Beyond Exponentially Fast Mixing in Average-Reward Reinforcement Learning via Multi-Level Monte Carlo Actor-Critic

Wesley A. Suttle [*1] **Amrit Singh Bedi** [*2] **Bhrij Patel** [2] **Brian M. Sadler** [1] **Alec Koppel** [3] **Dinesh Manocha** [2]

Abstract

Many existing reinforcement learning (RL) methods employ stochastic gradient iteration on the back end, whose stability hinges upon a hypothesis that the data-generating process mixes exponentially fast with a rate parameter that appears in the step-size selection. Unfortunately, this assumption is violated for large state spaces or settings with sparse rewards, and the mixing time is unknown, making the step size inoperable. In this work, we propose an RL methodology attuned to the mixing time by employing a multi-level Monte Carlo estimator for the critic, the actor, and the average reward embedded within an actor-critic (AC) algorithm. This method, which we call **M**ulti-level **A**ctor-**C**ritic (MAC), is developed specifically for infinite-horizon average-reward settings and neither relies on oracle knowledge of the mixing time in its parameter selection nor assumes its exponential decay; it is therefore readily applicable to applications with slower mixing times. Nonetheless, it achieves a convergence rate comparable to SOTA actor-critic algorithms. We experimentally show that these alleviated restrictions on the technical conditions required for stability translate to superior performance in practice for RL problems with sparse rewards.

1. Introduction

Modern machine learning (ML) techniques have enabled analyzing and making predictions from large-scale data. This is achieved through backpropagation in neural networks (Hinton et al., 2006), cloud processing of industrial data sets (McAfee et al., 2012), complex event simulators (Silver et al., 2016), and deep feature extraction (Krizhevsky et al., 2017), among other innovations. However, a crucial underlying aspect of these developments is whether training data is sufficiently informative. To put this in quantitative terms, most ML training mechanisms hinge upon training samples being independent and identically distributed (i.i.d.), which is often violated in real-world problems, such as natural language (Liu et al., 2021), financial markets (Heaton et al., 2016), and robotics (Gu et al., 2016), where data exhibits temporal dependence. Reinforcement learning (RL) algorithms, in particular, are limited by this constraint, as the data is inherently Markovian, owing to the fact that the RL problem is most commonly represented mathematically as a Markov Decision Process (MDP) (Sutton, 1988). For this reason, as well as the numerous applications of RL in recent years (Li, 2019), we focus on algorithms for RL methods when data exhibits Markovian dependence.

Under Markovian sampling, many convergence analyses of iterative methods for RL exist (Qiu et al., 2021b; Xu et al., 2020b) and typically consider a critical assumption about the rate at which the MDP's transition dynamics converge to the stationary distribution for a fixed policy. To establish rigorous analyses, restrictions are typically placed on the *mixing times* encountered during training: (1) prior oracle knowledge of mixing times is employed to determine an optimal step-size selection, as in (Duchi et al., 2012; Nagaraj et al., 2020); or (2) mixing times decay exponentially fast (Xu et al., 2020b; Wu et al., 2020; Qiu et al., 2021a; Chen & Zhao, 2022). In this work, we are interested in developing RL algorithms with performance certificates without the aforementioned conditions.

For instance, consider an RL problem where the agent must navigate through a continuous state space, such as a robot reaching a target location or a self-driving car traversing a complex road network. In these cases, the transition dynamics can be highly non-linear with sparse rewards, and the agent may have to explore many states before locating any rewards. In addition, if the environment's dynamics are highly random or there are many obstacles and the agent can get stuck in certain states for a long time, the total variation distance to the steady state decreases slowly, i.e., the

[*]Equal contribution [1]U.S. Army Research Laboratory, Adelphi, MD, USA. [2]Department of Computer Science, University of Maryland, College Park, USA. [3]JP Morgan Chase AI Research, USA. Correspondence to: Wesley A. Suttle <wesley.a.suttle.ctr@army.mil>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Table 1. This table compares the total sample complexity of actor-critic (AC) algorithms available in the literature. To our knowledge, this is the first AC algorithm with an explicit dependence on the underlying mixing time, defined as $\tau_{mix} := \max_{t\in[T]} \tau_{mix}^{\theta_t}$ where θ is the policy parameter (see Sec. 4 for details), that does not require the exponentially fast mixing assumption. We also remark that our proposed approach is oblivious to mixing time if we follow Dorfman & Levy (2022) to let $T_{\max} = T$ in Algorithm 1.

References	Sampling		Total complexity	Reward	Fast mixing
	Actor	Critic			
(Wang et al., 2019)	i.i.d.	i.i.d.	$\mathcal{O}(\epsilon^{-4})$	Discounted	Required
(Kumar et al., 2019)	i.i.d.	i.i.d.	$\mathcal{O}(\epsilon^{-4})$	Discounted	Required
(Qiu et al., 2021a)	i.i.d.	Markovian	$\tilde{\mathcal{O}}(\epsilon^{-3})$	Average	Required
(Xu et al., 2020b)	Markovian	Markovian	$\tilde{\mathcal{O}}(\epsilon^{-2})$	Discounted	Required
(Wu et al., 2020)	Markovian	Markovian	$\tilde{\mathcal{O}}(\epsilon^{-2.5})$	Average	Required
(Chen & Zhao, 2022)	Markovian	Markovian	$\tilde{\mathcal{O}}(\epsilon^{-2})$	Average	Required
This work	Markovian	Markovian	$\tilde{\mathcal{O}}(\tau_{\text{mix}}^2 \cdot \epsilon^{-2})$	Average	Not required

mixing rate for a given policy is slow. These issues often manifest in stationary MDPs that are simply weakly connected by a few distinct regions, which could be defined, e.g., by seasonality in data or distinct learning "tasks" comprised of similar states and sub-goals as detailed in Riemer et al. (2021). In summary, many RL environments exhibit a slower than exponential mixing rate due to high dimensionality, intrinsic volatility, sparse rewards, or that they contain distinct sub-tasks.

We seek RL methodologies attuned to environments that mix slowly, especially in the context of actor-critic (AC), due to the fact that it underlies much of modern deep RL (Konda & Tsitsiklis, 1999). As previously noted, existing results (cf. Table 1) hinge upon either i.i.d. (Kumar et al., 2019) or exponentially fast mixing (Xu et al., 2020b; Wu et al., 2020; Qiu et al., 2021a; Chen & Zhao, 2022). We therefore aim to develop a variant of actor-critic that does not possess these limitations. To do so, inspired by Dorfman & Levy (2022), we develop a multi-level Monte Carlo gradient estimator and adaptive learning rate for the average reward, actor, *and* critic, called Multi-level Monte Carlo Actor-Critic (MAC). We compare the sample complexity of different methods in Table 1. Our main contributions are:

- We develop a variant of multi-level Monte Carlo for the average reward, policy gradient, and temporal difference estimates, which together comprise Multi-level Monte Carlo Actor-Critic (MAC) algorithm.
- We establish the convergence rate dependence of the proposed MAC algorithm on the mixing time without any assumption on its decay rate, which alleviates the exponentially fast mixing assumptions of prior works.
- Despite the two-timescale nature of MAC, our use of a modified Adagrad stepsize in the actor allows us to obtain final sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-2})$, instead of the $\tilde{\mathcal{O}}(\epsilon^{-2.5})$ of previous two-timescale analyses.
- We perform initial proof of concept experiments and observe that MAC outperforms vanilla actor-critic for settings with sparse rewards.

1.1. Related Works

We provide a brief overview of the related works here. Please refer to Appendix A for more detailed context.

TD Learning. For discounted TD with Markovian samples, Bhandari et al. (2018) established finite-time convergence bounds which scale linearly with mixing time τ_{mix}. Dorfman & Levy (2022) then improved the rate to be proportional to the $\sqrt{\tau_{mix}}$ using a multi-level gradient estimator and adaptive learning rate. Qiu et al. (2021a) studied TD under the average reward setting, which also imposes exponentially fast mixing that manifests in an additional logarithmic term in the sample complexity. These results all hinge upon imposing restrictive conditions on mixing time.

Policy Gradient. More recently, the sample complexity of policy gradient methods has been established for a variety of settings: for tabular (Bhandari & Russo, 2019; Agarwal et al., 2020) and softmax policies (Mei et al., 2020), rates to global optimality have been proven. For general parameterized policies, early works focused on "policy improvement" bounds (Pirotta et al., 2013; 2015). More recently, rates towards stationarity (Bedi et al., 2022) and local extrema (Zhang et al., 2020) have been studied, and under special neural architectures, globally optimal solutions (Wang et al., 2019; Leahy et al., 2022) are achievable. This topic is an active area of work – we merely identify that these performance certificates all require the mixing rate going to null exponentially fast.

Actor-Critic. As previously mentioned, the stability of actor-critic initially focused on asymptotics (Borkar & Konda, 1997). More recently, non-asymptotic rates have been derived under i.i.d. assumptions (Kumar et al., 2019;

Wang et al., 2019) and more recently under a variety of different types of Markovian data – see Table 1. However, these results impose that any temporal correlation of data across time vanishes exponentially fast as quantified by the mixing rate. In this way, we are able to match (Chen & Zhao, 2022) but without these restrictions.

2. Problem Formulation

We consider a reinforcement learning problem with average reward criterion, which can be mathematically defined as a Markov Decision Process (MDP) given by the tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, \mathbb{P}, r)$. Here, $\mathcal{S}$ is a finite state space; $\mathcal{A}$ is a finite action space; $\mathbb{P}(\cdot \mid s, a)$ is a distribution that determines transition to the next state s', and $r : \mathcal{S} \times \mathcal{A} \to [0, r_{\max}]$ is a bounded reward function that informs the merit of selecting action a when starting in state s. A policy $\pi(\cdot \mid s)$ of an MDP maps the state s to the probability distribution over actions a. Formally, $\pi : \mathcal{S} \to \triangle_{|\mathcal{A}|}$, where $\triangle_{|\mathcal{A}|}$ is the probability simplex. In the average reward setting, we seek to find a policy π such that the long-term average reward is given by $J(\pi) := \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T} r(s_t, a_t)\right]$ is maximized. In practice, when the state space is large, it is difficult to search over a general class of policies since its parameterization scales with $|\mathcal{S}|$. Therefore, we restrict focus to the case that π is parameterized by a vector $\theta \in \mathbb{R}^d$, where d denotes the parameter dimension, which leads to the notion of a parameterized policy π_θ. Optimizing the average reward with respect to policy parameters θ is the main goal of this work, which we formalize as:

$$\max_{\theta} J(\pi_\theta) := \lim_{T\to\infty} \mathbb{E}_{s_{t+1}\sim\mathbb{P}(\cdot|s_t,a_t), a_t\sim\pi_\theta(\cdot|s_t)}\left[R_T\right], \tag{1}$$

where $R_T := \frac{1}{T}\sum_{t=0}^{T} r(s_t, a_t)$. Denote by d^{π_θ} the unique stationary state distribution induced by policy π_θ. Then we can also write $J(\pi_\theta) = \mathbb{E}_{s\sim d^{\pi_\theta}, a\sim\pi_\theta}[r(s, a)]$. It turns to be essential to further algorithm development to define the differential action-value (Q) function as

$$Q^{\pi_\theta}(s, a) = \mathbb{E}\left[\sum_{t=0}^{\infty} [r(s_t, a_t) - J(\pi_\theta)]\right], \tag{2}$$

such that $s_0 = s, a_0 = a$, and action $a \sim \pi_\theta$. This implies that we can write the differential state value function as

$$V^{\pi_\theta}(s) = \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}[Q^{\pi_\theta}(s, a)]. \tag{3}$$

We will drop the term "differential" in what follows and simply refer to Q^{π_θ} and V^{π_θ} as the state-action and state value functions. From (2) and (3), we can write the value of a state s via Bellman's equation (Puterman, 2014):

$$V^{\pi_\theta}(s) = \mathbb{E}[r(s, a) - J(\pi_\theta) + V^{\pi_\theta}(s')], \tag{4}$$

where the expectation is over $a \sim \pi_\theta(\cdot|s), s' \sim \mathbb{P}(\cdot|a, s)$. Next, we shift to defining the standard actor-critic framework to solve (1), in order to illuminate its merits and drawbacks in terms of the conditions it imposes on the state-action occupancy measure, i.e., the product measure associated with the expectations in (1) and (4).

2.1. Decay Rates of Mixing Times

It is inherent to RL that the data-generating mechanism is Markovian, which means that assumptions that trajectory data is independent and identically distributed cannot hold (Wang et al., 2019; Kumar et al., 2019; Qiu et al., 2021b). That is, the noise driving the estimation error of the algorithm updates is heteroscedastic. Because of this challenge, various technical conditions have been considered to quantify the degree of correlation in data across time, mostly inherited from the applied probability literature – see (Levin & Peres, 2017). Most prior stability and sample complexity results of RL algorithms for the average reward setting are defined in terms of the *mixing time*, which is the minimum time at which the transition dynamics are near the long-term steady-state distribution induced by a policy π_θ, as formalized next.

Definition 2.1 (ϵ-Mixing Time)**.** Let d^{π_θ} denote the stationary distribution of the Markov chain induced by π_θ. The ϵ-mixing time of this Markov chain is defined as

$$\tau^{\theta}_{mix}(\epsilon) := \inf\{t : \sup_{s\in\mathcal{S}} \|P^t(\cdot|s) - d^{\pi_\theta}(\cdot)\|_{TV} \leq \epsilon\}, \tag{5}$$

where $\|\cdot\|_{TV}$ is the total variation distance. The conventional mixing time is defined as $\tau^{\theta}_{mix} := \tau^{\theta}_{mix}(1/4)$.

Limitations of Previous Work. In all the earlier works from Table 1, a crucial assumption is regarding the exponentially fast decay rate of the mixing time. Specifically, all these works assume that there exist $\zeta > 0$ and $\rho \in (0, 1)$ such that, for all θ, it holds that $\sup_{s\in\mathcal{S}} \|P^t(\cdot|s) - d^{\pi_\theta}\|_{TV} \leq \zeta\rho^t$, which implies exponentially fast mixing for all induced Markov chains. Also, to proceed with the convergence analysis in the works mentioned in Table 1, knowledge of ζ and ρ is required for the optimal step size selection, which is usually unknown in practice. Moreover, there is a wide range of applications where polynomial decay rates have some fundamental role to play in defining RL algorithms that can generalize well across tasks - see (Riemer et al., 2021) for a detailed description.

Therefore, in this work, we are interested in going beyond the exponentially fast mixing requirements and seek to develop actor-critic algorithms which do not require access to mixing time values a priori for optimal performance. We present our proposed algorithm in the next section.

