Regret Analysis of Average-Reward Unichain MDPs via an Actor-Critic Approach

Swetha Ganesh
Purdue University, USA
ganesh49@purdue.edu

Vaneet Aggarwal
Purdue University, USA
vaneet@purdue.edu

Abstract

Actor-Critic methods are widely used for their scalability, yet existing theoretical guarantees for infinite-horizon average-reward Markov Decision Processes (MDPs) often rely on restrictive ergodicity assumptions. We propose NAC-B, a Natural Actor-Critic with Batching, that achieves order-optimal regret of $\tilde{O}(\sqrt{T})$ in infinite-horizon average-reward MDPs under the unichain assumption, which permits both transient states and periodicity. This assumption is among the weakest under which the classic policy gradient theorem remains valid for average-reward settings. NAC-B employs function approximation for both the actor and the critic, enabling scalability to problems with large state and action spaces. The use of batching in our algorithm helps mitigate potential periodicity in the MDP and reduces stochasticity in gradient estimates, and our analysis formalizes these benefits through the introduction of the constants C_{hit} and C_{tar}, which characterize the rate at which empirical averages over Markovian samples converge to the stationary distribution.

1 Introduction

Reinforcement Learning (RL) involves an agent interacting with an unknown environment to maximize long-term rewards. It has been successfully applied to diverse areas such as traffic engineering, resource allocation, and ride-sharing [18, 11, 5]. RL problems are commonly framed as episodic, discounted, or (infinite-horizon) average-reward; the average-reward setting is especially suited to real-world tasks due to its ability to better capture long-term behavior.

Our focus is on the average-reward Markov Decision Process (MDP) setting, where a key measure of algorithmic performance is expected regret. [7] established that for a broad class of MDPs, any algorithm incurs expected regret lower bounded by $\Omega(\sqrt{T})$, where T is the time horizon. Many existing regret analyses in this setting assume ergodicity, a strong assumption that simplifies analysis by ensuring fast mixing to a stationary distribution, but is often difficult to verify or justify in practice.

In the absence of ergodicity, existing algorithms are typically designed for model-based or tabular settings, with computational complexity that scales with $|\mathcal{S}|$ and $|\mathcal{A}|$, the cardinalities of the state and action spaces, respectively. These requirements are prohibitive for environments with large or continuous state and action spaces. Although linear MDPs reduce improves dependence on $|\mathcal{S}|$, they still rely on strong structural assumptions, namely, linearity of the transition dynamics and reward functions. Moreover, value-based methods in such settings often remain computationally expensive due to the need to maximize over all actions at each iteration.

To overcome these limitations, a promising alternative is the use of Policy Gradient (PG) methods, which directly optimize parameterized policies, often modeled by neural networks, via gradient descent. While PG methods have shown strong empirical performance, most theoretical analyses assume ergodicity. In contrast, our work develops PG methods that do not rely on ergodicity.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Algorithm	Regret	Ergodicity-free	General Policy
MDP-OOMD [38]	$\tilde{O}(\sqrt{T})$	No	No
Optimistic Q-learning [38]	$\tilde{O}(T^{2/3})$	Yes[(1)]	No
MDP-EXP2 [39]	$\tilde{O}(\sqrt{T})$	No	No
UCB-AVG [45]	$\tilde{O}(\sqrt{T})$	Yes[(1)]	No
PPG [9]	$\tilde{O}(T^{3/4})$	No	Yes
PHAPG [17]	$\tilde{O}(\sqrt{T})$	No	Yes
Optimistic Q-learning [3]	$\tilde{O}(\sqrt{T})$	Yes	No
γ-DC-LSCVI-UCB [20]	$\tilde{O}(\sqrt{T})$	Yes[(1)]	No
This work (Algorithm 1)	$\tilde{O}(\sqrt{T})$	Yes	Yes

Table 1: A comparison of regret bounds for model-free RL algorithms in infinite-horizon average-reward MDPs. By "General Policy," we refer to algorithms that employ a parameterized, policy-based approach where the parameter vector $\theta \in \mathbb{R}^d$, possibly with $d \ll |\mathcal{S}||\mathcal{A}|$. [(1)] These analyses consider the more general setting of weakly communicating MDPs or Bellman optimality.

Specifically, we make only a minimal assumption: each policy induces a unichain MDP, i.e., it has a single recurrent class. This relaxation enables analysis in more general and realistic environments.

1.1 Related works

Value-Based Approaches: Model-based algorithms like those in [7, 4, 21, 46], as well as recent model-free methods, achieve the optimal $\mathcal{O}(\sqrt{T})$ regret. However, most are tabular and value-based, maintaining Q-values for each state action pair, an approach that scales poorly to large or continuous spaces. Examples include optimistic Q-learning methods [38, 3, 45]. The γ-DC-LSCVI-UCB algorithm [20] improves scalability with respect to $|\mathcal{S}|$ but requires maximization over the entire action set at each iteration, which is still computationally expensive for large $|\mathcal{A}|$. Furthermore, its analysis assumes a linear MDP with known feature representations, whereas our setting is considerably more general and does not rely on such structural knowledge or assumptions.

Direct Policy Gradient Approaches: Unlike value-based methods, policy gradient (PG) approaches are well-suited to environments with large or continuous state and action spaces, and are relatively easy to implement. Despite their empirical success, theoretical analysis of PG methods has largely focused on ergodic MDPs. This is due to favorable properties in the ergodic setting, such as bounded hitting times and fast mixing, which simplify analysis and enable near-optimal regret bounds. These properties ensure that samples collected t_{mix} steps apart approximate independent draws from the stationary distribution. Algorithms like MDP-OOMD [38], PPG [9], and PHAPG [17] exploit these characteristics to construct low-bias value estimates. However, to the best of our knowledge, no prior work provides regret guarantees for PG methods in the more general unichain setting, where such strong mixing assumptions do not hold.

Actor-Critic Approaches: In contrast to direct policy gradient methods that estimate gradients from sampled trajectories, Actor-Critic (AC) methods use Temporal Difference (TD) based critics to aid policy gradient through bootstrapped value estimates. Further, direct methods scale poorly and rely on access to mixing or hitting times, typically unknown in practice. AC methods are more sample-efficient but introduce addtional bias, making theoretical analysis harder. Regret results for AC methods are limited and focus on global convergence (pseudo-regret), as seen in [29, 37, 16]. The MLMC-NAC algorithm in [16] achieves order-optimal convergence using multi-level Monte Carlo to reduce bias, but still assumes fast mixing. Extending such results to unichain MDPs, which lack exponential mixing, requires fundamentally different techniques and is the focus of this paper.

Unichain Analyses: A few recent works explicitly address the unichain setting [3, 23]. The Optimistic Q-learning algorithm in [3], discussed earlier, is value-based. The SPMD method in [23] takes a policy-based approach with order-optimal sample complexity but uses a tabular policy, leading to poor scalability with large action spaces. It also assumes access to a simulator (generator), which is often unrealistic. Moreover, both works rely on stronger assumptions than the classical unichain condition used in our analysis (see Remark 1).

Discounted to Average Techniques: The most common strategy for relaxing ergodicity assumptions in reinforcement learning is to reduce the average-reward problem to a discounted MDP. This reduction is widely adopted in works addressing weakly communicating or Bellman optimality settings [38, 45, 20, 46], where the discounted problem is solved with a discount factor γ close to 1. In particular, these works select γ satisfying $1/(1-\gamma) \simeq T^{\beta}$ for some $\beta > 0$. Notably, this approach requires very sharp bounds in DMDPs in terms of $1/(1-\gamma)$ whereas, existing guarantees for policy gradient methods in the discounted setting exhibit very poor dependence on this term, resulting in much weaker performance bounds if such Discounted to Average Techniques are used (see Appendix A). Additionally, these approaches require access to a simulator that allows sampling from a given distribution ρ at each iteration, which can be impractical.

1.2 Main contributions

We propose an actor-critic algorithm in Algorithm 1, Natural Actor Critic with Batching (NAC-B), that achieves order-optimal regret of $\tilde{O}(\sqrt{T})$ in average-reward Markov decision processes (MDPs) under the unichain assumption. Our key contributions are summarized as follows:

- **First Regret Guarantees for General Policy Gradients Beyond Ergodicity:** While policy gradient (PG) methods with general parameterizations have been analyzed under the restrictive ergodicity assumption, we extend this to the weaker unichain condition. This aligns with the assumption used in the foundational policy gradient theorem for average-reward MDPs [33]. Our work provides the first regret guarantees for parameterized PG methods under this setting.
- **Online Learning Without Simulator Access:** Unlike discounted-MDP reductions that require a simulator to reset to arbitrary initial distributions at each iteration, our algorithm operates entirely online in unichain MDPs. This enables learning in environments where simulator access is unavailable or costly (e.g., real-time network traffic engineering, robotics).
- **Technical Challenges:** Analyses in the ergodic setting rely crucially on the property of exponentially fast mixing. In contrast, the classical unichain setting does not generally admit analogous results. For example, in a periodic Markov decision process, the quantity $\lim_{t\to\infty}(P^{\pi_\theta})^t(s_0,\cdot)$ may not even exist. To address this challenge arising from periodicity, we employ large-batch averaging. Specifically, we show that the time-averaged distribution $\left\|\frac{1}{t}\sum_{i=1}^{t}(P^{\pi_\theta})^i(s_0,\cdot) - d^{\pi_\theta}(\cdot)\right\|_{\mathrm{TV}}$ converges to the stationary distribution at a rate of $\mathcal{O}(1/t)$ (see Lemma 2). This averaging approach enables us to show that the bias and variance of our estimators decay at a sufficiently fast rate, albeit not exponentially (see Lemma 3). An additional consequence of these results is that they facilitate the derivation of bounds on the value and Q-functions.

  Despite the reduction in bias and variance, transient states continue to pose analytical challenges. Intuitively, this is because they do not contribute meaningful long-term information. A more technical explanation is provided in Section 4.2. To isolate the impact of transient states, we prove a rapid entry into the recurrent class (Lemma 10). Combined with the strong Markov property, this allows us to restrict our analysis to the recurrent class.

2 Problem Formulation and Preliminaries

We study an infinite-horizon reinforcement learning problem with an average reward criterion, modeled as a Markov Decision Process (MDP). The MDP is represented by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, P, \rho)$, where $\mathcal{S}$ denotes the finite state space and $\mathcal{A}$ represents the finite action space. The reward function $r : \mathcal{S} \times \mathcal{A} \to [0,1]$ assigns a bounded reward to each state-action pair. The state transition function $P : \mathcal{S} \times \mathcal{A} \to \Delta^{|\mathcal{S}|}$ determines the probability distribution over the next state given the current state and action, where $\Delta^{|\mathcal{S}|}$ denotes the probability simplex over $\mathcal{S}$. The initial state distribution is given by $\rho : \mathcal{S} \to [0,1]$. A policy $\pi : \mathcal{S} \to \Delta^{|\mathcal{A}|}$ specifies a probability distribution over actions for each state. Given a policy π, the long-term average reward is defined as

$$J_\rho^\pi \triangleq \lim_{T\to\infty} \frac{1}{T}\,\mathbb{E}\left[\sum_{t=0}^{T-1} r(s_t,a_t)\middle| s_0 \sim \rho\right], \tag{1}$$

where the expectation is taken over trajectories generated by executing actions according to $a_t \sim \pi(\cdot|s_t)$ and transitioning states via $s_{t+1} \sim P(\cdot|s_t,a_t)$, for all $t \geq 0$. For simplicity, we drop the dependence on ρ whenever it is clear from the context. We focus on a parametrized policy class Π, where each policy is indexed by a parameter $\theta \in \Pi_\Theta$, with $\Pi_\Theta \subset \mathbb{R}^d$. Our goal is to solve the optimization problem: $\max_{\theta\in\Pi_\Theta} J^{\pi_\theta} \triangleq J(\theta)$.

A policy $\pi_\theta \in \Pi$ induces a transition function $P^{\pi_\theta} : \mathcal{S} \to \Delta^{|\mathcal{S}|}$, given as $P^{\pi_\theta}(s, s') = \sum_{a \in \mathcal{A}} P(s'|s, a)\pi_\theta(a|s), \quad \forall s, s' \in \mathcal{S}$. The corresponding stationary distribution is defined as:

Definition 1. *Let $d^{\pi_\theta} \in \Delta^{|\mathcal{S}|}$ denote the stationary distribution of the Markov chain induced by π_θ, given by*

$$d^{\pi_\theta}(s) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \Pr(s_t = s \mid s_0 \sim \rho, \pi_\theta). \tag{2}$$

We assume the following property for the MDP, which is defined as a unichain MDP:

Assumption 1. *The MDP $\mathcal{M}$ is such that, for every policy $\pi \in \Pi$, the induced Markov chain has a single recurrent class.*

Remark 1. *An MDP $\mathcal{M}$ satisfying this property is referred to as a unichain MDP [30]. This assumption does not require irreducibility, as transient states may be present, nor does it impose aperiodicity. Consequently, it is strictly weaker than the standard ergodicity assumptions.*

Under this condition, the stationary distribution d^{π_θ} is well-defined, independent of the initial distribution ρ, and satisfies the balance equation: $(P^{\pi_\theta})^\top d^{\pi_\theta} = d^{\pi_\theta}$.

This assumption forms the basis of the foundational policy gradient theorem for average-reward MDPs [33]. Although various alternative definitions of unichain MDPs appear in the literature, the notion employed here is classical and, in fact, weaker than several recent formulations. For example, [3] assume the existence of a state s_0 that is recurrent under all policies, a stronger condition than ours. Similarly, [23] adopt a mixing unichain assumption, which additionally requires aperiodicity. In contrast, our analysis does not depend on either of these stronger conditions.

Since we cannot rely on the mixing time commonly used in the analysis of ergodic MDPs, we consider two alternative quantities, C^θ_{hit} and C^θ_{tar}, which are more appropriate in the unichain setting. These are defined with respect to the Markov chain induced by a stationary policy π as follows:

Let $\mathcal{S}^\theta_R \subseteq \mathcal{S}$ denote the recurrent class under policy π_θ, and let $T_\theta := \inf\{t \geq 0 : s_t \in \mathcal{S}^\theta_R\}$ denote the first hitting time of the recurrent class. Then we define

$$C^\theta_{\mathrm{hit}} := \max_{s\in\mathcal{S}} \mathbb{E}^\theta_s[T_\theta], \tag{3}$$

as the worst-case expected time to enter the recurrent class when starting from any state $s \in \mathcal{S}$. Note that if there are no transient states, then $C^\theta_{\mathrm{hit}} = 0$.

Similarly, for any $s, s' \in \mathcal{S}$, let $T_{s'} := \inf\{t \geq 0 : s_t = s'\}$ be the first hitting time to state s'. Then we define

$$C^\theta_{\mathrm{tar}} := \sum_{s'\in\mathcal{S}} d^\pi(s')\mathbb{E}^\theta_s[T_{s'}], \tag{4}$$

as the expected time to reach a state drawn from d^{π_θ}, starting from a fixed state $s \in \mathcal{S}^\theta_R$. Notably, by the Random Hitting Target lemma (Corollary 2.14, [6]), this quantity is independent of the choice of s. We define $C_{\mathrm{hit}} := \sup_\theta C^\theta_{\mathrm{hit}}$ and $C_{\mathrm{tar}} := \sup_\theta C^\theta_{\mathrm{tar}}$ and $C := C_{\mathrm{hit}} + C_{\mathrm{tar}}$.

We write the average reward as:

$$J(\theta) = \mathbb{E}_{s\sim d^{\pi_\theta}, a\sim\pi_\theta(\cdot|s)}[r(s, a)] = (d^{\pi_\theta})^\top r^{\pi_\theta}, \text{ where } r^{\pi_\theta}(s) \triangleq \sum_{a\in\mathcal{A}} r(s, a)\pi_\theta(a|s),\ \forall s \in \mathcal{S} \tag{5}$$

The average reward $J(\theta)$ is also independent of the initial distribution, ρ. Furthermore, $\forall \theta \in \Pi_\Theta$, there exist a function $Q^{\pi_\theta} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that the following Bellman equation is satisfied $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$.

$$Q^{\pi_\theta}(s, a) = r(s, a) - J(\theta) + \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[V^{\pi_\theta}(s')\right] \tag{6}$$

where the state value function, $V^{\pi_\theta} : \mathcal{S} \to \mathbb{R}$ is defined as $V^{\pi_\theta}(s) = \sum_{a\in\mathcal{A}} \pi_\theta(a|s)Q^{\pi_\theta}(s, a),\ \forall s \in \mathcal{S}$ [30]. Note that if (6) is satisfied by Q^{π_θ}, then it is also satisfied by $Q^{\pi_\theta} + c$ for any arbitrary constant, c.

Additionally, we define the advantage function as $A^{\pi_\theta}(s, a) \triangleq Q^{\pi_\theta}(s, a) - V^{\pi_\theta}(s)$. If Assumption 1 holds, it is known that the policy gradient at θ, $\nabla_\theta J(\theta)$ can be expressed as follows [33]:

$$\nabla_\theta J(\theta) = \mathbb{E}_{(s,a)\sim\nu^{\pi_\theta}}[A^{\pi_\theta}(s, a)\nabla_\theta \log \pi_\theta(a|s)], \tag{7}$$

where $\nu^{\pi_\theta}(s,a) := d^{\pi_\theta}(s)\pi_\theta(a|s)$. This is commonly referred to as the Policy Gradient theorem and while the proof can be found in [33], we include it in Appendix B for completeness.

Natural Policy Gradient (NPG) methods update θ along the NPG direction ω_θ^* defined as

$$\omega_\theta^* = F(\theta)^\dagger \nabla_\theta J(\theta), \tag{8}$$

where $\dagger$ denotes the Moore-Penrose pseudoinverse and $F(\theta)$ is the Fisher matrix defined as

$$F(\theta) = \mathbb{E}\left[\nabla_\theta \log \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s)^\top\right] \tag{9}$$

where the expectation is taken over $(s,a) \sim \nu^{\pi_\theta}$. This yields the NPG update

$$\theta_{k+1} = \theta_k + \alpha\omega_k^*.$$

Let $J^* \triangleq \sup_{\boldsymbol{\theta}\in\Pi_\Theta} J(\theta)$. For a given MDP $\mathcal{M}$ and a time horizon T, the regret of an algorithm $\mathbb{A}$ is defined as follows.

$$\mathrm{Reg}_T(\mathbb{A}, \mathcal{M}) \triangleq \sum_{t=0}^{T-1} \left(J^* - r(s_t, a_t)\right) \tag{10}$$

where the action, $a_t,\ t \in \{0, 1, \cdots\}$ is chosen by following the algorithm $\mathbb{A}$ based on the trajectory up to time, t, and the state, s_{t+1} is obtained by following the state transition function, P. Wherever there is no confusion, we shall simplify the notation of regret to Reg_T.

