mathbb{E} \left[\frac{N^2}{\pi_\theta(a_t | s_t)} \right] \\ &\leq \frac{G^2}{|\tau| - N} \sum_{t=t_s+N}^{t_e} \mathbb{E} \left[\sum_{a \in \mathcal{A}} \pi_\theta(a | s_t) \mathbb{E} \left[\frac{N^2}{\pi_\theta(a | s_t)} \right] \right] \leq AN^2 G^2 \end{aligned} \quad (71)$$

where (a) uses Assumption 2 and (b) follows from the discussion in the proof of Lemma 1(a). Also,

$$\begin{aligned}
 & \mathbb{E} \|\nabla_\theta^2 \Phi(\theta, \tau)\|^2 \\
 &= \mathbb{E} \left\| \frac{1}{|\tau| - N} \sum_{t=t_s+N}^{t_e} \left\{ \Psi_t^{(1)}(\tau) \nabla_\theta^2 \log \pi_\theta(a_t|s_t) \right. \right. \\
 &\quad \left. \left. - \Psi_t^{(2)}(\tau) \left(\frac{\nabla_\theta^2 \log \pi_\theta(a_t|s_t)}{\pi_\theta(a_t|s_t)} - \frac{\nabla_\theta \log \pi_\theta(a_t|s_t) \nabla_\theta \log \pi_\theta(a_t|s_t)^\top}{\pi_\theta(a_t|s_t)} \right) \right\} \right\|^2 \\
 &= \frac{1}{|\tau| - N} \sum_{t=t_s+N}^{t_e} \mathbb{E} \left\| \Psi_t^{(1)}(\tau) \nabla_\theta^2 \log \pi_\theta(a_t|s_t) \right. \\
 &\quad \left. - \Psi_t^{(2)}(\tau) \left(\frac{\nabla_\theta^2 \log \pi_\theta(a_t|s_t)}{\pi_\theta(a_t|s_t)} - \frac{\nabla_\theta \log \pi_\theta(a_t|s_t) \nabla_\theta \log \pi_\theta(a_t|s_t)^\top}{\pi_\theta(a_t|s_t)} \right) \right\|^2 \\
 &\leq \frac{3}{|\tau| - N} \sum_{t=t_s+N}^{t_e} \left\{ B^2 \mathbb{E} \left[\left\| \Psi_t^{(1)}(\tau) \right\|^2 \right] + \mathbb{E} \left[\left\| \Psi_t^{(2)}(\tau) \right\|^2 \left(\frac{B^2}{\pi_\theta^2(a_t|s_t)} + \frac{G^4}{\pi_\theta^2(a_t|s_t)} \right) \right] \right\}
 \end{aligned} \tag{72}$$

It follows from the definitions of $\{\Psi_t^{(k)}(\tau)\}_{k \in \{0,1\}}$ that

$$\mathbb{E} \left[\left\| \Psi_t^{(1)}(\tau) \right\|^2 \right] \leq N^2 \text{ and } \mathbb{E} \left[\left\| \Psi_t^{(2)}(\tau) \right\|^2 \middle| s_t, a_t \right] \leq N^2 \pi_\theta(a_t|s_t) \tag{73}$$

Using the above inequalities, one can derive the relation stated below.

$$\begin{aligned}
 \mathbb{E} \|\nabla^2 \Phi(\theta, \tau)\|^2 &\leq 3B^2 N^2 + \frac{3}{|\tau| - N} \sum_{t=t_s+N}^{t_e} N^2 \mathbb{E} \left[\frac{B^2 + G^4}{\pi_\theta(a_t|s_t)} \right] \\
 &\leq 3B^2 N^2 + \frac{3}{|\tau| - N} \sum_{t=t_s+N}^{t_e} N^2 \mathbb{E} \left[\sum_{a \in \mathcal{A}} \pi_\theta(a|s_t) \left[\frac{B^2 + G^4}{\pi_\theta(a|s_t)} \right] \right] \\
 &\leq 3B^2 N^2 + 3AN^2(B^2 + G^4)
 \end{aligned} \tag{74}$$

Combining the above results, we finally obtain

$$\mathbb{E} \|B(\theta, \tau)\|^2 \leq 2AN^2 G^4 |\tau|^2 + 6B^2 N^2 + 6AN^2(B^2 + G^4) \tag{75}$$

E Proof of Lemma 4

The regret for Algorithm 1 can be decomposed as:

$$\text{Reg}_T = \sum_{t=0}^{T-1} (J^* - r(s_t, a_t)) = H \sum_{k=1}^K (J^* - J(\tilde{\theta}_k)) + \sum_{k=1}^K \sum_{t \in \mathcal{I}_k} (J(\tilde{\theta}_k) - r(s_t, a_t)) \tag{76}$$

where $\mathcal{I}_k \triangleq \{(k-1)H, \dots, kH-1\}$.

Whereas, the regret for Algorithm 3 is (expressed below) slightly different since the trajectories are sampled using the sequence $\{\theta_1, \hat{\theta}_1, \theta_2, \hat{\theta}_2, \dots\}$, instead of $\{\tilde{\theta}_1, \tilde{\theta}_2, \dots\}$:

$$\begin{aligned}
 \text{Reg}_T &= \sum_{t=0}^{T-1} (J^* - r(s_t, a_t)) = H \sum_{k=1}^K (J^* - J(\theta_k)) + \sum_{k=1}^K \sum_{t \in \mathcal{I}_k} (J(\theta_k) - r(s_t, a_t)) \\
 &\quad + H \sum_{k=1}^K (J^* - J(\hat{\theta}_k)) + \sum_{k=1}^K \sum_{t \in \mathcal{I}_k} (J(\hat{\theta}_k) - r(s_t, a_t)) \\
 &\stackrel{(a)}{=} \mathcal{O} \left(H \sum_{k=1}^K (J^* - J(\theta_k)) + \sum_{k=1}^K \sum_{t \in \mathcal{I}_k} (J(\theta_k) - r(s_t, a_t)) \right),
 \end{aligned} \tag{77}$$

where the proof of (a) in Equation (77) is provided in Section H. Thus, we can focus on the decomposition of the form in (76). The expectation of the second term in (76) can be expressed as follows,

$$\begin{aligned}
 \mathbb{E} \left[\sum_{k=1}^K \sum_{t \in \mathcal{I}_k} \left(J(\tilde{\theta}_k) - r(s_t, a_t) \right) \right] &\stackrel{(a)}{=} \mathbb{E} \left[\sum_{k=1}^K \sum_{t \in \mathcal{I}_k} \mathbb{E}_{s' \sim P(\cdot | s_t, a_t)} [V^{\pi_{\tilde{\theta}_k}}(s')] - Q^{\pi_{\tilde{\theta}_k}}(s_t, a_t) \right] \\
 &\stackrel{(b)}{=} \mathbb{E} \left[\sum_{k=1}^K \sum_{t \in \mathcal{I}_k} V^{\pi_{\tilde{\theta}_k}}(s_{t+1}) - V^{\pi_{\tilde{\theta}_k}}(s_t) \right] \\
 &= \mathbb{E} \left[\sum_{k=1}^K V^{\pi_{\tilde{\theta}_k}}(s_{kH}) - V^{\pi_{\tilde{\theta}_k}}(s_{(k-1)H}) \right] \\
 &= \mathbb{E} \left[\underbrace{\sum_{k=1}^{K-1} V^{\pi_{\tilde{\theta}_{k+1}}}(s_{kH}) - V^{\pi_{\tilde{\theta}_k}}(s_{kH})}_{T_3} \right] \\
 &\quad + \underbrace{\mathbb{E} [V^{\pi_{\tilde{\theta}_K}}(s_T) - V^{\pi_{\tilde{\theta}_0}}(s_0)]}_{T_4}
 \end{aligned} \tag{78}$$

where (a) follows from Bellman equation and (b) follows from the fact that $\mathbb{E}[V^{\pi_{\tilde{\theta}_k}}(s_{t+1})] = \mathbb{E}_{s' \sim P(\cdot | s_t, a_t)} [V^{\pi_{\tilde{\theta}_k}}(s')]$ and $\mathbb{E}[V^{\pi_{\tilde{\theta}_k}}(s_t)] = \mathbb{E}[Q^{\pi_{\tilde{\theta}_k}}(s_t, a_t)]$. The term T_3 in (78) can be bounded as in Wei et al. [2020], Bai et al. [2024]. This is given in Lemma 9 (as stated below). Furthermore, the term T_4 can be upper-bounded as $\mathcal{O}(t_{\text{mix}})$ since $V^\pi(\cdot) \leq 5t_{\text{mix}}$ Wei et al. [2020].