3. Actor-Critic Method

3.1. Elements of Actor-Critic

We start by providing a quick recap of the standard actor-critic (AC) algorithm in average reward RL settings. The AC algorithm operates by alternating updates between the actor and critic, which are respectively defined in terms of gradient updates to policy parameters θ and estimates of the value function $V^{\pi_\theta}(s)$ based on the fixed point recursion implied by Bellman's equation (4). To do so, we proceed by writing down a gradient ascent iteration for the maximization in (1) given by

$$\theta_{t+1} = \theta_t + \alpha_t \nabla_\theta J(\pi_{\theta_t}), \tag{6}$$

where α_t is the actor learning rate. From the policy gradient (PG) Theorem (Williams, 1992; Sutton et al., 1999), it is well-known that $\nabla_\theta J(\pi_{\theta_t})$ takes the explicit form:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{(s,a,s')\sim\Gamma_\theta}\left[\delta^{\pi_\theta} \cdot \nabla_\theta \log \pi_\theta(a|s)\right], \tag{7}$$

with the *temporal difference* (TD) δ^{π_θ} defined as (Sutton, 1988):

$$\delta^{\pi_\theta} := r(s,a) - J(\pi_\theta) + V^{\pi_\theta}(s') - V^{\pi_\theta}(s), \tag{8}$$

and $\Gamma_\theta := s \sim d^{\pi_\theta}, a \sim \pi_\theta, s' \sim \mathbb{P}(\cdot|s,a)$ is the short notation for the joint distribution. We note that there are two parts in the expression of PG in (7): the score function $\nabla_\theta \log \pi_\theta(a|s)$, which comes from the policy parameterization, and the TD term δ^{π_θ}, obtained by rearranging the $V^{\pi_\theta}(s)$ term in (4) to the other side of the expression and grouping expectations.

Critic update: We restrict focus to the case that the value function $V^{\pi_\theta}(s)$, is estimated by the inner product between a given feature map $\phi(s)$ and a weight vector ω, which can be shown to be exact under some special cases such as linear MDPs where the assumption of realizability is met (Tsitsiklis & Van Roy, 1997; Bhandari et al., 2018; Dorfman & Levy, 2022; Qiu et al., 2021a). Hence, we can write $V_\omega(s) = \langle \phi(s), \omega \rangle$ where $V_\omega(s)$ denotes the estimator to $V^{\pi_\theta}(s)$ in terms of parameters $\omega \in \mathbb{R}^m$ and feature map $\phi : \mathcal{S} \to \mathbb{R}^m$ of state s to m-dimensional space such that $\|\phi(s)\| \leq 1$ for all $s \in \mathcal{S}$. TD learning is used to find ω, which minimizes error $G(\omega)$ defined by

$$\min_{\omega\in\Omega} G(\omega) := \sum_{s\in S} d^{\pi_\theta}(s)(V^{\pi_\theta}(s) - V_\omega(s))^2, \tag{9}$$

where we take $\Omega \subset \mathbb{R}^m$ to be a norm-ball of sufficiently large radius $R_\omega > 0$ about the origin. The TD(0) update for the critic parameter ω is given as

$$\begin{aligned}\omega_{t+1} =& \Pi_\Omega\big[\omega_t + \beta_t\big(r(s_t,a_t) - J(\pi_{\theta_t}) + \langle \phi(s_{t+1}), \omega_t\rangle \\ &- \langle \phi(s_t), \omega_t\rangle\big)\phi(s_t)\big],\end{aligned} \tag{10}$$

Algorithm 1 **M**ulti-level Monte Carlo **A**ctor-**C**ritic (MAC)

Initialize: Policy parameter θ_0, actor step size α_t, critic step size β_t, average reward tracking step size γ_t, initial state $s_0^0 \sim \mu_0(\cdot)$, maximum rollout length $T_{\max}$.
for $t = 0$ **to** $T-1$ **do**
  Set $s_t^1 = s_t^0$
  Sample level length $j_t \sim \text{Geom}(1/2)$
  for $i = 1, \ldots, 2^{j_t}$ **do**
    Take action $a_t^i \sim \pi_{\theta_t}(\cdot|s_t^i)$
    Collect next state $s_t^{i+1} \sim P(\cdot|s_t^i, a_t^i)$
    Receive reward $r_t^i = r(s_t^i, a_t^i)$
  end for
  Evaluate MLMC gradient f_t^{MLMC}, h_t^{MLMC}, and g_t^{MLMC} via (13), (15), (16)
  Update parameters following (17)
  Set $s_{t+1}^0 = s_t^{2^{j_t}}$
end for

where β_t is the critic learning rate. We remark that the critic update in (11) requires that we know $J(\pi_{\theta_t})$ (time-averaged reward), which is typically not available. We can replace this unknown quantity with a recursive estimate for the average reward obtained by $\eta_{t+1} = \eta_t - \gamma_t(\eta_t - r(s_t, a_t))$, where γ_t is the average reward learning rate.

Finally, we can write vanilla actor-critic updates as

$$\begin{aligned}\eta_{t+1} &= \eta_t + \gamma_t \cdot f_t && \text{(reward tracking)}\\ \omega_{t+1} &= \Pi_\Omega\big[\omega_t + \beta_t \cdot g_t\big], && \text{(critic update)}\\ \theta_{t+1} &= \theta_t + \alpha_t \cdot \delta^{\pi_{\theta_t}} \cdot h_t, && \text{(actor update)}\end{aligned} \tag{11}$$

where we have

$$\begin{aligned} f_t &= r(s_t,a_t) - \eta_t,\\ g_t &= \big(r(s_t,a_t) - \eta_t + \langle \phi(s_{t+1}) - \phi(s_t), \omega_t\rangle\big)\phi(s_t),\\ h_t &= \delta^{\pi_{\theta_t}} \cdot \nabla_\theta \log \pi_{\theta_t}(a_t|s_t),\\ \delta^{\pi_{\theta_t}} &= r(s_t,a_t) - \eta_t + \langle \phi(s_{t+1}) - \phi(s_t), \omega_t\rangle. \end{aligned} \tag{12}$$

As previously mentioned, in existing works the stability of (11)-(12) is only ensured under the exponentially fast mixing condition, which can preclude cases with reward sparsity or large state spaces. We therefore develop an augmentation of actor-critic that alleviates this restriction in the following subsection.

3.2. Multi-level Monte Carlo Actor-Critic

Recent work by Dorfman & Levy (2022) has developed the use of Multi-level Monte Carlo techniques together with AdaGrad step-size selection to develop a gradient estimator for Markovian data in stochastic optimization settings. We build upon these techniques in putting forth an MLMC gradient estimator for the actor, critic, and reward tracking.

Specifically, we propose to replace the stochastic gradients f_t, g_t, and h_t in (11) with the following MLMC gradients. For each t, we let $J_t \sim \text{Geom}(1/2)$, then we collect a trajectory $\mathcal{T}_t := \{s_t^i, a_t^i, r_t^i, s_t^{i+1}\}_{i=1}^{2^{J_t}}$ by interacting with the environment using policy parameter vector θ_t. Our MLMC policy gradient estimator is then given by

$$h_t^{MLMC} = h_t^0 + \begin{cases} 2^{J_t}(h_t^{J_t} - h_t^{J_t-1}), & \text{if } 2^{J_t} \leq T_{\max} \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

with $h_t^j = \frac{1}{2^j}\sum_{i=1}^{2^j} h(\theta_t; s_t^i, a_t^i)$ aggregating 2^j gradients:

$$h(\theta_t; s_t^i, a_t^i) = \delta_i^{\pi_{\theta_t}} \cdot \nabla_\theta \log \pi_{\theta_t}(a_t^i|s_t^i), \tag{14}$$
$$\delta_i^{\pi_{\theta_t}} = r(s_t^i, a_t^i) - \eta_t + \langle \phi(s_{t+1}^i) - \phi(s_t^i), \omega_t \rangle.$$

Based on (13) and (14), we can write analogous MLMC estimators f_t^{MLMC} and g_t^{MLMC} for the reward tracking and critic, respectively:

$$f_t^{MLMC} = f_t^0 + \begin{cases} 2^{J_t}(f_t^{J_t} - f_t^{J_t-1}), & \text{if } 2^{J_t} \leq T_{\max} \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

with $f_t^j = \frac{1}{2^j}\sum_{i=1}^{2^j} f(\eta_t; s_t^i, a_t^i) = \frac{1}{2^j}\sum_{i=1}^{2^j}\big(r(s_t^i, a_t^i) - \eta_t\big)$; and

$$g_t^{MLMC} = g_t^0 + \begin{cases} 2^{J_t}(g_t^{J_t} - g_t^{J_t-1}), & \text{if } 2^{J_t} \leq T_{\max} \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

with $g_t^j = \frac{1}{2^j}\sum_{i=1}^{2^j} g(\eta_t; s_t^i, a_t^i) = \frac{1}{2^j}\sum_{i=1}^{2^j}\big(r(s_t^i, a_t^i) - \eta_t + \langle \phi(s_{t+1}^i) - \phi(s_t^i), \omega_t \rangle\big)\,\phi(s_t^i)$.

Overall, the proposed multi-level Monte Carlo actor-critic (MAC) takes the form

$$\begin{aligned} \eta_{t+1} &= \eta_t + \gamma_t \cdot f_t^{MLMC} && \text{(reward tracking)} \\ \omega_{t+1} &= \Pi_\Omega\big[\omega_t + \beta_t \cdot g_t^{MLMC}\big], && \text{(critic update)} \\ \theta_{t+1} &= \theta_t + \eta_t \cdot \delta^{\pi_{\theta_t}} \cdot h_t^{MLMC}, && \text{(actor update)} \end{aligned} \tag{17}$$

We summarize the proposed algorithm in Algorithm 1.

Remark 1. The multi-level gradients (13), (15), and (16) used in MAC require a stochastic number of samples for each estimate. This differs from the classic scheme (11), where we only need one sample (s_t, a_t, s_{t+1}) to evaluate the actor and critic gradients. The original motivation for using geometric sampling (i.e., $J_t \sim \text{Geom}(1/2)$) in MLMC estimators is that it allows us to obtain an unbiased gradient estimate averaged over $\mathcal{O}(T_{\max})$ samples while using only $\mathcal{O}(\log T_{\max})$ samples in expectation. This is made precise for our setting in equation (21) of Proposition 4.5. The reason that only an expected $\mathcal{O}(\log T_{max})$ samples are required is that, though $J_t \sim \text{Geom}(1/2)$, the maximum value j is allowed to take in the MLMC estimators (13), (15), (16) is $j_{max} = \lfloor \log T_{max} \rfloor$, implying that the expected number of samples used is actually $\mathcal{O}\left(\sum_{j=1}^{j_{max}} P(J_t = j) 2^j\right) = \mathcal{O}(\log T_{max})$. Importantly for our case, the structure of the MLMC estimator also allows us to quantify the effect of mixing time on its squared norm, as we will show in Proposition 4.5, equation (22). Surprisingly, using these features of MLMC estimators, we can accommodate Markovian sampling in our actor-critic error analysis without resorting to the exponentially fast mixing assumption. This is a critical innovation in our analysis.

4. Non-asymptotic Convergence Analysis

In this section we provide convergence rate and sample complexity results for Algorithm 1. We extend the MLMC analysis of Dorfman & Levy (2022) to the actor-critic setting, where we combine it with the two-timescale finite-time analysis of Wu et al. (2020) to obtain non-asymptotic convergence guarantees for MAC (cf. Algorithm 1). Salient features of our approach: **(1)** it avoids uniform ergodicity assumptions required in previous finite-time analyses (Zou et al., 2019; Wu et al., 2020; Chen & Zhao, 2022); **(2)** it explicitly characterizes convergence rate dependence on the mixing times encountered during training; **(3)** it (i) clarifies the trade-offs between mixing times and MLMC rollout length $T_{\max}$, and (ii) extends the standard analysis to handle additional sources of bias in the MLMC estimator, both of which were missing from the analysis of Dorfman & Levy (2022); **(4)** it leverages modified Adagrad stepsizes to avoid the slower convergence rates of previous two-timescale analyses (Wu et al., 2020) (cf. Theorem 4.8).

The rest of this section is structured as follows. We first outline standard assumptions (cf. Sec. 4.1) from the literature and provide some preliminary results. Second, we analyze the policy gradient norm (cf. Sec. 4.2) associated with Algorithm 1, which provides a preliminary convergence rate and characterizes its dependence on the error arising from the critic estimation procedure, the MLMC bias resulting from the choice of $T_{\max}$ and mixing times encountered, and the bias inherent in using function approximation for the critic. Third, we analyze the convergence (cf. Sec. 4.3) of the critic estimation error, characterizing its dependence on the MLMC bias and its convergence rate. Finally, we combine the actor and critic analyses to provide our main convergence rate and sample complexity (cf. Theorem 4.8) results for MAC. To keep the exposition clear, we provide simplified versions of our main results and omit proofs in this section. Mathematically precise statements and detailed proofs of all results are presented in the appendix.

4.1. Assumptions and Propositions

The algorithmic setting considered in this paper is that of actor-critic with linear function approximation, where the critic updates correspond to using TD(0) (Sutton & Barto, 2018) to estimate the state value function. Specifically, we assume that, for a given critic parameter $\omega \in \mathbb{R}^k$ and state s, our critic approximator is of the form $V_\omega(s) = \phi(s)^T\omega$ [cf. (9), where $\phi : \mathcal{S} \to \mathbb{R}^k$ is a given feature mapping that we assume satisfies $\sup_s \|\phi(s)\| \leq 1$.

As discussed in Ch. 9 of (Sutton & Barto, 2018), for a fixed policy parameter θ, TD(0) with linear function approximation will converge to the minimum of the mean squared projected Bellman error (MSPBE), which satisfies

$$A_\theta \omega = b_\theta, \tag{18}$$

$$A_\theta = \mathbb{E}_{s\sim\mu_\theta, a\sim\pi_\theta, s'\sim p(\cdot|s,a)}\left[\phi(s)(\phi(s) - \phi(s'))^T\right],$$
$$b_\theta = \mathbb{E}_{s\sim\mu_\theta, a\sim\pi_\theta}\left[(r(s,a) - J(\theta))\phi(s)\right].$$

In what follows, we will use $\omega^*(\theta)$ to denote the fixed point satisfying Eq. (18) for a given θ. We will also use $\omega_t^* = \omega^*(\theta_t)$ to denote the fixed point associated with policy parameter vector θ_t at time t. For a given feature mapping ϕ, we define the worst-case approximation error to be

$$\mathcal{E}_{app} = \sup_\theta \sqrt{\mathbb{E}_{s\sim\mu_\theta}\left[\phi(s)^T\omega^*(\theta) - V^{\pi_\theta}(s)\right]^2}, \tag{19}$$

which we assume to be finite. Intuitively, $\mathcal{E}_{app}$ quantifies the quality of the feature mapping: when the features are well-designed, $\mathcal{E}_{app}$ will be small or even 0, while poorly designed features will tend to have higher worst-case error.

Analyses of TD learning typically assume positive definiteness of the matrices A_θ to ensure the solvability of the MSPBE minimization problem and uniqueness of its solutions (Bhandari et al., 2018; Zou et al., 2019; Qiu et al., 2021b), which we subsequently impose via Assumption 4.1.

Assumption 4.1. There exist $\lambda > 0$ such that, for all θ, the matrix A_θ is positive definite, its eigenvalues are all bounded and have norm greater than or equal to λ.