Algorithm 1 Natural Actor-Critic with Batching

Input: Initial parameters θ_0, $\{\omega_0^k\}$, and $\{\xi_0^k\}$, policy stepsize α, critic parameters (β, c_β), NPG stepsize γ, initial state $s_0 \sim \rho$, outer loop size K, inner loop size H, batch-size B
for $k = 0, 1, \cdots, K-1$ **do**
  ▷ Average reward and critic estimation
  for $h = 0, 1, \cdots, H-1$ **do**
    $s_k^0 \leftarrow s_0$
    for $b = 1, 2, \cdots, B$ **do**
      Take action $a_b^{kh} \sim \pi_{\theta_k}(\cdot|s_b^{kh})$
      Collect next state $s_{b+1}^{kh} \sim P(\cdot|s_b^{kh}, a_b^{kh})$
      Receive reward $r(s_b^{kh}, a_b^{kh})$
    end for
    Update the combined average reward and critic estimate $\xi_h^k = [\eta_h^k, \zeta_h^k]^\top$ using (22)-(24)
    $s_0 \leftarrow s_B^{kh}$
  end for
  $\xi_k \leftarrow \xi_H^k$
  ▷ Natural Policy Gradient estimation
  for $h = 0, 1, \cdots, H-1$ **do**
    $s_k^0 \leftarrow s_0$
    for $b = 1, 2, \cdots, B$ **do**
      Take action $a_b^{kh} \sim \pi_{\theta_k}(\cdot|s_b^{kh})$
      Collect next state $s_{b+1}^{kh} \sim P(\cdot|s_b^{kh}, a_b^{kh})$
      Receive reward $r(s_b^{kh}, a_b^{kh})$
    end for
    Update NPG estimate ω_h^k using (25)-(29)
    $s_0 \leftarrow s_B^{kh}$
  end for
  ▷ Policy update
  $s_k^0 \leftarrow s_0$
  Update policy parameter θ_k using $\theta_{k+1} \leftarrow \theta_k + \alpha\omega_k$
  $s_0 \leftarrow s_B^k$
end for

3 Proposed Algorithm

We propose a *Natural Actor–Critic with Batching* algorithm (Algorithm 1), which runs for K outer iterations (or *epochs*) of natural policy gradient updates. Each outer iteration includes H inner iterations to estimate the average reward and value function, followed by another H iterations to estimate the NPG direction. We first present the necessary preliminaries, followed by the detailed description of the algorithm.

Preliminaries: For any fixed θ_k, it is known that ω_k^* is the solution to the optimization problem

$$\omega_k^* = \arg\min_\omega L_{\nu^{\pi_{\theta_k}}}(\omega, \theta_k),$$

where $L_{\nu^{\pi_\theta}}(\omega, \theta)$ is defined as follows

$$L_{\nu^{\pi_\theta}}(\omega, \theta) = \frac{1}{2}\mathbb{E}\left[\left(A^{\pi_\theta}(s,a) - \omega^\top \nabla_\theta \log \pi_\theta(a|s)\right)^2\right], \tag{11}$$

where the expectation is taken over $(s,a) \sim \nu^{\pi_\theta}$. It can be seen that $L_{\nu^{\pi_\theta}}(\omega, \theta)$ is a quadratic function in ω with gradient given by $F(\theta)\omega - \nabla J(\theta)$. Since neither $F(\theta_k)$ nor $\nabla J(\theta_k)$ is available in closed form, we estimate them from samples and perform stochastic gradient descent on $L_{\nu^{\pi_{\theta_k}}}(\omega, \theta_k)$ to obtain an approximate NPG direction ω_k.

We estimate $F(\theta_k)$ using the outer product $[\nabla \log \pi_{\theta_k}(a|s) \nabla \log \pi_{\theta_k}(a|s)^\top]$ computed from samples $(s,a) \sim \nu^{\pi_{\theta_k}}$. However, estimating $\nabla J(\theta_k)$ is more challenging, as it requires access to the advantage function, which is not directly observable. We note that $A^{\pi_{\theta_k}}(s,a)$ can be expressed using $J(\theta_k)$ and $V^{\pi_{\theta_k}}$, as follows from the definition of $A^{\pi_{\theta_k}}(s,a)$ and Bellman's equation. Consequently, we focus next on estimating these quantities.

Notice that $\eta_k = J(\theta_k)$ is the solution to the following equation:

$$R_{\theta_k}(\eta_k) = 0, \quad \text{where} \quad R_{\theta_k}(\eta) = \mathbb{E}_{(s,a)\sim\nu^{\pi_{\theta_k}}}[\eta - r(s,a)]. \tag{12}$$

Given access to $R_{\theta_k}(\eta)$, the solution to the above can be computed via the following iterative update:

$$\eta_{h+1}^k = \eta_h^k - c_\beta \beta R_{\theta_k}(\eta_h^k), \tag{13}$$

where β and c_β are step-size parameters.

To approximate the value function $V^{\pi_{\theta_k}}(\cdot)$, we employ a linear critic $\hat{V}(\zeta_{\theta_k},\cdot) = \zeta_{\theta_k}^\top \phi(\cdot)$, where $\zeta_{\theta_k} \in \mathbb{R}^m$ denotes the critic parameter and $\phi(s) \in \mathbb{R}^m$ is a feature mapping satisfying $|\phi(s)| \leq 1$ for all $s \in \mathcal{S}$. We also assume that the feature vectors $\phi(s)$ are linearly independent. The parameter $\zeta_{\theta_k} \in \mathbb{R}^m$ is obtained by solving the following optimization problem:

$$\min_{\zeta\in\mathbb{R}^m} E(\theta_k,\zeta) := \frac{1}{2}\mathbb{E}_{s\sim d^{\pi_{\theta_k}}}\left(V^{\pi_{\theta_k}}(s) - \hat{V}(\zeta,s)\right)^2. \tag{14}$$

This formulation enables the use of a gradient-based iterative procedure to compute ζ_{θ_k}. Specifically, the gradient of $E(\theta_k,\zeta)$ is given by:

$$\begin{aligned}\nabla_\zeta E(\theta_k,\zeta) &= \mathbb{E}_{s\sim d^{\pi_{\theta_k}}}\left[\left(V^{\pi_{\theta_k}}(s) - \hat{V}(\zeta,s)\right)\nabla_\zeta \hat{V}(\zeta,s)\right] \\ &= \mathbb{E}_{s\sim d^{\pi_{\theta_k}}}\left[\left(V^{\pi_{\theta_k}}(s) - \zeta^\top\phi(s)\right)\phi(s)\right]\end{aligned} \tag{15}$$

The above quantity can be approximated as follows

$$\nabla_\zeta E(\theta_k,\zeta) = \mathbb{E}_{s\sim d^{\pi_{\theta_k}}}\left[\left(V^{\pi_{\theta_k}}(s) - \zeta^\top\phi(s)\right)\phi(s)\right] \tag{16}$$

$$= \mathbb{E}_{(s,a)\sim\nu^{\pi_{\theta_k}},s'\sim P(\cdot|s,a)}\left[\left(r(s,a) - J(\theta_k) + V^{\pi_{\theta_k}}(s') - \zeta^\top\phi(s)\right)\phi(s)\right] \tag{17}$$

$$\approx \mathbb{E}_{(s,a)\sim\nu^{\pi_{\theta_k}},s'\sim P(\cdot|s,a)}\left[\left(r(s,a) - J(\theta_k) + \zeta^\top\phi(s') - \zeta^\top\phi(s)\right)\phi(s)\right] \tag{18}$$

$$:= \widetilde{\nabla}_\zeta E(\theta_k,\zeta) \tag{19}$$

Using this gradient, ζ_{θ_k} can be computed iteratively via:

$$\zeta_{h+1}^k = \zeta_h^k - \beta\widetilde{\nabla}_\zeta E(\theta_k,\zeta_h^k). \tag{20}$$

The updates in (13), and (20) can be combined into one update for $\xi_h^k = [(\eta_h^k)^\top, (\zeta_h^k)^\top]^\top$ as follows

$$\xi_{h+1}^k = \xi_h^k - \beta[c_\beta R_\theta(\eta_h^k)^\top, \widetilde{\nabla}_\zeta E(\theta_k,\zeta_h^k)^\top]^\top \tag{21}$$

The joint update rule for η_h^k and ζ_h^k in (21) forms the basis of the critic step in Algorithm 1, where sample averages from the trajectory are used in place of the intractable expectations.

Algorithm 1 operates using a single trajectory of total length $T = 2KHB$. For each outer iteration k, the algorithm collects a contiguous segment of $2BH$ samples by following the current policy π_{θ_k}, starting from the final state of the previous iteration. The first BH samples from π_{θ_k} are used to perform H iterations of the critic update subroutine, and the next BH samples are used for H iterations of the NPG update.

Critic subroutine: For each $h \in 0,\ldots,H-1$, we use a sample averaged version of (21) to obtain the update

$$\xi_{h+1}^k = \xi_h^k - \beta\left[\frac{1}{B}\sum_{b=1}^B v_k(z_b^{kh};\xi_h^k)\right], \tag{22}$$

where β is the critic learning rate, B is the batch size, $z_b^{kh} = (s_b^{kh}, a_b^{kh}, s_{b+1}^{kh})$ is a sampled transition, and $v(z_b^{kh};\xi_h^k)$ is given by:

$$v_k(z_b^{kh};\xi_h^k) = A_v(\theta_k, z_b^{kh})\xi_h^k - b_v(\theta_k, z_b^{kh}), \tag{23}$$

where the matrices are defined as

$$A_v(\theta_k, z_b^{kh}) = \begin{bmatrix} c_\beta & 0 \\ \phi(s_b^{kh}) & \phi(s_b^{kh}) \left[\phi(s_b^{kh}) - \phi(s_{b+1}^{kh})\right]^\top \end{bmatrix}, \; b_v(\theta_k, z_b^{kh}) = \begin{bmatrix} c_\beta r(s_b^{kh}, a_b^{kh}) \\ r(s_b^{kh}, a_b^{kh})\phi(s_b^{kh}) \end{bmatrix} \tag{24}$$

The resulting update follows standard constructions in Temporal Difference learning (e.g., [44]). After H critic updates, we take $\eta_k := \eta_H^k$ and $\zeta_k := \zeta_H^k$ as the final estimates of the average reward and critic parameters, giving $\hat{V}(\zeta_k, \cdot) = \zeta_k^\top \phi(\cdot)$ as the value estimate.

Natural Policy Gradient estimation: The *advantage estimate* at sample z_b^{kh} is computed from the Bellman equation using η_k and $\hat{V}(\zeta_k, \cdot)$ as follows

$$\hat{A}^{\pi_{\theta_k}}(z_b^{kh}; \xi_k) = r(s_b^{kh}, a_b^{kh}) - \eta_k + \zeta_k^\top \left[\phi(s_{b+1}^{kh}) - \phi(s_b^{kh})\right]. \tag{25}$$

Based on the above estimate, we construct an approximate policy gradient:

$$b_u(\theta_k, \xi_k, z_b^{kh}) = \hat{A}^{\pi_{\theta_k}}(z_b^{kh}; \xi_k) \nabla_\theta \log \pi_{\theta_k}(a_b^{kh} | s_b^{kh}), \tag{26}$$

and refine the natural gradient estimate ω_k via H iterations of the update:

$$\omega_{h+1}^k = \omega_h^k + \gamma \left[\frac{1}{B} \sum_{b=1}^{B} u_k(z_b^{kh}; \xi_k)\right], \tag{27}$$

where γ is the NPG learning rate. The update direction $u_k(z_b^{kh}; \xi_k)$ is given by

$$u_k(z_b^{kh}; \xi_k) = A_u(\theta_k, z_b^{kh})\omega_h^k - b_u(\theta_k, \xi_k, z_b^{kh}), \tag{28}$$

with

$$A_u(\theta_k, z_b^{kh}) = \nabla_\theta \log \pi_{\theta_k}(a_b^{kh} | s_b^{kh}) \nabla_\theta \log \pi_{\theta_k}(a_b^{kh} | s_b^{kh})^\top. \tag{29}$$

Set $\omega_k := \omega_H^k$. The policy is updated as: $\theta_{k+1} = \theta_k + \alpha\omega_k$, where α is the policy learning rate.

4 Regret Guarantee for Unichain MDP

4.1 Assumptions

Let $A_v(\theta) := \mathbb{E}_\theta [A_v(\theta, z)]$, and $b_v(\theta) := \mathbb{E}_\theta [b_v(\theta, z)]$ where $A_v(\theta, z)$, $b_v(\theta, z)$ are defined in (24) and (26) and the expectation $\mathbb{E}_\theta$ is computed over the distribution of $z = (s, a, s')$ where $(s, a) \sim \nu^{\pi_\theta}$, $s' \sim P(\cdot|s, a)$. For arbitrary policy parameter θ, we denote $\xi_\theta^* = [A_v(\theta)]^\dagger b_v(\theta) = [\eta_\theta^*, \zeta_\theta^*]^\top$. Using these notations, below we state some assumptions related to the critic analysis.

Definition 2. *We define the critic approximation error,* ϵ_{app} *as follows.*

$$\epsilon_{\text{app}} = \sup_\theta \inf_\zeta \left\{ \frac{1}{2} \mathbb{E}_{s \sim d^{\pi_\theta}} \left(V^{\pi_\theta}(s) - \hat{V}(\zeta, s) \right)^2 \right\}. \tag{30}$$

The definition 2 is widely adopted in the literature [32, 16, 40, 29, 12], although there may be minor variations in notation, such as the inclusion of a square root. This definition is closely linked to the representational capacity of the chosen feature map, with ϵ_{app} characterizing the resulting approximation error. A well-constructed feature map can yield a small or even vanishing ϵ_{app}.

Assumption 2. *Let* $M_\theta := \mathbb{E}_\theta \left[\phi(s) \left(\phi(s) - \phi(s')\right)^\top\right]$*, where the expectation* $\mathbb{E}_\theta$ *is taken over* $s \sim d^{\pi_\theta}$ *and* $s' \sim P^{\pi_\theta}(s, \cdot)$*. Then, for all* θ*, there exists a constant* $\lambda > 0$ *such that for all* $x \in \ker(M_\theta)^\perp$ *(i.e., the subspace orthogonal to the kernel of* M_θ*), the following inequality holds:*

$$x^\top M_\theta x \geq \lambda \|x\|^2. \tag{31}$$

Comments on Assumption 2: We emphasize that this condition is substantially weaker than the commonly imposed requirement of strict positive-definiteness of M_θ, which appears in nearly all Actor-Critic works [16, 29, 37, 28, 32]. In Section 4.2, we explain why the strict positive definiteness of the critic matrix cannot be guaranteed in the unichain case, as opposed to the ergodic case, and discuss the resulting analytical challenges.

We will now state some assumptions related to the policy parameterization. Before proceeding, we define the policy approximation error.

Definition 3. *Define ϵ_{bias} as the least upper bound on the transferred compatible function approximation error, $L_{\nu^{\pi^*}}(\omega^*_\theta;\theta)$, i.e.,*

$$\epsilon_{\text{bias}} \coloneqq \sup_\theta L_{\nu^{\pi^*}}(\omega^*_\theta;\theta)$$

where ω^*_θ denotes the exact Natural Policy Gradient direction at θ (8), π^* indicates the optimal policy, and L_ν is defined in (11).

The term ϵ_{bias} is a standard quantity in the literature on parameterized PG methods [1, 14, 26, 25, 32, 37, 16, 17, 40], and it reflects the expressivity of the chosen policy class. For example, when the policy class is expressive enough to represent any stochastic policy, such as with softmax parameterization, we have $\epsilon_{\text{bias}} = 0$ [2]. In contrast, under more restrictive parameterizations that do not cover all stochastic policies, we may have $\epsilon_{\text{bias}} > 0$. Nevertheless, this bias is often negligible when using rich neural network parameterizations [36].

Assumption 3. *For all $\theta, \theta_1, \theta_2$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$, the following statements hold.*

(a) $\|\nabla_\theta \log \pi_\theta(a|s)\| \le G_1$ $\quad(b)$ $\|\nabla_\theta \log \pi_{\theta_1}(a|s) - \nabla_\theta \log \pi_{\theta_2}(a|s)\| \le G_2\|\theta_1-\theta_2\|$

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.*

Comments on Assumptions 3-4: We emphasize that these are standard in the PG literature [24, 14, 26, 25, 32, 37, 16, 17]. Assumption 3 stipulates that the score function is both bounded and Lipschitz-continuous, an assumption frequently used in error decomposition and in bounding the policy gradient. Assumption 4 requires that the eigenvalues of the Fisher information matrix are uniformly bounded away from zero, a condition commonly employed to establish global complexity guarantees for PG methods. Notably, Assumptions 3-4 have recently been verified for a range of policy classes, including Gaussian and Cauchy distributions having parameterized means with clipping [24, 14].

4.2 Main Result

The following theorem gives the regret bound for the proposed algorithm.

Theorem 1 (Main Result)**.** *Consider Algorithm 1 and suppose Assumptions 1–4 hold. Let J be L smooth and set $K = \Theta(\sqrt{T}/(\log T))$, $B = \Theta(\sqrt{T})$ and $H = \Theta(\log T)$. Then, for a suitable choice of learning parameters, the expected regret satisfies*

$$\mathbb{E}[\text{Reg}_T] \le \tilde{\mathcal{O}}\left(T(\sqrt{\epsilon_{\text{bias}}} + \sqrt{\epsilon_{\text{app}}}) + \sqrt{T}(C^2 C_{\text{tar}} + C_{\text{tar}})\right), \tag{32}$$

where $C \coloneqq C_{\text{tar}} + C_{\text{hit}}$.

The choice of learning rates along with the regret bound containing all problem-specific constants is provided in Appendix H. We now provide a brief overview of the proof. We first begin with a regret decomposition obtained using standard techniques:

Lemma 1. *Consider the setting in Theorem 1. Then, the expected regret of Algorithm 1 satisfies*

$$\begin{aligned}\mathbb{E}[\text{Reg}_T] \le\, & T\sqrt{\epsilon_{\text{bias}}} + HBG_1\sum_{k=0}^{K-1}\mathbb{E}\left\|(\mathbb{E}_k[\omega_k] - \omega^*_k)\right\| + \alpha G_2 HB\sum_{k=0}^{K-1}\mathbb{E}\left\|\omega_k - \omega^*_k\right\|^2\\ &+ \frac{\alpha G_2 HB}{\mu^2}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla_\theta J(\theta_k)\right\|^2 + \frac{HB}{\alpha}\mathbb{E}_{s\sim d^{\pi^*}}[\text{KL}(\pi^*(\cdot|s)\|\pi_{\theta_0}(\cdot|s))] + 2C(K+1).\end{aligned} \tag{33}$$

The proof of the above is given in Appendix G. This regret decomposition crucially depends on the bias, $\mathbb{E}\left\|(\mathbb{E}_k[\omega_k] - \omega^*_k)\right\|$, and second-order error of the NPG estimate, $\mathbb{E}\left\|\omega_k - \omega^*_k\right\|^2$. These in turn dependent on the quality of the critic estimates. Hence the main focus of our proof consists of bounding these terms. The analysis of remaining terms then follows from these bounds; details are provided in Appendix H.