Lemma 9. *If Assumptions 1 and 2 hold, then the following inequalities are true $\forall k, \forall (s, a) \in \mathcal{S} \times \mathcal{A}$*

$$\begin{aligned}
 (a) \quad &|\pi_{\theta_{k+1}}(a|s) - \pi_{\theta_k}(a|s)| \leq \mathcal{O}(G\pi_{\bar{\theta}_k}(a|s)\|\theta_{k+1} - \theta_k\|) \\
 (b) \quad &\sum_{k=1}^{K-1} \mathbb{E} |J(\theta_{k+1}) - J(\theta_k)| \leq \mathcal{O}(Gt_{\text{mix}} \sum_{k=1}^K \|\theta_{k+1} - \theta_k\|) \\
 (c) \quad &\sum_{k=1}^K \mathbb{E} |V^{\pi_{\theta_{k+1}}}(s) - V^{\pi_{\theta_k}}(s)| \leq \mathcal{O}(Gt_{\text{mix}}^2 \log^2 T \sum_{k=1}^K \|\theta_{k+1} - \theta_k\|),
 \end{aligned}$$

where $\bar{\theta}_k$ is some convex combination of θ_k and θ_{k+1} .

Lemma 9 can be understood as providing stability insights into our algorithm. It is similar to Lemma 7 in Wei et al. [2020]. Essentially, it asserts that the policy parameters undergo updates in a manner that reduces the average difference between consecutive average reward and value functions as the number of iterations, denoted by k , increases.

Proof of Lemma 9. Using Taylor's expansion, we can write the following $\forall (s, a) \in \mathcal{S} \times \mathcal{A}, \forall k$.

$$\begin{aligned}
 |\pi_{\theta_{k+1}}(a|s) - \pi_{\theta_k}(a|s)| &= |(\theta_{k+1} - \theta_k)^\top \nabla_\theta \pi_{\bar{\theta}}(a|s)| \\
 &= \pi_{\bar{\theta}_k}(a|s) |(\theta_{k+1} - \theta_k)^\top \nabla_\theta \log \pi_{\bar{\theta}_k}(a|s)| \\
 &\leq \pi_{\bar{\theta}_k}(a|s) \|\theta_{k+1} - \theta_k\| \|\nabla_\theta \log \pi_{\bar{\theta}_k}(a|s)\| \stackrel{(a)}{\leq} G\pi_{\bar{\theta}_k}(a|s) \|\theta_{k+1} - \theta_k\|
 \end{aligned} \tag{79}$$

where $\bar{\theta}_k$ is some convex combination of θ_k and θ_{k+1} and (a) follows from Assumption 2. This concludes the first statement.

Lemma 10. [Wei et al., 2020, Lemma 15] *For two policies π and π' , the difference of the objective function J is*

$$J^\pi - J^{\pi'} = \sum_s \sum_a d^\pi(s) (\pi(a|s) - \pi'(a|s)) Q^{\pi'}(s, a) \tag{80}$$

Applying (79) and Lemma 10, we obtain,

$$\begin{aligned}
 \sum_{k=1}^K \mathbb{E} |J(\theta_{k+1}) - J(\theta_k)| &= \sum_{k=1}^K \mathbb{E} \left| \sum_{s,a} d^{\pi_{\theta_{k+1}}}(s) (\pi_{\theta_{k+1}}(a|s) - \pi_{\theta_k}(a|s)) Q^{\pi_{\theta_k}}(s, a) \right| \\
 &\leq \sum_{k=1}^K \mathbb{E} \left[\sum_{s,a} d^{\pi_{\theta_{k+1}}}(s) |\pi_{\theta_{k+1}}(a|s) - \pi_{\theta_k}(a|s)| |Q^{\pi_{\theta_k}}(s, a)| \right] \\
 &\leq G \sum_{k=1}^K \mathbb{E} \left[\sum_{s,a} d^{\pi_{\theta_{k+1}}}(s) \pi_{\bar{\theta}_k}(a|s) \|\theta_{k+1} - \theta_k\| |Q^{\pi_{\theta_{k+1}}}(s, a)| \right] \\
 &\leq G \sum_{k=1}^K \mathbb{E} \left[\underbrace{\sum_{s,a} d^{\pi_{\theta_{k+1}}}(s) \pi_{\bar{\theta}_k}(a|s) \gamma_k}_{=1} \cdot 6t_{\text{mix}} \right] \\
 &= 6Gt_{\text{mix}} \sum_{k=1}^K \|\theta_{k+1} - \theta_k\|.
 \end{aligned} \tag{81}$$

Next, recall that for any policy π_θ , $r^{\pi_\theta}(s) := \sum_a \pi_\theta(a|s)r(s, a)$. Note that, for any policy parameter θ , and any state $s \in \mathcal{S}$, the following holds.

$$\begin{aligned}
 V^{\pi_\theta}(s) &= \sum_{t=0}^{\infty} \langle (P^{\pi_\theta})^t(s, \cdot) - d^{\pi_\theta}, r^{\pi_\theta} \rangle = \sum_{t=0}^{N-1} \langle (P^{\pi_\theta})^t(s, \cdot), r^{\pi_\theta} \rangle - NJ(\theta) \\
 &\quad + \sum_{t=N}^{\infty} \langle (P^{\pi_\theta})^t(s, \cdot) - d^{\pi_\theta}, r^{\pi_\theta} \rangle.
 \end{aligned} \tag{82}$$

Define the following quantity

$$\delta^{\pi_\theta}(s, N) := \sum_{t=N}^{\infty} \|(P^{\pi_\theta})^t(s, \cdot) - d^{\pi_\theta}\|_1. \tag{83}$$

for any policy π_θ and state s . With $N = 7t_{\text{mix}}(\log_2 T)$, $\delta^{\pi_\theta}(s, T) \leq \frac{1}{T^6}$. Combining this result with the fact that the reward function is bounded in $[0, 1]$, we obtain,

$$\begin{aligned}
 &\sum_{k=1}^K \mathbb{E} |V^{\pi_{\theta_{k+1}}}(s) - V^{\pi_{\theta_k}}(s)| \\
 &\leq \sum_{k=1}^K \mathbb{E} \left| \sum_{t=0}^{N-1} \langle (P^{\pi_{\theta_{k+1}}})^t(s, \cdot) - (P^{\pi_{\theta_k}})^t(s, \cdot), r^{\pi_{\theta_{k+1}}} - r^{\pi_{\theta_k}} \rangle \right| + \sum_{k=1}^K \mathbb{E} \left| \sum_{t=0}^{N-1} \langle (P^{\pi_{\theta_k}})^t(s, \cdot), r^{\pi_{\theta_{k+1}}} - r^{\pi_{\theta_k}} \rangle \right| \\
 &\quad + N \sum_{k=1}^K \mathbb{E} |J(\theta_{k+1}) - J(\theta_k)| + \frac{2K}{T^5} \\
 &\stackrel{(a)}{\leq} \sum_{k=1}^K \sum_{t=0}^{N-1} \mathbb{E} \|(P^{\pi_{\theta_{k+1}}})^t - (P^{\pi_{\theta_k}})^t\|_\infty \|r^{\pi_{\theta_{k+1}}} - r^{\pi_{\theta_k}}\|_\infty + \sum_{k=1}^K \sum_{t=0}^{N-1} \mathbb{E} \|r^{\pi_{\theta_{k+1}}} - r^{\pi_{\theta_k}}\|_\infty \\
 &\quad + 6Gt_{\text{mix}} \sum_{k=1}^K \gamma_k + \frac{2K}{T^5}
 \end{aligned} \tag{84}$$

where (a) follows from (81) and substituting $N = 7t_{\text{mix}}(\log_2 T)$. For the first term, note that,

$$\begin{aligned}
 & \left\| ((P^{\pi_{\theta_{k+1}}})^t - (P^{\pi_{\theta_k}})^t) r^{\pi_{\theta_{k+1}}} \right\|_{\infty} \\
 & \leq \left\| P^{\pi_{\theta_{k+1}}} ((P^{\pi_{\theta_{k+1}}})^{t-1} - (P^{\pi_{\theta_k}})^{t-1}) r^{\pi_{\theta_{k+1}}} \right\|_{\infty} + \left\| (P^{\pi_{\theta_{k+1}}} - P^{\pi_{\theta_k}}) (P^{\pi_{\theta_k}})^{t-1} r^{\pi_{\theta_{k+1}}} \right\|_{\infty} \\
 & \stackrel{(a)}{\leq} \left\| ((P^{\pi_{\theta_{k+1}}})^{t-1} - (P^{\pi_{\theta_k}})^{t-1}) r^{\pi_{\theta_{k+1}}} \right\|_{\infty} + \max_s \|P^{\pi_{\theta_{k+1}}}(s, \cdot) - P^{\pi_{\theta_k}}(s, \cdot)\|_1
 \end{aligned} \tag{85}$$

Inequality (a) holds since every row of $P^{\pi_{\theta_k}}$ sums to 1 and $\|(P^{\pi_{\theta_k}})^{t-1} r^{\pi_{\theta_{k+1}}}\|_{\infty} \leq 1$. Moreover, invoking (79), we get,

$$\begin{aligned}
 \max_s \|P^{\pi_{\theta_{k+1}}}(s, \cdot) - P^{\pi_{\theta_k}}(s, \cdot)\|_1 &= \max_s \left| \sum_{s'} \sum_a (\pi_{\theta_{k+1}}(a|s) - \pi_{\theta_k}(a|s)) P(s'|s, a) \right| \\
 &\leq G \|\theta_{k+1} - \theta_k\| \max_s \left| \sum_{s'} \sum_a \pi_{\theta_k}(a|s) P(s'|s, a) \right| \\
 &\leq G \|\theta_{k+1} - \theta_k\|.
 \end{aligned}$$