As indicated in our description of the algorithm in the previous section, we execute a projection onto a norm-ball with radius $R_\omega > 0$, denoted by set Ω, in our critic update step [cf. (17)]. As mentioned in (Wu et al., 2020), given Assumption 4.1, we can simply take $R_\omega = 2R/\lambda$, since $\|b_\theta\| \leq 2R$ by the boundedness of rewards, and $\|A_\theta^{-1}\| \leq 1/\lambda$.

In order to establish an ascent-type condition on the policy gradient, we require some regularity conditions which have been considered in recent analyses of model-free RL methods (Papini et al., 2018; Kumar et al., 2019; Zhang et al., 2020; Xu et al., 2020a), as detailed next.

Assumption 4.2. Let $\{\pi_\theta\}_{\theta\in\mathbb{R}^d}$ denote our parameterized policy class. There exist $B, K, L > 0$ such that

1. $\|\nabla \log \pi_\theta(a|s)\| \leq B$, for all $\theta \in \mathbb{R}^d$,
2. $\|\nabla \log \pi_\theta(a|s) - \nabla \log \pi_{\theta'}(a|s)\| \leq K\|\theta - \theta'\|$, for all $\theta, \theta' \in \mathbb{R}^d$,
3. $|\pi_\theta(a|s) - \pi_{\theta'}(a|s)| \leq L\|\theta - \theta'\|$, for all $\theta, \theta' \in \mathbb{R}^d$.

Finally, for our last major assumption we impose a condition on the ergodicity coefficients of the family of state transition kernels $\{P_\theta\}$ induced by the policy class $\{\pi_\theta\}$, where $P_\theta(s'|s) = \int_{\mathcal{A}} \pi_\theta(a|s)p(s'|s,a)da$. For a fixed transition kernel P, defined its ergodicity coefficient to be $\kappa(P) := \sup_{s,s'} \|P(\cdot|s) - P(\cdot|s')\|_{TV}$ (Mitrophanov, 2005). Furthermore, for a given $k \in \mathbb{N}$ and fixed P, let P^k denote the induced k-step transition kernel.

Assumption 4.3. For every θ, there exists $k \in \mathbb{N}$ such that the ergodicity coefficient $\kappa(P_\theta^k)$ satisfies $\kappa(P_\theta^k) < 1$.

In prior works, related quantities are assumed to go to null exponentially fast (uniform ergodicity) in finite-time analyses of average-reward actor-critic (Wu et al., 2020; Qiu et al., 2021b; Chen & Zhao, 2022) and related RL methods (Melo et al., 2008; Bhandari et al., 2018; Zou et al., 2019) (Theorem 3.1 of (Mitrophanov, 2005) establishes a correspondence). In our case, we merely require it to be upper-bounded by a constant, meaning that the degree of non-stationarity of the transition dynamics cannot be arbitrarily large, and at worst has bounded drift with time. This allows us to better accommodate large state spaces comprised of distinct regions, for example.

We are now ready to provide two important propositions that will be important in the core analysis to follow.

Proposition 4.4. *Under Assumptions 4.1-4.3, there exists $L_\omega > 0$ s.t. $\|\omega^*(\theta) - \omega^*(\theta')\| \leq L_\omega\|\theta - \theta'\|$, for all θ, θ'.*

Please refer Lemma D.2 in the appendix for the proof of Proposition 4.4. The next proposition is a generalization of Lemma 3.1 from Dorfman & Levy (2022), adapted to our actor-critic setting, that explicitly characterizes the computational cost associated with MLMC rollout length $T_{\max}$.

Before stating our main results, we first establish a result characterizing the mean and variance of the MLMC gradient estimators $f_t^{MLMC}, g_t^{MLMC}, h_t^{MLMC}$ used in the MAC updates defined in (17). Since the core result is the same for all three estimators, we formulate and derive the result for a general MLMC estimator l_t^{MLMC}. We note that l_t^{MLMC} can be replaced by any one of $f_t^{MLMC}, g_t^{MLMC}, h_t^{MLMC}$ and the result will hold. To prepare to state the result, let a policy parameter θ_t be given and sample $J_t \sim \text{Geom}(1/2)$. Fix $T_{\max} \in \mathbb{N}$ such that $T_{\max} \geq \tau_{mix}^{\theta_t}$. Fix a trajectory $z_t = \{z_t^i = (s_t^i, a_t^i, r_t^i, s_t^{i+1})\}_{i\in[2^{J_t}]}$ generated by following policy π_{θ_t} starting from $s_t^0 \sim \mu_0(\cdot)$. Let $\nabla L(x) := \mathbb{E}_{z\sim\mu_{\theta_t},\pi_{\theta_t}}[l(x,z)]$ be a gradient that we wish to estimate over z_t where $x \in \mathcal{K} \subset \mathbb{R}^k$ is the parameter of the estimator

l, e.g., x could be $x_t = \theta_t, \eta_t$, or ω_t. The MLMC estimator (cf. (13), (15), (16)) thus becomes

$$l_t^{MLMC} = l_t^0 + \begin{cases} 2^{J_t}(l_t^{J_t} - l_t^{J_t-1}), & \text{if } 2^{J_t} \le T_{\max}, \\ 0, & \text{otherwise.} \end{cases} \quad (20)$$

We are ready to present our result for the MLMC estimator in Proposition 4.5.

Proposition 4.5. *Let $j_{\max} = \lfloor \log T_{\max} \rfloor$. Fix x_t measurable w.r.t. the σ-algebra $\mathcal{F}_{t-1} = \sigma(\theta_k, \eta_k, \omega_k; k \le t-1)$, where $x_t \in \{\theta_t, \eta_t, \omega_t\}$. Assume $T_{max} \ge \tau_{mix}^{\theta_t}$, $\|\nabla L(x)\| \le G_L$, for all $x \in \mathcal{K}$, and $\|l_t^N\| \le G_L$, for all $N \in [T_{\max}]$. Then*

$$\mathbb{E}_{t-1}\left[l_t^{MLMC}\right] = \mathbb{E}_{t-1}\left[l_t^{j_{\max}}\right], \quad (21)$$

$$\mathbb{E}\left[\|l_t^{MLMC}\|^2\right] \le \widetilde{\mathcal{O}}\left(G_L^2 \tau_{mix}^{\theta_t} \log T_{\max}\right). \quad (22)$$

We provide the proof of Proposition 4.5 with a detailed description of the statement in Lemma B.3 in the appendix.

Remark 2. We note that the corresponding result in Dorfman & Levy (2022) hides the logarithmic dependence of the second moment bound (22) on the MLMC rollout length $T_{\max}$, subsuming it into the $\widetilde{\mathcal{O}}(\cdot)$ order notation. When $T_{\max}$ is allowed to grow with time, e.g., by setting $T_{\max} = T$ as in Dorfman & Levy (2022), the true impact of using MLMC is not accurately accounted for. Furthermore, a finite value for $T_{\max}$ must be used in practice, so it is important to understand its true effect. We rigorously characterize its effect with Proposition 4.5. Nonetheless, it is important to note that our main results, including Theorems 4.6, 4.7, and 4.8 (and their detailed analogues in the appendix, which retain additional information on the effects of T_{max} and problem-dependent constants), all still hold with T_{max} replaced by T. In particular, when $T_{max} = T$, our results all hold without the assumption that $T_{max} \ge \tau_{mix}^{\theta_t}$.

In addition, Proposition 4.5, its precursor results (see Lemmas B.1, B.2 in appendix), and our extensions of it (see Lemmas C.1, D.3, C.2, D.4 in appendix) are the critical tools that allow us to smoothly accommodate Markovian sampling and reveal the dependence on mixing times encountered in the analysis. Equation (21) is used at many points in the analysis to tie the behavior of our MLMC estimates to that of the lower-bias estimators $f_t^{j_{\max}}, g_t^{j_{\max}}, h_t^{j_{\max}}$, while equation (22) renders the dependence on $\log T_{\max}$ and mixing time explicitly, and allows us to avoid uniform ergodicity assumptions. These innovations allow us to derive the improved actor and critic convergence analyses presented next.

4.2. Convergence of the Actor

In this section, we take the first step towards establishing convergence of Algorithm 1 by providing a bound on the average policy gradient norm. This result explicitly characterizes the actor convergence in terms of its dependence on the average reward tracking and critic estimation error, mixing times encountered during training, MLMC rollout length $T_{\max}$, and the function approximation bias $\mathcal{E}_{app}$. We present our first main result in Theorem 4.6.

Theorem 4.6. *Assume $J(\theta)$ is L-smooth, $\sup_\theta |J(\theta)| \le M$, and $\|\nabla J(\theta)\|, \|h_t^{MLMC}\| \le G_H$, for all θ, t. Assume also that $T_{max} \ge \tau_{mix}^{\theta_t}$, for each t. Let $\alpha_t = \alpha_t' / \sqrt{\sum_{t=1}^T \|h_t^{MLMC}\|^2}$, where $\{\alpha_t'\}$ is an auxiliary stepsize sequence with $\alpha_t' \le 1$, for all $t \ge 1$. Then*

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right] \le \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \mathcal{O}\left(\frac{1}{T}\sum_{t=1}^T \mathcal{E}(t)\right) + \widetilde{\mathcal{O}}\left(\max_{t\in[T]} \tau_{mix}^{\theta_t} \frac{\log T_{\max}}{T_{\max}}\right) + \mathcal{O}\left(\mathcal{E}_{app}\right), \quad (23)$$

where $\mathcal{E}(t) = \mathbb{E}\left[\|\eta_t - \eta_t^*\|^2\right] + \mathbb{E}\left[\|\omega_t - \omega_t^*\|^2\right]$. and $\eta_t^* = J(\theta_t)$.

We provide a more detailed statement of Theorem 4.6 and a complete proof in Theorem C.4 in the appendix. In addition to the $\mathcal{O}\left(T^{-1/2}\right)$ term and the inherent $\mathcal{O}\left(\mathcal{E}_{app}\right)$ bias term, this bound depends on the average value of the critic error via $\mathcal{E}(t)$ and Markovian sampling through $\max_{t\in[T]} \tau_{mix}^{\theta_t} \frac{\log T_{\max}}{T_{\max}}$. As we will see in Theorem 4.7 in the following subsection, the $\mathcal{E}(t)$ term dies to 0 at a favorable rate. The presence of the Markovian sampling term, however, marks the point where our work departs significantly from previous work.

Remark 3. Interestingly, we note that the right-hand side of (23) no longer depends upon the step size rate as in Wu et al. (2020, Theorem 4.5) due to the use of our modified Adagrad stepsize in the actor update. This allows us to derive an improved overall sample complexity in Theorem 4.8.

An important consequence of Theorem 4.6 is that the level of bias resulting from Markov sampling can be controlled by choosing $T_{\max}$ appropriately. When the maximum mixing time likely to be encountered during training – captured here by the term $\max_{t\in[T]} \tau_{mix}^{\theta_t}$, is small – it makes sense to choose $T_{\max}$ to be relatively small as well. When mixing times are long, on the other hand, choosing $T_{\max}$ accordingly keeps the Markovian sampling bias manageable.

4.3. Convergence of the Critic

We turn next to characterizing the convergence of the critic error term arising in bound (23) of Theorem 4.6. Similar to that theorem, the resulting bound expresses critic convergence in terms of mixing times encountered during training as well as MLMC rollout length $T_{\max}$. This result is also where our actor-critic scheme explicitly becomes two-timescale due to our choice of stepsize sequences.

Theorem 4.7. *Assume* $\beta_t = \gamma_t = (1+t)^{-\nu}, \alpha_t = \alpha'_t/\sqrt{\sum_{k=1}^{t} \|h_t^{MLMC}\|^2}$, *and* $\alpha'_t = (1+t)^{-\sigma}$, *where* $0 < \nu < \sigma < 1$. *Assume also that* $T_{max} \geq \tau_{mix}^{\theta_t}$, *for each* t. *Then*

$$\frac{1}{T}\sum_{t=1}^{T} \mathcal{E}(t) \leq \mathcal{O}\left(T^{\nu-1}\right) + \mathcal{O}\left(T^{-2(\sigma-\nu)}\right) + \widetilde{\mathcal{O}}\left(\max_{t\in[T]} \tau_{mix}^{\theta_t} \log T_{\max}\right) \mathcal{O}\left(T^{-\nu}\right) + \widetilde{\mathcal{O}}\left(\max_{t\in[T]} \tau_{mix}^{\theta_t} \frac{\log T_{\max}}{T_{\max}}\right). \tag{24}$$

For the proof of Theorem 4.7, refer to Theorems D.1 and D.5) in the appendix. Unlike the actor bound (23), the only term in (24) that does not diminish with T is the Markovian sampling term containing $\max_{t\in[T]} \tau_{mix}^{\theta_t} \frac{\log T_{\max}}{T_{\max}}$. As in the actor case, this bias can be controlled via the proper selection of $T_{\max}$. As we will see in the final result of this section, this Markovian sampling term will ultimately be absorbed into the analogous term from Theorem 4.6.

4.4. Convergence Rate and Sample Complexity

We now present our main result characterizing the convergence rate of Algorithm 1 in terms of only the total number of iterations, mixing times encountered and $T_{\max}$ used during training, and the function approximation bias $\mathcal{E}_{app}$. We present the result in Theorem 4.8 next, which follows directly from Theorems 4.6 and 4.7.

Theorem 4.8. *(Convergence Rate) Under the assumptions of Theorems 4.6 and 4.7 and with selection* $\sigma = 0.75$ *and* $\nu = 0.5$, *we have*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right] \leq \mathcal{O}\left(\mathcal{E}_{app}\right) + \widetilde{\mathcal{O}}\left(\frac{\tau_{mix}\log T_{\max}}{\sqrt{T}}\right) + \widetilde{\mathcal{O}}\left(\frac{\tau_{mix}\log T_{\max}}{T_{\max}}\right), \tag{25}$$

where $\tau_{mix} := \max_{t\in[T]} \tau_{mix}^{\theta_t}$.

The proof of Theorem 4.8 is provided in Appendix E. The result in Theorem 4.8 provides an explicit dependence of the final convergence rate on the maximum mixing time τ_{mix} encountered during training as well as rollout length $T_{\max}$. The first term is $\mathcal{O}\left(\mathcal{E}_{app}\right)$ on the right-hand side of (25) is unavoidable due to the use of linear function approximation for the critic, but can be kept small or even driven to zero with appropriate feature selection. The second term shows the dependence on the mixing rate and shows that we recover the original i.i.d. rates if $\tau_{mix} = 1$. The last term on the right-hand side of (25) is interesting because that is the final bias we are incurring due to the use of finite length rollout trajectories $T_{\max}$. Importantly, if we take $T_{\max} = T$ as in Dorfman & Levy (2022), we recover the rate $\mathcal{O}\left(\mathcal{E}_{app}\right) + \widetilde{\mathcal{O}}\left(\frac{\tau_{mix}\log T_{\max}}{\sqrt{T}}\right)$.