Properties of Markovian Sample Average: Analyses in the ergodic setting often rely on the property of exponentially fast mixing. In contrast, even irreducible but periodic Markov chains do not exhibit analogous exponential mixing behavior. A common strategy to address periodicity is to apply time-averaging, for which convergence results are available in the literature for irreducible, periodic chains [31]. In what follows, we extend such results to Markov chains with a single recurrent class.

Lemma 2. *Let Assumption 1 hold and consider any $\theta \in \Pi_\Theta$. Then, the following bound holds:*

$$\left\| \frac{1}{t} \sum_{i=1}^{t} (P^{\pi_\theta})^i(s_0, \cdot) - d^{\pi_\theta}(\cdot) \right\|_{\mathrm{TV}} \leq \frac{C}{t}, \quad \forall s_0 \in \mathcal{S}. \tag{34}$$

Using the result above, we show in the following lemma that employing a batch size of B reduces both the bias and variance of the estimate by a factor of $1/B$. Accordingly, in Algorithm 1, we choose a large batch size $B = \Theta(\sqrt{T})$ to obtain an order-optimal dependence on T in the regret bound.

Lemma 3. *Let Assumption 1 hold, and let $f : \mathcal{S} \to \mathbb{R}^d$ satisfy $\|f(s)\| \leq C_f$ for all $s \in \mathcal{S}$, for some constant $C_f > 0$. Then, the following bounds hold:*

$$(i) \ \left\| \mathbb{E}\left[\frac{1}{B} \sum_{i=1}^{B} f(s_i) \right] - \mu \right\| \leq \frac{\sqrt{d} C_f C}{B} \quad (ii) \ \mathbb{E} \left\| \left[\frac{1}{B} \sum_{i=1}^{B} f(s_i) \right] - \mu \right\|^2 \leq \frac{C_f^2 + 2\sqrt{d} C_f^2 C}{B}$$

where $\mathbb{E}$ denotes the expectation of the Markov chain $\{s_i\}$ induced by π_θ starting from any $s_0 \in \mathcal{S}$ and $\mu = \mathbb{E}_{s \sim d^{\pi_\theta}}[f(s)]$.

Another consequence of Lemma 2 is in providing bounds for the value and Q functions. In the ergodic case, it is known that the functions V^π and Q^π are $\mathcal{O}(t_{\mathrm{mix}})$. In the unichain case, using the above result, we show that these functions are still bounded, by $\mathcal{O}(C)$ instead. The proof of these results are given in Appendix C. We next describe the error analysis of the critic and NPG subroutines in Algorithm 1.

Challenges due to transient states: We observe that the critic and NPG updates can be interpreted as stochastic linear updates with Markovian noise. From the preceding lemmas, it follows that the bias and variance of the noise in these averaged update directions can be made sufficiently small. Nevertheless, the analysis still presents certain challenges.

We show that the kernel of the critic matrix $A_v(\theta)$ can change depending on the policy. More specifically, the kernel will depend on the set of transient states (see Lemma 13). This is unlike in the ergodic case, where the kernel remains the same for all policies. As a result, strict positive definiteness of the critic matrix cannot, in general, be ensured by using a carefully selected set of feature vectors.

Although a prior work in the ergodic case has considered the case where the critic matrix is singular [44], our approach differs significantly, in addition to incorporating the sample averaging analysis discussed in the previous section. In particular, our regret guarantees require a sharp bias bound for the critic updates, $\|\mathbb{E}[\xi_h] - \xi^*\|$, in addition to the standard second-order error bound $\mathbb{E}[\|\xi_h - \xi^*\|^2]$. The analysis in [44] only addresses the second-order error. Furthermore, to establish our convergence guarantees, we require the following condition:

$$\ker(A_v(\theta)) \subseteq \ker(A_v(z, \theta)). \tag{35}$$

Equation (35) holds in the ergodic case, even if the critic matrix is singular, but generally does not hold in unichain settings (see Appendix E). However, it is satisfied when the Markov chain is restricted to its recurrent class. To address this, we show that, with high probability, the chain induced by π_{θ_k} enters the recurrent class, $\mathcal{S}_R^{\theta_k}$, very early, allowing us to focus on the recurrent class.

More specifically, recall that in each outer iteration k, Algorithm 1 collects a contiguous segment of $2BH$ samples under π_{θ_k}. We define the event $\mathcal{E}_B^k := \{T_{\theta_k} \leq B\}$, where T_{θ_k} is the first hitting time of $\mathcal{S}_R^{\theta_k}$ for this segment. Lemma 10 shows that $\Pr((\mathcal{E}_B^k)^c) \leq 2^{-\lfloor B/2C_{\mathrm{hit}} \rfloor}$, which implies that the chain reaches the recurrent class within the first B steps with overwhelmingly high probability. We perform our analysis conditioned on this event.

Let $\mathcal{E}_B := \bigcap_{k=1}^{K} \mathcal{E}_B^k$. We decompose the expected regret based on whether $\mathcal{E}_B$ holds:

$$\mathbb{E}[\mathrm{Reg}_T] = \mathbb{E}[\mathrm{Reg}_T | \mathcal{E}_B] \Pr(\mathcal{E}_B) + \mathbb{E}[\mathrm{Reg}_T | \mathcal{E}_B^c] \Pr(\mathcal{E}_B^c).$$

Since the instantaneous regret is at most 1, the total regret is bounded above by T, and applying a union bound over K yields

$$\mathbb{E}[\mathrm{Reg}_T] \leq \mathbb{E}[\mathrm{Reg}_T | \mathcal{E}_B] + TK 2^{-\lfloor B/2C_{\mathrm{hit}} \rfloor}.$$

This decomposition allows us to isolate the high-probability event $\mathcal{E}_B$, under which all policies quickly enter their recurrent classes. By choosing $B = \sqrt{T}$, the second term becomes negligible for large T, and we may therefore restrict our analysis to the case where $\mathcal{E}_B$ holds.

Analysis under event $\mathcal{E}_B$: By standard Markov chain theory, T_θ is a stopping time hence, the Strong Markov Property (Theorem 1.2.5 of [27]) applies at T_θ. In particular, conditioned on s_T, the post-T_θ trajectory is independent of the pre-T_θ path and evolves as a Markov chain initialized at s_T under the original kernel P^{π_θ}. As a result, it suffices to analyze the critic and NPG subroutines with the samples restricted to the irreducible class. In Appendix D, we analyze a generic stochastic linear recursion and show that these subroutines can be viewed as special cases of this recursion.

Theorem 2. *Consider Algorithm 1 under the assumptions of Theorem 1. Conditioned on the event* $\mathcal{E}_B$*, the critic estimate* ξ_k *satisfies*

$$\mathbb{E}\left[\|\Pi(\xi_k - \xi_k^*)\|^2 | \mathcal{E}_B\right] \leq \tilde{\mathcal{O}}\left(e^{-c_1 H}\|\xi_0 - \xi^*\|^2 + \frac{C_{\text{tar}}\sqrt{m}}{\lambda^4 B}\right)$$

and

$$\|\Pi(\mathbb{E}[\xi_k|\mathcal{E}_B] - \xi_k^*)\|^2 \leq \tilde{\mathcal{O}}\left(e^{-c_1 H}\|\xi_0 - \xi^*\|^2 + \frac{C_{\text{tar}}^2 m}{\lambda^6 B^2}\right)$$

where $c_1 = \frac{\lambda^3}{16}$, $\xi_k^* = [A_v(\theta_k)]^\dagger b_v(\theta_k) = [\eta_k^*, \zeta_k^*]^\top$ *and* Π *denotes the projection onto the space* $\ker(A_v(\theta))^\perp$.

The details regarding the verification of all conditions of the generic linear recursion used in Theorem 4 for the critic update is provided in Appendix E.

Theorem 3. *Consider Algorithm 1 under the assumptions of Theorem 1. Conditioned on the event* $\mathcal{E}_B$*, the NPG estimate* ω_k *satisfies*

$$\mathbb{E}\left[\|\omega_k - \omega_k^*\|^2 | \mathcal{E}_B\right] \leq \mathcal{O}\left(e^{-c_2 H}\|\omega_0 - \omega_k^*\|^2 + \frac{\sqrt{d}G_1^4 C_{\text{tar}} C^2}{\mu^2 B} + \frac{G_1^2\,\mathbb{E}[\|\Pi(\xi_k - \xi_k^*)\|^2|\mathcal{E}_B]}{\mu^2} + \frac{G_1^2 \epsilon_{\text{app}}}{\mu^2}\right)$$

and

$$\|\,\mathbb{E}[\omega_k|\mathcal{E}_B] - \omega_k^*\|^2 \leq \mathcal{O}\left(e^{-c_2 H}\|\omega_0 - \omega_k^*\|^2 + \frac{dG_1^8 C_{\text{tar}}^2}{\mu^6 B^2} + \frac{G_1^2\|\Pi(\mathbb{E}[\xi_k|\mathcal{E}_B] - \xi_k^*)\|^2}{\mu^2} + \frac{G_1^2 \epsilon_{\text{app}}}{\mu^2}\right)$$

where $c_2 = \frac{\mu^2}{4(1+4C)G_1}$.

Similar to the critic update, the details regarding the verification of all conditions of the generic linear recursion used in Theorem 4 for the NPG update is provided in Appendix F.

5 Conclusion

We have presented NAC-B, a Natural Actor-Critic algorithm with batching, that achieves $\tilde{O}(\sqrt{T})$ regret in infinite-horizon average-reward MDPs under the unichain assumption. Unlike prior work that relies on stronger ergodicity conditions, NAC-B operates under a weaker structural assumption that allows for both transient states and periodic behavior, thereby broadening the applicability of policy gradient methods in average-reward settings. Our approach leverages function approximation for scalability and introduces batching to mitigate issues arising from periodicity and high variance in gradient estimates. Future works include extending this analysis to other settings such as constrained MDPs [8, 42], neural critics [15] and relaxing the unichain assumption, though we anticipate this to be a challenging task (See Remark 3).

References

[1] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, pages 64–66, 2020.

[2] 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.

[3] P. Agrawal and S. Agrawal. Optimistic q-learning for average reward and episodic reinforcement learning. *arXiv preprint arXiv:2407.13743*, 2024.

[4] S. Agrawal and R. Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. *Advances in Neural Information Processing Systems*, 30, 2017.

[5] A. O. Al-Abbasi, A. Ghosh, and V. Aggarwal. Deeppool: Distributed model-free algorithm for ride-sharing using deep reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems*, 20(12):4714–4727, 2019.

[6] D. Aldous and J. A. Fill. *Reversible Markov Chains and Random Walks on Graphs*. 2002. URL `https://www.stat.berkeley.edu/~aldous/RWG/book.html`.

[7] P. Auer, T. Jaksch, and R. Ortner. Near-optimal regret bounds for reinforcement learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21, 2008.

[8] Q. Bai, W. Mondal, and V. Aggarwal. Learning general parameterized policies for infinite horizon average reward constrained mdps via primal-dual policy gradient algorithm. *Advances in Neural Information Processing Systems*, 37:108566–108599, 2024.

[9] Q. Bai, W. U. Mondal, and V. Aggarwal. Regret analysis of policy gradient algorithm for infinite horizon average reward markov decision processes. In *AAAI Conference on Artificial Intelligence*, 2024.

[10] R. Bhatia. *Matrix Analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997.

[11] C.-L. Chen, H. Zhou, J. Chen, M. Pedramfar, T. Lan, Z. Zhu, C. Zhou, P. Mauri Ruiz, N. Kumar, H. Dong, and V. Aggarwal. Learning-based two-tiered online optimization of region-wide datacenter resource allocation. *IEEE Transactions on Network and Service Management*, 22(1): 572–581, 2025.

[12] X. Chen and L. Zhao. Finite-time analysis of single-timescale actor-critic. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[13] Y. Ding, J. Zhang, and J. Lavaei. On the global optimum convergence of momentum-based policy gradient. In *International Conference on Artificial Intelligence and Statistics*, pages 1910–1934, 2022.

[14] 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, 2023.

[15] S. Ganesh, J. Chen, W. U. Mondal, and V. Aggarwal. Order-optimal global convergence for actor-critic with general policy and neural critic parametrization. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025.

[16] S. Ganesh, W. U. Mondal, and V. Aggarwal. Order-optimal global convergence for average reward reinforcement learning via actor-critic approach. In *The 42nd International Conference on Machine Learning (ICML)*, 2025.

[17] S. Ganesh, W. U. Mondal, and V. Aggarwal. Order-optimal regret with novel policy gradient approaches in infinite horizon average reward MDPs. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

[18] 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.

[19] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, 4 edition, 2013.

[20] K. Hong and A. Tewari. A computationally efficient algorithm for infinite-horizon average-reward linear MDPs. In *Forty-second International Conference on Machine Learning*, 2025.

[21] T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *The Journal of Machine Learning Research*, 11:1563–1600, 2010.

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

[23] T. Li, F. Wu, and G. Lan. Stochastic first-order methods for average-reward markov decision processes. *Mathematics of Operations Research*, 0(0), 2024. doi: 10.1287/moor.2023.1427. To appear.

[24] 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.

[25] S. Masiha, S. Salehkaleybar, N. He, N. Kiyavash, and P. Thiran. Stochastic second-order methods provably beat sgd for gradient-dominated functions. *arXiv preprint arXiv:2205.12856v1*, 2022.

[26] 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*, pages 3097–3105, 2024.

[27] J. Norris. *Markov Chains*. Cambridge University Press, 1998.

[28] P. Panda and S. Bhatnagar. Two-timescale critic-actor for average reward mdps with function approximation. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

[29] B. Patel, W. A. Suttle, A. Koppel, V. Aggarwal, B. M. Sadler, A. S. Bedi, and D. Manocha. Global optimality without mixing time oracles in average-reward rl via multi-level actor-critic. In *International Conference on Machine Learning*, 2024.

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

[31] G. O. Roberts and J. S. Rosenthal. Shift-coupling and convergence rates of ergodic averages. *Communications in Statistics. Stochastic Models*, 13(1):147–165, 1997. doi: 10.1080/15326349708807418.

[32] W. A. Suttle, A. Bedi, B. Patel, B. M. Sadler, A. Koppel, and D. Manocha. Beyond exponentially fast mixing in average-reward reinforcement learning via multi-level monte carlo actor-critic. In *International Conference on Machine Learning*, pages 33240–33267, 2023.

[33] 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.

[34] J. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997. doi: 10.1109/9.580874.

[35] J. Wang, M. Wang, and L. F. Yang. Near sample-optimal reduction-based policy learning for average reward mdp. *arXiv preprint arXiv:2212.00603*, 2022.

[36] 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.

[37] Y. Wang, Y. Wang, Y. Zhou, and S. Zou. Non-asymptotic analysis for single-loop (natural) actor-critic with compatible function approximation. In *International Conference on Machine Learning*, 2024.

[38] 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, 2020.

[39] C.-Y. Wei, M. J. Jahromi, H. Luo, and R. Jain. Learning infinite-horizon average-reward mdps with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3007–3015, 2021.

[40] P. Xu, F. Gao, and Q. Gu. Sample efficient policy gradient methods with recursive variance reduction. In *International Conference on Learning Representations*, 2019.

[41] P. Xu, F. Gao, and Q. Gu. An improved convergence analysis of stochastic variance-reduced policy gradient. In *Uncertainty in Artificial Intelligence*, pages 541–551, 2020.

[42] Y. Xu, S. Ganesh, W. U. Mondal, Q. Bai, and V. Aggarwal. Global convergence for average reward constrained mdps with primal-dual actor critic algorithm. *Advances in Neural Information Processing Systems*, 2025.

[43] R. Yuan, R. M. Gower, and A. Lazaric. A general sample complexity analysis of vanilla policy gradient. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 3332–3380, 2022.

[44] S. Zhang, Z. Zhang, and S. T. Maguluri. Finite sample analysis of average-reward td learning and q-learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 1230–1242, 2021.

[45] Z. Zhang and Q. Xie. Sharper model-free reinforcement learning for average-reward markov decision processes. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195, pages 5476–5477, 12–15 Jul 2023.

[46] M. Zurek and Y. Chen. Span-based optimal sample complexity for weakly communicating and general average reward MDPs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims are demonstrated in the key results (mainly in Section 4.2), which match the description in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The assumptions given in the paper give the limitations of this work. Further, future work direction in the conclusions describe another limitation of this work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We have provided the assumptions used in the work in Section 4.1 or referenced in the theorem statements. All claims are backed with complete proofs in the Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: The experimental results are not provided, since the focus of the paper is on theoretical sample complexity analysis.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: The paper does not include experiments requiring code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: Our work is theoretical, and hence there are no experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All the points mentioned in the NeurIPS Code of Ethics are taken into consideration.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There are no specific societal impacts of this work beyond that the algorithms may have impact in the reinforcement learning algorithms in different applications.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: There is no release of data or models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: There is no code, data, models, etc. involved.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: There is no code, data, models, etc. involved in the paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does neither involve crowd-sourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: We do not use LLM in our paper.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Discussion on the Discounted to Average Method

We note that while there are several sample complexity analyses for discounted policy gradient methods, most treat the term $(1-\gamma)^{-1}$ as a constant. As a result, most results exhibit an overly pessimistic dependence on this factor, since it is not the main focus of these works. However, achieving an optimal dependence on $(1-\gamma)^{-1}$ is essential to obtain optimal regret or convergence guarantees when using the discounted-to-average reduction. This reduction is particularly sensitive to the $(1-\gamma)^{-1}$ factor.

We now compare the sample complexity, a closely related notion to regret, of existing policy gradient algorithms when applied to the average-reward MDP (AMDP) setting. We focus on sample complexity since it aligns with the typical form of guarantees in the discounted setting and facilitates direct comparison. To translate results from the discounted MDP (DMDP) to the AMDP setting, we use Theorem 1 from [35], which was also employed in [46] to establish optimal sample complexity for average-reward MDPs under a generative model. This reduction bound shows that if the DMDP with discount factor γ is solved to an accuracy of $\epsilon_\gamma = \mathcal{O}((1-\gamma)^{-1}\epsilon)$, the resulting policy will be of $\mathcal{O}(\epsilon)$ accuracy in the original AMDP.