Plugging the above result into (85) and using a recursive argument, we get,

$$\begin{aligned}
 \left\| ((P^{\pi_{\theta_{k+1}}})^t - (P^{\pi_{\theta_k}})^t) r^{\pi_{\theta_{k+1}}} \right\|_{\infty} &\leq \sum_{t'=1}^t \max_s \|P^{\pi_{\theta_{k+1}}}(s, \cdot) - P^{\pi_{\theta_k}}(s, \cdot)\|_1 \\
 &\leq \sum_{t'=1}^t G \|\theta_{k+1} - \theta_k\| \leq tG \|\theta_{k+1} - \theta_k\|
 \end{aligned}$$

Finally, we have

$$\sum_{k=1}^K \sum_{t=0}^{N-1} \mathbb{E} \left\| ((P^{\pi_{\theta_{k+1}}})^t - (P^{\pi_{\theta_k}})^t) r^{\pi_{\theta_{k+1}}} \right\|_{\infty} \leq \sum_{k=1}^K \sum_{t=0}^{N-1} tG \gamma_k \leq \mathcal{O} \left(GN^2 \sum_{k=1}^K \|\theta_{k+1} - \theta_k\| \right). \tag{86}$$

Moreover, notice that,

$$\begin{aligned}
 \sum_{k=1}^K \sum_{t=0}^{N-1} \mathbb{E} \|r^{\pi_{\theta_{k+1}}} - r^{\pi_{\theta_k}}\|_{\infty} &\leq \sum_{k=1}^K \sum_{t=0}^{N-1} \mathbb{E} \left[\max_s \left| \sum_a (\pi_{\theta_{k+1}}(a|s) - \pi_{\theta_k}(a|s)) r(s, a) \right| \right] \\
 &\leq GN \sum_{k=1}^K \|\theta_{k+1} - \theta_k\|.
 \end{aligned} \tag{87}$$

Note that for Algorithm 1, $\|\tilde{\theta}_{k+1} - \tilde{\theta}_k\| \leq \mathcal{O}(\gamma_k/\eta_k) = \mathcal{O}\left(\frac{G}{\mu k^{1/5}}\right)$. Substituting in Lemma 9 with the regret decomposition in (76) yields

$$\mathbb{E}[\text{Reg}_T] = H \sum_{k=1}^K \left(J^* - \mathbb{E}[J(\tilde{\theta}_k)] \right) + \mathcal{O} \left(G^2 \mu^{-1} t_{\text{mix}}^2 (\log^2 T) K^{4/5} \right).$$

Whereas, for Algorithm 3, $\|\theta_{k+1} - \theta_k\| = \gamma_k = \frac{6G}{\mu(k+2)}$. It similarly follows from the decomposition in (77) that

$$\mathbb{E}[\text{Reg}_T] = H \sum_{k=1}^K \left(J^* - \mathbb{E}[J(\theta_k)] \right) + \mathcal{O} \left(G^2 \mu^{-1} t_{\text{mix}}^2 (\log^3 T) \right). \tag{88}$$

□

F Proof of Lemma 5

By the L -smoothness of \bar{J} (where L is defined in Lemma 2) and using the update rule for θ_k , we get

$$\begin{aligned} -\bar{J}(\theta_{k+1}) &\leq -\bar{J}(\theta_k) - \langle \nabla_{\theta} \bar{J}(\theta_k), \theta_{k+1} - \theta_k \rangle + \frac{L}{2} \|\theta_{k+1} - \theta_k\|^2 \\ &= -\bar{J}(\theta_k) - \gamma_k \frac{\langle \nabla_{\theta} \bar{J}(\theta_k), d_k \rangle}{\|d_k\|} + \frac{L\gamma_k^2}{2} \end{aligned}$$

Now let us bound the second term in the above inequality. Define $\hat{e}_k = d_k - \nabla_{\theta} \bar{J}(\theta_k)$. We consider two cases. First, if $\|\hat{e}_k\| \leq \frac{1}{2} \|\nabla_{\theta} \bar{J}(\theta_k)\|$, then

$$\begin{aligned} -\frac{\langle \nabla_{\theta} \bar{J}(\theta_k), d_k \rangle}{\|d_k\|} &= \frac{-\|\nabla_{\theta} \bar{J}(\theta_k)\|^2 - \langle \nabla_{\theta} \bar{J}(\theta_k), \hat{e}_k \rangle}{\|d_k\|} \\ &\leq \frac{-\|\nabla_{\theta} \bar{J}(\theta_k)\|^2 + \|\nabla_{\theta} \bar{J}(\theta_k)\| \|\hat{e}_k\|}{\|d_k\|} \\ &\leq \frac{-\|\nabla_{\theta} \bar{J}(\theta_k)\|^2 + \frac{1}{2} \|\nabla_{\theta} \bar{J}(\theta_k)\|^2}{\|d_k\|} \\ &\leq -\frac{\|\nabla_{\theta} \bar{J}(\theta_k)\|^2}{2(\|\nabla_{\theta} \bar{J}(\theta_k)\| + \|\hat{e}_k\|)} \\ &\leq -\frac{1}{3} \|\nabla_{\theta} \bar{J}(\theta_k)\| \end{aligned}$$

Otherwise, if $\|\hat{e}_k\| \geq \frac{1}{2} \|\nabla_{\theta} \bar{J}(\theta_k)\|$, we have

$$\begin{aligned} -\frac{\langle \nabla_{\theta} \bar{J}(\theta_k), d_k \rangle}{\|d_k\|} &\leq \|\nabla_{\theta} \bar{J}(\theta_k)\| \\ &= -\frac{1}{3} \|\nabla_{\theta} \bar{J}(\theta_k)\| + \frac{4}{3} \|\nabla_{\theta} \bar{J}(\theta_k)\| \\ &\leq -\frac{1}{3} \|\nabla_{\theta} \bar{J}(\theta_k)\| + \frac{8}{3} \|\hat{e}_k\| \end{aligned}$$

Combining the two cases gives

$$-\bar{J}(\theta_{k+1}) \leq -\bar{J}(\theta_k) - \frac{\gamma_k}{3} \|\nabla_{\theta} \bar{J}(\theta_k)\| + \frac{8\gamma_k}{3} \|\hat{e}_k\| + \frac{L\gamma_k^2}{2} \quad (89)$$

Let $e_k := d_k - \nabla_{\theta} J(\theta_k)$. Using Lemma 2, 3 and the fact that $\|\theta_k - \theta_0\| \leq T, \forall k \leq K$, we obtain

$$-J(\theta_{k+1}) \leq -J(\theta_k) - \frac{\gamma_k}{3} \|\nabla_{\theta} J(\theta_k)\| + \frac{8\gamma_k}{3} \|e_k\| + \frac{L\gamma_k^2}{2} + \mathcal{O}\left(\frac{AGt_{\text{mix}} \log T}{T^3}\right) \quad (90)$$

Now adding J^* and taking expectation on both sides and using Gradient Domination Lemma gives

$$\begin{aligned} &J^* - \mathbb{E}[J(\theta_{k+1})] \\ &\leq \mathbb{E}\left[J^* - J(\theta_k) - \frac{\mu\gamma_k}{3G}(J^* - J(\theta_k) - \sqrt{\epsilon_{\text{bias}}}) + \frac{8\gamma_k}{3} \|e_k\| + \frac{L\gamma_k^2}{2} + \mathcal{O}\left(\frac{AGt_{\text{mix}} \log T}{T^3}\right)\right] \\ &= \left(1 - \frac{\mu\gamma_k}{3G}\right)(J^* - \mathbb{E}[J(\theta_k)]) + \frac{\mu\gamma_k}{3G} \sqrt{\epsilon_{\text{bias}}} + \frac{8\gamma_k}{3} \mathbb{E}\|e_k\| + \frac{L\gamma_k^2}{2} + \mathcal{O}\left(\frac{AGt_{\text{mix}} \log T}{T^3}\right) \end{aligned}$$

We unroll the above recursion with the help of the following lemma.

Lemma 11. [Fatkhullin et al., 2023, Lemma 12] Let a be a positive real, ξ a positive integer and let $\{r_t\}_{t \geq 0}$ be a non-negative sequence satisfying for every integer $t \geq 0$

$$r_{t+1} - r_t \leq -a\alpha_t r_t + \nu_t,$$

where $\{\alpha_t\}_{t \geq 0}, \{\beta_t\}_{t \geq 0}$ are non-negative sequences and $a\alpha_t \leq 1$ for all t . Then for $\alpha_t = \frac{2}{a(t+\xi)}$ we have for every integers $t_0, T \geq 1$

$$r_T \leq \frac{(t_0 + \xi - 1)^2 r_{t_0}}{(T + \xi - 1)^2} + \frac{\sum_{t=0}^{T-1} \nu_t (t + \xi)^2}{(T + \xi - 1)^2}.$$

Since γ_k is chosen such that $\frac{\mu\gamma_k}{3G} = \frac{2}{(k+2)}$, we can invoke Lemma 11 to obtain

$$J^* - \mathbb{E}[J(\theta_K)] \leq \frac{J^* - J(\theta_0)}{(K+1)^2} + \frac{\sum_{k=0}^{K-1} \nu_k (k+2)^2}{(K+1)^2},$$

where $\nu_k := \frac{\mu\gamma_k}{3G} \sqrt{\epsilon_{\text{bias}}} + \frac{8\gamma_k}{3} \mathbb{E} \|e_k\| + \frac{L\gamma_k^2}{2} + \mathcal{O}\left(\frac{AGt_{\text{mix}} \log T}{T^3}\right)$.