We conclude this section with a sample complexity result for Algorithm 1.

Corollary 4.9. *Let us consider* $T_{\max} = \sqrt{T}$ *and* $\mathcal{E}_{app} \leq \epsilon$. *Absorbing the logarithmic terms in the* $\tilde{\mathcal{O}}$ *notation, it holds that to achieve* $\min_{1\leq t\leq T} \mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right] \leq \epsilon$, *we need* $T \geq \tilde{\mathcal{O}}\left(\frac{\tau_{mix}^2}{\epsilon^2}\right)$.

The proof of Corollary 4.9 follows directly from the statement of Theorem 4.8. We remark that, even for fast mixing settings where we can ignore the dependence on τ_{mix}^2 in Corollary 4.9, our proposed algorithm achieves sample complexity $\tilde{\mathcal{O}}\left(\frac{1}{\epsilon^2}\right)$, which improves upon the state of the art result of $\tilde{\mathcal{O}}\left(\frac{1}{\epsilon^{2.5}}\right)$ in Wu et al. (2020). This improvement is due to the use of Adagrad step size in the actor update.

Remark 4. It is interesting to note that the analysis presented in this section recovers results for the simplified i.i.d. sampling setting: since mixing occurs immediately, $\max_{t\in[T]} \tau_{mix}^{\theta_t} = 1$, so we can simply choose $T_{max} = 1$. At the other extreme, when mixing is very slow we intuitively expect that single- or few-sample estimates of the policy gradient like those considered in (Wu et al., 2020; Xu et al., 2020b; Qiu et al., 2021b; Chen & Zhao, 2022) will be highly inaccurate due to the failure of the fast mixing condition of Assumption 4.2 of (Wu et al., 2020) and Assumption 2 of (Xu et al., 2020b), for example, making a larger number of samples imperative. Theorems 4.6, 4.7, and 4.8 are the first results to shed light on this trade-off.

5. Experiments

In this section, we perform preliminary proof-of-concept experiments to evaluate the performance of the proposed MAC algorithm and compare it against vanilla actor-critic. While we concede that numerous enhancements to actor-critic have been considered, based on Nesterov acceleration (Kumar et al., 2019), parallelization (Asynchronous Advantage Actor-Critic (Mnih et al., 2016)), and offline processing of prior trajectory information (Soft Actor-Critic (Haarnoja et al., 2018)), our focus is on revealing the experimental dependence of actor-critic's stability on the environment's mixing rate. Therefore, for carefully controlled experimentation, we only compare against vanilla actor-critic as detailed in Sec. 3.1. We consider an $n \times n$ grid with a starting position at the top left and a goal at the bottom right. There are five actions: stay, up, down, left, and right. An action that results in the goal state gives the agent a $+1$ reward and $+0$ for all other states. In Figure 1, we report algorithm performance in terms of mean reward returns over 5 trials with 95% confidence intervals.

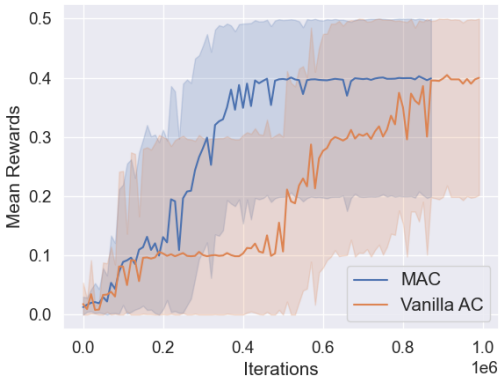
We compare MAC against vanilla actor-critic with a standard gradient estimator. In practice, we use a constant learning rate for the actor, critic, and reward estimation. For comparison, we ran vanilla actor-critic for 1 million iterations setting its constant rollout length to the largest integer under the average rollout length of MAC. For $T_{\max} = 8$, the average rollout length is 3.42, so the rollout length for vanilla AC is 3. Thus, 3 million samples were observed for the vanilla actor-critic. To have a similar number of observed samples, we ran MAC for 877192 iterations. Similarly when $T_{\max} = 16$, the average rollout length is 4.26. Therefore, we ran MAC for 936768 iterations. The details table of hyperparameters is provided in Appendix F. In Figure 1 (a) we set $n = 6$ and $T_{\max} = 8$ for MAC. For MAC and vanilla actor-critic, we set the learning rate for actor, critic, and reward estimation to .01. In Figure 1 (b), $n = 10$ and $T_{\max} = 16$ and learning rate is .005. We observe that for both experiments, MAC converges faster to the maximum reward than vanilla actor-critic, showing MLMC's advantage over a standard gradient estimator.

6. Conclusions and Limitations

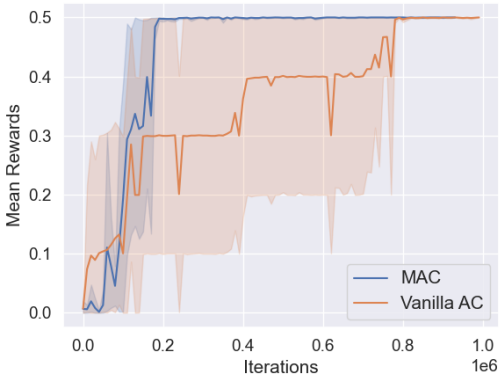
In this work, we proposed a new, multi-level Monte Carlo-based actor-critic algorithm. In our analysis of this scheme, we established for the first time an explicit dependence of the convergence rate of an actor-critic algorithm on the mixing times of the underlying Markov transitions induced by the policies encountered during training. Furthermore, our use of multi-level Monte Carlo estimators also allowed us to remove the fast mixing assumptions of previous works, extending the applicability of actor-critic algorithms to slower-mixing problems frequently encountered in robotics, finance, etc. As a limitation, our current dependence on mixing time may not be the sharpest possible. One can likely further improve the dependence on mixing time from linear to sublinear, which is a valid scope of future research. In addition, how best to choose $T_{\max}$ in our algorithm remains an open question. Developing adaptive selection methods based on bounding mixing times or avoiding mixing time estimation altogether (e.g., via the techniques in Zahavy et al. (2020)) is an important direction for future work.

7. Acknowledgements

Bedi and Manocha would like to acknowledge the support from Army Cooperative Agreement W911NF-21-2-0076 and Amazon Research Award 2022. Suttle would like to acknowledge the support of the Army Research Laboratory (ARL) Distinguished Postdoctoral Fellowship and ARL Cooperative Agreement W911NF-22-2-0003.



(a)



(b)

Figure 1. (a) Mean Rewards over 3 million samples with $T_{\max} = 8$ for MAC and rollout = 3 for vanilla actor-critic with 6×6 grid. (b) Mean Rewards over 4 million samples with $T_{\max} = 16$ for MAC and rollout = 4 for vanilla actor-critic with 10×10 grid.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, pp. 64–66. PMLR, 2020. 2, 12

Bedi, A. S., Chakraborty, S., Parayil, A., Sadler, B. M., Tokekar, P., and Koppel, A. On the hidden biases of policy mirror ascent in continuous action spaces. In *International Conference on Machine Learning*, pp. 1716–1731. PMLR, 2022. 2, 12

Bertsekas, D. P. Approximate policy iteration: A survey and some new methods. *Journal of Control Theory and Applications*, 9(3):310–335, 2011. 12

Bhandari, J. and Russo, D. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019. 2, 12

Bhandari, J., Russo, D., and Singal, R. A finite time analysis of temporal difference learning with linear function approximation. *CoRR*, abs/1806.02450, 2018. 2, 4, 6, 12

Borkar, V. S. and Konda, V. R. The actor-critic algorithm as multi-time-scale stochastic approximation. *Sadhana*, 22: 525–543, 1997. 2, 12, 13

Borkar, V. S. and Meyn, S. P. The ode method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38 (2):447–469, 2000. 12

Chen, X. and Zhao, L. Finite-time analysis of single-timescale actor-critic, 2022. 1, 2, 3, 5, 6, 8, 13

Dorfman, R. and Levy, K. Y. Adapting to mixing time in stochastic optimization with Markovian data. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5429–5446. PMLR, 17–23 Jul 2022. 2, 4, 5, 6, 7, 8, 12, 13, 14, 15, 16

Duchi, J. C., Agarwal, A., Johansson, M., and Jordan, M. I. Ergodic mirror descent. *SIAM Journal on Optimization*, 22(4):1549–1578, 2012. 1

Gu, S., Holly, E., Lillicrap, T., and Levine, S. Deep reinforcement learning for robotic manipulation. *arXiv preprint arXiv:1610.00633*, 1, 2016. 1

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. PMLR, 2018. 8

Heaton, J., Polson, N. G., and Witte, J. Deep portfolio theory. *arXiv preprint arXiv:1605.07230*, 2016. 1

Hinton, G. E., Osindero, S., and Teh, Y.-W. A fast learning algorithm for deep belief nets. *Neural computation*, 18 (7):1527–1554, 2006. 1

Konda, V. and Tsitsiklis, J. Actor-critic algorithms. In Solla, S., Leen, T., and Müller, K. (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999. 2, 12

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 1

Kumar, H., Koppel, A., and Ribeiro, A. On the sample complexity of actor-critic method for reinforcement learning with function approximation. *arXiv preprint arXiv:1910.08412*, 2019. 2, 3, 6, 8, 13

Leahy, J.-M., Kerimkulov, B., Siska, D., and Szpruch, L. Convergence of policy gradient for entropy regularized mdps with neural network approximation in the mean-field regime. In *International Conference on Machine Learning*, pp. 12222–12252. PMLR, 2022. 2, 12

Levin, D. A. and Peres, Y. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017. 3

Li, Y. Reinforcement learning applications. *arXiv preprint arXiv:1908.06973*, 2019. 1

Liu, M., Ho, S., Wang, M., Gao, L., Jin, Y., and Zhang, H. Federated learning meets natural language processing: A survey. *arXiv preprint arXiv:2107.12603*, 2021. 1

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., and Barton, D. Big data: the management revolution. *Harvard business review*, 90(10):60–68, 2012. 1

Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pp. 6820–6829. PMLR, 2020. 2, 12

Melo, F. S., Meyn, S. P., and Ribeiro, M. I. An analysis of reinforcement learning with function approximation. In *Proceedings of the 25th international conference on Machine learning*, pp. 664–671, 2008. 6

Mitrophanov, A. Y. Sensitivity and convergence of uniformly ergodic markov chains. *Journal of Applied Probability*, 42(4):1003–1014, 2005. 6, 23

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pp. 1928–1937. PMLR, 2016. 8

Nagaraj, D., Wu, X., Bresler, G., Jain, P., and Netrapalli, P. Least squares regression with markovian data: Fundamental limits and algorithms. *Advances in neural information processing systems*, 33:16666–16676, 2020. 1

Papini, M., Binaghi, D., Canonaco, G., Pirotta, M., and Restelli, M. Stochastic variance-reduced policy gradient. In *International conference on machine learning*, pp. 4026–4035. PMLR, 2018. 6

Pirotta, M., Restelli, M., and Bascetta, L. Adaptive step-size for policy gradient methods. *Advances in Neural Information Processing Systems*, 26, 2013. 2, 12

Pirotta, M., Restelli, M., and Bascetta, L. Policy gradient in lipschitz markov decision processes. *Machine Learning*, 100:255–283, 2015. 2, 12

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014. 3

Qiu, S., Yang, Z., Ye, J., and Wang, Z. On finite-time convergence of actor-critic algorithm. *IEEE Journal on Selected Areas in Information Theory*, 2(2):652–664, 2021a. 1, 2, 4, 12

Qiu, S., Yang, Z., Ye, J., and Wang, Z. On finite-time convergence of actor-critic algorithm. *IEEE Journal on Selected Areas in Information Theory*, 2(2):652–664, 2021b. 1, 3, 6, 8

Riemer, M., Raparthy, S. C., Cases, I., Subbaraj, G., Touzel, M. P., and Rish, I. Continual learning in environments with polynomial mixing times. *arXiv preprint arXiv:2112.07066*, 2021. 2, 3

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016. 1

Sutton, R. Learning to predict by the method of temporal differences. *Machine Learning*, 3:9–44, 08 1988. 1, 4, 12, 23

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018. 6, 23

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999. 4, 12

Tadić, V. On the convergence of temporal-difference learning with linear function approximation. *Machine learning*, 42:241–267, 2001. 12

Tsitsiklis, J. and Van Roy, B. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997. 4, 12

Wang, L., Cai, Q., Yang, Z., and Wang, Z. Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*, 2019. 2, 3, 12, 13

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992. 4, 12

Wu, Y. F., Zhang, W., Xu, P., and Gu, Q. A finite-time analysis of two time-scale actor-critic methods. *Advances in Neural Information Processing Systems*, 33:17617–17628, 2020. 1, 2, 5, 6, 7, 8, 19

Xu, P., Gao, F., and Gu, Q. An improved convergence analysis of stochastic variance-reduced policy gradient. In *Uncertainty in Artificial Intelligence*, pp. 541–551. PMLR, 2020a. 6

Xu, T., Wang, Z., and Liang, Y. Improving sample complexity bounds for (natural) actor-critic algorithms. *Advances in Neural Information Processing Systems*, 33: 4358–4369, 2020b. 1, 2, 8

Zahavy, T., Cohen, A., Kaplan, H., and Mansour, Y. Unknown mixing times in apprenticeship and reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence*, pp. 430–439. PMLR, 2020. 9

Zhang, K., Koppel, A., Zhu, H., and Basar, T. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6):3586–3612, 2020. 2, 6, 12

Zou, S., Xu, T., and Liang, Y. Finite-sample analysis for sarsa with linear function approximation. *Advances in neural information processing systems*, 32, 2019. 5, 6, 23

Appendix

Table of Contents

A. Detailed Context of Related Works

Actor-critic by Konda & Tsitsiklis (1999) comprises algorithms that alternate between value function estimation (critic) and policy search updates (actor), which may be seen as a form of policy iteration (Bertsekas, 2011) that incorporates stochastic approximation (Borkar & Konda, 1997). We discuss each facet separately, before launching into their fusion.

TD Learning To evaluate the policy update direction, an estimate of the value function is required. To compute this estimate, stochastic fixed point iterations are considered to solve Bellman's equation Sutton (1988), whose stability under linear function approximation was established in Tsitsiklis & Van Roy (1997). Since then, a plethora of works has studied the stability properties of TD-based policy evaluation. Initially, their asymptotic convergence was prioritized (Tadić, 2001), but more recently, non-asymptotic results have gained salience. For discounted TD with Markovian samples, Bhandari et al. (2018) established finite-time convergence bounds which scale linearly with mixing time τ_{mix}. Dorfman & Levy (2022) then improved the rate to be proportional to the $\sqrt{\tau_{mix}}$ using a multi-level gradient estimator and adaptive learning rate. Qiu et al. (2021a) studied TD under the average reward setting, which also imposes exponentially fast mixing that manifests in an additional logarithmic term in the sample complexity. These results all hinge upon imposing restrictive conditions on the mixing time.