Table 2 summarizes the sample complexity bounds for discounted policy gradient methods. The best guarantee among these is $\mathcal{O}(\epsilon^{-5})$, which remains significantly worse than the optimal $\mathcal{O}(\epsilon^{-2})$ rate. This is akin to a regret bound of order $\mathcal{O}(T^{4/5})$. Although some actor-critic methods, such as that of [41], offer improved sample complexity guarantees in terms of $(1-\gamma)^{-1}$, they assume ergodicity even in the discounted setting and are therefore not considered.

Table 2: Overview of global optimality convergence for discounted PG algorithms. DMDP Sample Complexity: Number of samples to achieve $J^*_\gamma - \mathbb{E}\, J_\gamma(\theta_T) \leq \frac{\varepsilon+\sqrt{\varepsilon_{\text{bias}}}}{1-\gamma}$. AMDP Sample Complexity: Number of samples to achieve $J^* - \mathbb{E}\, J(\theta_T) \leq \varepsilon + \sqrt{\varepsilon_{\text{bias}}}$.

Algorithm	DMDP Sample Complexity	AMDP Sample Complexity
Vanilla-PG [43]	$(1-\gamma)^{-2}\varepsilon^{-3}$	ε^{-5}
Natural-PG [24]	$(1-\gamma)^{-4}\varepsilon^{-3}$	ε^{-7}
SRVR-PG [24]	$(1-\gamma)^{-2.5}\varepsilon^{-3}$	$\varepsilon^{-5.5}$
STORM-PG [13]	$(1-\gamma)^{-9}\varepsilon^{-3}$	ε^{-12}
SCRN [25]	$(1-\gamma)^{-5}\varepsilon^{-2.5}$	$\varepsilon^{-7.5}$
N-PG-IGT [14]	$(1-\gamma)^{-5}\varepsilon^{-2.5}$	$\varepsilon^{-7.5}$
(N)-HARPG [14]	$(1-\gamma)^{-4}\varepsilon^{-2}$	ε^{-6}
ANPG [26]	$(1-\gamma)^{-4}\varepsilon^{-2}$	ε^{-6}

B Policy Gradient Theorem

Lemma 4 (Policy Gradient theorem, [33]). *Let Assumption 1 hold and assume a differentiable policy parameterization. Then, the following holds for all parameters $\theta \in \Theta$.*

$$\nabla J(\theta) = \mathbb{E}_{(s,a)\sim\eta^{\pi_\theta}}\left[Q^{\pi_\theta}(s,a)\nabla \log \pi_\theta(a|s)\right] \tag{36}$$

Proof of Lemma 4. We begin by noting that

$$V^{\pi_\theta}(s) = \sum_a \pi_\theta(a|s) Q^{\pi_\theta}(s,a). \tag{37}$$

By differentiating both sides of the above equation with respect to θ, we obtain

$$\nabla V^{\pi_\theta}(s) = \sum_a \nabla\pi_\theta(a|s)Q^{\pi_\theta}(s,a) + \pi_\theta(a|s)\nabla Q^{\pi_\theta}(s,a) \tag{38}$$

Recall that

$$Q^{\pi_\theta}(s,a) = r(s,a) + \sum_{s'} P(s'|s,a)V^{\pi_\theta}(s') - J(\theta) \tag{39}$$

In a similar manner to (38), taking the derivative of both sides of the above equation with respect to θ gives

$$\nabla Q^{\pi_\theta}(s,a) = \sum_{s'} P(s'|s,a)\nabla V^{\pi_\theta}(s') - \nabla J(\theta). \tag{40}$$

Substituting (40) in (38) yields

$$\nabla V^{\pi_\theta}(s) = \sum_a \left[\nabla \pi_\theta(a|s) Q^{\pi_\theta}(s,a) + \pi_\theta(a|s)\left(\sum_{s'} P(s'|s,a)\nabla V^{\pi_\theta}(s') - \nabla J(\theta)\right)\right] \tag{41}$$

Re-arranging the above equation, we arrive at the following expression for $\nabla J(\theta)$

$$\nabla J(\theta) = \sum_a \left[\nabla \pi_\theta(a|s) Q^{\pi_\theta}(s,a) + \sum_{s'} \pi_\theta(a|s) P(s'|s,a)\nabla V^{\pi_\theta}(s')\right] - \nabla V^{\pi_\theta}(s) \tag{42}$$

Taking expectation over $s \sim d^{\pi_\theta}$ on both sides, we obtain

$$\begin{aligned}\nabla J(\theta) = &\sum_{s,a} \left[d^{\pi_\theta}(s)\nabla \pi_\theta(a|s) Q^{\pi_\theta}(s,a) + \sum_{s'} d^{\pi_\theta}(s)\pi_\theta(a|s) P(s'|s,a)\nabla V^{\pi_\theta}(s')\right] \\ &- \sum_s d^{\pi_\theta}(s)\nabla V^{\pi_\theta}(s)\end{aligned} \tag{43}$$

Since $\sum_s d^{\pi_\theta}(s) P^{\pi_\theta}(s,s') = d^{\pi_\theta}(s')$, we arrive at the following expression.

$$\nabla J(\theta) = \sum_{s,a} [d^{\pi_\theta}(s)\nabla \pi_\theta(a|s) Q^{\pi_\theta}(s,a)] \tag{44}$$

□

C General Properties of Unichain Markov Decision Processes

In this section, we provide detailed proofs of Lemmas 2 and 3 from the main paper, along with bounds on the value functions (Lemma 7) and show that the hitting time of the recurrent class remains bounded by B with a high-probability (Lemma 10). We begin with a few useful preliminary results.

Lemma 5 (Strong Markov Property, Theorem 1.2.5 of [27])**.** *Let $\{s_t\}_{t\geq 0}$ be a discrete-time Markov chain on a countable state space with transition kernel P. If T is a (possibly infinite) stopping time, then for all $s \in \mathcal{S}$ and any bounded measurable function f,*

$$\mathbb{E}\left[f(s_{T+n}) \mid \mathcal{F}_T\right] = \mathbb{E}\left[f(s_n) \mid s_T\right] \quad \textit{a.s. on } \{T < \infty\}, \quad \textit{for all } n \geq 0.$$

The following intermediate result provides a bound on the convergence rate of the empirical state distribution to the stationary distribution when restricted to the recurrent class.

Lemma 6. *Let Assumption 1 hold and consider any $\theta \in \Theta$. Then, the following bound holds:*

$$\left\| \frac{1}{t}\sum_{i=1}^{t} (P^{\pi_\theta})^i(s_0,\cdot) - d^{\pi_\theta}(\cdot) \right\|_{\mathrm{TV}} \leq \frac{C_{\mathrm{tar}}}{t}, \quad \forall s_0 \in \mathcal{S}_R^\theta. \tag{45}$$

Proof. Since $\mathcal{S}_R^\theta$ is closed and communicating, the Markov chain restricted to $\mathcal{S}_R^\theta$ remains a Markov chain and is, in fact, irreducible. The result then follows from Corollary 3 of [31], which applies to (possibly periodic) irreducible Markov chains. □

C.1 Proof of Lemma 2

Note that

$$\begin{aligned}\left\| \frac{1}{t}\sum_{i=1}^{t} (P^{\pi_\theta})^i(s_0,\cdot) - d^{\pi_\theta}(\cdot) \right\|_{\mathrm{TV}} &= \frac{1}{2}\left\| \frac{1}{t}\sum_{i=1}^{t} (P^{\pi_\theta})^i(s_0,\cdot) - d^{\pi_\theta}(\cdot) \right\|_1 \\ &= \frac{1}{2t}\left\| \sum_{i=1}^{t} [(P^{\pi_\theta})^i(s_0,\cdot) - d^{\pi_\theta}(\cdot)] \right\|_1\end{aligned} \tag{46}$$

Thus, it suffices to show that the quantity

$$\left\|\sum_{i=1}^{t}[(P^{\pi_\theta})^i(s_0,\cdot)-d^{\pi_\theta}(\cdot)]\right\|_1$$

is bounded by C, for all $t \geq 1$. Note that this term can be broken down in the following manner.

$$\begin{aligned}
&\left\|\sum_{i=1}^{t}(P^{\pi_\theta})^i(s_0,\cdot)-d^{\pi_\theta}(\cdot)\right\|_1\\
&=\sum_{s\in\mathcal{S}}\left|\sum_{i=1}^{t}(P^{\pi_\theta})^i(s_0,s)-d^{\pi_\theta}(s)\right|\\
&\overset{(a)}{=}\underbrace{\sum_{s\in\mathcal{S}_R^\theta}\left|\sum_{i=1}^{t}(P^{\pi_\theta})^i(s_0,s)-d^{\pi_\theta}(s)\right|}_{T_1}+\underbrace{\sum_{s\in(\mathcal{S}_R^\theta)^c}\sum_{i=1}^{t}(P^{\pi_\theta})^i(s_0,s)}_{T_2},
\end{aligned}\tag{47}$$

where (a) follows since $d^{\pi_\theta}(s)=0$ for all $s\in(\mathcal{S}_R^\theta)^c$. Let $\mathbb{E}_{s_0}^\theta$ denote the expectation over the Markov chain generated from π_θ starting from state s_0. Consider term T_2 and observe that

$$\begin{aligned}
\sum_{s\in(\mathcal{S}_R^\theta)^c}\sum_{i=1}^{t}(P^{\pi_\theta})^i(s_0,s)&=\mathbb{E}_{s_0}^\theta\left[\sum_{i=1}^{t}\sum_{s\in(\mathcal{S}_R^\theta)^c}\mathbf{1}(s_i=s)\right]=\mathbb{E}_{s_0}^\theta\left[\sum_{i=1}^{t}\mathbf{1}(s_i\in(\mathcal{S}_R^\theta)^c)\right]\\
&=\mathbb{E}_{s_0}^\theta[\min\{t,T_\theta-1\}]\leq\mathbb{E}_{s_0}^\theta[T_\theta]\leq C_{\text{hit}}.
\end{aligned}\tag{48}$$

We adopt the convention that the sum $\sum_{i=j}^{k}s_i$ is defined to be zero whenever $j>k$, regardless of the values of the summands s_i. The term T_2 can be bounded as follows.

$$\begin{aligned}
&\sum_{s\in\mathcal{S}_R^\theta}\left|\sum_{i=1}^{t}[(P^{\pi_\theta})^i(s_0,s)-d^{\pi_\theta}(s)]\right|=\sum_{s\in\mathcal{S}_R^\theta}\left|\mathbb{E}_{s_0}^\theta\left[\sum_{i=1}^{t}[\mathbf{1}(s_i=s)-d^{\pi_\theta}(s)]\right]\right|\\
&\leq\mathbb{E}_\theta\left[\sum_{s\in\mathcal{S}_R^\theta}\left|\mathbb{E}_{T_\theta}\left[\sum_{i=1}^{t}[\mathbf{1}(s_i=s)-d^{\pi_\theta}(s)]\right]\right|\right]\\
&\leq\mathbb{E}_\theta\left[\underbrace{\sum_{s\in\mathcal{S}_R^\theta}\left|\mathbb{E}_{T_\theta}\left[\sum_{i=1}^{T_\theta-1}[\mathbf{1}(s_i=s)-d^{\pi_\theta}(s)]\right]\right|}_{T_3}+\underbrace{\sum_{s\in\mathcal{S}_R^\theta}\mathbb{E}_{T_\theta}\left[\sum_{i=T_\theta}^{t}[\mathbf{1}(s_i=s)-d^{\pi_\theta}(s)]\right]}_{T_4}\right],
\end{aligned}$$

where $\mathbb{E}_{T_\theta}$ denotes the conditional expectation given the hitting time T_θ, and $\mathbb{E}_\theta$ denotes the expectation over the random variable T_θ. Finally, consider term T_3

$$\mathbb{E}_\theta\left[\sum_{s\in\mathcal{S}_R^\theta}\left|\mathbb{E}_{T_\theta}\left[\sum_{i=1}^{T_\theta-1}[\mathbf{1}(s_i=s)-d^{\pi_\theta}(s)]\right]\right|\right]\overset{(a)}{=}\mathbb{E}_\theta\left[\sum_{s\in\mathcal{S}_R^\theta}\mathbb{E}_{T_\theta}\left[\sum_{i=1}^{T_\theta-1}d^{\pi_\theta}(s)\right]\right]\leq\mathbb{E}_\theta[T_\theta]\leq C_{\text{hit}},$$

where (a) follows since, by definition, no state $s\in\mathcal{S}_R^\theta$ is visited before T_θ. T_4 can be bounded as

$$\begin{aligned}
\mathbb{E}_\theta\left[\sum_{s\in\mathcal{S}_R^\theta}\left|\mathbb{E}_{T_\theta}\left[\sum_{i=T_\theta}^{t}[\mathbf{1}(s_i=s)-d^{\pi_\theta}(s)]\right]\right|\right]&\overset{(a)}{=}\mathbb{E}_\theta\left[\left\|\sum_{i=1}^{t-T_\theta+1}[(P^{\pi_\theta})^i(s_{T_\theta+1},s)-d^{\pi_\theta}(s)]\right\|_1\right]\\
&\overset{(b)}{\leq}2C_{\text{tar}},
\end{aligned}$$

where (a) follows from the Strong Markov property and (b) follows from the fact that $s_{T_\theta}\in\mathcal{S}_R^\theta$ and using Lemma 6. Combining the bounds on T_1, T_2, T_3, and T_4 gives

$$\left\|\sum_{i=1}^{t}[(P^{\pi_\theta})^i(s_0,\cdot)-d^{\pi_\theta}(\cdot)]\right\|_1\leq 2(C_{\text{tar}}+C_{\text{hit}}),\tag{49}$$

which yields the desired result using (46).

C.2 Value function bounds

Recall that the value function V^π and the action-value function Q^π are not uniquely defined for average-reward infinite-horizon MDPs, even in the ergodic case. We therefore consider V^{π_θ} such that

$$\sum_{s\in\mathcal{S}} d^{\pi_\theta}(s)V^{\pi_\theta}(s) = 0. \tag{50}$$

Under this condition, $V^{\pi_\theta}(s)$ and $Q^{\pi_\theta}(s,a)$ are given by

$$\begin{aligned} V^{\pi_\theta}(s) &= \lim_{N\to\infty} \frac{1}{N}\sum_{T=1}^{N} \mathbb{E}_\theta\left[\sum_{t=0}^{T}(r(s_t,a_t) - J(\theta))\Big| s_0 = s\right] \\ Q^{\pi_\theta}(s,a) &= \lim_{N\to\infty} \frac{1}{N}\sum_{T=1}^{N} \mathbb{E}_\theta\left[\sum_{t=0}^{T}(r(s_t,a_t) - J(\theta))\Big| s_0 = s, a_0 = a\right] \end{aligned} \tag{51}$$

where $\mathbb{E}_\theta[\cdot]$ denotes expectation with respect to trajectories generated by the policy π_θ. These definitions employ the Cesàro limit to account for potential periodicity, which was absent in the ergodic case.

An equivalent expression for $V^{\pi_\theta}(s)$ is

$$V^{\pi_\theta}(s) = \lim_{N\to\infty} \frac{1}{N}\sum_{T=1}^{N}\left[\sum_{t=0}^{T}\left[(P^{\pi_\theta})^t(s,\cdot) - d^{\pi_\theta}\right]^\top r^{\pi_\theta}\right]. \tag{52}$$

We now provide bounds for V^π, Q^π, and the advantage function $A^\pi = Q^\pi - V^\pi$ under the condition in (50).

Lemma 7. *For all $\theta \in \Theta$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$, the following bounds hold:*

(i) $V^{\pi_\theta}(s) \leq 2(C_{\text{hit}} + C_{\text{tar}})$

(ii) $Q^{\pi_\theta}(s,a) \leq (1 + 2(C_{\text{hit}} + C_{\text{tar}}))$

(iii) $A^{\pi_\theta}(s,a) \leq (1 + 4(C_{\text{hit}} + C_{\text{tar}}))$

Proof. Using Hölder's inequality and the bound on the reward function, we obtain the following

$$\left(\sum_{t=0}^{T}[(P^{\pi_\theta})^t(s,\cdot) - d^{\pi_\theta}]\right)^\top r^\pi \leq \left\|\sum_{t=0}^{T}[(P^{\pi_\theta})^t(s,\cdot) - d^{\pi_\theta}]\right\|_1 \|r^\pi\|_\infty \leq \left\|\sum_{t=0}^{T}[(P^{\pi_\theta})^t(s,\cdot) - d^{\pi_\theta}]\right\|_1.$$

Combining the above bound with (49) concludes (i). Parts (ii) and (iii) now follow easily, since both $r(s,a)$ and, consequently, J are confined to the interval $[0,1]$, which gives

$$Q^{\pi_\theta}(s,a) = r(s,a) - J^{\pi_\theta} + \mathbb{E}_{s'\sim P(\cdot|s,a)}[V^{\pi_\theta}(s')] \leq (1 + 2(C_{\text{hit}} + C_{\text{tar}})) \tag{53}$$

and

$$A^{\pi_\theta}(s,a) = Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s) \leq (1 + 4(C_{\text{hit}} + C_{\text{tar}})). \tag{54}$$

□

C.3 Proof of Lemma 3

For clarity and ease of reference, we re-state and prove parts (i) and (ii) of Lemma 3 separately. We also provide a sharper bound for the case where the Markov chain is restricted to the irreducible class.