G Proof of Theorems 4 and 1

We first provide our results assuming that the initial distribution for trajectories sampled at each iteration is ρ . Under this setting, the proof methods used are inspired by the Lemma 7 of Fatkhullin et al. [2023]. However, here the gradient is biased unlike the discounted case and the resulting terms arising from the bias must be bounded.

For ease of exposition, we let $\hat{e}_k = d_k - \nabla_{\theta} J(\theta_k)$, $e_k := g(\theta_k, \tau_k) - \nabla_{\theta} J(\tilde{\theta}_k)$,

$$S_k := \nabla_{\theta} J(\theta_{k-1}) - \nabla_{\theta} J(\theta_k) + \nabla^2 J(\theta_k)(\theta_{k-1} - \theta_k)$$

and

$$Z_k := \nabla_{\theta} J(\tilde{\theta}_k) - \nabla_{\theta} J(\theta_k) + \nabla^2 J(\theta_k)(\tilde{\theta}_k - \theta_k).$$

We have

$$\|S_k\| \leq L_h \|\theta_k - \theta_{k-1}\|^2 = L_h \gamma_{k-1}^2 \quad (91)$$

and

$$\|Z_k\| \leq L_h \|\tilde{\theta}_k - \theta_k\|^2 = L_h \frac{(1 - \eta_k)^2}{\eta_k^2} \|\theta_k - \theta_{k-1}\|^2 \leq L_h \frac{(1 - \eta_k)^2}{\eta_k^2} \gamma_{k-1}^2. \quad (92)$$

Using the update rule of the sequence $\{d_k\}_{k \geq 1}$ in Algorithm 1, we obtain the following recursion

$$\begin{aligned} \hat{e}_k &= d_k - \nabla_{\theta} J(\theta_k) \\ &= (1 - \eta_k) d_{k-1} + \eta_k g(\theta_k, \tau_k) - \nabla_{\theta} J(\theta_k) \\ &= (1 - \eta_k) (d_{k-1} - \nabla_{\theta} J(\theta_{k-1})) + (1 - \eta_k) \nabla_{\theta} J(\theta_{k-1}) + \eta_k g(\theta_k, \tau_k) - \nabla_{\theta} J(\theta_k) \\ &= (1 - \eta_k) \hat{e}_{k-1} + (1 - \eta_k) \nabla_{\theta} J(\theta_{k-1}) + \eta_k g(\theta_k, \tau_k) - \nabla_{\theta} J(\theta_k) \\ &= (1 - \eta_k) \hat{e}_{k-1} + \eta_k (g(\theta_k, \tau_k) - \nabla_{\theta} J(\tilde{\theta}_k)) + \eta_k \nabla_{\theta} J(\tilde{\theta}_k) + (1 - \eta_k) \nabla_{\theta} J(\theta_{k-1}) - \nabla_{\theta} J(\theta_k) \\ &= (1 - \eta_k) \hat{e}_{k-1} + \eta_k e_k + \eta_k \nabla_{\theta} J(\tilde{\theta}_k) + (1 - \eta_k) \nabla_{\theta} J(\theta_{k-1}) - \nabla_{\theta} J(\theta_k) \\ &= (1 - \eta_k) \hat{e}_{k-1} + \eta_k e_k + \eta_k (\nabla_{\theta} J(\tilde{\theta}_k) - \nabla_{\theta} J(\theta_k)) + (1 - \eta_k) (\nabla_{\theta} J(\theta_{k-1}) - \nabla_{\theta} J(\theta_k)) \\ &= (1 - \eta_k) \hat{e}_{k-1} + \eta_k e_k + \eta_k (S_k - \nabla^2 J(\theta_k)(\theta_{k-1} - \theta_k)) + (1 - \eta_k) (Z_k - \nabla^2 J(\theta_k)(\tilde{\theta}_k - \theta_k)) \\ &= (1 - \eta_k) \hat{e}_{k-1} + \eta_t e_k + (1 - \eta_k) S_k + \eta_t Z_k - (\eta_k \nabla^2 J(\theta_k)(\theta_{k-1} - \theta_k) \\ &\quad + (1 - \eta_k) \nabla^2 J(\theta_k)(\tilde{\theta}_k - \theta_k)) \\ &= (1 - \eta_k) \hat{e}_{k-1} + \eta_t e_k + (1 - \eta_k) S_k + \eta_t Z_k, \end{aligned} \quad (93)$$

where the last term of (93) is 0 since $\tilde{\theta}_k = \theta_k + \frac{1 - \eta_k}{\eta_k} (\theta_k - \theta_{k-1})$. This simplification provides the main motivation behind the update rule of $\tilde{\theta}_k$.

Let $\zeta_{k,K} := \prod_{j=k}^{K-1} (1 - \eta_{j+1})$ (with $\zeta_{K,K} = 1$). Unrolling the above recursion yields

$$\hat{e}_K = \zeta_{0,K} \hat{e}_0 + \sum_{k=0}^{K-1} \eta_{k+1} \zeta_{k+1,K} e_{k+1} + \sum_{k=0}^{K-1} (1 - \eta_{k+1}) \zeta_{k+1,K} S_{k+1} + \sum_{k=0}^{K-1} \eta_{k+1} \zeta_{k+1,K} Z_{k+1}.$$

Taking the norm and expectation on both sides give

$$\begin{aligned} & \mathbb{E} \|\hat{e}_K\| \\ &= \mathbb{E} \left\| \zeta_{0,K} \hat{e}_0 + \sum_{k=0}^{K-1} \eta_{k+1} \zeta_{k+1,K} e_{k+1} + \sum_{k=0}^{K-1} (1 - \eta_{k+1}) \zeta_{k+1,K} S_{k+1} + \sum_{k=0}^{K-1} \eta_{k+1} \zeta_{k+1,K} Z_{k+1} \right\| \\ &\leq \mathbb{E} \|\zeta_{0,K} \hat{e}_0\| + \mathbb{E} \left\| \sum_{k=0}^{K-1} \eta_{k+1} \zeta_{k+1,K} e_{k+1} \right\| + \sum_{k=0}^{K-1} (1 - \eta_{k+1}) \zeta_{k+1,K} \mathbb{E} \|S_{k+1}\| \\ &\quad + \sum_{k=0}^{K-1} \eta_{k+1} \zeta_{k+1,K} \mathbb{E} \|Z_{k+1}\| \\ &\leq \mathbb{E} \|\zeta_{0,K} \hat{e}_0\| + \left(\mathbb{E} \left\| \sum_{k=0}^{K-1} \eta_{k+1} \zeta_{k+1,K} e_{k+1} \right\|^2 \right)^{1/2} + \sum_{k=0}^{K-1} (1 - \eta_{k+1}) \zeta_{k+1,K} L_h \gamma_{k-1}^2 \\ &\quad + \sum_{k=0}^{K-1} \zeta_{k+1,K} L_h \frac{(1 - \eta_k)^2}{\eta_k} \gamma_{k-1}^2 \\ &\leq \mathbb{E} \|\zeta_{0,K} \hat{e}_0\| + \left(\mathbb{E} \left\| \sum_{k=0}^{K-1} \eta_{k+1} \zeta_{k+1,K} e_{k+1} \right\|^2 \right)^{1/2} + 2L_h \sum_{k=0}^{K-1} \zeta_{k+1,K} \frac{\gamma_{k-1}^2}{\eta_k}. \end{aligned}$$

We now focus on the second term in the RHS of the above line.

$$\begin{aligned} & \mathbb{E} \left\| \sum_{k=0}^{K-1} \eta_{k+1} \zeta_{k+1,K} e_{k+1} \right\|^2 \\ &= \sum_{k=0}^{K-1} \mathbb{E} \|\eta_{k+1} \zeta_{k+1,K} e_{k+1}\|^2 + 2 \sum_{i=0}^{K-1} \sum_{j=0}^i \mathbb{E} \langle \eta_{i+1} \zeta_{i+1,K} e_{i+1}, \eta_{j+1} \zeta_{j+1,K} e_{j+1} \rangle \\ &= \sum_{k=0}^{K-1} (\eta_{k+1} \zeta_{k+1,K})^2 \mathbb{E} \|e_{k+1}\|^2 + 2 \sum_{i=0}^{K-1} \sum_{j=0}^i (\eta_{i+1} \zeta_{i+1,K} \eta_{j+1} \zeta_{j+1,K}) \mathbb{E} \langle e_{i+1}, e_{j+1} \rangle. \end{aligned}$$