Policy Gradient With a value function estimate in hand, one can multiple this quantity together with the gradient of the log-likelihood of a policy, i.e., the score function, to evaluate an estimate of the policy gradient (Williams, 1992; Sutton et al., 1999). Then, gradient ascent steps are taken with respect to policy parameters. The convergence of policy gradient has been studied extensively. Similar to TD, early work (Borkar & Meyn, 2000) focused on asymptotic stability via tools from dynamical systems (Borkar & Meyn, 2000). More recently, its sample complexity has been established for a variety of settings: for tabular (Bhandari & Russo, 2019; Agarwal et al., 2020) and softmax policies (Mei et al., 2020), rates to global optimality exist. For general parameterized policies, early works focused on "policy improvement" bounds (Pirotta et al., 2013; 2015), and more recently, rates towards stationarity (Bedi et al., 2022) and local extrema (Zhang et al., 2020) have been studied, and under special neural architectures, globally optimal solutions (Wang et al., 2019; Leahy et al., 2022) are achievable. This topic is an active area of work, and covering all related sub-topics is beyond our scope. We merely identify that these performance certificates all hinge upon the mixing rate of the induced Markov chain going to null exponentially fast.

Actor-Critic As previously mentioned, the stability of actor-critic was initially focused on asymptotics (Borkar & Konda, 1997). More recently, its non-asymptotic rate has been derived under i.i.d. assumptions (Kumar et al., 2019; Wang et al., 2019), and more recently under a variety of different types of Markovian data – see Table 1. However, these results impose that any temporal correlation of data across time vanishes exponentially fast as quantified by the mixing rate. In this way, we are able to match (Chen & Zhao, 2022) but without this restriction.

B. Preliminaries

Before proceeding with our analysis of Algorithm 1, we need some preliminary results and assumptions.

B.1. Preliminary Results

The statements of the results in this section have been adapted from (Dorfman & Levy, 2022) to fit the setting considered in our paper. Except in the case of Lemma B.3, their proofs follow directly from that work. First, we need the following concentration bound concerning gradient estimation from Markovian data.

Lemma B.1. *Lemma A.5, (Dorfman & Levy, 2022). Fix $K, N \in \mathbb{N}$ such that $N \geq 2K$. Let a policy parameter $\theta_t \in \Theta$ be given, and fix a trajectory $z_t = \{z_t^i = (s_t^i, a_t^i, r_t^i, s_t^{i+1})\}_{i\in[N]}$ generated by following policy π_{θ_t} starting from $s_t^0 \sim \mu_0(\cdot)$. Let $\nabla L(x)$ be a gradient that we wish to estimate over z_t, where $\mathbb{E}_{z\sim\mu_{\theta_t},\pi_{\theta_t}}[l(x,z)] = \nabla L(x)$, and $x \in \mathcal{K} \subset \mathbb{R}^k$ is the parameter of the estimator l, i.e., $x_t = \theta_t, \eta_t$, or ω_t. Finally, assume that $\|l(x,z)\|, \|\nabla L(x)\| \leq G_L$, for all $x \in \mathcal{K}, z \in \mathcal{S}\times\mathcal{A}\times\mathbb{R}\times\mathcal{S}$. Then, for every $\delta > Nd_{mix}(K)$ and every $x_t \in \mathcal{K}$ measurable w.r.t. the σ-algebra $\mathcal{F}_{t-1} = \sigma(\theta_k, \eta_k, \omega_k, z_k; k \leq t-1)$, we have*

$$\mathbb{P}_{t-1}\left(\left\|\frac{1}{N}\sum_{i=1}^{N} l(x_t, z_t^i) - \nabla L(x_t)\right\| \leq 12G_L\sqrt{\frac{K}{N}}\left(1+\sqrt{\log(K/\tilde{\delta})}\right) + \frac{6GK}{N}\right) \geq 1-\delta, \tag{26}$$

where $\tilde{\delta} = \delta - Nd_{mix}(K)$.

We will use this result to facilitate our analyses of each of the MLMC estimators $f_t^{MLML}, g_t^{MLMC}, l_t^{MLMC}$ used in Algorithm 1. We also need the following error bound, which follows from Lemma B.1.

Lemma B.2. *Lemma A.6, (Dorfman & Levy, 2022). Let $\nabla L, l, z_t$ be as in Lemma B.1. Define $l_t^N = \frac{1}{N}\sum_{i=1}^{N} l(x_t, z_t^i)$. Fix $T_{max} \in \mathbb{N}$ and let $K = \tau_{max}^{\theta_t}\lceil 2\log T_{max}\rceil$. Then, for every $N \in [T_{max}]$ and every $x_t \in \mathcal{K}$ measurable w.r.t. $\mathcal{F}_{t-1}$,*

$$\mathbb{E}\left[\|l_t^N - \nabla L(x_t)\|\right] \leq O\left(G_L\sqrt{\log KN}\sqrt{\frac{K}{N}}\right), \tag{27}$$

$$\mathbb{E}\left[\|l_t^N - \nabla L(x_t)\|^2\right] \leq O\left(G_L^2\log(KN)\frac{K}{N}\right). \tag{28}$$

The following important result establishes key properties of MLMC estimators. It is an extension of Lemma 3.1 from (Dorfman & Levy, 2022), clarifying the effect of using rollout length T_{max} in the MLMC estimator.

Lemma B.3. *Let $\nabla L, l, z_t$ be as in Lemma B.1. Let $J_t \sim$ Geom$(1/2)$. Define the MLMC estimator*

$$l_t^{MLMC} = l_t^0 + \begin{cases} 2^{J_t}(l_t^{J_t} - l_t^{J_t-1}), & \text{if } 2^{J_t} \geq T_{max}, \\ 0, & \text{otherwise.} \end{cases} \tag{29}$$

Let $j_{max} = \lfloor \log T_{max} \rfloor$. Fix x_t measurable w.r.t. $\mathcal{F}_{t-1}$. Assume $T_{max} \geq \tau_{mix}^{\theta_t}$, $\|\nabla L(x)\| \leq G_L$, for all $x \in \mathcal{K}$, and $\|l_t^N\| \leq G_L$, for all $N \in [T_{max}]$. Then

$$\mathbb{E}_{t-1}\left[l_t^{MLMC}\right] = \mathbb{E}_{t-1}\left[l_t^{j_{max}}\right], \tag{30}$$

$$\mathbb{E}\left[\|l_t^{MLMC}\|^2\right] \leq \widetilde{\mathcal{O}}\left(G_L^2\tau_{mix}^{\theta_t}\log T_{max}\right). \tag{31}$$

Proof. For brevity, let $l_t := l_t^{MLMC}$. To show (30), we simply recall that $l_t = l_t^0 + 2^{J_t}\left(l_t^{J_t} - l_t^{J_t-1}\right)$ and note that

$$\mathbb{E}_{t-1}\left[l_t\right] = \mathbb{E}_{t-1}\left[l_t^0\right] + \sum_{i=1}^{j_{max}} P(J_t = j)2^j \mathbb{E}_{t-1}\left[l_t^j - l_t^{j-1}\right] = \mathbb{E}_{t-1}\left[l_t^{j_{max}}\right]. \tag{32}$$

For (31), first note that by Cauchy-Schwarz and boundedness of l_t^j, for all $j \in [T_{max}]$, we know that

$$\mathbb{E}\left[\|l_t\|^2\right] \leq 2\mathbb{E}\left[\|l_t - l_t^0\|^2\right] + 2G_L^2. \tag{33}$$

Now, since $l_t = l_t^0 + 2^{J_t}\left(l_t^{J_t} - l_t^{J_t-1}\right)$,

$$\begin{aligned}
\mathbb{E}\left[\|l_t - l_t^0\|^2\right] &= \sum_{j=1}^{j_{max}} P(J_t = j)\mathbb{E}\left[\left\|2^j\left(l_t^j - l_t^{j-1}\right)\right\|^2\right] && (34)\\
&= \sum_{j=1}^{j_{max}} 2^j \mathbb{E}\left[\left\|\left(l_t^j - l_t^{j-1}\right)\right\|^2\right] && (35)\\
&\leq \sum_{j=1}^{j_{max}} 2^j \left(2\mathbb{E}\left[\left\|l_t^j - \nabla J(\theta_t)\right\|^2\right] + 2\mathbb{E}\left[\left\|l_t^{j-1} - \nabla J(\theta_t)\right\|^2\right]\right) && (36)\\
&\overset{(a)}{\leq} \sum_{j=1}^{j_{max}} 2^j \left(\widetilde{\mathcal{O}}\left(\frac{1}{2^j} G_L^2 \tau_{mix}^{\theta_t} \log(T_{max})\right)\right) && (37)\\
&= \sum_{j=1}^{j_{max}} \widetilde{\mathcal{O}}\left(G_L^2 \tau_{mix}^{\theta_t} \log T_{max}\right) && (38)\\
&= \widetilde{\mathcal{O}}\left(G_L^2 \tau_{mix}^{\theta_t} \log T_{max}\right), && (39)
\end{aligned}$$

where (a) follows from Lemma B.2 and (39) holds by the definition of j_{max}. Combining (33) with (39) gives the result. □

Finally, we will use the following result to manipulate the AdaGrad stepsizes in the final result of this section.

Lemma B.4. *Lemma 4.2, (Dorfman & Levy, 2022). For any non-negative real numbers* $\{a_i\}_{i\in[n]}$,

$$\sum_{i=1}^{n} \frac{a_i}{\sqrt{\sum_{j=1}^{i} a_j}} \leq 2\sqrt{\sum_{i=1}^{n} a_i}. \tag{40}$$

B.2. Assumptions

We will also need the following assumptions.

Assumption B.5. The objective $J(\theta)$ is L-Lipschitz in θ. There exists G_H such that $\|\nabla J(\theta)\| \leq G_H$, for all θ.

Assumption B.6. The critic update includes a projection onto the ball of radius R_ω about the origin.

Assumption B.7. For each θ, the matrix $A_\theta = \mathbb{E}_{s\sim\mu_\theta, a\sim\pi_\theta, s'\sim p(\cdot|s,a)}\left[\phi(s)(\phi(s) - \phi(s'))^T\right]$ is positive definite.

C. Convergence Analysis of Actor

In this section, we provide a bound on the average policy gradient norm achieved by Algorithm 1, leveraging the MLMC analysis machinery of (Dorfman & Levy, 2022) to reveal dependence on the worst-case mixing time encountered during training. Combined with the error analysis of Section D, this forms the core of our analysis of Algorithm 1. The analysis

largely follows that of (Dorfman & Levy, 2022), with key modifications to accommodate the *average reward estimation*, *critic estimation*, and *critic function approximation bias* inherent in the average-reward actor-critic setting.

As the first step in our actor analysis, we prove a version of Lemma B.2 that incorporates average reward estimation error and critic error. Before starting the result and its proof, we develop some notation to facilitate the exposition. Let

$$\nabla J_t^i = \left(r_t^i - \eta_t + \langle \phi(s_t^{i+1}), \omega_t \rangle - \langle \phi(s_t^i), \omega_t \rangle\right) \nabla \log \pi_{\theta_t}\left(a_t^i | s_t^i\right), \tag{41}$$

$$\nabla J_t^{i,\eta} = \left(r_t^i - \eta_t^* + \langle \phi(s_t^{i+1}), \omega_t \rangle - \langle \phi(s_t^i), \omega_t \rangle\right) \nabla \log \pi_{\theta_t}\left(a_t^i | s_t^i\right), \tag{42}$$

$$\nabla J_t^{i,\eta,\omega} = \left(r_t^i - \eta_t^* + \langle \phi(s_t^{i+1}), \omega_t^* \rangle - \langle \phi(s_t^i), \omega_t^* \rangle\right) \nabla \log \pi_{\theta_t}\left(a_t^i | s_t^i\right), \tag{43}$$

$$\nabla J_t^{i,\eta,V} = \left(r_t^i - \eta_t^* + V_{\theta_t}(s_t^{i+1}) - V_{\theta_t}(s_t^i)\right) \nabla \log \pi_{\theta_t}\left(a_t^i | s_t^i\right), \tag{44}$$

where $\eta_t^* = J(\theta_t)$ and ω_t^* is the limiting point of TD(0) applied to evaluating the policy π_{θ_t}. Notice that

$$\nabla J_t^i - \nabla J(\theta_t) = \big(\underbrace{\nabla J_t^i - \nabla J_t^{i,\eta}}_{(a)}\big) + \big(\underbrace{\nabla J_t^{i,\eta} - \nabla J_t^{i,\eta,\omega}}_{(b)}\big) + \big(\underbrace{\nabla J_t^{i,\eta,\omega} - \nabla J_t^{i,\eta,V}}_{(c)}\big) + \big(\underbrace{\nabla J_t^{i,\eta,V} - \nabla J(\theta_t)}_{(d)}\big), \tag{45}$$

where

$$\text{(a): } \nabla J_t^i - \nabla J_t^{i,\eta} = (\eta_t^* - \eta_t) \nabla \log \pi_{\theta_t}\left(a_t^i | s_t^i\right) \tag{46}$$

$$\text{(b): } \nabla J_t^{i,\eta} - \nabla J_t^{i,\eta,w} = \langle \phi(s_t^{i+1}) - \phi(s_t^i), \omega_t - \omega_t^* \rangle \nabla \log \pi_{\theta_t}\left(a_t^i | s_t^i\right) \tag{47}$$

$$\text{(c): } \nabla J_t^{i,\eta,w} - \nabla J_t^{i,\eta,V} = \left[\left(\langle \phi(s_t^{i+1}), \omega_t^* \rangle - V_{\theta_t}(s_t^{i+1})\right) - \left(\langle \phi(s_t^i), \omega_t^* \rangle - V_{\theta_t}(s_t^i)\right)\right] \nabla \log \pi_{\theta_t}\left(a_t^i | s_t^i\right) \tag{48}$$

and, since $\mathbb{E}_{\mu_{\theta_t}, \pi_{\theta_t}}\left[\nabla J_t^{i,\eta,V}\right] = \nabla J(\theta_t)$, (d) is the error between $\nabla J(\theta_t)$ and the ideal policy gradient estimator. Define

$$\mathcal{E}_{\text{app}} := \sup_{s,\theta} |\langle \phi(s), \omega(\theta) - V_\theta(s) \rangle|, \quad C := \sup_{s,s'} \|\phi(s) - \phi(s')\|, \tag{49}$$

and let $B > 0$ be such that

$$\sup_{\theta,a,s} \|\nabla \log \pi_\theta(a|s)\| \leq B. \tag{50}$$

Lemma C.1. *Assume $\|\nabla J(\theta)\|, \|\nabla J_t^{i,\eta,V}\| \leq G_H$, for all θ, s_t^i, a_t^i. Fix $T_{max} \in \mathbb{N}$ and let $K = \tau_{max}^{\theta_t} \lceil 2 \log T_{max} \rceil$. Define $h_t^N = \frac{1}{N} \sum_{i=1}^N \nabla J_t^i$, for $N \in [T_{max}]$. Then, for all $N \in [T_{max}]$ and θ_t measurable w.r.t. $\mathcal{F}_{t-1}$,*

$$\mathbb{E}\left[\|h_t^N - \nabla J(\theta_t)\|\right] \leq O\left(G_H \sqrt{\log KN} \sqrt{\frac{K}{N}}\right) + \mathcal{E}_1(t) + 2B\mathcal{E}_{app}, \tag{51}$$

$$\mathbb{E}\left[\|h_t^N - \nabla J(\theta_t)\|^2\right] \leq O\left(G_H^2 \log(KN) \frac{K}{N}\right) + \mathcal{E}_2(t) + 16B^2\mathcal{E}_{app}, \tag{52}$$

where

$$\mathcal{E}_1(t) = B\mathbb{E}\left[\|\eta_t - \eta_t^*\|\right] + BC\mathbb{E}\left[\|\omega_t - \omega_t^*\|\right], \tag{53}$$

$$\mathcal{E}_2(t) = 4B^2\mathbb{E}\left[\|\eta_t - \eta_t^*\|^2\right] + 4B^2C^2\mathbb{E}\left[\|\omega_t - \omega_t^*\|^2\right]. \tag{54}$$