Lemma 8 (Bias of Markovian Sample Average)**.** *Let Assumption 1 hold, and let $f : \mathcal{S} \to \mathbb{R}^d$ satisfy $\|f(s)\| \leq C_f$ for all $s \in \mathcal{S}$, for some constant $C_f > 0$. Then, the following bound holds:*

$$\left\|\mathbb{E}\left[\frac{1}{B}\sum_{i=1}^{B} f(s_i)\right] - \mu\right\| \leq \frac{\sqrt{d}C_f C}{B}$$

where $\mathbb{E}$ denotes the expectation of the Markov chain $\{s_i\}$ induced by π_θ starting from any $s_0 \in \mathcal{S}$ and $\mu = \mathbb{E}_{s\sim d^{\pi_\theta}}[f(s)]$. Furthermore, if $s_1 \in \mathcal{S}_R^\theta$, then

$$\left\|\mathbb{E}\left[\frac{1}{B}\sum_{i=1}^{B} f(s_i)\right] - \mu\right\| \leq \frac{\sqrt{d}C_f C_{\text{tar}}}{B}.$$

Proof. Let $f_j : \mathcal{S} \to \mathbb{R}$ denote the j^{th} co-ordinate of f, i.e., $f(s) = [f_1(s), f_2(s), \cdots, f_d(s)]^\top$. We express the j^{th} co-ordinate of $\mathbb{E}\left[\frac{1}{B}\sum_{i=1}^{B} f(s_i)\right]$, $\mathbb{E}\left[\frac{1}{B}\sum_{i=1}^{B} f_j(s_i)\right]$, as

$$\mathbb{E}\left[\frac{1}{B}\sum_{i=1}^{B} f_j(s_i)\right] = p^\top\left[\frac{1}{B}\sum_{i=1}^{B} P^i\right] f_j, \tag{55}$$

Similarly, μ_j can be written as

$$\mu_j = \mathbb{E}_{s\sim d}[f_j(s)] = (d^{\pi_\theta})^\top f_j. \tag{56}$$

Then, by using Hölder's inequality for each j

$$\left|\mathbb{E}_{s_1\sim p}\left[\frac{1}{B}\sum_{i=1}^{B} f_j(s_i)\right] - \mu_j\right|^2 = \left\|p^\top\left[\frac{1}{B}\sum_{i=1}^{B} P^i\right] - (d^{\pi_\theta})^\top\right\|_1^2 \|f_j\|_\infty^2 \overset{(a)}{\leq} \frac{C_f^2 C^2}{B^2}, \tag{57}$$

where (a) follows from Lemma 2. If $s_1 \in \mathcal{S}_R^\theta$, then Lemma 6 can be applied instead, replacing C with C_{tar}. The result then follows by noting that

$$\left\|\mathbb{E}_{s_1\sim p}\left[\frac{1}{B}\sum_{i=1}^{B} f(s_i)\right] - \mu\right\| = \sqrt{\sum_{j=1}^{d}\left|\mathbb{E}_{s_1\sim p}\left[\frac{1}{B}\sum_{i=1}^{B} f_j(s_i)\right] - \mu_j\right|^2}. \tag{58}$$

□

Lemma 9 (Variance of Markovian Sample Average)**.** *Let Assumption 1 hold, and let $f : \mathcal{S} \to \mathbb{R}^d$ satisfy $\|f(s)\| \leq C_f$ for all $s \in \mathcal{S}$, for some constant $C_f > 0$. Then, the following bounds hold:*

$$\mathbb{E}\left\|\left[\frac{1}{B}\sum_{i=1}^{B} f(s_i)\right] - \mu\right\|^2 \leq \frac{C_f^2 + 2\sqrt{d}C_f^2 C}{B}$$

where $\mathbb{E}$ denotes the expectation of the Markov chain $\{s_i\}$ induced by π_θ starting from any $s_0 \in \mathcal{S}$ and $\mu = \mathbb{E}_{s\sim d^{\pi_\theta}}[f(s)]$. Furthermore, if $s_1 \in \mathcal{S}_R^\theta$, then

$$\mathbb{E}\left\|\left[\frac{1}{B}\sum_{i=1}^{B} f(s_i)\right] - \mu\right\|^2 \leq \frac{C_f^2 + 2\sqrt{d}C_f^2 C_{\text{tar}}}{B}$$

Observe that

$$\begin{aligned}
\text{Var}\left(\frac{1}{B}\sum_{i=1}^{B} f(s_i)\right) &= \mathbb{E}\left\|\frac{1}{B}\sum_{i=1}^{B} f(s_i) - \mu\right\|^2 \\
&= \frac{\sum_{i=1}^{B}\mathbb{E}\|f(s_i) - \mu\|^2 + 2\sum_{i=1}^{B}\sum_{j=i+1}^{B}\mathbb{E}\langle f(s_i) - \mu, f(s_j) - \mu\rangle}{B^2} \\
&= \frac{\sum_{i=1}^{B}\mathbb{E}\|f(s_i) - \mu\|^2 + 2\sum_{i=1}^{B}\mathbb{E}\langle f(s_i) - \mu, \mathbb{E}_i[\sum_{j=i+1}^{B}(f(s_j) - \mu)]\rangle}{B^2} \\
&\leq \frac{\sum_{i=1}^{B}\mathbb{E}\|f(s_i) - \mu\|^2 + 2\sum_{i=1}^{B}\mathbb{E}\|f(s_i) - \mu\|\|\mathbb{E}_i[\sum_{j=i+1}^{B}(f(s_j) - \mu)]\|}{B^2} \\
&\overset{(a)}{\leq} \frac{\sum_{i=1}^{B}\mathbb{E}\|f(s_i) - \mu\|^2 + 2\sum_{i=1}^{B}\sqrt{d}C_f C_{\text{tar}}\,\mathbb{E}\|f(s_i) - \mu\|}{B^2} \\
&\leq \frac{C_f^2 + 2\sqrt{d}C_f^2 C}{B},
\end{aligned} \tag{59}$$

where $\mathbb{E}_i$ denotes the conditional expectation given s_i and (a) follows from Lemma 8. If $s_1 \in \mathcal{S}_R^\theta$, then C can be replaced by C_{tar}.

C.4 Bound on the Hitting time of the Recurrent Class

Lemma 10. *Let Assumption 1 hold, and let $\{s_t\}_{t\geq 0}$ be the state sequence induced by policy π_θ. Then, for any $B \geq 1$, the probability of hitting $\mathcal{S}_R^\theta$ within B steps satisfies*

$$\Pr(T_\theta \leq B) \geq 1 - 2^{-\lfloor B/2C_{\mathrm{hit}} \rfloor}.$$

Proof. By Markov's inequality,

$$\Pr(T_\theta > 2C_{\mathrm{hit}}) \leq \frac{\mathbb{E}[T_\theta]}{2C_{\mathrm{hit}}} \leq \frac{1}{2},$$

so $\Pr(T_\theta \leq 2C_{\mathrm{hit}}) \geq \frac{1}{2}$.

Now partition the time horizon $[0, B]$ into $N := \lfloor B/2C_{\mathrm{hit}} \rfloor$ disjoint intervals of length $2C_{\mathrm{hit}}$. By the Markov property and the uniform bound on the expected hitting time, the probability of not hitting $\mathcal{S}_R^\theta$ in each interval is at most $\frac{1}{2}$, regardless of the starting state. Therefore,

$$\Pr(T_\theta > B) \leq \left(\frac{1}{2}\right)^N = 2^{-\lfloor B/2C_{\mathrm{hit}} \rfloor}.$$

Taking the complement gives the desired bound:

$$\Pr(T_\theta \leq B) \geq 1 - 2^{-\lfloor B/2C_{\mathrm{hit}} \rfloor}.$$

□

D Analysis of a General Linear Recursion

Recall that the Critic and NPG subroutines in the algorithm can be viewed as instances of stochastic linear recursions. To analyze their behavior in a unified framework, we consider a general stochastic linear recursion that captures the essential structure of both subroutines. This approach allows us to derive bounds that apply to each case as a special instance. Specifically, we study the recursion of the form:

$$x_{h+1} = x_h - \bar{\beta}(\widehat{P}_h x_h - \widehat{q}_h), \tag{60}$$

where $\widehat{P}_h$ and $\widehat{q}_h$ are noisy estimates of $P \in \mathbb{R}^{n\times n}$ and $q \in \mathbb{R}^n$, respectively, and $h \in \{0, \ldots, H-1\}$. Assume that the following bounds hold for all h:

$$(A_1) \quad \mathbb{E}_h\left[\|\widehat{P}_h - P\|^2\right] \leq \sigma_P^2, \quad (A_2) \quad \|\mathbb{E}_h\left[\widehat{P}_h\right] - P\|^2 \leq \delta_P^2,$$

$$(A_3) \quad \mathbb{E}_h\left[\|\widehat{q}_h - q\|^2\right] \leq \sigma_q^2, \quad (A_4) \quad \|\mathbb{E}_h\left[\widehat{q}_h\right] - q\|^2 \leq \delta_q^2,$$

and

$$(A_5) \quad \|\mathbb{E}\left[\widehat{q}_h\right] - q\|^2 \leq \bar{\delta}_q^2.$$

Here, $\mathbb{E}_h$ denotes the conditional expectation given the history up to step h. Since $\mathbb{E}[\widehat{q}_h] = \mathbb{E}[\mathbb{E}_h[\widehat{q}_h]]$, we have $\bar{\delta}_q^2 \leq \delta_q^2$. Additionally, assume:

$$(A_6) \quad \|P\| \leq \Lambda_P, \quad (A_7) \quad \|q\| \leq \Lambda_q,$$

and

$$(A_8) \quad x^\top P x \geq \lambda_P \|x\|^2, \quad (A_9) \quad \ker(P) \subseteq \ker(\widehat{P}_h)$$

for all $x \in \ker(P)^\perp$ and $h \geq 1$. Let $x^* = P^\dagger q$, where $P^\dagger$ is the Moore-Penrose pseudoinverse of P and Π denote the projection operator onto the space $\ker(P)^\perp$, i.e., the orthogonal complement of the kernel (null space) of P.

While a general linear analysis was conducted in [16], our approach differs significantly due to the fact that the critic matrix is not strictly positive definite. This necessitates the additional condition (A_9), which was not required in the ergodic setting. Despite the lack of strict positive definiteness, we still achieve the same convergence rate.

Theorem 4. *Consider the recursion* (60) *and suppose* $(A_1) - (A_9)$ *hold. Further, let* $\delta_P \le \lambda_P/8$ *and* $\bar{\beta} = \frac{\lambda_P}{\Lambda_P}$*. Then, the following bounds hold for all* $H \ge 1$*:*

$$(a) \quad \mathbb{E}\left[\|\Pi(x_H - x^*)\|^2\right] \le \mathcal{O}\left(\exp\left(-\frac{H\lambda_P^2}{4\Lambda_q}\right) R_0^2 + \sigma_P^2 \lambda_P^{-2} \Lambda_q^2 \Lambda_P^{-1} + \sigma_q^2 \Lambda_P^{-1} + \delta_P^2 \lambda_P^{-4} \Lambda_q^2 + \lambda_P^{-2} \delta_q^2\right).$$

and

$$(b) \quad \|\Pi(\mathbb{E}[x_H] - x^*)\|^2 \le \mathcal{O}\left(\exp\left(-\frac{H\lambda_P^2}{4\Lambda_q}\right) R_0^2 + \delta_P^2 R_0^2 \lambda_P^{-2} + \delta_P^2 \Lambda_P^2 \lambda_P^{-4} + \bar{\delta}_q^2 \lambda_P^{-2}\right),$$

where $R_0 = \|\Pi(x_0 - x_*)\|^2$

D.1 Proof of Theorem 4*(a)*

Before proceeding, we provide some useful results with their proofs.

Lemma 11. *Consider a matrix* A *such that* $x^\top A x \ge \lambda \|x\|^2$ *for all* $x \in (\ker(A))^\perp$*. Then, we have*

$$\|A^\dagger\| \le \frac{1}{\lambda}$$

Proof. It is known that the operator norm of $A^\dagger$ is $\|A^\dagger\| = 1/\sigma_r$, where σ_r is the smallest nonzero singular value of A (see Theorem 5.5.1 in [19]).

Note that $x^\top A x = x^\top A^\top x$, which in turn implies $x^\top A x = (1/2) x^\top (A + A^\top) x$. The condition $x^\top A x \ge \lambda \|x\|^2$ implies that $x^\top \frac{(A+A^\top)}{2} x \ge \lambda \|x\|^2$ for all $x \in (\ker(A))^\perp$. Since $\frac{(A+A^\top)}{2}$ is a symmetric matrix, this implies that all non-zero eigenvalues of $\frac{(A+A^\top)}{2}$ are bounded below by λ. Now using Prop. III.5.1 in [10], it follows that $\sigma_r \ge \lambda$ which concludes the proof.

□

Lemma 12. *Let* Π *denote the projection operator onto the space* $\ker(P)^\perp$*, i.e., the orthogonal complement of the kernel (null space) of* P*. Then the following properties hold:*

(i) $\Pi^\top \Pi = \Pi^2 = \Pi$*,*

(ii) If $\widetilde{P}$ *is such that* $\ker(P) \subseteq \ker(\widetilde{P})$*, then* $\widetilde{P}\Pi = \widetilde{P}$*.*

Proof. The first property holds since Π is a projection operator onto a linear subspace. Such operators are both symmetric and idempotent, which implies $\Pi^\top = \Pi$ and $\Pi^2 = \Pi$, hence $\Pi^\top \Pi = \Pi$.

For the second property, let $\Pi^\perp$ denote the projection operator onto $\ker(P)$. By the orthogonal decomposition of the space, we have

$$\Pi + \Pi^\perp = I,$$

where I is the identity operator. For any x, it follows that

$$\widetilde{P}x = \widetilde{P}(\Pi x + \Pi^\perp x) = \widetilde{P}\Pi x,$$

where the last equality holds since $\Pi^\perp x \in \ker(P) \subseteq \ker(\widetilde{P})$ and thus $\widetilde{P}\Pi^\perp x = 0$. This establishes the result. □

Coming back to the proof, consider (60) and apply the projection operator Π to both sides. Using the linearity of Π, we obtain the following:

$$\begin{aligned}
\Pi(x_{h+1} - x^*) &= \Pi\left(x_h - x^* - \bar{\beta}(\widehat{P}_h x_h - \widehat{q}_h)\right) \\
&= \Pi\left(x_h - x^* - \bar{\beta}(P x_h - q + M_{h+1})\right) \\
&= \Pi\left(x_h - x^* - \bar{\beta} P(x_h - x^*) + \bar{\beta} M_{h+1}\right) \\
&= \Pi(x_h - x^*) - \bar{\beta}\Pi P(x_h - x^*) + \bar{\beta}\Pi M_{h+1},
\end{aligned} \tag{61}$$

where $M_{h+1} = (\widehat{P}_h - P)x_h - (\widehat{q}_h - q)$.

Taking the square norm on both sides of the above equation and then taking expectation we obtain

$$\begin{aligned}
&\mathbb{E}\,\|\Pi(x_{h+1} - x^*)\|^2 \\
&= \mathbb{E}\,\|\Pi(x_h - x^*) - \bar{\beta}\Pi P(x_h - x^*) + \bar{\beta}\Pi M_{h+1}\|^2 \\
&= \underbrace{\mathbb{E}\,\|\Pi(x_h - x^*) - \bar{\beta}\Pi P(x_h - x^*)\|^2}_{T_1} + \underbrace{\beta^2\,\mathbb{E}\,\|\Pi M_{h+1}\|^2}_{T_2} \\
&\quad + \underbrace{2\,\mathbb{E}\langle \Pi(x_h - x^*) - \bar{\beta}\Pi P(x_h - x^*), \bar{\beta}\Pi M_{h+1}\rangle}_{T_3}
\end{aligned} \tag{62}$$

We now analyze term T_1

$$\begin{aligned}
&\mathbb{E}\,\|\Pi(x_h - x^*) - \bar{\beta}\Pi P(x_h - x^*)\|^2 \\
&\quad = \mathbb{E}\left[\|\Pi(x_h - x^*)\|^2 - 2\bar{\beta}\langle \Pi(x_h - x^*), \Pi P(x_h - x^*)\rangle + \bar{\beta}^2\|\Pi P(x_h - x^*)\|^2\right] \\
&\quad = \mathbb{E}\left[\|\Pi(x_h - x^*)\|^2 - 2\bar{\beta}(x_h - x^*)^\top \Pi^\top \Pi P(x_h - x^*) + \bar{\beta}^2\|\Pi P(x_h - x^*)\|^2\right] \\
&\quad \overset{(a)}{=} \mathbb{E}\left[\|\Pi(x_h - x^*)\|^2 - 2\bar{\beta}(x_h - x^*)^\top \Pi^\top P\Pi(x_h - x^*) + \bar{\beta}^2\|\Pi P(x_h - x^*)\|^2\right] \\
&\quad = \mathbb{E}\left[\|\Pi(x_h - x^*)\|^2 - 2\bar{\beta}(\Pi(x_h - x^*))^\top P(\Pi(x_h - x^*)) + \bar{\beta}^2\|\Pi P(x_h - x^*)\|^2\right] \\
&\quad \overset{(b)}{=} (1 - \lambda_P\bar{\beta} + \bar{\beta}^2\Lambda_P)\,\mathbb{E}\,\|\Pi(x_h - x^*)\|^2,
\end{aligned} \tag{63}$$

where (a) follows from (A_9) and Lemma 12 and (b) follows from (A_6) and (A_8). Now, consider term T_3

$$\begin{aligned}
&\mathbb{E}\langle \Pi(x_h - x^*) - \bar{\beta}\Pi P(x_h - x^*), \bar{\beta}\Pi M_{h+1}\rangle \\
&\quad = \mathbb{E}\langle \Pi(x_h - x^*) - \bar{\beta}\Pi P(x_h - x^*), \bar{\beta}\,\mathbb{E}[\Pi M_{h+1}|x_h]\rangle \\
&\quad \overset{(a)}{\leq} \frac{\bar{\beta}\lambda_P}{2}\,\mathbb{E}\,\|\Pi(x_h - x^*) - \bar{\beta}\Pi P(x_h - x^*)\|^2 + \frac{2\bar{\beta}}{\lambda_P}\,\mathbb{E}\,\|\Pi\,\mathbb{E}[M_{h+1}|x_h]\|^2
\end{aligned} \tag{64}$$

where (a) follows from the linearity of Π, which allows it to be exchanged with the conditional expectation, and from applying Young's inequality. Adding T_1 and T_3 gives

$$\begin{aligned}
&\mathbb{E}\,\|\Pi(x_h - x^*) - \bar{\beta}\Pi P(x_h - x^*)\|^2 + \mathbb{E}\langle \Pi(x_h - x^*) - \bar{\beta}\Pi P(x_h - x^*), \bar{\beta}\Pi M_{h+1}\rangle \\
&\leq \left(1 + \frac{\bar{\beta}\lambda_P}{2}\right)\mathbb{E}\,\|\Pi(x_h - x^*) - \bar{\beta}\Pi P(x_h - x^*)\|^2 + \frac{2\bar{\beta}}{\lambda_P}\|\Pi\,\mathbb{E}[M_{h+1}|x_h]\|^2 \\
&= \left(1 + \frac{\bar{\beta}\lambda_P}{2}\right)\left(1 - \lambda_P\beta + \beta^2\Lambda_P\right)\mathbb{E}\,\|\Pi(x_h - x^*)\|^2 + \frac{2\bar{\beta}}{\lambda_P}\|\Pi\,\mathbb{E}[M_{h+1}|x_h]\|^2 \\
&= \left(1 - \frac{\bar{\beta}\lambda_P}{2} + \frac{\bar{\beta}^2(2\Lambda_P - \lambda_P^2)}{2} + \frac{\bar{\beta}^3\Lambda_P\lambda_P}{2}\right)\mathbb{E}\,\|\Pi(x_h - x^*)\|^2 + \frac{2\bar{\beta}}{\lambda_P}\|\Pi\,\mathbb{E}[M_{h+1}|x_h]\|^2
\end{aligned} \tag{65}$$