Define for every integer $t \geq 1$ the σ -field $\mathcal{F}_t := \sigma(\{\tilde{\theta}_0, \tilde{\xi}_0, \dots, \tilde{\xi}_{t-1}\})$ where $\tilde{\xi}_s \sim p(\cdot | \pi_{\tilde{\theta}_s})$ for every $0 \leq s \leq t-1$. Notice that for any integers $t_2 > t_1 \geq 1$ we have

$$\begin{aligned} |\mathbb{E} \langle e_{t_1}, e_{t_2} \rangle| &= |\mathbb{E}[\mathbb{E}[\langle e_{t_1}, e_{t_2} \rangle | \mathcal{F}_{t_2}]]| = |\mathbb{E}[\langle e_{t_1}, \mathbb{E}[e_{t_2} | \mathcal{F}_{t_2}] \rangle]| \\ &\leq \mathbb{E}[\|e_{t_1}\| \|\mathbb{E}[e_{t_2} | \mathcal{F}_{t_2}]\|] \leq \sigma_g \beta_g. \end{aligned} \tag{94}$$

We state some useful lemmas for bounding terms arising from the step-sizes:

Lemma 12. [Fatkhullin et al., 2023, Lemma 14] Let $q \in [0, 1]$ and let $\eta_t = \left(\frac{2}{t+2}\right)^q$ for every integer t . Then for every integer t and any integer $T \geq 1$ we have

$$\eta_t(1 - \eta_{t+1}) \leq \eta_{t+1} \quad \text{and} \quad \prod_{t=0}^{T-1} (1 - \eta_{t+1}) \leq \eta_T. \tag{95}$$

Lemma 13. [Fatkhullin et al., 2023, Lemma 15] Let $q \in [0, 1)$, $p \geq 0$, $\gamma_0 > 0$ and let $\eta_t = \left(\frac{2}{t+2}\right)^q$, $\gamma_t = \gamma_0 \left(\frac{2}{t+2}\right)^p$ for every integer t . Then for any integers t and $T \geq 1$, it holds

$$\sum_{t=0}^{T-1} \gamma_t \prod_{\xi=t+1}^{T-1} (1 - \eta_\xi) \leq C \gamma_T \eta_T^{-1},$$

where $C = C(p, q) := 2^{p-q}(1-q)^{-1} t_0 \exp\left(2^q(1-q)t_0^{1-q}\right) + 2^{2p+1-q}(1-q)^{-2}$ and $t_0 := \max\left\{\left(\frac{p}{(1-q)2^q}\right)^{\frac{1}{1-q}}, 2\left(\frac{p-q}{(1-q)^2}\right)^{\frac{1}{1-q}}\right\}$.

We then obtain the following

$$\begin{aligned} & \mathbb{E} \left\| \sum_{k=0}^{K-1} \eta_{k+1} \zeta_{k+1, K} e_{k+1} \right\|^2 \\ & \leq \sum_{k=0}^{K-1} (\eta_{k+1} \zeta_{k+1, K})^2 \sigma_g^2 + 2 \sum_{i=0}^{K-1} \sum_{j=0}^i (\eta_{i+1} \zeta_{i+1, K} \eta_{j+1} \zeta_{j+1, K}) \cdot \sigma_g \beta_g \\ & \leq \sigma_g^2 \cdot \sum_{k=0}^{K-1} \eta_{k+1}^2 \zeta_{k+1, K} + 2 \sigma_g \beta_g \cdot \sum_{i=0}^{K-1} \sum_{j=0}^i \eta_K^2 \\ & \stackrel{(a)}{=} \mathcal{O}(\sigma_g^2 \cdot \eta_K + \sigma_g \beta_g \cdot K^2 \eta_K^2) = \mathcal{O}\left(\frac{AG^2 t_{\text{mix}}^2 (\log T)^2}{K^{4/5}} + \frac{A^2 G^2 t_{\text{mix}}^2 (\log T)^2 \cdot K^{2/5}}{T^4}\right). \end{aligned}$$

where (a) follows from Lemmas 12 and 13. We have

$$\begin{aligned} \mathbb{E} \|d_K - \nabla_\theta J(\theta_K)\| &= \mathbb{E} \|\hat{e}_K\| \\ &\leq \mathbb{E} \|\zeta_{0, K} \hat{e}_0\| + \left(\mathbb{E} \left\| \sum_{k=0}^{K-1} \eta_{k+1} \zeta_{k+1, K} e_{k+1} \right\|^2 \right)^{1/2} + 2L_h \sum_{k=0}^{K-1} \zeta_{k+1, K} \frac{\gamma_{k-1}^2}{\eta_k} \\ &\leq \eta_K \mathbb{E} \|\hat{e}_0\| + \left(\mathbb{E} \left\| \sum_{k=0}^{K-1} \eta_{k+1} \zeta_{k+1, K} e_{k+1} \right\|^2 \right)^{1/2} + 2L_h \cdot C \gamma_K^2 \eta_K^{-2} \\ &= \mathcal{O}\left(\frac{AG t_{\text{mix}} \log T}{K^{2/5}} + \frac{G^2 L_h}{\mu^2 K^{2/5}}\right). \end{aligned}$$

Note that

$$\begin{aligned} \frac{\sum_{k=0}^{K-1} \frac{\mu \gamma_k}{3G} \cdot \sqrt{\epsilon_{\text{bias}}} \cdot (k+2)^2}{(K+1)^2} &= \frac{\sum_{k=0}^{K-1} 2\sqrt{\epsilon_{\text{bias}}} \cdot (k+2)}{(K+1)^2} \\ &= \frac{(K^2 + 3K) \cdot \sqrt{\epsilon_{\text{bias}}}}{(K+1)^2} \\ &\leq \sqrt{\epsilon_{\text{bias}}}. \end{aligned}$$

Also,

$$\begin{aligned} \frac{\sum_{k=0}^{K-1} \frac{L \gamma_k^2}{2} \cdot (k+2)^2}{(K+1)^2} &= \frac{\sum_{k=0}^{K-1} \frac{L(6G)^2}{2\mu^2(k+2)^2} \cdot (k+2)^2}{(K+1)^2} \\ &= \frac{\sum_{k=0}^{K-1} \frac{18G^2 L}{\mu^2}}{(K+1)^2} \\ &\leq \frac{18G^2 L}{\mu^2(K+1)}. \end{aligned}$$

Since $\mathbb{E} \|\hat{e}_k\| = \mathcal{O}\left(\frac{G t_{\text{mix}} \log T}{k^{2/5}} + \frac{G^2 L_h}{\mu^2 k^{2/5}}\right)$, we have

$$\begin{aligned} \frac{\sum_{k=0}^{K-1} \frac{8\gamma_k}{3} \mathbb{E} \|\hat{e}_k\| \cdot (k+2)^2}{(K+1)^2} &= \frac{\sum_{k=0}^{K-1} \frac{16G}{\mu} \mathbb{E} \|\hat{e}_k\| \cdot (k+2)}{(K+1)^2} \\ &\leq \frac{\sum_{k=0}^{K-1} \mathcal{O}\left(\frac{AG^2 t_{\text{mix}} \log T \cdot k^{2/5}}{\mu} + \frac{G^3 L_h \cdot k^{2/5}}{\mu^3}\right)}{(K+1)^2} \\ &\leq \mathcal{O}\left(\frac{AG^2 t_{\text{mix}} \log T \cdot K^{-2/5}}{\mu} + \frac{G^3 L_h \cdot K^{-2/5}}{\mu^3}\right). \end{aligned}$$

It follows from Lemma 5 that for all $K \geq 1$

$$J^* - \mathbb{E}[J(\theta_K)] \leq \sqrt{\epsilon_{\text{bias}}} + \mathcal{O}\left(\frac{AG^2 t_{\text{mix}} \log T \cdot K^{-2/5}}{\mu} + \frac{G^3 L_h \cdot K^{-2/5}}{\mu^3}\right).$$

Note that $\tilde{\theta}_k - \theta_k = \frac{(1-\eta_k)\gamma_{k-1}}{\eta_k} \frac{d_{k-1}}{\|d_{k-1}\|}$. Using Theorem 3, we obtain

$$\begin{aligned} J^* - J(\tilde{\theta}_k) &\leq (J^* - J(\theta_k)) + \mathcal{O}\left(-\frac{\gamma_{k-1}}{\eta_k} \|\nabla_{\theta} J(\theta_k)\| + \frac{\gamma_{k-1}}{\eta_k} \|d_{k-1} - \nabla_{\theta} J(\theta_k)\| + \frac{L\gamma_{k-1}^2}{\eta_k^2} + \frac{AG t_{\text{mix}} \log T}{T^3}\right) \\ &\leq (J^* - J(\theta_k)) + \mathcal{O}\left(\frac{\gamma_{k-1}}{\eta_k} (\|e_{k-1}\| + \|\nabla_{\theta} J(\theta_{k-1}) - \nabla_{\theta} J(\theta_k)\|) + \frac{L\gamma_{k-1}^2}{\eta_k^2} + \frac{AG t_{\text{mix}} \log T}{T^3}\right) \end{aligned} \quad (96)$$