Proof. First notice that

$$\left\|h_t^N - \nabla J(\theta_t)\right\| \leq \left\|\frac{1}{N}\sum_{i=1}^N \nabla J_t^{i,\eta,V} - \nabla J(\theta_t)\right\| + \left\|\frac{1}{N}\sum_{i=1}^N \nabla J_t^i - \nabla J_t^{i,\eta}\right\| \tag{55}$$

$$+ \left\|\frac{1}{N}\sum_{i=1}^N \nabla J_t^{i,\eta} - \nabla J_t^{i,\eta,\omega}\right\| + \left\|\frac{1}{N}\sum_{i=1}^N \nabla J_t^{i,\eta,\omega} - \nabla J_t^{i,\eta,V}\right\| \tag{56}$$

$$\leq \left\|\frac{1}{N}\sum_{i=1}^N \nabla J_t^{i,\eta,V} - \nabla J(\theta_t)\right\| + B\|\eta_t - \eta_t^*\| + BC\|\omega_t - \omega_t^*\| + 2B\mathcal{E}_{app}. \tag{57}$$

As a consequence, we also have

$$\left\|h_t^N - \nabla J(\theta_t)\right\|^2 \leq 4\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla J_t^{i,\eta,V} - \nabla J(\theta_t)\right\|^2 + 4B^2\|\eta_t - \eta_t^*\|^2 + 4B^2C^2\|\omega_t - \omega_t^*\|^2 + 16B^2\mathcal{E}_{app}^2. \tag{58}$$

Taking expectations and applying Lemma B.2 with $x_t = \theta_t$, $l(\theta_t, z_t^i) = \nabla J_t^{i,\eta,V}$, $\nabla L(\theta_t) = \nabla J(\theta_t)$ yields the result. □

We next prove a key result regarding the bias and second moment of our policy gradient estimate. It is a generalization of Lemma 3.1 in (Dorfman & Levy, 2022) building on our Lemma C.1.

Lemma C.2. *Let $j_{max} = \lfloor \log T_{max} \rfloor$ in Algorithm 1. Fix θ_t measurable w.r.t. $\mathcal{F}_{t-1}$. Assume $T_{max} \geq \tau_{mix}^{\theta_t}$, $\|\nabla J(\theta)\| \leq G_H$, for all θ, and $\|h_t^N\| \leq G_H$, for all $N \in [T_{max}]$. Then*

$$\mathbb{E}_{t-1}\left[h_t^{MLMC}\right] = \mathbb{E}_{t-1}\left[h_t^{j_{max}}\right], \tag{59}$$

$$\mathbb{E}\left[\|h_t^{MLMC}\|^2\right] \leq \widetilde{\mathcal{O}}\left(G_H^2\tau_{mix}^{\theta_t}\log T_{max}\right) + 8\log(T_{max})T_{max}\left(\mathcal{E}_2(t) + 16B^2\mathcal{E}_{app}^2\right). \tag{60}$$

Proof. For brevity, let $h_t := h_t^{MLMC}$. Equation (59) follows directly from Lemma B.3. For (60), first note that by Cauchy-Schwarz and boundedness of h_t^j, for all $j \in [T_{max}]$, we know that

$$\mathbb{E}\left[\|h_t\|^2\right] \leq 2\mathbb{E}\left[\|h_t - h_t^0\|^2\right] + 2G_H^2. \tag{61}$$

Now, since $h_t = h_t^0 + 2^{J_t}\left(h_t^{J_t} - h_t^{J_t-1}\right)$,

$$\mathbb{E}\left[\|h_t - h_t^0\|^2\right] = \sum_{j=1}^{j_{max}} P(J_t = j)\mathbb{E}\left[\left\|2^j\left(h_t^j - h_t^{j-1}\right)\right\|^2\right] \tag{62}$$

$$= \sum_{j=1}^{j_{max}} 2^j\mathbb{E}\left[\left\|\left(h_t^j - h_t^{j-1}\right)\right\|^2\right] \tag{63}$$

$$\leq \sum_{j=1}^{j_{max}} 2^j\left(2\mathbb{E}\left[\left\|h_t^j - \nabla J(\theta_t)\right\|^2\right] + 2\mathbb{E}\left[\left\|h_t^{j-1} - \nabla J(\theta_t)\right\|^2\right]\right). \tag{64}$$

Next, we can write

$$\mathbb{E}\left[\|h_t - h_t^0\|^2\right] \overset{\text{(a)}}{\leq} \sum_{j=1}^{j_{max}} 2^j\left(\widetilde{\mathcal{O}}\left(\frac{1}{2^j}G_H^2\tau_{mix}^{\theta_t}\log(T_{max})\right) + 4\mathcal{E}_2(t) + 16B^2\mathcal{E}_{app}^2\right) \tag{65}$$

$$= \sum_{j=1}^{j_{max}}\left(\widetilde{\mathcal{O}}\left(G_H^2\tau_{mix}^{\theta_t}\log T_{max}\right) + 4\cdot 2^j\left[\mathcal{E}_2(t) + 16B^2\mathcal{E}_{app}^2\right]\right) \tag{66}$$

$$\overset{\text{(b)}}{\leq} \log T_{max}\left(\widetilde{\mathcal{O}}\left(G_H^2\tau_{mix}^{\theta_t}\log T_{max}\right) + 4T_{max}\left[\mathcal{E}_2(t) + 16B^2\mathcal{E}_{app}^2\right]\right) \tag{67}$$

$$= \widetilde{\mathcal{O}}\left(G_H^2\tau_{mix}^{\theta_t}\log T_{max}\right) + 4\log(T_{max})T_{max}\left[\mathcal{E}_2(t) + 16B^2\mathcal{E}_{app}^2\right], \tag{68}$$

where (a) follows from Lemma C.1 and (b) holds by the definition of j_{max}. Combining (61) with (68) gives the result. □

Before proceeding to the final policy gradient norm bound of our actor analysis, we need one additional auxiliary result.

Lemma C.3. *Assume $J(\theta)$ is L-smooth. Let $\Delta_t = \sup_\theta J(\theta) - J(\theta_t)$ and $\Delta_{max}^T = \max_{t\in[T]}\Delta_t$. Then*

$$\sum_{t=1}^{T}\|\nabla J(\theta_t)\|^2 \leq \frac{\Delta_{max}^T}{\alpha_T} + \frac{L}{2}\sum_{t=1}^{T}\alpha\|h_t^{MLMC}\|^2 + \sum_{t=1}^{T}\langle\nabla J(\theta_t) - h_t^{MLMC}, \nabla J(\theta_t)\rangle. \tag{69}$$

Proof. Once again, write $h_t := h_t^{MLMC}$ for brevity. We first have

$$J(\theta_{t+1}) \geq J(\theta_t) + \alpha_t \nabla J(\theta_t)^T h_t - \frac{L\alpha_t^2}{2}\|h_t\|^2 \tag{70}$$

$$= J(\theta_t) + \alpha_t \|\nabla J(\theta_t)\|^2 - \alpha_t \langle \nabla J(\theta_t) - h_t, \nabla J(\theta_t)\rangle - \frac{L\alpha_t^2}{2}\|h_t\|^2, \tag{71}$$

where the first equality holds from the smoothness of $J(\theta)$ and the fact that $\theta_{t+1} = \theta_t + \alpha_t h_t$. Rearranging gives

$$\|\nabla J(\theta_t)\|^2 \leq \frac{J(\theta_{t+1}) - J(\theta_t)}{\alpha_t} + \frac{L\alpha_t}{2}\|h_t\|^2 + \langle \nabla J(\theta_t) - h_t, \nabla J(\theta_t)\rangle, \tag{72}$$

and summing yields

$$\sum_{t=1}^T \|\nabla J(\theta_t)\|^2 \leq \sum_{t=1}^T \frac{\Delta_t - \Delta_{t+1}}{\alpha_t} + \frac{L}{2}\sum_{t=1}^T \alpha_t \|h_t\|^2 + \sum_{t=1}^T \langle \nabla J(\theta_t) - h_t, \nabla J(\theta_t)\rangle \tag{73}$$

$$\leq \sum_{t=1}^T \frac{\Delta_{max}^T}{\alpha_T} + \frac{L}{2}\sum_{t=1}^T \alpha_t \|h_t\|^2 + \sum_{t=1}^T \langle \nabla J(\theta_t) - h_t, \nabla J(\theta_t)\rangle. \tag{74}$$

□

We are now ready to prove the main result of this section.

Theorem C.4. *Assume $J(\theta)$ is L-smooth, $\sup_\theta |J(\theta)| \leq M$, and $\|\nabla J(\theta)\|, \|h_t^{MLMC}\| \leq G_H$, for all θ, t. Assume also that $T_{max} \geq \tau_{mix}^{\theta_t}$, for each t. Let $\alpha_t = \alpha_t' / \sqrt{\sum_{t=1}^T \|h_t^{MLMC}\|^2}$, where $\{\alpha_t'\}$ is an auxiliary stepsize sequence with $\alpha_t' \leq 1$, for all $t \geq 1$. Then*

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right] \leq \widetilde{\mathcal{O}}\left((M+L)G_H \frac{1}{\sqrt{T}}\sqrt{\max_{t\in[T]} \tau_{mix}^{\theta_t} \log T_{max}}\right) \tag{75}$$

$$+ \frac{2M+L}{T}\sqrt{\sum_{t=1}^T 8\log(T_{max}) T_{max}\left(\mathcal{E}_2(t) + 16B^2\mathcal{E}_{app}^2\right)} \tag{76}$$

$$+ \widetilde{\mathcal{O}}\left(G_H^2 \max_{t\in[T]} \tau_{mix}^{\theta_t} \frac{\log T_{max}}{T_{max}}\right) + \frac{1}{T}\sum_{t=1}^T \mathcal{E}_2(t) + 16B^2\mathcal{E}_{app}^2. \tag{77}$$

Proof. Again let $h_t := h_t^{MLMC}$. We have

$$\sum_{t=1}^T \|\nabla J(\theta_t)\|^2 \overset{(a)}{\leq} \Delta_{max}\sqrt{\sum_{t=1}^T \|h_t\|^2} + \frac{L}{2}\sum_{t=1}^T \frac{\alpha_t'\|h_t\|^2}{\sqrt{\sum_{k=1}^t \|h_k\|^2}} + \sum_{t=1}^T \langle \nabla J(\theta_t) - h_t, \nabla J(\theta_t)\rangle \tag{78}$$

$$\overset{(b)}{\leq} \Delta_{max}\sqrt{\sum_{t=1}^T \|h_t\|^2} + \frac{L}{2}\sum_{t=1}^T \frac{\|h_t\|^2}{\sqrt{\sum_{k=1}^t \|h_k\|^2}} + \sum_{t=1}^T \langle \nabla J(\theta_t) - h_t, \nabla J(\theta_t)\rangle \tag{79}$$

$$\overset{(c)}{\leq} (\Delta_{max} + L)\sqrt{\sum_{t=1}^T \|h_t\|^2} + \sum_{t=1}^T \langle \nabla J(\theta_t) - h_t, \nabla J(\theta_t)\rangle, \tag{80}$$

where (a) follows from Lemma C.3, inequality (b) by the definition of α_t, and (c) is by Lemma B.4. This implies that

$$\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right] \overset{(a)}{\leq} \mathbb{E}\left[(\Delta_{max}+L)\sqrt{\sum_{t=1}^{T}\|h_t\|^2}\right] + \sum_{t=1}^{T}\mathbb{E}\left[\langle \nabla J(\theta_t) - h_t^{j_{max}}, \nabla J(\theta_t)\rangle\right] \tag{81}$$

$$\overset{(b)}{\leq} \mathbb{E}\left[(\Delta_{max}+L)\sqrt{\sum_{t=1}^{T}\|h_t\|^2}\right] + \sum_{t=1}^{T}\mathbb{E}\left[\|\nabla J(\theta_t) - h_t^{j_{max}}\| \cdot \|\nabla J(\theta_t)\|\right] \tag{82}$$

$$\overset{(c)}{\leq} \mathbb{E}\left[(\Delta_{max}+L)\sqrt{\sum_{t=1}^{T}\|h_t\|^2}\right] + \sum_{t=1}^{T}\left(\mathbb{E}\left[\|\nabla J(\theta_t) - h_t^{j_{max}}\|^2\right]\right)^{1/2}\left(\mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right]\right)^{1/2} \tag{83}$$

$$\overset{(d)}{\leq} \mathbb{E}\left[(\Delta_{max}+L)\sqrt{\sum_{t=1}^{T}\|h_t\|^2}\right] + \left(\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla J(\theta_t) - h_t^{j_{max}}\|^2\right]\right)^{1/2}\left(\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right]\right)^{1/2}, \tag{84}$$

where (a) follows from the law of total expectation, the fact that θ_t, θ_t^* are deterministic conditioned on $\mathcal{F}_{t-1}$, and Lemma C.2, (b) follows by Cauchy-Schwarz, and (b) and (c) by applications of Hölder's inequality. Define

$$A(T) = \mathbb{E}\left[(\Delta_{max}+L)\sqrt{\sum_{t=1}^{T}\|h_t\|^2}\right], \tag{85}$$

$$B(T) = \frac{1}{4}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla J(\theta_t) - h_t^{j_{max}}\|^2\right], \tag{86}$$

$$C(T) = \sum_{t=1}^{T}\mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right]. \tag{87}$$