Consider the term $\Pi\,\mathbb{E}[M_{h+1}|x_h]$ in the above sum.

$$\begin{aligned}
\Pi\,\mathbb{E}[M_{h+1}|x_h] &= \Pi(\mathbb{E}[(\widehat{P}_h - P)x_h - (\widehat{q}_h - q)|x_h]) \\
&= \Pi\,\mathbb{E}[(\widehat{P}_h - P)|x_h]x_h - \Pi\,\mathbb{E}[(\widehat{q}_h - q)|x_h] \\
&= \Pi\,\mathbb{E}[(\widehat{P}_h - P)|x_h](x_h - x^*) + \Pi\,\mathbb{E}[(\widehat{P}_h - P)|x_h]x^* - \Pi\,\mathbb{E}[(\widehat{q}_h - q)|x_h] \\
&\overset{(a)}{=} \Pi\,\mathbb{E}[(\widehat{P}_h - P)|x_h]\Pi(x_h - x^*) + \Pi\,\mathbb{E}[(\widehat{P}_h - P)|x_h]x^* - \Pi\,\mathbb{E}[(\widehat{q}_h - q)|x_h]
\end{aligned}$$

where (a) follows from (A_9) and Lemma 12. Using the above bound, we obtain the following using triangle inequality

$$\begin{aligned}
\mathbb{E}\,\|\Pi\,\mathbb{E}[M_{h+1}|x_h]\|^2 &\leq 3\,\mathbb{E}[\|\Pi\,\mathbb{E}[(\widehat{P}_h - P)|x_h]\Pi(x_h - x^*)\|^2] + 3\,\mathbb{E}[\|\Pi\,\mathbb{E}[(\widehat{P}_h - P)|x_h]x^*\|^2] \\
&\quad + 3\,\mathbb{E}[\|\Pi\,\mathbb{E}[(\widehat{q}_h - q)|x_h]\|^2] \\
&\leq 3\,\mathbb{E}[\|\,\mathbb{E}[(\widehat{P}_h - P)|x_h]\|^2\|\Pi(x_h - x^*)\|^2] + 3\,\mathbb{E}[\|\,\mathbb{E}[(\widehat{P}_h - P)|x_h]\|^2\|x^*\|^2] \\
&\quad + 3\,\mathbb{E}[\|\,\mathbb{E}[(\widehat{q}_h - q)|x_h]\|^2] \\
&\overset{(a)}{\leq} 3\delta_P^2\,\mathbb{E}\,\|\Pi(x_h - x^*)\|^2 + 3\delta_P^2\lambda_P^{-2}\Lambda_q^2 + 3\delta_q^2
\end{aligned}$$

where, (a) follows from (A_2) and (A_4). Similarly, we analyze the term T_2 below.

$$\begin{aligned}
&\bar{\beta}^2\,\mathbb{E}\,\|\Pi M_{h+1}\|^2 \\
&\quad\le \bar{\beta}^2\,\mathbb{E}\,\|M_{h+1}\|^2 \\
&\quad= \bar{\beta}^2\,\mathbb{E}\,\|(\widehat{P}_h - P)x_h - (\widehat{q}_h - q)\|^2 \\
&\quad= \bar{\beta}^2\,\mathbb{E}\,\|(\widehat{P}_h - P)\Pi(x_h - x^*) + (\widehat{P}_h - P)\Pi x^* - (\widehat{q}_h - q)\|^2 \\
&\quad= 3\bar{\beta}^2\left(\mathbb{E}[\mathbb{E}[\|\widehat{P}_h - P\|^2|x_h]\|\Pi(x_h - x^*)\|^2] + \mathbb{E}\,\|\widehat{P}_h - P\|^2\|x^*\|^2 + \mathbb{E}\,\|\widehat{q}_h - q\|^2\right) \\
&\quad\overset{(a)}{\le} 3\bar{\beta}^2\sigma_P^2\,\mathbb{E}[\|\Pi(x_h - x^*)\|^2] + 3\bar{\beta}^2\sigma_P^2\lambda_P^{-2}\Lambda_q^2 + 3\bar{\beta}^2\sigma_q^2
\end{aligned} \tag{66}$$

where (a) follows from (A_1) and (A_3). Combining bounds obtained for T_1, T_2 and T_3 with (62), we obtain the following bound.

$$\begin{aligned}
&\mathbb{E}\,\|\Pi(x_{h+1} - x^*)\|^2 \\
&\le \left(1 - \frac{\bar{\beta}(\lambda_P - 12\delta_P^2\lambda_P^{-1})}{2} + \frac{\bar{\beta}^2(3\sigma_P^2 + 2\Lambda_P - \lambda_P^2)}{2} + \frac{\bar{\beta}^3\Lambda_P\lambda_P}{2}\right)\mathbb{E}\,\|\Pi(x_h - x^*)\|^2 \\
&+ 3\bar{\beta}^2\sigma_P^2\lambda_P^{-2}\Lambda_q^2 + 3\bar{\beta}^2\sigma_q^2 + 6\bar{\beta}\delta_P^2\lambda_P^{-3}\Lambda_q^2 + 6\bar{\beta}\lambda_P^{-1}\delta_q^2
\end{aligned} \tag{67}$$

Suppose $12\delta_P^2 \le \lambda_P^2/2$ and $\bar{\beta}$ is small enough such that $\bar{\beta} \le 1$ and $\bar{\beta} \le \frac{\lambda_P}{2\left((3\sigma_P^2 + 2\Lambda_P - \lambda_P^2) + \Lambda_P\lambda_P\right)}$. Then, the above bound gives us

$$\mathbb{E}\,\|\Pi(x_{h+1} - x^*)\|^2 \le \left(1 - \frac{\bar{\beta}\lambda_P}{4}\right)\mathbb{E}\,\|\Pi(x_h - x^*)\|^2 + \epsilon_h, \tag{68}$$

where

$$\epsilon_h := 3\bar{\beta}^2\sigma_P^2\lambda_P^{-2}\Lambda_q^2 + 3\bar{\beta}^2\sigma_q^2 + 6\bar{\beta}\delta_P^2\lambda_P^{-3}\Lambda_q^2 + 6\bar{\beta}\lambda_P^{-1}\delta_q^2. \tag{69}$$

Unrolling the recursion yields

$$\mathbb{E}\,\|\Pi(x_{h+1} - x^*)\|^2 \le \left(1 - \frac{\bar{\beta}\lambda_P}{4}\right)^H \mathbb{E}\,\|\Pi(x_1 - x^*)\|^2 + \sum_{h=1}^{H}\left(1 - \frac{\bar{\beta}\lambda_P}{4}\right)^{H-h+1}\epsilon_h \tag{70}$$

$$\le \exp\left(-\frac{H\bar{\beta}\lambda_P}{4}\right)\mathbb{E}\,\|\Pi(x_1 - x^*)\|^2 + \frac{4\epsilon_h}{\bar{\beta}\lambda_P} \tag{71}$$

Setting $\bar{\beta} = \frac{\lambda_P}{\Lambda_P}$ in the above equation yields the final result.

D.2 Proof of Theorem 4*(b)*

We analyze the squared bias term $\|\Pi(\mathbb{E}[x_h] - x^*)\|^2$, where x_h denotes the iterate at step h, x^* is the optimal solution, and Π is a projection operator. Recall that the update rule for x_{h+1} is can be written as:

$$x_{h+1} = x_h - \bar{\beta}P(x_h - x^*) + \bar{\beta}M_{h+1},$$

Taking expectation and applying the projection operator Π on both sides, we obtain:

$$\|\Pi(\mathbb{E}[x_{h+1}] - x^*)\|^2 = \|\Pi\mathbb{E}[(x_h - x^*) - \bar{\beta}P(x_h - x^*) + \bar{\beta}M_{h+1}]\|^2.$$

Expanding the norm using the properties of inner products yields:

$$\begin{aligned}
\|\Pi(\mathbb{E}[x_{h+1}] - x^*)\|^2 &= \|\Pi\mathbb{E}[(x_h - x^*) - \bar{\beta}P(x_h - x^*)]\|^2 + \bar{\beta}^2\|\Pi\mathbb{E}[M_{h+1}]\|^2 \\
&\quad - 2\langle\Pi\mathbb{E}[(x_h - x^*) - \bar{\beta}P(x_h - x^*)], \bar{\beta}\Pi\mathbb{E}[M_{h+1}]\rangle.
\end{aligned} \tag{72}$$

Using Young's inequality, we bound the cross-term as follows:

$$\begin{aligned}
-2\langle\Pi\mathbb{E}[(x_h - x^*) - \bar{\beta}P(x_h - x^*)], \bar{\beta}\Pi\mathbb{E}[M_{h+1}]\rangle &\le \frac{\bar{\beta}\lambda_P}{2}\|\Pi\mathbb{E}[(x_h - x^*) - \bar{\beta}P(x_h - x^*)]\|^2 \\
&\quad + \frac{2\bar{\beta}}{\lambda_P}\|\Pi\mathbb{E}[M_{h+1}]\|^2.
\end{aligned}$$

Substituting this bound back, we find

$$\|\Pi(\mathbb{E}[x_{h+1}]-x^*)\|^2 \leq \left(1+\frac{\bar{\beta}\lambda_P}{2}\right)\|\Pi\mathbb{E}[(x_h-x^*)-\bar{\beta}P(x_h-x^*)]\|^2+\left(\frac{2\bar{\beta}}{\lambda_P}+\bar{\beta}^2\right)\|\Pi\mathbb{E}[M_{h+1}]\|^2.$$

Next, we analyze the term $\|\Pi\mathbb{E}[(x_h - x^*) - \bar{\beta}P(x_h - x^*)]\|^2$. Using the linearity of expectation, we have

$$\|\Pi\mathbb{E}[(x_h - x^*) - \bar{\beta}P(x_h - x^*)]\|^2 = \|\Pi(\mathbb{E}[x_h] - x^*) - \bar{\beta}\Pi P(\mathbb{E}[x_h] - x^*)\|^2.$$

Expanding this norm

$$\begin{aligned}\|\Pi(\mathbb{E}[x_h] - x^*) - \bar{\beta}P(\mathbb{E}[x_h] - x^*)\|^2 = \|\Pi(\mathbb{E}[x_h] - x^*)\|^2 + \bar{\beta}^2\|P(\mathbb{E}[x_h] - x^*)\|^2 \\ - 2\bar{\beta}\langle\Pi(\mathbb{E}[x_h] - x^*), P\Pi(\mathbb{E}[x_h] - x^*)\rangle.\end{aligned}$$

Since $\Pi(\mathbb{E}[x_h] - x^*) \in \ker(P)^\perp$, we can use A_6 and A_8 to obtain

$$\|\Pi(\mathbb{E}[x_h] - x^*) - \bar{\beta}P(\mathbb{E}[x_h] - x^*)\|^2 \leq (1 - \bar{\beta}\lambda_P + \bar{\beta}^2\Lambda_P)\|\Pi(\mathbb{E}[x_h] - x^*)\|^2.$$

Finally, we bound the noise term $\|\Pi\mathbb{E}[M_{h+1}]\|^2$. Using the conditional expectation property, we expand

$$\|\Pi\mathbb{E}[M_{h+1}]\|^2 = \|\Pi\mathbb{E}[\mathbb{E}[M_{h+1}|x_h]]\|^2.$$

Decomposing M_{h+1} into its components

$$M_{h+1} = (\widehat{P}_h - P)(x_h - x^*) + (\widehat{P}_h - P)x^* + (\widehat{q}_h - q),$$

we bound each term separately. Using A_2 and A_5, we find

$$\|\Pi\mathbb{E}[M_{h+1}]\|^2 \leq 3\delta_P^2 R_0^2 + 3\delta_P^2\Lambda_P^2\lambda_P^{-2} + 3\bar{\delta}_q^2,$$

where $R_0^2 = \mathbb{E}[\|\Pi(x_h - x^*)\|^2]$.

Combining all results, the squared bias term satisfies the recurrence:

$$\begin{aligned}\|\Pi(\mathbb{E}[x_{h+1}] - x^*)\|^2 &\leq \left(1+\frac{\bar{\beta}\lambda_P}{2}\right)(1 - \bar{\beta}\lambda_P + \bar{\beta}^2\Lambda_P)\|\Pi(\mathbb{E}[x_h] - x^*)\|^2 + \bar{\epsilon}_h \\ &= \left(1-\frac{\bar{\beta}\lambda_P}{2}+\frac{\bar{\beta}^2(2\Lambda_P-\lambda_P^2)}{2}+\frac{\bar{\beta}^3\Lambda_P\lambda_P}{2}\right)\|\Pi(\mathbb{E}[x_h] - x^*)\|^2 + \bar{\epsilon}_h\end{aligned}$$

where

$$\bar{\epsilon}_h = \left(\frac{2\bar{\beta}}{\lambda_P}+\bar{\beta}^2\right)\left(3\delta_P^2 R_0^2 + 3\delta_P^2\Lambda_P^2\lambda_P^{-2} + 3\bar{\delta}_q^2\right).$$

Suppose β is such that $\bar{\beta} < (2\Lambda_P)^{-1}$. Then, the above bound gives us

$$\|\Pi(\mathbb{E}[x_{h+1}] - x^*)\|^2 \leq \left(1-\frac{\bar{\beta}\lambda_P}{4}\right)\|\Pi(\mathbb{E}[x_h] - x^*)\|^2 + \bar{\epsilon}_h \tag{73}$$

Unrolling this recursion over h steps, we obtain:

$$\|\Pi(\mathbb{E}[x_H] - x^*)\|^2 \leq \left(1-\frac{\bar{\beta}\lambda_P}{4}\right)^H\|\Pi(x_0 - x^*)\|^2 + \frac{\bar{\epsilon}_h}{\bar{\beta}\lambda_P} \tag{74}$$

Setting $\bar{\beta} = \frac{\lambda_P}{\Lambda_P}$ in the above equation yields the final result.

E Proof of Theorem 2

Recall the definition of Z_θ

$$Z_\theta \coloneqq \left\{z \,\middle|\, [\Phi z](i) = [\Phi z](j) \quad \forall i,j \in \mathcal{S}_R^\theta\right\}. \tag{75}$$

We now state and prove a few useful results for the critic analysis. Recall that $M_\theta \coloneqq \mathbb{E}_\theta\left[\phi(s)\left(\phi(s)-\phi(s')\right)^\top\right]$ constitutes a submatrix of the critic matrix $A_v(\theta)$. We begin by characterizing the kernel of M_θ in Lemma 13.

Lemma 13. *Let Assumption 1 hold, and suppose Φ is full rank. Then* $\ker(M_\theta) = Z_\theta$.

Proof. Let $f : \mathcal{S} \to \mathbb{R}$ be an arbitrary function. Consider the quantity

$$\mathbb{E}\left[(f(s) - f(s'))^2\right],$$

where the expectation is over $s \sim d^{\pi_\theta}$ and $s' \sim P^{\pi_\theta}(\cdot \mid s)$ is the next state under the transition kernel induced by policy π_θ. This expression captures the expected squared one-step difference in the values of f along the Markov chain.

It is known that if the chain is irreducible and f is non-constant, then $\mathbb{E}_{s\sim d^{\pi_\theta}}\left[(f(s) - f(s'))^2\right] > 0$ [34]. We extend this result to the setting where the chain induced by π_θ has a single recurrent class (possibly with transient states).

Suppose f is non-constant on this recurrent class. Then, we claim that $\mathbb{E}_{s\sim d^{\pi_\theta}}\left[(f(s) - f(s'))^2\right] > 0$. It suffices to show that there exist states s, s' within the recurrent class such that $P^{\pi_\theta}(s, s') > 0$ and $f(s) \neq f(s')$. Suppose, for the sake of contradiction, that no such pair exists. Then, the recurrent class could be partitioned into disjoint subsets, each corresponding to a distinct constant value of f, with no transitions between subsets. This, however, contradicts the fact that the recurrent class is closed and all states within it are communicating. Hence, such a pair (s, s') must exist, which implies that the expected squared difference is strictly positive since

$$\mathbb{E}_{s\sim d^{\pi_\theta}}\left[(f(s) - f(s'))^2\right] \geq d^{\pi_\theta}(s)P^{\pi_\theta}(s, s')(f(s) - f(s'))^2 > 0. \tag{76}$$

Separately, note that $\mathbb{E}_{s\sim d^{\pi_\theta}}[(f(s) - f(s'))^2]$ can also be written as follows

$$\begin{aligned}\mathbb{E}_{s\sim d^{\pi_\theta}}[(f(s) - f(s'))^2] &= \mathbb{E}_{s\sim d^{\pi_\theta}}[f(s)^2 - 2f(s)f(s') + f(s')^2] \\ &\overset{(a)}{=} 2\,\mathbb{E}_{s\sim d^{\pi_\theta}}[f(s)(f(s) - f(s'))],\end{aligned} \tag{77}$$

where (a) follows since $\sum_{s\in\mathcal{S}} d^{\pi_\theta}(s)P^{\pi_\theta}(s, s') = d^{\pi_\theta}(s')$ Taking $f = \Phi z$, it follows that $\mathbb{E}_{s\sim d^{\pi_\theta}}[([\Phi z](s) - [\Phi z](s'))^2] = z^\top M_\theta z$, which concludes the proof. □

The following result provides a lower bound on $\xi^\top A_v(\theta)\xi$ using Assumption 2.