Recall that $\mathbb{E} \|\hat{e}_k\| = \mathcal{O}\left(\frac{G t_{\text{mix}} \log T}{k^{2/5}} + \frac{G^2 L_h}{\mu^2 k^{2/5}}\right)$. Separately, invoking Lemmas 2(b) and (c), we obtain $\|\nabla_{\theta} J(\theta_{k-1}) - \nabla_{\theta} J(\theta_k)\| \leq \mathcal{O}\left(L\|\theta_{k-1} - \theta_k\| + \frac{\sqrt{A}G t_{\text{mix}} \log T}{T^4}\right) \leq \mathcal{O}\left(L\gamma_{k-1} + \frac{\sqrt{A}G t_{\text{mix}} \log T}{T^4}\right)$. Taking expectation in (96) and substituting these bounds yields

$$\begin{aligned} J^* - \mathbb{E}[J(\tilde{\theta}_k)] &\leq (J^* - \mathbb{E}[J(\theta_k)]) \\ &+ \mathcal{O}\left(\frac{\gamma_{k-1}G t_{\text{mix}} \log T}{\eta_k k^{2/5}} + \frac{\gamma_{k-1}G^2 L_h}{\eta_k \mu^2 k^{2/5}} + \frac{L\gamma_{k-1}^2}{\eta_k} + \frac{\gamma_{k-1}\sqrt{A}G t_{\text{mix}} \log T}{\eta_k T^4} + \frac{L\gamma_k^2}{\eta_k^2} + \frac{AG t_{\text{mix}} \log T}{T^3}\right). \end{aligned} \quad (97)$$

It follows that

$$\begin{aligned} J^* - \mathbb{E}[J(\tilde{\theta}_K)] &\leq \sqrt{\epsilon_{\text{bias}}} + \mathcal{O}\left(\frac{AG^2 t_{\text{mix}} \log T \cdot K^{-2/5}}{\mu} + \frac{G^3 L_h \cdot K^{-2/5}}{\mu^3} + \frac{(\sqrt{A}BG t_{\text{mix}} \log T + \sqrt{A}G^2 t_{\text{mix}}^2 t_{\text{hit}} (\log T)^2) K^{-2/5}}{\mu}\right). \end{aligned}$$

It follows that

$$\begin{aligned} H \sum_{K=1}^{T/H} (J^* - \mathbb{E}[J(\tilde{\theta}_K)]) &\leq T\sqrt{\epsilon_{\text{bias}}} + \mathcal{O}\left(\frac{HG^2 A t_{\text{mix}} \log T \cdot (T/H)^{3/5}}{\mu} + \frac{HG^3 L_h \cdot (T/H)^{3/5}}{\mu^3} \right. \\ &\quad \left. + \frac{H(\sqrt{A}BG t_{\text{mix}} \log T + \sqrt{A}G^2 t_{\text{mix}}^2 t_{\text{hit}} (\log T)^2)(T/H)^{3/5}}{\mu}\right). \end{aligned}$$

Ignoring constants and logarithmic terms, the above term is $\tilde{\mathcal{O}}(H^{2/5}T^{3/5})$. Using the regret bound in Lemma 4, it follows that

$$\mathbb{E}[\text{Reg}_T] \leq \mathcal{O}(H^{2/5}T^{3/5} + T^{4/5}H^{-4/5}). \quad (98)$$

It can be seen that $H = \tilde{O}(T^{1/6})$ balances the two terms to provide a regret of order $\tilde{O}(T^{2/3})$. Thus, the regret bound follows for the constructed MDP. To convert the above results to our actual setting, where the starting state of each trajectory is determined by the previous trajectory, we use a similar argument used in Lemma 2 where we obtain results under an imaginary MDP first and then translate it into the real MDP. Here, we will consider an imaginary setup, where the state distribution becomes $(P^{\pi_{\theta_i}})^N \rho$ at every iteration i after completion of the buffer trajectory of length N . We also let $f(X) := \sum_{k=1}^K (J^* - J(\theta_k))$, where $X := (\theta_0, \theta_1, \tau_1, \tau_2, \dots, \tau_K)$.

Let \mathbb{E}' and \Pr' denote expectation and probability under this setup. Since $f(X)$ is non-negative, notice that:

$$\frac{\mathbb{E}[f(X)]}{\mathbb{E}'[f(X)]} = \frac{\sum_X f(X) \Pr(X)}{\sum_X f(X) \Pr'(X)} \leq \max_X \frac{\Pr(X)}{\Pr'(X)}. \quad (99)$$

Observe that given θ_0 and θ_1

$$\frac{\Pr(X)}{\Pr'(X)} = \frac{\Pr(\tau_1|\theta_0, \theta_1) \cdot \Pr(\tau_2|\theta_0, \theta_1, \tau_1) \cdots \Pr(\tau_K|\theta_0, \theta_1, \tau_1, \dots, \tau_{K-1})}{\Pr'(\tau_1|\theta_0, \theta_1) \cdot \Pr'(\tau_2|\theta_0, \theta_1, \tau_1) \cdots \Pr'(\tau_K|\theta_0, \theta_1, \tau_1, \dots, \tau_{K-1})} \quad (100)$$

$$\stackrel{(a)}{=} \frac{\Pr(\tau_1|\theta_0, \theta_1) \cdot \Pr(\tau_2|\theta_1, \tau_1) \cdot \Pr(\tau_3|\theta_2, \tau_2) \cdots \Pr(\tau_K|\theta_{K-1}, \tau_{K-1})}{\Pr'(\tau_1|\theta_0, \theta_1) \cdot \Pr'(\tau_2|\theta_1, \tau_1) \cdot \Pr'(\tau_3|\theta_2, \tau_2) \cdots \Pr'(\tau_K|\theta_{K-1}, \tau_{K-1})}, \quad (101)$$

where (a) follows the observation that τ_i is only a function of θ_i and τ_{i-1} , while θ_i is completely determined by θ_{i-1} and τ_{i-1} . We have

$$\frac{\Pr(\tau_i|\theta_{i-1}, \tau_{i-1})}{\Pr'(\tau_i|\theta_{i-1}, \tau_{i-1})} = \frac{(P^{\pi_{\theta_i}})^N(s, s')}{(P^{\pi_{\theta_i}})^N \rho(s')} \leq \max_{s'} 1 + \frac{(P^{\pi_{\theta_i}})^N(s, s') - (P^{\pi_{\theta_i}})^N \rho(s')}{(P^{\pi_{\theta_i}})^N \rho(s')} \quad (102)$$

$$\stackrel{(a)}{\leq} \max_{s'} 1 + \frac{2}{T^6 d^{\pi_{\theta_i}}(s')} \leq 1 + \frac{2t_{\text{hit}}}{T^6} \leq 1 + \frac{2}{T^5}, \quad (103)$$

where (a) follows from Lemma 16 and the fact that $T \geq 2t_{\text{hit}}$. We then obtain

$$\frac{\Pr(X)}{\Pr'(X)} \leq \left(1 + \frac{2}{T^5}\right)^K \leq e^{\frac{2K}{T^5}} \leq e^{\frac{2}{T^4}} \leq \left(1 + \frac{4}{T^4}\right), \quad (104)$$

and the result follows. The same argument also holds for Algorithm 3.

H Proof of Theorems 5 and 2

Let $\mathcal{V}_k := g(\tau_k, \theta_k) - \nabla_{\theta} \bar{J}(\theta_k)$ and $\mathcal{W}_k := \nabla_{\theta} \bar{J}(\theta_{k-1}) - \nabla_{\theta} \bar{J}(\theta_k) + B(\hat{\tau}_k, \hat{\theta}_k)(\theta_k - \theta_{k-1})$. These quantities are defined in such a way that they are both zero-mean with bounded variance, which was also the motivating factor for our choice of \bar{J} . It is easy to see that $\mathbb{E}[\mathcal{V}_k] = 0$ (from the definition of \bar{J}) and $\mathbb{E}[\|\mathcal{V}_k\|^2] \leq \sigma_g^2$ (from Lemma 2). To see that $\mathbb{E}[\mathcal{W}_k] = 0$, observe that

$$\begin{aligned} \mathbb{E}[\mathcal{W}_k] &= \mathbb{E}[\nabla_{\theta} \bar{J}(\theta_{k-1}) - \nabla_{\theta} \bar{J}(\theta_k) + B(\hat{\tau}_k, \hat{\theta}_k)(\theta_k - \theta_{k-1})] \\ &= \mathbb{E}[\mathbb{E}[\nabla_{\theta} \bar{J}(\theta_{k-1}) - \nabla_{\theta} \bar{J}(\theta_k) + B(\hat{\tau}_k, \hat{\theta}_k)(\theta_k - \theta_{k-1}) \mid \theta_{k-1}, \theta_k, \hat{\theta}_k]] \\ &= \mathbb{E}[\nabla_{\theta} \bar{J}(\theta_{k-1}) - \nabla_{\theta} \bar{J}(\theta_k) + \nabla^2 \bar{J}(\hat{\theta}_k)(\theta_k - \theta_{k-1})] \\ &= \mathbb{E}[\nabla_{\theta} \bar{J}(\theta_{k-1}) - \nabla_{\theta} \bar{J}(\theta_k)] + \mathbb{E}\left[\int_0^1 \nabla^2 \bar{J}(q\theta_k + (1-q)\theta_{k-1})(\theta_k - \theta_{k-1}) dq\right] = 0. \end{aligned} \quad (105)$$

The variance bound for \mathcal{W}_k can be obtained as:

$$\begin{aligned} \mathbb{E}[\|\mathcal{W}_k\|^2] &= \mathbb{E}\left\|\nabla_{\theta} \bar{J}(\theta_{k-1}) - \nabla_{\theta} \bar{J}(\theta_k) + B(\hat{\tau}_k, \hat{\theta}_k)(\theta_k - \theta_{k-1})\right\|^2 \\ &\leq 2\mathbb{E}\left\|\nabla_{\theta} \bar{J}(\theta_{k-1}) - \nabla_{\theta} \bar{J}(\theta_k)\right\|^2 + 2\mathbb{E}\left\|B(\hat{\tau}_k, \hat{\theta}_k)(\theta_k - \theta_{k-1})\right\|^2 \\ &\stackrel{(a)}{\leq} 4L^2\gamma_{k-1}^2, \end{aligned} \quad (106)$$

where (a) follows using the fact that \bar{J} is L -smooth and the bound on the Hessian estimate variance (both implied by Lemma 2(c)).