The foregoing inequality becomes

$$C(T) \leq A(T) + 2\sqrt{B(T)}\sqrt{C(T)} \tag{88}$$

Consider the following chain of implications:

$$C(T) \leq A(T) + 2\sqrt{B(T)}\sqrt{C(T)} \implies \left(\sqrt{C(T)} - \sqrt{B(T)}\right)^2 \leq A(T) + B(T) \tag{89}$$

$$\implies \sqrt{C(T)} - \sqrt{B(T)} \leq \sqrt{A(T)} + \sqrt{B(T)} \tag{90}$$

$$\implies \sqrt{C(T)} \leq \sqrt{A(T)} + 2\sqrt{B(T)} \tag{91}$$

$$\implies C(T) \leq 2A(T) + 8B(T). \tag{92}$$

We therefore have

$$\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right] \leq 2\mathbb{E}\left[(\Delta_{max}+L)\sqrt{\sum_{t=1}^{T}\|h_t\|^2}\right] + 2\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla J(\theta_t) - h_t^{j_{max}}\|^2\right] \tag{93}$$

$$\tag{94}$$

Now,

$$\mathbb{E}\left[(\Delta_{max}+L)\sqrt{\sum_{t=1}^{T}\|h_t\|^2}\right] \overset{(a)}{\leq} (2M+L)\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\|h_t\|^2\right]} \tag{95}$$

$$\overset{(b)}{\leq} (2M+L)\sqrt{\widetilde{\mathcal{O}}\left(TG_H^2\max_{t\in[T]}\tau_{mix}^{\theta_t}\log T_{max}\right)+\sum_{t=1}^{T}8\log(T_{max})T_{max}\left(\mathcal{E}_2(t)+16B^2\mathcal{E}_{app}^2\right)} \tag{96}$$

$$\overset{(c)}{\leq} \widetilde{\mathcal{O}}\left((M+L)G_H\sqrt{T\max_{t\in[T]}\tau_{mix}^{\theta_t}\log T_{max}}\right)+(2M+L)\sqrt{8\sum_{t=1}^{T}\log(T_{max})T_{max}\left(\mathcal{E}_2(t)+16B^2\mathcal{E}_{app}^2\right)}, \tag{97}$$

where (a) follows by the fact that $\Delta_{\max}\leq 2M$ and Jensen's inequality, (b) is from Lemma C.2, and (c) follows since $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$. Furthermore, by the second-order bound of Lemma C.1 we have

$$\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla J(\theta_t)-h_t^{j_{max}}\|^2\right]\leq\widetilde{\mathcal{O}}\left(TG_H^2\tau_{mix}^{\theta_t}\frac{\log T_{max}}{T_{max}}\right)+\sum_{t=1}^{T}\mathcal{E}_2(t)+T16B^2\mathcal{E}_{app}^2. \tag{98}$$

Combining these expressions and dividing by T completes the proof.

□

D. Average Reward Tracking and Critic Error Analyses

In this section we bound the error arising from the average reward tracking and critic estimation. Combined with the actor gradient norm bound of Section C, this will complete the analysis of Algorithm 1. Our analysis broadly follows that of (Wu et al., 2020), with key modifications leveraging our novel MLMC machinery to handle Markovian sampling in a more streamlined manner.

D.1. Average Reward Tracking Analysis

The main result of this subsection is the following bound on the average reward tracking error.

Theorem D.1. *Assume* $\gamma_t=(1+t)^{-\nu}, \alpha=\alpha_t'/\sqrt{\sum_{k=1}^{t}\|h_t\|^2}$, *and* $\alpha_t'=(1+t)^{-\sigma}$, *where* $0<\nu<\sigma<1$. *Furthermore, assume* $\sup_{s,a}|r(s,a)|\leq R$. *Then*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[(\eta_t-\eta_t^*)^2\right]\leq\mathcal{O}\left(T^{\nu-1}\right)+\mathcal{O}\left(T^{-2(\sigma-\nu)}\right) \tag{99}$$

$$+\widetilde{\mathcal{O}}\left(\max_{t\in[T]}\tau_{mix}^{\theta_t}\log T_{max}\right)\mathcal{O}\left(T^{-\nu}\right) \tag{100}$$

$$+\widetilde{\mathcal{O}}\left(\sqrt{\max_{t\in[T]}\tau_{mix}^{\theta_t}\frac{\log T_{max}}{T_{max}}}\right). \tag{101}$$

Proof. First, recall that the average reward tracking update is given by

$$\eta_{t+1}=\eta_t-\gamma_t f_t, \tag{102}$$

where for brevity we set $f_t:=f_t^{\text{MLMC}}$. We can rewrite the tracking error term $(\eta_{t+1}-\eta_{t+1}^*)^2$ as

$$(\eta_{t+1}-\eta_{t+1}^*)^2=(\eta_{t+1}-\eta_t^*+\eta_t^*-\eta_{t+1}^*)^2 \tag{103}$$

$$=(\eta_t-\gamma_t f_t-\eta_t^*+\eta_t^*-\eta_{t+1}^*)^2. \tag{104}$$

Expanding the squares and regrouping terms yields

$$(\eta_{t+1} - \eta^*_{t+1})^2 = (\eta_t - \eta^*_t)^2 - 2\gamma_t(\eta_t - \eta^*_t)f_t + 2(\eta_t - \eta^*_t)(\eta^*_t - \eta^*_{t+1}) - 2\gamma_t(\eta^*_t - \eta^*_{t+1})f_t + (\eta^*_t - \eta^*_{t+1})^2 + \gamma_t^2(f_t)^2 \tag{105}$$

$$= (\eta_t - \eta^*_t)^2 - 2\gamma_t(\eta_t - \eta^*_t)f_t + 2(\eta_t - \eta^*_t)(\eta^*_t - \eta^*_{t+1}) + (\eta^*_t - \eta^*_{t+1} - \gamma_t f_t)^2. \tag{106}$$

Next, we utilize the bound $(a+b)^2 \leq 2a^2 + 2b^2$ to upper bound the last term in the right hand side of (106) to obtain

$$(\eta_{t+1} - \eta^*_{t+1})^2 \leq (\eta_t - \eta^*_t)^2 - 2\gamma_t(\eta_t - \eta^*_t)f_t + 2(\eta_t - \eta^*_t)(\eta^*_t - \eta^*_{t+1}) + 2(\eta^*_t - \eta^*_{t+1})^2 + 2(\gamma_t f_t)^2. \tag{107}$$

Now notice that the function whose gradient we are estimating with f_t is simply the strongly convex function $F(\eta_t) = \frac{1}{2}(\eta_t - \eta^*_t)^2 = \frac{1}{2}(\eta_t - J(\theta_t))^2$. Clearly $F'(\eta_t) = \eta_t - J(\theta_t)$ is Lipschitz in η_t and F has strong convexity parameter $m_F = 1$. Adding and subtracting $2\gamma_t(\eta_t - \eta^*_t)F'(\eta_t)$ in the above expression gives

$$(\eta_{t+1} - \eta^*_{t+1})^2 \leq (\eta_t - \eta^*_t)^2 - 2\gamma_t(\eta_t - \eta^*_t)F'(\eta_t) + 2\gamma_t(\eta_t - \eta^*_t)(F'(\eta_t) - f_t) + 2(\eta_t - \eta^*_t)(\eta^*_t - \eta^*_{t+1}) + 2(\eta^*_t - \eta^*_{t+1})^2 + 2(\gamma_t f_t)^2. \tag{108}$$

From the strong convexity of F with $m_F = 1$, we can write

$$(\eta_{t+1} - \eta^*_{t+1})^2 \leq (\eta_t - \eta^*_t)^2 - 2\gamma_t(\eta_t - \eta^*_t)^2 + 2\gamma_t(\eta_t - \eta^*_t)(F'(\eta_t) - f_t) + 2(\eta_t - \eta^*_t)(\eta^*_t - \eta^*_{t+1}) + 2(\eta^*_t - \eta^*_{t+1})^2 + 2(\gamma_t f_t)^2 \tag{109}$$

$$= (1 - 2\gamma_t)(\eta_t - \eta^*_t)^2 + 2\gamma_t(\eta_t - \eta^*_t)(F'(\eta_t) - f_t) + 2(\eta_t - \eta^*_t)(\eta^*_t - \eta^*_{t+1}) + 2(\eta^*_t - \eta^*_{t+1})^2 + 2(\gamma_t f_t)^2. \tag{110}$$

Taking expectations and summing yields

$$\sum_{t=1}^T \mathbb{E}[(\eta_t - \eta^*_t)^2] \leq \underbrace{\sum_{t=1}^T \frac{1}{2\gamma_t}\mathbb{E}[(\eta_t - \eta^*_t)^2 - (\eta_t - \eta^*_t)^2]}_{I_1} + \underbrace{\sum_{t=1}^T \mathbb{E}[(\eta_t - \eta^*_t)(F'(\eta_t) - f_t)]}_{I_2} + \underbrace{\sum_{t=1}^T \frac{1}{\gamma_t}\mathbb{E}[(\eta_t - \eta^*_t)(\eta^*_t - \eta^*_{t+1})]}_{I_3} + \underbrace{\sum_{t=1}^T \frac{1}{\gamma_t}\mathbb{E}[(\eta^*_t - \eta^*_{t+1})^2]}_{I_4} + \underbrace{\sum_{t=1}^T \gamma_t \mathbb{E}[(f_t)^2]}_{I_5}. \tag{111}$$

We next provide intermediate bounds for all the terms I_1, I_2, I_3, I_4 and I_5 in the right hand side of (111). We will subsequently manipulate these intermediate bounds to obtain the final bound of Theorem D.1.

Bound on I_1: By rearranging terms in I_1, we get

$$I_1 = \sum_{t=1}^T \frac{1}{2\gamma_t}\mathbb{E}[(\eta_t - \eta^*_t)^2 - (\eta_t - \eta^*_t)^2]$$

$$= \frac{1}{2\gamma_1}\mathbb{E}[(\eta_1 - \eta^*_1)^2] + \sum_{t=2}^T \left(\frac{1}{2\gamma_t} - \frac{1}{2\gamma_{t-1}}\right)\mathbb{E}[(\eta_t - \eta^*_t)^2] - \frac{1}{2\gamma_T}\mathbb{E}[(\eta_{T+1} - \eta^*_{T+1})^2] \tag{112}$$

$$\leq \frac{R^2}{\gamma_T}, \tag{113}$$

where we use the fact that $(\eta_t - \eta^*_t)^2 \leq 2R^2$.

Bound on I_2: For I_2, first notice that $\eta_t, \eta^*_t = J(\theta_t)$ are deterministic conditioned on $\mathcal{F}_{t-1}$ from Lemma B.1. This means we can rewrite the expectation in I_2 as

$$I_2 = \sum_{t=1}^T \mathbb{E}[\mathbb{E}_{t-1}[(\eta_t - \eta^*_t)(F'(\eta_t) - f_t)]] = \sum_{t=1}^T \mathbb{E}[(\eta_t - \eta^*_t)(F'(\eta_t) - \mathbb{E}_{t-1}[f_t])], \tag{114}$$

where $\mathbb{E}_{t-1}[\ldots]$ denotes expectation conditioned on $\mathcal{F}_{t-1}$. From C.2 we know that $\mathbb{E}_{t-1}[f_t] = \mathbb{E}_{t-1}[f_t^{j_{\max}}]$, hence we can write the expression in (114) as

$$I_2 = \sum_{t=1}^{T} \mathbb{E}[(\eta_t - \eta_t^*)(F'(\eta_t) - \mathbb{E}_{t-1}[f_t^{j_{\max}})]] = \sum_{t=1}^{T} \mathbb{E}[\mathbb{E}_{t-1}[(\eta_t - \eta_t^*)(F'(\eta_t) - f_t^{j_{\max}})]] \tag{115}$$

$$= \sum_{t=1}^{T} \mathbb{E}[(\eta_t - \eta_t^*)(F'(\eta_t) - f_t^{j_{\max}})]. \tag{116}$$

Taking absolute values, then applying the triangle, Jensen, and Cauchy-Schwarz inequalities, we can upper bound (116) by

$$|I_2| = \left| \sum_{t=1}^{T} \mathbb{E}[(\eta_t - \eta_t^*)(F'(\eta_t) - f_t^{j_{\max}})] \right| \leq \sum_{t=1}^{T} \mathbb{E}\left[\left| (\eta_t - \eta_t^*)(F'(\eta_t) - f_t^{j_{\max}}) \right| \right]$$

$$\leq \sum_{t=1}^{T} \mathbb{E}\left[\left| (\eta_t - \eta_t^*) \right| \cdot \left| (F'(\eta_t) - f_t^{j_{\max}}) \right| \right]. \tag{117}$$

We know that $|\eta_t - \eta_t^*| \leq 2R$ by assumption, implying

$$|I_2| \leq 2R \sum_{t=1}^{T} \mathbb{E}\left[\left| (F'(\eta_t) - f_t^{j_{\max}}) \right| \right]. \tag{118}$$

By Lemma B.2 with $x_t = \eta_t, \nabla L(x_t) = \nabla F(\eta_t)$ and $l(x_t, z_t) = f_t$, and the fact that the Lipschitz constant of $\nabla F(\eta_t)$ is 1, we obtain the following upper bound on I_2:

$$|I_2| \leq 2R \sum_{t=1}^{T} \widetilde{\mathcal{O}}\left(\sqrt{\tau_{\text{mix}}^{\theta_t} \frac{\log T_{\max}}{T_{\max}}} \right). \tag{119}$$

Bound on I_3: By Hölder's inequality,

$$|I_3| = \left| \sum_{t=1}^{T} \frac{1}{\gamma_t} \mathbb{E}[(\eta_t - \eta_t^*)(\eta_t^* - \eta_{t+1}^*)] \right| \leq \left(\sum_{t=1}^{T} \mathbb{E}\left[(\eta_t - \eta_t^*)^2 \right] \right)^{1/2} \left(\sum_{t=1}^{T} \frac{1}{\gamma_t^2} \mathbb{E}\left[(\eta_t^* - \eta_{t+1}^*)^2 \right] \right)^{1/2}. \tag{120}$$

Notice that $|\eta_t^* - \eta_{t+1}^*| = |J(\theta_t) - J(\theta_{t+1})| \leq L|\theta_t - \theta_{t+1}| \leq LG_H\alpha_t$ due to the Lipschitz continuity of $J(\theta)$ in θ and boundedness of $\|\nabla J(\theta)\|$ from Assumption B.5. This implies

$$|I_3| \leq \left(\sum_{t=1}^{T} \mathbb{E}\left[(\eta_t - \eta_t^*)^2 \right] \right)^{1/2} \left(L^2 G_H^2 \sum_{t=1}^{T} \frac{\alpha_t^2}{\gamma_t^2} \right)^{1/2}. \tag{121}$$

Bound on I_4: Similarly, due to Assumption B.5 we have

$$I_4 = \sum_{t=1}^{T} \frac{1}{\gamma_t} \mathbb{E}[(\eta_t^* - \eta_{t+1}^*)^2] \leq L^2 G_H^2 \sum_{t=1}^{T} \frac{\alpha^2}{\gamma_t}. \tag{122}$$

Bound on I_5: Finally, by Lemma B.3 and taking $G_F = 2R$ without loss of generality, we have

$$I_5 = \sum_{t=1}^{T} \gamma_t \mathbb{E}[(f_t)^2] \leq \sum_{t=1}^{T} \gamma_t \widetilde{\mathcal{O}}\left(R^2 \tau_{\text{mix}}^{\theta_t} \log T_{\max} \right). \tag{123}$$