Lemma 14. *For a large enough c_β, Assumption 2 implies that $\xi^\top A_v(\theta)\xi \geq (\lambda/2)\|\xi\|^2$ for all $\xi = [\eta, \zeta]^\top$ such that $\zeta \in \ker(M_\theta)^\perp$, for all $\theta \in \Theta$.*

Proof of Lemma 14. Recall that $A_v(\theta) = \mathbb{E}_\theta[A_v(z)]$ where $\mathbb{E}_\theta$ denotes expectation over the distribution of $z = (s, a, s')$ where $(s, a) \sim \nu^{\pi_\theta}$, $s' \sim P(\cdot|s, a)$. Hence, for any $\xi = [\eta, \zeta]$ such that $\zeta \in \ker(M_\theta)^\perp$, we have

$$\begin{aligned}\xi^\top A_v(\theta)\xi &= c_\beta\eta^2 + \eta\zeta^\top\,\mathbb{E}_\theta\left[\phi(s)\right] + \zeta^\top\,\mathbb{E}_\theta\left[\phi(s)\left[\phi(s) - \phi(s')\right]^\top\right]\zeta \\ &\overset{(a)}{\geq} c_\beta\eta^2 - |\eta|\,\|\zeta\| + \lambda\,\|\zeta\|^2 \\ &\geq \|\xi\|^2\left\{\min_{u\in[0,1]} c_\beta u - \sqrt{u(1-u)} + \lambda(1-u)\right\} \overset{(b)}{\geq} (\lambda/2)\,\|\xi\|^2\end{aligned} \tag{78}$$

where (a) is a consequence of Assumption 2 and the fact that $\|\phi(s)\| \leq 1, \forall s \in \mathcal{S}$. Finally, (b) is satisfied when $c_\beta \geq \lambda + \sqrt{\frac{1}{\lambda^2} - 1}$. This concludes the proof of Lemma 14. □

Given event $\mathcal{E}_B$, it follows that all states visited after the critic update, s_b^{kh}, where $h \geq 1$, $s_b^{kh} \in \mathcal{S}_R^{\theta_k}$. For $s \in \mathcal{S}_R^{\theta_k}$, we have the following result.

Lemma 15. *Let $s \in \mathcal{S}_R^\theta$ and $s' \sim P^{\pi_\theta}(\cdot \mid s)$ is the next state under the transition kernel induced by policy π_θ. Then $\ker(A_v(\theta)) \subset \ker(A_v(\theta, z))$, where $z = (s, a, s')$.*

Proof. Let $z \in \ker(A_v(\theta))$. Then $z = [0,\, \zeta^\top]^\top$, where $\zeta \in \ker(M_\theta) = Z_\theta$. It follows that

$$A_v(\theta)z = \begin{bmatrix} 0 \\ \phi(s)(\phi(s)-\phi(s'))^\top \zeta \end{bmatrix}.$$

By the definition of Z_θ, we have

$$\phi(s)(\phi(s)-\phi(s'))^\top \zeta = \phi(s)(\phi(s)^\top \zeta - \phi(s')^\top \zeta) = 0,$$

since $s, s' \in \mathcal{S}_R^\theta$.

□

In particular, (A_8) and (A_9) for the critic update with $h \geq 1$ follow from Lemmas 14 and 15, respectively. For any $z = (s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, we have the following.

$$\|A_v(\theta, z)\| \leq |c_\beta| + \|\phi(s)\| + \left\|\phi(s)(\phi(s)-\phi(s'))^\top\right\| \overset{(a)}{\leq} c_\beta + 3 = \mathcal{O}(c_\beta), \tag{79}$$

$$\|b_v(\theta, z)\| \leq |c_\beta r(s,a)| + \|r(s,a)\phi(s)\| \overset{(b)}{\leq} c_\beta + 1 = \mathcal{O}(c_\beta) \tag{80}$$

where (a), (b) hold since $|r(s,a)| \leq 1$ and $\|\phi(s)\| \leq 1, \forall (s,a) \in \mathcal{S} \times \mathcal{A}$. This verifies (A_6) and (A_7). Using Lemmas 8 and 9 with the above bounds, we obtain

$$\left\|\mathbb{E}\left[\frac{1}{B}\sum_{i=1}^{B} A_v(\theta, z_i)\right] - A_v(\theta)\right\| \leq \mathcal{O}\left(\frac{\sqrt{m} c_\beta C_{\text{tar}}}{B}\right) \tag{81}$$

and

$$\mathbb{E}\left\|\left[\frac{1}{B}\sum_{i=1}^{B} A_v(\theta, z_i)\right] - A_v(\theta)\right\|^2 \leq \mathcal{O}\left(\frac{c_\beta^2 + \sqrt{m} c_\beta^2 C_{\text{tar}}}{B}\right). \tag{82}$$

Similarly,

$$\left\|\mathbb{E}\left[\frac{1}{B}\sum_{i=1}^{B} b_v(\theta, z_i)\right] - b_v(\theta)\right\| \leq \mathcal{O}\left(\frac{\sqrt{m} c_\beta C_{\text{tar}}}{B}\right) \tag{83}$$

and

$$\mathbb{E}\left\|\left[\frac{1}{B}\sum_{i=1}^{B} b_v(\theta, z_i)\right] - b_v(\theta)\right\|^2 \leq \mathcal{O}\left(\frac{c_\beta^2 + \sqrt{m} c_\beta^2 C_{\text{tar}}}{B}\right). \tag{84}$$

These yield conditions (A_1)–(A_4) for the critic update, and condition (A_5) follows by setting $\bar{\delta}_q = \delta_q$. Since all conditions (A_1)–(A_9) are now verified, we can apply Theorem 4 with the critic learning rates $c_\beta = \lambda + \sqrt{\frac{1}{\lambda^2} - 1}$ and $\beta = \frac{\lambda^2}{2}$ to obtain the following bounds.

$$\mathbb{E}\left[\|\Pi(\xi_H - \xi^*)\|^2\right] \leq \mathcal{O}\left(\exp\left(-\tfrac{H\lambda^3}{16}\right)\|\xi_1 - \xi^*\|^2 + \frac{C_{\text{tar}}\sqrt{m}}{\lambda^4 B} + \frac{C_{\text{tar}}\sqrt{m}}{\lambda B} + \frac{C_{\text{tar}}^2 m}{\lambda^6 B^2} + \frac{C_{\text{tar}}^2 m}{\lambda^2 B^2}\right).$$

and

$$\|\Pi(\mathbb{E}[x_H] - x^*)\|^2 \leq \mathcal{O}\left(\exp\left(-\tfrac{H\lambda^3}{16}\right)\|\xi_1 - \xi^*\|^2 + \frac{C_{\text{tar}}^2 m \|\xi_1 - \xi^*\|^2}{\lambda^2 B^2} + \frac{C_{\text{tar}}^2 m}{\lambda^6 B^2} + \frac{C_{\text{tar}}^2 m}{\lambda^2 B^2}\right).$$

F Proof of Theorem 3

We note that condition (A_9) holds trivially for the NPG update, since $\ker(A_u(\theta)) = \{0\}$. Furthermore, by Assumption 4, condition (A_8) is readily satisfied with $\lambda_P = \mu$. Finally, from Assumption 3, we have

$$\|A_u(\theta, z)\| = \|\nabla \log \pi_\theta(a|s) \nabla \log \pi_\theta(a|s)^\top\| \le \|\nabla \log \pi_\theta(a|s)\|^2 \le G_1^2, \tag{85}$$

for all $z = (s, a, s')$. Combining this bound with Lemmas 8 and 9, it follows that

$$\left\| \mathbb{E}\left[\frac{1}{B}\sum_{i=1}^{B} A_u(\theta, z_i)\right] - A_u(\theta) \right\| \le \mathcal{O}\left(\frac{\sqrt{d}G_1^2 C_{\mathrm{tar}}}{B}\right) \tag{86}$$

and

$$\mathbb{E}\left\| \left[\frac{1}{B}\sum_{i=1}^{B} A_u(\theta, z_i)\right] - A_u(\theta) \right\|^2 \le \mathcal{O}\left(\frac{G_1^4 + \sqrt{d}G_1^4 C_{\mathrm{tar}}}{B}\right). \tag{87}$$

The above bounds yield (A_6), (A_2) and (A_1), respectively. To obtain (A_7), we combine the bound on the Advantage function from Lemma 7 with Assumption 3 to bound the policy gradient as follows

$$\begin{aligned} \|\nabla J(\theta)\| &= \|\,\mathbb{E}[A^{\pi_\theta}(s,a)\nabla \log \pi_\theta(a|s)]\| \le \|A^{\pi_\theta}(s,a)\|\|\nabla \log \pi_\theta(a|s)\| \\ &\le (1 + 4(C_{\mathrm{hit}} + C_{\mathrm{tar}}))G_1. \end{aligned} \tag{88}$$

To prove the other statements, recall the definition of $b_u(\theta_k, \xi_k, \cdot)$ from (28). Let $\mathbb{E}_{u,z}$ denote the expectation over $\{z_i\}_{i=1}^B$, given the entire history prior to z_1 (including ξ_k), $\mathbb{E}_u$ be denote the expectation given the entire history prior to z_1 with $s_1 \sim d^{\pi_{\theta_k}}$ and $\mathbb{E}_v$ denote the expectation over the entire history prior to z_1. Observe the following relations for arbitrary θ_k, ξ_k.

$$\begin{aligned} &\mathbb{E}_u\left[\frac{1}{B}\sum_{i=1}^{B} b_u(\theta_k, \xi_k, z_i)\right] - \nabla_\theta J(\theta_k) \\ &= \mathbb{E}_u\left[\frac{1}{B}\sum_{i=1}^{B}\Big\{r(s_i,a_i) - \eta_k + \langle \phi(s_{i+1}) - \phi(s_i), \zeta_k\rangle\Big\}\nabla_\theta \log_{\pi_{\theta_k}}(a_i|s_i)\right] - \nabla_\theta J(\theta_k) \\ &\overset{(a)}{=} \underbrace{\mathbb{E}_u\left[\frac{1}{B}\sum_{i=1}^{B}\Big\{\eta_k^* - \eta_k + \langle \phi(s_{i+1}) - \phi(s_i), \zeta_k - \zeta_k^*\rangle\Big\}\nabla_\theta \log_{\pi_{\theta_k}}(a_i|s_i)\right]}_{T_0} + \\ &\quad + \underbrace{\mathbb{E}_u\left[\frac{1}{B}\sum_{i=1}^{B}\Big\{(\langle \phi(s_i), \zeta_k^*\rangle - V^{\pi_{\theta_k}}(s_i)) + (V^{\pi_{\theta_k}}(s_{i+1}) - \langle \phi(s_{i+1}), \zeta_k^*\rangle)\Big\}\nabla_\theta \log_{\pi_{\theta_k}}(a_i|s_i)\right]}_{T_1} \\ &\quad + \underbrace{\mathbb{E}_u\left[\frac{1}{B}\sum_{i=1}^{B}\Big\{r(s_i,a_i) - \eta_k^* + V^{\pi_{\theta_k}}(s_i) - V^{\pi_{\theta_k}}(s_{i+1})\Big\}\nabla_\theta \log_{\pi_{\theta_k}}(a_i|s_i)\right] - \nabla_\theta J(\theta_k)}_{T_2} \end{aligned}$$

We have $T_2 = 0$ as a consequence of the Bellman's equation (6). Whereas,

$$\|T_1\|^2 \le 2\,\mathbb{E}_{d^{\pi_{\theta_k}}}\, |\langle \phi(s_i), \zeta_k^*\rangle - V^{\pi_{\theta_k}}(s_i)|^2 G_1^2 \le 4G_1^2 \epsilon_{\mathrm{app}}.$$

Finally, note that

$$\begin{aligned} T_0 &= \mathbb{E}_u\left[\frac{1}{B}\sum_{i=1}^{B}\Big\{\eta_k^* - \eta_k + \langle \phi(s_{i+1}) - \phi(s_i), \zeta_k - \zeta_k^*\rangle\Big\}\nabla_\theta \log_{\pi_{\theta_k}}(a_i|s_i)\right] \\ &= \mathbb{E}_u\left[\frac{1}{B}\sum_{i=1}^{B}\Big\{\eta_k^* - \eta_k + \langle \phi(s_{i+1}) - \phi(s_i), \Pi(\zeta_k - \zeta_k^*)\rangle\Big\}\nabla_\theta \log_{\pi_{\theta_k}}(a_i|s_i)\right] \end{aligned}$$

which yields

$$\|T_0\|^2 \le G_1^2 \left(\|\eta_k - \eta_k^*\|^2 + \|\Pi(\zeta_k - \zeta_k^*)\|^2\right) = G_1^2 \|\Pi(\xi_k - \xi_k^*)\|^2. \tag{89}$$

Moreover, from Lemmas 8 and 9, we have

$$\left\|\mathbb{E}_{u,z}\left[\frac{1}{B}\sum_{i=1}^{B} b_u(\theta_k, \xi_k, z_i)\right] - \mathbb{E}_u\left[\frac{1}{B}\sum_{i=1}^{B} b_u(\theta_k, \xi_k, z_i)\right]\right\| \le \mathcal{O}\left(\frac{\sqrt{d}G_1^2\|\Pi\xi_k\|^2 C_{\mathrm{tar}}}{B}\right) \tag{90}$$

and

$$\mathbb{E}_{u,z}\left\|\left[\frac{1}{B}\sum_{i=1}^{B} b_u(\theta_k, \xi_k, z_i)\right] - \mathbb{E}_u\left[\frac{1}{B}\sum_{i=1}^{B} b_u(\theta_k, \xi_k, z_i)\right]\right\|^2 \le \mathcal{O}\left(\frac{\sqrt{d}G_1^2\|\Pi\xi_k\|^2 C_{\mathrm{tar}}}{B}\right). \tag{91}$$

It follows that

$$\begin{aligned}&\mathbb{E}\left\|\frac{1}{B}\sum_{i=1}^{B} b_u(\theta_k, \xi_k, z_i) - \nabla_\theta J(\theta_k)\right\|^2\\ &\quad\le \mathcal{O}\left(\frac{\sqrt{d}G_1^2\,\mathbb{E}\,\|\Pi\xi_k\|^2 C_{\mathrm{tar}}}{B} + G_1^2\,\mathbb{E}\,\|\Pi(\xi_k - \xi_k^*)\|^2 + G_1^2\epsilon_{\mathrm{app}}\right)\end{aligned} \tag{92}$$

and

$$\begin{aligned}&\left\|\mathbb{E}_{u,z}\left[\frac{1}{B}\sum_{i=1}^{B} b_u(\theta_k, \xi_k, z_i)\right] - \nabla_\theta J(\theta_k)\right\|^2\\ &\quad\le \mathcal{O}\left(\frac{dG_1^4\,\mathbb{E}\,\|\Pi\xi_k\|^2 C_{\mathrm{tar}}^2}{B^2} + G_1^2\,\mathbb{E}\,\|\Pi(\xi_k - \xi_k^*)\|^2 + G_1^2\epsilon_{\mathrm{app}}\right)\end{aligned} \tag{93}$$

Conditions (A_3) and (A_4) now follow. In contrast to the critic analysis, we employ a sharper bound for (A_5), which is necessary to obtain order-optimal regret. We derive this below.

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{B}\sum_{i=1}^{B} b_u(\theta_k, \xi_k, z_i)\right] - \nabla_\theta J(\theta_k)\\ &= \mathbb{E}_u\left[\mathbb{E}_v\left[\frac{1}{B}\sum_{i=1}^{B} b_u(\theta_k, \xi_k, z_i)\right]\right] - \nabla_\theta J(\theta_k)\\ &\overset{(a)}{=} \underbrace{\mathbb{E}_u\left[\frac{1}{B}\sum_{i=1}^{B}\left\{\eta_k^* - \mathbb{E}_v[\eta_k] + \langle\phi(s_{i+1}) - \phi(s_i), \mathbb{E}_v[\zeta_k] - \zeta_k^*\rangle\right\}\nabla_\theta \log_{\pi_{\theta_k}}(a_i|s_i)\right]}_{T_0} +\\ &+ \underbrace{\mathbb{E}_u\left[\frac{1}{B}\sum_{i=1}^{B}\left\{(\langle\phi(s_i), \zeta_k^*\rangle - V^{\pi_{\theta_k}}(s_i)) + (V^{\pi_{\theta_k}}(s_{i+1}) - \langle\phi(s_{i+1}), \zeta_k^*\rangle)\right\}\nabla_\theta \log_{\pi_{\theta_k}}(a_i|s_i)\right]}_{T_1}\\ &+ \underbrace{\mathbb{E}_u\left[\frac{1}{B}\sum_{i=1}^{B}\left\{r(s_i, a_i) - \eta_k^* + V^{\pi_{\theta_k}}(s_i) - V^{\pi_{\theta_k}}(s_{i+1})\right\}\nabla_\theta \log_{\pi_{\theta_k}}(a_i|s_i)\right] - \nabla_\theta J(\theta_k)}_{T_2},\end{aligned} \tag{94}$$

With the above decomposition, we obtain

$$\begin{aligned}&\left\|\mathbb{E}\left[\frac{1}{B}\sum_{i=1}^{B} b_u(\theta_k, \xi_k, z_i)\right] - \nabla_\theta J(\theta_k)\right\|^2\\ &\quad\le \mathcal{O}\left(\frac{dG_1^4\,\mathbb{E}\,\|\Pi\xi_k\|^2 C_{\mathrm{tar}}^2}{B^2} + G_1^2\|\Pi(\mathbb{E}[\xi_k] - \xi_k^*)\|^2 + G_1^2\epsilon_{\mathrm{app}}\right)\end{aligned} \tag{95}$$

This concludes the verification of condition (A_5). Note that the bound involves the term $\|\Pi(\mathbb{E}[\xi_k] - \xi_k^*)\|^2$, rather than $\mathbb{E}[\|\Pi(\xi_k - \xi_k^*)\|^2]$, making it significantly sharper. We can now invoke Theorem 4 by setting the NPG step-size $\gamma := \frac{\mu}{G_1^2}$ to derive bounds on both the second-order error and the bias of the NPG estimate ω_k, as follows.