We can now obtain a recursion for $d_k - \nabla_\theta \bar{J}(\theta_k)$ using the update rule of the sequence (d_k) in terms of \mathcal{V}_k and \mathcal{W}_k introduced earlier. We have

$$\begin{aligned}
 d_k - \nabla_\theta \bar{J}(\theta_k) &= (1 - \eta_k) \left(d_{k-1} + B(\hat{\tau}_k, \hat{\theta}_k)(\theta_k - \theta_{k-1}) \right) + \eta_k g(\tau_k, \theta_k) - \nabla_\theta \bar{J}(\theta_k) \\
 &= (1 - \eta_k) d_{k-1} + \eta_k g(\tau_k, \theta_k) + (1 - \eta_k) B(\hat{\tau}_k, \hat{\theta}_k)(\theta_k - \theta_{k-1}) - \nabla_\theta \bar{J}(\theta_k) \\
 &= (1 - \eta_k) (d_{k-1} - \nabla_\theta \bar{J}(\theta_{k-1})) + \eta_k g(\tau_k, \theta_k) \\
 &\quad + (1 - \eta_k) B(\hat{\tau}_k, \hat{\theta}_k)(\theta_k - \theta_{k-1}) - \nabla_\theta \bar{J}(\theta_k) + (1 - \eta_k) \nabla_\theta \bar{J}(\theta_{k-1}) \\
 &= (1 - \eta_k) (d_{k-1} - \nabla_\theta \bar{J}(\theta_{k-1})) + \eta_k (g(\tau_k, \theta_k) - \nabla_\theta \bar{J}(\theta_k)) \\
 &\quad + (1 - \eta_k) \left(\nabla_\theta \bar{J}(\theta_{k-1}) - \nabla_\theta \bar{J}(\theta_k) + B(\hat{\tau}_k, \hat{\theta}_k)(\theta_k - \theta_{k-1}) \right) \\
 &= (1 - \eta_k) (d_{k-1} - \nabla_\theta \bar{J}(\theta_{k-1})) + \eta_k \mathcal{V}_k + (1 - \eta_k) \mathcal{W}_k.
 \end{aligned}$$

With the decomposition mentioned above, we can derive a recursive upper bound on the norm of V_k as follows:

$$\begin{aligned}
 V_k &= \mathbb{E} \|d_k - \nabla_\theta \bar{J}(\theta_k)\|^2 \\
 &= \mathbb{E} \|(1 - \eta_k) (d_{k-1} - \nabla_\theta \bar{J}(\theta_{k-1})) + \eta_k \mathcal{V}_k + (1 - \eta_k) \mathcal{W}_k\|^2 \\
 &\stackrel{(a)}{=} (1 - \eta_k)^2 \mathbb{E} \|d_{k-1} - \nabla_\theta \bar{J}(\theta_{k-1})\|^2 + \mathbb{E} \|\eta_k \mathcal{V}_k + (1 - \eta_k) \mathcal{W}_k\|^2 \\
 &\leq (1 - \eta_k) V_{k-1} + 2\eta_k^2 \mathbb{E} \|\mathcal{V}_k\|^2 + 2\mathbb{E} \|\mathcal{W}_k\|^2 \\
 &\leq (1 - \eta_k) V_{k-1} + 2\eta_k^2 \sigma_g^2 + 8\gamma_{k-1}^2 L^2
 \end{aligned}$$

where (a) can be inferred by noticing that the conditional expectation of the random variable $\mathbb{E}[\eta_k \mathcal{V}_k + (1 - \eta_k) \mathcal{W}_k | \theta_{k-1}] = 0$. Unrolling this recursion using Lemma 11 with $t_0 = 0$, $a = 1$ and $\xi = 2$, we obtain

$$V_K = \mathbb{E} \|d_K - \nabla_\theta \bar{J}(\theta_K)\|^2 \leq \frac{V_0}{(K+1)^2} + \frac{\sum_{k=0}^{K-1} \nu_k (k+2)^2}{(K+1)^2}, \quad (107)$$

where $\nu_k = 2\eta_k^2 \sigma_g^2 + 8\gamma_{k-1}^2 L^2$. Since $\gamma_k = \frac{6G}{\mu(k+2)}$ and $\eta_k = \frac{2}{k+2}$, we have $\nu_k (k+2)^2 = \mathcal{O}\left(\sigma_g^2 + \frac{G^2 L^2}{\mu^2}\right)$ and with this we have

$$\frac{\sum_{k=0}^{K-1} \nu_k (k+2)^2}{(K+1)^2} \leq \mathcal{O}\left(\frac{\sigma_g^2}{K} + \frac{G^2 L^2}{K\mu^2}\right). \quad (108)$$

Thus, for all $K \geq 1$

$$\mathbb{E} \|d_K - \nabla_\theta \bar{J}(\theta_K)\|^2 \leq \mathcal{O}\left(\frac{\mathbb{E} \|d_0 - \nabla_\theta \bar{J}(\theta_0)\|^2}{K^2} + \frac{\sigma_g^2}{K} + \frac{G^2 L^2}{K\mu^2}\right). \quad (109)$$

It follows that

$$\begin{aligned}
 \mathbb{E} \|d_K - \nabla_\theta \bar{J}(\theta_K)\| &\leq \left(\mathbb{E} \|d_K - \nabla_\theta \bar{J}(\theta_K)\|^2 \right)^{1/2} \\
 &\leq \mathcal{O}\left(\frac{\mathbb{E} \|d_0 - \nabla_\theta \bar{J}(\theta_0)\|}{K} + \frac{\sigma_g}{\sqrt{K}} + \frac{GL}{\sqrt{K}\mu}\right).
 \end{aligned} \quad (110)$$

Note that

$$\begin{aligned}
 \mathbb{E} \|d_K - \nabla_\theta J(\theta_K)\| &= \mathbb{E} \|d_K - \nabla_\theta \bar{J}(\theta_K) + \nabla_\theta \bar{J}(\theta_K) - \nabla_\theta J(\theta_K)\| \\
 &\leq \mathbb{E} \|d_K - \nabla_\theta \bar{J}(\theta_K)\| + \|\nabla_\theta \bar{J}(\theta_K) - \nabla_\theta J(\theta_K)\| \\
 &= \mathbb{E} \|d_K - \nabla_\theta \bar{J}(\theta_K)\| + \|\mathbb{E}[g(\theta_K, \tau)] - \nabla_\theta J(\theta_K)\| \\
 &\stackrel{(a)}{\leq} \mathcal{O}\left(\frac{\sigma_g}{\sqrt{K}} + \frac{GL}{\sqrt{K}\mu} + \frac{AGt_{\text{mix}} \log T}{T^4}\right),
 \end{aligned} \quad (111)$$

where (a) follows from (110) and Lemma 2.

From Lemma 5, we have for every integer $K \geq 1$:

$$J^* - \mathbb{E}[J(\theta_K)] \leq \frac{J^* - J(\theta_0)}{(K+1)^2} + \frac{\sum_{k=1}^K \nu_k (k+2)^2}{(K+1)^2},$$

where $\nu_k := \frac{\mu\gamma_k}{3G} \cdot \sqrt{\epsilon_{\text{bias}}} + \frac{8\gamma_k}{3} \mathbb{E} \|d_k - \nabla_{\theta} J(\theta_k)\| + \frac{L\gamma_k^2}{2}$.

Observe that since $\gamma_k = \frac{6G}{\mu(k+2)}$ we have

$$\frac{\sum_{k=1}^K \frac{8\gamma_k}{3} \mathbb{E} \|d_k - \nabla_{\theta} J(\theta_k)\| (k+2)^2}{(K+1)^2} = \mathcal{O} \left(\frac{G\sigma_g}{\mu\sqrt{K}} + \frac{G^2L}{\mu^2\sqrt{K}} + \frac{AG^2t_{\text{mix}} \log T}{\mu T^4} \right). \quad (112)$$

It follows that

$$J^* - \mathbb{E}[J(\theta_K)] \leq \mathcal{O} \left(\frac{G\sigma_g}{\mu\sqrt{K}} + \frac{G^2L}{\mu^2\sqrt{K}} + \frac{AG^2t_{\text{mix}} \log T}{\mu T^4} \right). \quad (113)$$

Proof of Equation (77): In order to get obtain the final regret decomposition, we show that

$$H \sum_{k=1}^K (J^* - J(\hat{\theta}_k)) = \mathcal{O} \left(H \sum_{k=1}^K (J^* - J(\theta_k)) \right), \quad (114)$$

while the rest of the proof follows from Section E. From the L -smoothness of \bar{J} , we have

$$-\bar{J}(\hat{\theta}_{k+1}) \leq -\bar{J}(\theta_k) - \langle \nabla_{\theta} \bar{J}(\theta_k), \hat{\theta}_{k+1} - \theta_k \rangle + \frac{L}{2} \|\hat{\theta}_{k+1} - \theta_k\|^2. \quad (115)$$

Note that $\hat{\theta}_{k+1} = q_{k+1}\theta_{k+1} + (1-q_{k+1})\theta_k$, which implies $\hat{\theta}_{k+1} - \theta_k = q_{k+1}(\theta_{k+1} - \theta_k)$. Substituting this in (115), adding J^* to both sides and taking expectation conditioned on θ_k and θ_{k+1} yields

$$\mathbb{E}[J^* - \bar{J}(\hat{\theta}_{k+1}) \mid \theta_k, \theta_{k+1}] \leq J^* - \bar{J}(\theta_k) - \frac{1}{2} \langle \nabla_{\theta} \bar{J}(\theta_k), \theta_{k+1} - \theta_k \rangle + \frac{L}{6} \|\theta_{k+1} - \theta_k\|^2.$$

Utilizing arguments similar to (90), we obtain

$$J^* - \mathbb{E} J(\hat{\theta}_{k+1}) \leq J^* - \mathbb{E} J(\theta_k) - \frac{\gamma_k}{3} \mathbb{E} \|\nabla_{\theta} J(\theta_k)\| + \frac{8\gamma_k}{3} \mathbb{E} \|\hat{e}_k\| + \frac{L\gamma_k^2}{2} \quad (116)$$

$$+ \mathcal{O} \left(\frac{AGt_{\text{mix}} \log T}{T^3} \right). \quad (117)$$

Taking into account the above bound and the inequality $\|\theta_{k+1} - \hat{\theta}_k\| \leq \|\theta_{k+1} - \theta_k\| + \|\theta_k - \hat{\theta}_k\| \leq 2\gamma_k$, and by replacing the bounds for σ_g and L in (113), we derive the subsequent bound for the expected regret of Algorithm 3:

$$\begin{aligned} \mathbb{E}[\text{Reg}_T] &\leq T\sqrt{\epsilon_{\text{bias}}} + \mathcal{O} \left(\frac{\sqrt{AG^2t_{\text{mix}} \log T}}{\mu} \cdot \sqrt{T} + \frac{\sqrt{AG^4t_{\text{hit}}t_{\text{mix}}^2(\log T)^{3/2}}}{\mu^2} \cdot \sqrt{T} \right. \\ &\quad \left. + \frac{\sqrt{A}(BG + G^3)t_{\text{mix}} \log T}{\mu^2} \cdot \sqrt{T} \right). \end{aligned} \quad (118)$$

I Auxillary lemmas

Lemma 14. [Wei et al., 2020, Lemma 14] For ergodic MDPs with mixing time t_{mix} , the following holds $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$, any policy π and $\forall g \in \{r, c\}$.

$$(a) |V_g^{\pi}(s)| \leq 5t_{\text{mix}}, \quad (b) |Q_g^{\pi}(s, a)| \leq 6t_{\text{mix}}$$

Lemma 15. [Wei et al., 2020, Corollary 13.1] For an ergodic MDP with mixing time t_{mix} , we have

$$\|(P^\pi)^t(s, \cdot) - d^\pi\|_1 \leq 2 \cdot 2^{-t/t_{\text{mix}}}, \quad (119)$$

for all $\pi, s \in \mathcal{S}$ and $t \geq 2t_{\text{mix}}$.

Lemma 16. Let $N = 7t_{\text{mix}} \log_2 T$. For an ergodic MDP with mixing time $t_{\text{mix}} < T/4$, define

$$\delta^\pi(s, T) := \sum_{t=N}^{\infty} \|(P^\pi)^t(s, \cdot) - d^\pi\|_1. \quad (120)$$

We have the following bound $\forall \pi \in \Pi, s \in \mathcal{S}$.

$$\delta^\pi(s, T) \leq \frac{1}{T^6} \quad (121)$$

Proof. From Lemma 15, we have

$$\delta^\pi(s, T) \leq \sum_{t=N}^{\infty} 2 \cdot 2^{-t/t_{\text{mix}}} \leq \frac{2 \cdot 2^{-N/t_{\text{mix}}}}{1 - 2^{-1/t_{\text{mix}}}} \leq \frac{4t_{\text{mix}}}{\ln 2} \cdot 2^{-N/t_{\text{mix}}} \leq \frac{1}{T^6}. \quad (122)$$

The second inequality follows from the fact that $1 - 2^{-x} \geq \frac{1}{2}x \ln 2$, $\forall x \in (0, 1)$. \square

Lemma 17. [Wei et al., 2020, Lemma 16] Let $\mathcal{I} = \{t_1 + 1, t_1 + 2, \dots, t_2\}$ be a certain period of an epoch k of Algorithm 2 with length N . Then for any s , the probability that the algorithm never visits s in \mathcal{I} is upper bounded by

$$\left(1 - \frac{3d^{\pi_{\theta_k}}(s)}{4}\right)^{\lfloor \frac{|\mathcal{I}|}{N} \rfloor} \quad (123)$$

Lemma 18 (Lemma 4, [Bai et al., 2024]). The difference in the performance for policies π and π' is

$$J^\pi - J^{\pi'} = \mathbb{E}_{s \sim d^\pi} \mathbb{E}_{a \sim \pi(\cdot|s)} [A^{\pi'}(s, a)] \quad (124)$$

Lemma 19 (Gradient domination lemma). Let Assumptions 2, 3, and 4 hold. The following inequality holds for any $\theta \in \Theta$.

$$J^* - J(\theta) \leq \sqrt{\epsilon_{\text{bias}}} + \frac{G}{\mu} \|\nabla_\theta J(\theta)\| \quad (125)$$

Proof. From Lemma 18, we arrive at

$$J^* - J(\theta) = \mathbb{E}_{s \sim d^{\pi^*}} \mathbb{E}_{a \sim \pi^*(\cdot|s)} [A^{\pi_\theta}(s, a)] \quad (126)$$

Moreover, we obtain the following from Assumption 3.

$$\begin{aligned} \epsilon_{\text{bias}} &\geq \mathbb{E}_{s \sim d^{\pi^*}} \mathbb{E}_{a \sim \pi^*(\cdot|s)} \left[\left(A^{\pi_\theta}(s, a) - \nabla_\theta \log \pi_\theta(a|s) \cdot \omega_\theta^* \right)^2 \right] \\ &\geq \left(\mathbb{E}_{s \sim d^{\pi^*}} \mathbb{E}_{a \sim \pi^*(\cdot|s)} \left[A^{\pi_\theta}(s, a) - \nabla_\theta \log \pi_\theta(a|s) \cdot \omega_\theta^* \right] \right)^2 \end{aligned} \quad (127)$$

From (126) and (127), one can deduce

$$\begin{aligned} \sqrt{\epsilon_{\text{bias}}} &\geq \mathbb{E}_{s \sim d^{\pi^*}} \mathbb{E}_{a \sim \pi^*(\cdot|s)} \left[A^{\pi_\theta}(s, a) - \nabla_\theta \log \pi_\theta(a|s) \cdot \omega_\theta^* \right] \\ &= (J^* - J(\theta)) - \mathbb{E}_{s \sim d^{\pi^*}} \mathbb{E}_{a \sim \pi^*(\cdot|s)} \left[\nabla_\theta \log \pi_\theta(a|s) \cdot \omega_\theta^* \right] \end{aligned} \quad (128)$$

Rearranging the above inequality yields

$$\begin{aligned}
 J^* - J(\theta) &\leq \sqrt{\epsilon_{\text{bias}}} + \mathbb{E}_{s \sim d^{\pi^*}} \mathbb{E}_{a \sim \pi^*(\cdot|s)} \left[\nabla_{\theta} \log \pi_{\theta}(a|s) \cdot \omega_{\theta}^* \right] \\
 &\leq \sqrt{\epsilon_{\text{bias}}} + \mathbb{E}_{s \sim d^{\pi^*}} \mathbb{E}_{a \sim \pi^*(\cdot|s)} \left[\|\nabla_{\theta} \log \pi_{\theta}(a|s)\| \|\omega_{\theta}^*\| \right] \\
 &\leq \sqrt{\epsilon_{\text{bias}}} + G \|\omega_{\theta}^*\|
 \end{aligned} \tag{129}$$

where the last inequality follows from Assumption 2. Note that

$$\|\omega_{\theta}^*\| = \|F(\theta)^{\dagger} \nabla_{\theta} J(\theta)\| \stackrel{(a)}{\leq} \mu^{-1} \|\nabla_{\theta} J(\theta)\|$$

where (a) is a consequence of Assumption 4. It follows that

$$J^* - J(\theta) \leq \sqrt{\epsilon_{\text{bias}}} + \frac{G}{\mu} \|\nabla_{\theta} J(\theta)\| \tag{130}$$

This concludes the proof of Lemma 19. □