$$
\begin{aligned}
&\mathbb{E}\left[\|\omega_k - \omega_k^*\|^2\right] \\
&\leq \mathcal{O}\Bigg(\exp\left(-\tfrac{H\mu^2}{4(1+4C)G_1}\right)\|\omega_0 - \omega_k^*\|^2 + \frac{\sqrt{d}G_1^6 C_{\mathrm{tar}}}{\mu^2 B} + \mathbb{E}\,\|\Pi(\xi_k - \xi_k^*)\|^2 + \epsilon_{\mathrm{app}} \\
&\quad + \frac{\sqrt{d}G_1^4 C_{\mathrm{tar}} C^2}{\mu^2 B} + \frac{dG_1^4 C^2 C_{\mathrm{tar}}^2}{\mu^4 B^2} + \frac{dG_1^8 C_{\mathrm{tar}}^2}{\mu^2 B^2} + G_1^2\mu^{-2}\,\mathbb{E}\,\|\Pi(\xi_k - \xi_k^*)\|^2 + G_1^2\mu^{-2}\epsilon_{\mathrm{app}} \Bigg) \\
&\overset{(a)}{\leq} \mathcal{O}\Bigg(\exp\left(-\tfrac{H\mu^2}{4(1+4C)G_1}\right)\|\omega_0 - \omega_k^*\|^2 + \frac{\sqrt{d}G_1^6 C_{\mathrm{tar}}}{\mu^2 B} + G_1^2\mu^{-2}\exp\left(-\tfrac{H\lambda^3}{16}\right)\|\xi_1 - \xi^*\|^2 \\
&\quad + G_1^2\mu^{-2}\frac{C_{\mathrm{tar}}\sqrt{m}}{\lambda^4 B} + G_1^2\mu^{-2}\frac{C_{\mathrm{tar}}\sqrt{m}}{\lambda B} + G_1^2\mu^{-2}\frac{C_{\mathrm{tar}}^2 m}{\lambda^6 B^2} + G_1^2\mu^{-2}\frac{C_{\mathrm{tar}}^2 m}{\lambda^2 B^2} + G_1^2\mu^{-2}\epsilon_{\mathrm{app}} \\
&\quad + \frac{\sqrt{d}G_1^4 C_{\mathrm{tar}} C^2}{\mu^2 B} + \frac{dG_1^4 C^2 C_{\mathrm{tar}}^2}{\mu^4 B^2} + \frac{dG_1^8 C_{\mathrm{tar}}^2}{\mu^2 B^2} + G_1^2\mu^{-2}\epsilon_{\mathrm{app}} \Bigg),
\end{aligned} \tag{96}
$$

where (a) follows using the second-order bound in Theorem 2 and

$$
\begin{aligned}
\|\mathbb{E}[\omega_k] - \omega_k^*\|^2 &\leq \mathcal{O}\Bigg(\exp\left(-\tfrac{H\mu^2}{4(1+4C)G_1}\right)\|\omega_0 - \omega_k^*\|^2 + \frac{dG_1^8 C_{\mathrm{tar}}^2}{\mu^4 B^2} + \frac{dG_1^8 C_{\mathrm{tar}}^2}{\mu^6 B^2} + \frac{dG_1^6 C_{\mathrm{tar}}^2}{\mu^4 B^2} \\
&\quad + G_1^2\mu^{-2}\|\Pi(\mathbb{E}[\xi_k] - \xi_k^*)\|^2 + G_1^2\mu^{-2}\epsilon_{\mathrm{app}} \Bigg) \\
&\overset{(b)}{\leq} \mathcal{O}\Bigg(\exp\left(-\tfrac{H\mu^2}{4(1+4C)G_1}\right)\|\omega_0 - \omega_k^*\|^2 + \frac{dG_1^8 C_{\mathrm{tar}}^2}{\mu^4 B^2} + \frac{dG_1^8 C_{\mathrm{tar}}^2}{\mu^6 B^2} + \frac{dG_1^6 C_{\mathrm{tar}}^2}{\mu^4 B^2} \\
&\quad + G_1^2\mu^{-2}\exp\left(-\tfrac{H\lambda^3}{16}\right)\|\xi_1 - \xi^*\|^2 + G_1^2\mu^{-2}\frac{C_{\mathrm{tar}}^2 m\|\xi_1 - \xi^*\|^2}{\lambda^2 B^2} \\
&\quad + G_1^2\mu^{-2}\frac{C_{\mathrm{tar}}^2 m}{\lambda^6 B^2} + G_1^2\mu^{-2}\frac{C_{\mathrm{tar}}^2 m}{\lambda^2 B^2} + G_1^2\mu^{-2}\epsilon_{\mathrm{app}} \Bigg),
\end{aligned} \tag{97}
$$

where (b) follows using the bias bound in Theorem 2 and

G Proof of Lemma 1

The regret can be decomposed as follows, following the standard approach used in [38, 9]:

$$
\begin{aligned}
\mathrm{Reg}_T &= \sum_{t=0}^{T-1}\left(J^* - r(s_t, a_t)\right) \\
&= HB\sum_{k=1}^{K}\left(J^* - J(\theta_k)\right) + \sum_{k=1}^{K}\sum_{t\in\mathcal{I}_k}\left(J(\theta_k) - r(s_t, a_t)\right) \\
&= HB\sum_{k=1}^{K}\left(J^* - J(\theta_k)\right) + \mathbb{E}\left[\sum_{k=1}^{K} V^{\pi_{\theta_{k+1}}}(s_{kH}) - V^{\pi_{\theta_k}}(s_{kH})\right] + \mathbb{E}\left[V^{\pi_{\theta_K}}(s_T) - V^{\pi_{\theta_0}}(s_0)\right].
\end{aligned} \tag{98}
$$

Since $0 \le V^{\pi}(s) \le 2C$ for all $\pi \in \Pi$ and $s \in \mathcal{S}$ from Lemma 7, it follows that

$$\mathbb{E}[\mathrm{Reg}_T] \le HB \sum_{k=1}^{K} (J^* - \mathbb{E}[J(\theta_k)]) + 2C(K+1). \tag{99}$$

We now decompose the term $\sum_{k=1}^{K}(J^* - \mathbb{E}[J(\theta_k)])$. There are several existing results for this. We use Lemma 1 from [16], re-stated below, since it is sharper than most existing results.

Lemma 16. *Consider any policy update rule of form*

$$\theta_{k+1} = \theta_k + \alpha\omega_k. \tag{100}$$

If 3 holds, then the following inequality is satisfied for any K.

$$\begin{aligned} J^* - \frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}[J(\theta_k)] \le & \sqrt{\epsilon_{\text{bias}}} + \frac{G_1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left\|(\mathbb{E}_k\left[\omega_k\right] - \omega_k^*)\right\| + \frac{\alpha G_2}{2K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\omega_k\right\|^2 \\ & + \frac{1}{\alpha K}\mathbb{E}_{s\sim d^{\pi^*}}[\mathrm{KL}(\pi^*(\cdot|s)\|\pi_{\theta_0}(\cdot|s))] \end{aligned} \tag{101}$$

where $\mathrm{KL}(\cdot\|\cdot)$ is the Kullback-Leibler divergence, ω_k^* is the NPG direction $F(\theta_k)^{-1}\nabla J(\theta_k)$, π^* is the optimal policy, J^* is the optimal value of the function $J(\cdot)$, and $\mathbb{E}_k[\cdot]$ denotes conditional expectation given the history up to epoch k.

Using the above result, we obtain the following.

$$\begin{aligned} HB\sum_{k=0}^{K-1}(J^* - \mathbb{E}[J(\theta_k)]) \le & T\sqrt{\epsilon_{\text{bias}}} + \frac{BHG_1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|(\mathbb{E}_k\left[\omega_k\right] - \omega_k^*)\right\| + \frac{\alpha G_2 BH}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\omega_k\right\|^2 \\ & + \frac{BH}{\alpha}\mathbb{E}_{s\sim d^{\pi^*}}[\mathrm{KL}(\pi^*(\cdot|s)\|\pi_{\theta_0}(\cdot|s))] \end{aligned} \tag{102}$$

The term containing $\mathbb{E}\left\|\omega_k\right\|^2$ can be further decomposed as

$$\begin{aligned} \frac{\alpha G_2 BH}{K}\mathbb{E}\left\|\omega_k\right\|^2 & \le \frac{2\alpha G_2 BH}{K}\mathbb{E}\left\|\omega_k - \omega_k^*\right\|^2 + \frac{2\alpha G_2 BH}{K}\mathbb{E}\left\|\omega_k^*\right\|^2 \\ & \overset{(a)}{\le} \frac{2\alpha G_2 BH}{K}\mathbb{E}\left\|\omega_k - \omega_k^*\right\|^2 + \frac{2\alpha G_2 BH}{\mu^2 K}\mathbb{E}\left\|\nabla_\theta J(\theta_k)\right\|^2 \end{aligned} \tag{103}$$

where (a) follows from Assumption 4 and the definition that $\omega_k^* = F(\theta_k)^{-1}\nabla_\theta J(\theta_k)$. The result now follows by substituting (102) and (103) in (99).

H Proof of Theorem 1

We shall now derive a bound for $\sum_{k=0}^{K-1}\|\nabla_\theta J(\theta_k)\|^2$. Given that the function J is L-smooth, we obtain:

$$
\begin{aligned}
&J(\theta_{k+1})\\
&\geq J(\theta_k)+\langle\nabla_\theta J(\theta_k),\theta_{k+1}-\theta_k\rangle-\frac{L}{2}\|\theta_{k+1}-\theta_k\|^2\\
&=J(\theta_k)+\alpha\left\langle\nabla_\theta J(\theta_k),\omega_k\right\rangle-\frac{\alpha^2L}{2}\|\omega_k\|^2\\
&=J(\theta_k)+\alpha\left\langle\nabla_\theta J(\theta_k),\omega_k^*\right\rangle+\alpha\left\langle\nabla_\theta J(\theta_k),\omega_k-\omega_k^*\right\rangle-\frac{\alpha^2L}{2}\|\omega_k-\omega_k^*+\omega_k^*\|^2\\
&\overset{(a)}{\geq} J(\theta_k)+\alpha\left\langle\nabla_\theta J(\theta_k),F(\theta_k)^{-1}\nabla_\theta J(\theta_k)\right\rangle+\alpha\left\langle\nabla_\theta J(\theta_k),\omega_k-\omega_k^*\right\rangle\\
&\qquad-\alpha^2L\|\omega_k-\omega_k^*\|^2-\alpha^2L\|\omega_k^*\|^2\\
&\overset{(b)}{\geq} J(\theta_k)+\frac{\alpha}{G_1^2}\|\nabla_\theta J(\theta_k)\|^2+\alpha\left\langle\nabla_\theta J(\theta_k),\omega_k-\omega_k^*\right\rangle-\alpha^2L\|\omega_k-\omega_k^*\|^2-\alpha^2L\|\omega_k^*\|^2\\
&=J(\theta_k)+\frac{\alpha}{2G_1^2}\|\nabla_\theta J(\theta_k)\|^2+\frac{\alpha}{2G_1^2}\|\nabla_\theta J(\theta_k)+G_1^2(\omega_k-\omega_k^*)\|^2-\left(\frac{\alpha G_1^2}{2}+\alpha^2L\right)\|\omega_k-\omega_k^*\|^2\\
&\qquad-\alpha^2L\|\omega_k^*\|^2\\
&\geq J(\theta_k)+\frac{\alpha}{2G_1^2}\|\nabla_\theta J(\theta_k)\|^2-\left(\frac{\alpha G_1^2}{2}+\alpha^2L\right)\|\omega_k-\omega_k^*\|^2-\alpha^2L\|F(\theta_k)^{-1}\nabla_\theta J(\theta_k)\|^2\\
&\overset{(c)}{\geq} J(\theta_k)+\left(\frac{\alpha}{2G_1^2}-\frac{\alpha^2L}{\mu^2}\right)\|\nabla_\theta J(\theta_k)\|^2-\left(\frac{\alpha G_1^2}{2}+\alpha^2L\right)\|\omega_k-\omega_k^*\|^2
\end{aligned}
\tag{104}
$$

where (a) use the Cauchy-Schwarz inequality and the definition that $\omega_k^*=F(\theta_k)^{-1}\nabla_\theta J(\theta_k)$. Relations (b), and (c) are consequences of Assumption 3(a) and 4 respectively. Summing the above inequality over $k\in\{0,\cdots,K-1\}$, rearranging the terms and substituting $\alpha=\frac{\mu^2}{4G_1^2L}$, we obtain

$$
\begin{aligned}
\frac{\mu^2}{16G_1^4L}\left(\sum_{k=0}^{K-1}\|\nabla_\theta J(\theta_k)\|^2\right)&\leq J(\theta_K)-J(\theta_0)+\left(\frac{\mu^2}{8L}+\frac{\mu^4}{16G_1^4L}\right)\left(\sum_{k=0}^{K-1}\|\omega_k-\omega_k^*\|^2\right)\\
&\overset{(a)}{\leq} 2+\left(\frac{\mu^2}{8L}+\frac{\mu^4}{16G_1^4L}\right)\left(\sum_{k=0}^{K-1}\|\omega_k-\omega_k^*\|^2\right)
\end{aligned}
\tag{105}
$$

where (a) uses the fact that $J(\cdot)$ is absolutely bounded above by 1. Inequality (105) can be simplified as follows.

$$
\begin{aligned}
&\left(\sum_{k=0}^{K-1}\mathbb{E}\,\|\nabla_\theta J(\theta_k)\|^2\right)\\
&\qquad\leq\frac{32LG_1^4\mu^2}{\mu^4}+\left(2G_1^4+\mu^2\right)\left(\sum_{k=0}^{K-1}\mathbb{E}\,\|\omega_k-\omega_k^*\|^2\right)\\
&\qquad\leq\frac{32LG_1^4\mu^2}{\mu^4}+\left(2G_1^4+\mu^2\right)K\Bigg(\exp\left(-\tfrac{H\mu^2}{4(1+4C)G_1}\right)\|\omega_0-\omega_k^*\|^2\\
&\qquad+G_1^2\mu^{-2}\exp\left(-\tfrac{H\lambda^3}{16}\right)\|\xi_1-\xi^*\|^2+G_1^2\mu^{-2}\frac{C_{\rm tar}\sqrt{m}}{\lambda^4B}+G_1^2\mu^{-2}\frac{C_{\rm tar}^2m}{\lambda^6B^2}\\
&\qquad+\frac{\sqrt{d}G_1^4C_{\rm tar}C^2}{\mu^2B}+\frac{dG_1^4C^2C_{\rm tar}^2}{\mu^4B^2}+\frac{dG_1^8C_{\rm tar}^2}{\mu^2B^2}+G_1^2\mu^{-2}\epsilon_{\rm app}\Bigg)
\end{aligned}
\tag{106}
$$

.

Furthermore, using $\alpha = \frac{\mu^2}{4G_1^2 L}$, we find that $HB\sum_{k=0}^{K-1}(J^* - \mathbb{E}[J(\theta_k)])$ can be written as follows.

$$
\begin{aligned}
&\mathbb{E}[\mathrm{Reg}_T] \\
&\leq T\sqrt{\epsilon_{\mathrm{bias}}} + TG_1\,\mathbb{E}\,\|(\mathbb{E}_k[\omega_k] - \omega_k^*)\| + \frac{\mu^2 G_2 T}{4G_1^2 L}\,\mathbb{E}\,\|\omega_k\|^2 \\
&+ \frac{4G_1^2 L\sqrt{T}\log T}{\mu^2}\,\mathbb{E}_{s\sim d^{\pi^*}}[\mathrm{KL}(\pi^*(\cdot|s)\|\pi_{\theta_0}(\cdot|s))] \\
&\leq T\sqrt{\epsilon_{\mathrm{bias}}} + TG_1\,\mathbb{E}\,\|(\mathbb{E}_k[\omega_k] - \omega_k^*)\| + \frac{\mu^2 G_2 T}{2G_1^2 L}\,\mathbb{E}\,\|\omega_k - \omega_k^*\|^2 + \frac{G_2 T}{2G_1^2 L}\,\mathbb{E}\,\|\nabla_\theta J(\theta_k)\|^2 \\
&+ \frac{4G_1^2 L\sqrt{T}\log T}{\mu^2}\,\mathbb{E}_{s\sim d^{\pi^*}}[\mathrm{KL}(\pi^*(\cdot|s)\|\pi_{\theta_0}(\cdot|s))] + C(K+1).
\end{aligned}
$$

Using the bound on $\mathbb{E}\|\nabla_\theta J(\theta_k)\|^2$ derived earlier, along with the second-order error and bias of the NPG estimates, including all constants, from (96) and (97), we obtain the following regret bound, omitting all non-dominant terms.

$$
\begin{aligned}
&\mathbb{E}[\mathrm{Reg}_T] \\
&\leq \mathcal{O}\Bigg(G_2 G_1^2\sqrt{T}(\log T)\left[\frac{\mu^2 + L}{\mu^2 L}\right] + \frac{G_2\sqrt{T}}{G_1^2 L}\left[\frac{G_1^2 C_{\mathrm{tar}}\sqrt{m}}{\mu^2\lambda^4} + \frac{\sqrt{d}G_1^4 C_{\mathrm{tar}} C^2}{\mu^2}\right] \\
&+ \sqrt{T}\left[\frac{\sqrt{d}G_1^5 C_{\mathrm{tar}}}{\mu^3} + \frac{C_{\mathrm{tar}} G_1^2\sqrt{m}R_0}{\mu\lambda} + \frac{G_1^2 C_{\mathrm{tar}}\sqrt{m}}{\mu\lambda^3}\right] + \frac{TG_1^2\sqrt{\epsilon_{\mathrm{app}}}}{\mu} + T\sqrt{\epsilon_{\mathrm{bias}}}\Bigg).
\end{aligned}
\tag{107}
$$

The regret bound above is obtained by setting the learning rates as follows: policy update rate $\alpha = \frac{\mu^2}{4G_1^2 L}$, critic learning rates $(\beta, c_\beta) = \left(\frac{\lambda^2}{2}, \lambda + \sqrt{\frac{1}{\lambda^2} - 1}\right)$ and NPG estimation rate $\gamma = \frac{\mu}{G_1^2}$.

Remark 2 (Scalability of C_{hit} and C_{tar})**.** *The constants C_{hit} and C_{tar} reflect the difficulty of exploration under a given policy class and may scale with structural properties of the MDP such as state-space size or connectivity. Such dependence is inherent, as any algorithm must visit all states to avoid constant per-iteration regret, and even in ergodic MDPs the mixing time t_{mix} can scale with the state space. Our result extends existing ergodic-MDP guarantees to the more general unichain setting without degrading these constants. Specifically, while state-of-the-art actor–critic methods achieve complexity $O(t_{\mathrm{mix}}^3\sqrt{T})$ [16], our bound $O(C^3\sqrt{T})$ involves a constant C that characterizes the Cesàro mixing time [22, Sec. 6.6], which satisfies $C \leq 7t_{\mathrm{mix}}$ for ergodic chains [22, Ex. 6.11]. In certain Markov chains, such as the biased random walk on the n-cycle [22, Ex. 24.2], the Cesàro mixing time is only $O(n)$ versus the standard mixing time which is $\Theta(n^2)$, illustrating that averaging distributions over time accelerates convergence compared to single-step mixing.*

Remark 3 (On the Unichain Assumption)**.** *To the best of our knowledge, no existing policy gradient methods (with general policy parametrization) have theoretical guarantees for average-reward, infinite-horizon MDPs beyond the unichain case. Our work is the first to establish such guarantees under the general unichain assumption, which is the weakest known condition under which the policy gradient theorem holds [33]. While value-based methods admit guarantees beyond this setting, policy gradient approaches require fundamentally different analyses, making existing techniques inapplicable. A possible, though still conjectural, direction for relaxing the unichain assumption is to restrict policy search to unichain-inducing policies by maintaining strong exploration initially and annealing it as the algorithm converges. Notably, weakly communicating MDPs, the most general class solvable from a single stream of experience, always admit an optimal unichain policy [30].*