Combining the foregoing and recalling that $\gamma_t = (1+t)^{-\nu}, \alpha'_t = (1+t)^{-\sigma}$, $0 < \nu < \sigma < 1$, and $\alpha_t \leq \alpha'_t$, we get

$$\sum_{t=1}^{T} \mathbb{E}[(\eta_t - \eta_t^*)^2] \leq 2R^2(1+T)^{\nu} + 2TR\widetilde{\mathcal{O}}\left(\sqrt{\max_{t\in[T]} \tau_{\text{mix}}^{\theta_t} \frac{\log T_{\max}}{T_{\max}}}\right) \tag{124}$$

$$+ LG_H \left(\sum_{t=1}^{T} \mathbb{E}[(\eta_t - \eta_t^*)^2]\right)^{\frac{1}{2}} \left(\sum_{t=1}^{T} (1+t)^{-2(\sigma-\nu)}\right)^{\frac{1}{2}} \tag{125}$$

$$+ L^2 G_H^2 \sum_{t=1}^{T} (1+t)^{(\nu-2\sigma)} + \widetilde{\mathcal{O}}\left(\max_{t\in[T]} \tau_{\text{mix}}^{\theta_t} \log T_{\max}\right) \sum_{t=1}^{T} (1+t)^{-\nu} \tag{126}$$

$$\leq 2R^2(1+T)^{\nu} + \left[L^2 G_H^2 + \widetilde{\mathcal{O}}\left(\max_{t\in[T]} \tau_{\text{mix}}^{\theta_t} \log T_{\max}\right)\right] \sum_{t=1}^{T} (1+t)^{-\nu} \tag{127}$$

$$+ 2TR\widetilde{\mathcal{O}}\left(\sqrt{\max_{t\in[T]} \tau_{\text{mix}}^{\theta_t} \frac{\log T_{\max}}{T_{\max}}}\right) \tag{128}$$

$$+ \left(\sum_{t=1}^{T} \mathbb{E}[(\eta_t - \eta_t^*)^2]\right)^{\frac{1}{2}} \left(L^2 G_H^2 \sum_{t=1}^{T} (1+t)^{-2(\sigma-\nu)}\right)^{\frac{1}{2}}, \tag{129}$$

where the second inequality follows from the fact that $\nu - 2\sigma < -\nu$.

We now manipulate the foregoing inequality to obtain the desired bound. Define

$$A(T) = \sum_{t=1}^{T} \mathbb{E}[(\eta_t - \eta_t^*)^2], \tag{130}$$

$$B(T) = \frac{L^2 G_H^2}{4} \sum_{t=1}^{T} (1+t)^{-2(\sigma-\nu)}, \tag{131}$$

$$C(T) = 2R^2(1+T)^{\nu} + \left[L^2 G_H^2 + \widetilde{\mathcal{O}}\left(\max_{t\in[T]} \tau_{\text{mix}}^{\theta_t} \log T_{\max}\right)\right] \sum_{t=1}^{T} (1+t)^{-\nu} \tag{132}$$

$$+ 2TR\widetilde{\mathcal{O}}\left(\sqrt{\max_{t\in[T]} \tau_{\text{mix}}^{\theta_t} \frac{\log T_{\max}}{T_{\max}}}\right) \tag{133}$$

We can thus rewrite the foregoing inequality as

$$A(T) \leq C(T) + 2\sqrt{A(T)}\sqrt{B(T)}. \tag{134}$$

This expression is equivalent to

$$\left(\sqrt{A(T)} - \sqrt{B(T)}\right)^2 \leq C(T) + B(T), \tag{135}$$

which in turn gives the following chain of implications:

$$\left(\sqrt{A(T)} - \sqrt{B(T)}\right)^2 \leq C(T) + B(T) \implies \sqrt{A(T)} - \sqrt{B(T)} \leq \sqrt{C(T)} + \sqrt{B(T)} \tag{136}$$

$$\implies \sqrt{A(T)} \leq \sqrt{C(T)} + 2\sqrt{B(T)} \tag{137}$$

$$\implies A(T) \leq 2C(T) + 4B(T). \tag{138}$$

As a result, we have shown that

$$\sum_{t=1}^{T}\mathbb{E}[(\eta_t-\eta_t^*)^2] \leq 4R^2(1+T)^{\nu} + \left[2L^2G_H^2 + \widetilde{\mathcal{O}}\left(\max_{t\in[T]}\tau_{\text{mix}}^{\theta_t}\log T_{\max}\right)\right]\sum_{t=1}^{T}(1+t)^{-\nu} \tag{139}$$

$$+ 4TR\widetilde{\mathcal{O}}\left(\sqrt{\max_{t\in[T]}\tau_{\text{mix}}^{\theta_t}\frac{\log T_{\max}}{T_{\max}}}\right) \tag{140}$$

$$+ L^2G_H^2\sum_{t=1}^{T}(1+t)^{-2(\sigma-\nu)}. \tag{141}$$

Using the bound $\sum_{t=1}^{T}(1+t)^{-\xi} \leq \int_0^{t+1}x^{-\xi}dx = (t+1)^{1-\xi}/(1-\xi)$, this implies

$$\sum_{t=1}^{T}\mathbb{E}[(\eta_t-\eta_t^*)^2] \leq O(T^{\nu}) + \widetilde{\mathcal{O}}\left(\max_{t\in[T]}\tau_{\text{mix}}^{\theta_t}\log T_{\max}\right)O(T^{1-\nu}) + O(T^{1-2(\sigma-\nu)}) \tag{142}$$

$$+ T\widetilde{\mathcal{O}}\left(\sqrt{\max_{t\in[T]}\tau_{\text{mix}}^{\theta_t}\frac{\log T_{\max}}{T_{\max}}}\right) \tag{143}$$

Dividing by T completes the proof. □

Notice that, for $\sigma = 0.75$ and $\nu = 0.5$, this result becomes

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[(\eta_t-\eta_t^*)^2] \leq \widetilde{\mathcal{O}}\left(\max_{t\in[T]}\tau_{\text{mix}}^{\theta_t}\log T_{\max}\right)O\left(\frac{1}{\sqrt{T}}\right) + \widetilde{\mathcal{O}}\left(\sqrt{\max_{t\in[T]}\tau_{\text{mix}}^{\theta_t}\frac{\log T_{\max}}{T_{\max}}}\right). \tag{144}$$

D.2. Critic Error Analysis

In this subsection we provide a bound on the critic estimation error term $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\omega_t-\omega_t^*\|^2\right]$ appearing in the main actor analysis bound in Theorem C.4. To get started, we recall some facts about the TD(0) algorithm (Sutton, 1988). As discussed in Ch. 9 of (Sutton & Barto, 2018), for a fixed policy parameter, θ, TD(0) with linear function approximation will converge to the minimum of the mean squared projected Bellman error (MSPBE), which satisfies

$$A_\theta\omega = b_\theta, \tag{145}$$

$$A_\theta = \mathbb{E}_{s\sim\mu_\theta,a\sim\pi_\theta,s'\sim p(\cdot|s,a)}\left[\phi(s)(\phi(s)-\phi(s'))^T\right], \tag{146}$$

$$b_\theta = \mathbb{E}_{s\sim\mu_\theta,a\sim\pi_\theta}\left[(r(s,a)-J(\theta))\phi(s)\right]. \tag{147}$$

The target critic parameter ω_t^* at iteration t of our Algorithm 1 is thus given by $\omega_t^* = A_{\theta_t}^{-1}b_{\theta_t}$. From the definition of g_t^{MLMC}, the critic update $\omega_{t+1} = \omega_t + \beta_t g_t^{MLMC}$ is clearly an attempt to use an MLMC estimator to approximately perform the ideal update $\omega_{t+1} = \omega_t + \beta_t(b_{\theta_t} - A_{\theta_t}\omega_t)$. We can thus view $\nabla G(\omega_t) = b_{\theta_t} - A_{\theta_t}\omega_t$ as the gradient of the true critic objective $G(\omega_t)$ corresponding to using least squares minimization to solve the equation $A_\theta\omega = b_\theta$.

Our task in this section is to characterize the average error that arises when using critic parameters $\{\omega_t\}$ generated by Algorithm 1 to track the ideal parameters $\{\omega_t^*\}$. Before we provide the main result of this section, we need three useful lemmas and an assumption. The first result ensures that the optimal critic parameter is Lipschitz in θ.

Lemma D.2. *Define $P_\theta(s'|s) = \int_{\mathcal{A}}p(s'|s,a)\pi_\theta(a|s)da$, for each θ. Assume that, for all θ, the ergodicity coefficient $\kappa(P_\theta)$ of P_θ satisfies $\kappa(P_\theta) < 1$. Then there exists L_ω such that, for all θ, θ', $\omega^*(\theta) = A_\theta^{-1}b_\theta$ and $\omega^*(\theta') = A_{\theta'}^{-1}b_{\theta'}$ satisfy $\|\omega^*(\theta)-\omega^*(\theta')\| \leq L_\omega\|\theta-\theta'\|$.*

Proof. The result follows by applying the same reasoning as that for Lemma A.3 in (Zou et al., 2019) to the bound from Theorem 3.3 in (Mitrophanov, 2005). □

The next result is an extension of Lemma B.2 to our MLMC critic gradient estimator.

Lemma D.3. *Assume $\|\nabla G(\omega)\| \leq G_G$, for all ω such that $\|\omega\| \leq R_\omega$. Define $D = \sup_s \|\phi(s)\|$. Fix $T_{max} \in \mathbb{N}, \theta_t$ measurable with respect to $\mathcal{F}_{t-1}$, and let $K = \tau^{\theta_t}_{max}\lceil 2T_{max}\rceil$. Define $g_t^N = \frac{1}{N}\sum_{i=1}^N \delta_t^i \phi(s_t^i)$, for $N \in [T_{max}]$, where $\delta_t^i = r_t^i - \eta_t + (\phi(s_t^{i+1}) - \phi(s_t^i))^T\omega_t$. Then, for all $N \in [T_{max}]$,*

$$\mathbb{E}\left[\left\|g_t^N - \nabla G(\omega_t)\right\|\right] \leq O\left(G_G\sqrt{\log KN}\sqrt{\frac{K}{N}}\right) + D\mathbb{E}\left[|\eta_t - \eta_t^*|\right], \tag{148}$$

$$\mathbb{E}\left[\left\|g_t^N - \nabla G(\omega_t)\right\|^2\right] \leq O\left(G_G^2\log(KN)\frac{K}{N}\right) + D^2\mathbb{E}\left[(\eta_t - \eta_t^*)^2\right]. \tag{149}$$

Proof. Define

$$\delta_t^{i,\eta} = r_t^i - \eta_t^* + (\phi(s_t^{i+1}) - \phi(s_t^i))^T\omega_t, \tag{150}$$

$$g_t^{N,\eta} = \frac{1}{N}\sum_{i=1}^N \delta_t^{i,\eta}\phi(s_t^i). \tag{151}$$

Clearly

$$\left\|g_t^N - \nabla G(\omega_t)\right\| \leq \left\|g_t^N - g_t^{N,\eta}\right\| + \left\|g_t^{N,\eta} - \nabla G(\omega_t)\right\| \tag{152}$$

$$= \left\|\frac{1}{N}\sum_{i=1}^N \delta_t^i\phi(s_t^i) - \delta_t^{i,\eta}\phi(s_t^i)\right\| + \left\|\frac{1}{N}\sum_{i=1}^N \delta_t^{i,\eta}\phi(s_t^i) - \nabla G(\omega_t)\right\|. \tag{153}$$

Notice that the first term can be bounded by $D|\eta_t - \eta_t^*|$ and that Lemma B.2 applies to the second term. The remainder of the proof is analogous to that of Lemma C.1. □

Next, we need a critic version of Lemma B.3.

Lemma D.4. *Let $j_{max} = \lfloor \log T_{max}\rfloor$ and fix θ_t measurable w.r.t. $\mathcal{F}_{t-1}$. Assume $T_{max} \geq \tau^{\theta_t}_{mix}$ and $\|\nabla G(\omega)\| \leq G_G$, for all ω such that $\|\omega\| \leq R_\omega$. Then*

$$\mathbb{E}_{t-1}\left[g_t\right] = \mathbb{E}_{t-1}\left[g_t^{j_{max}}\right] \tag{154}$$

$$\mathbb{E}\left[\|g_t\|^2\right] \leq \widetilde{\mathcal{O}}\left(G_G^2\tau^{\theta_t}_{mix}\log T_{max}\right) + 8\log(T_{max})T_{max}D^2\mathbb{E}\left[(\eta_t - \eta_t^*)^2\right]. \tag{155}$$

Proof. The claim follows from Lemma D.3 by the same argument as that used in the proof of Lemma C.2. □

We now provide the main result of this section. The analysis is a modification of that used for the average reward tracking setting.

Theorem D.5. *Assume $\beta_t = (1+t)^{-\nu}, \alpha_t = \alpha_t'/\sqrt{\sum_{k=1}^t \|h_t\|^2}$, and $\alpha_t' = (1+t)^{-\sigma}$, where $0 < \nu < \sigma < 1$. Assume without loss of generality that $\alpha_t \leq \alpha_t'$, for all t. Furthermore, assume that Assumptions B.6 and B.7 hold. Then*

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\|\omega_t - \omega_t^*\|^2\right] \leq \mathcal{O}\left(T^{\nu-1}\right) + \mathcal{O}\left(T^{-2(\sigma-\nu)}\right) \tag{156}$$

$$+ \widetilde{\mathcal{O}}\left(\max_{t\in[T]}\tau^{\theta_t}_{mix}\log T_{max}\right)\mathcal{O}\left(T^{-\nu}\right) \tag{157}$$

$$+ \widetilde{\mathcal{O}}\left(\max_{t\in[T]}\tau^{\theta_t}_{mix}\frac{\log T_{max}}{T_{max}}\right). \tag{158}$$

Proof. By Assumption B.7 and the fact that $\nabla^2 G(\omega) = -A_\theta$, $G(\omega)$ is strongly concave. Let m denote its strong concavity parameter, so that $\langle \nabla G(\omega) - \nabla G(\omega'), \omega - \omega'\rangle \leq -m\|\omega - \omega'\|^2$, for all ω, ω'. Recall that $\omega_{t+1} = \Pi_{R_\omega}(\omega_t + \beta g_t)$, where we use $g_t = g_t^{MLMC}$ for brevity. We have

$$\|\omega_{t+1} - \omega_{t+1}^*\|^2 = \|\Pi_{R_\omega}(\omega_t + \beta g_t) - \omega_{t+1}^*\|^2 \leq \|\omega_t + \beta g_t - \omega_{t+1}^*\|^2, \tag{159}$$

where the inequality holds since $\|\omega_{t+1}^*\| \leq R_\omega$ by definition, so projection can only reduce the distance. Furthermore,

$$\begin{aligned}
\|w_{t+1} - w_{t+1}^*\|^2 &\leq \|w_t - \beta_t h_t - w_t^* + w_t^* - w_{t+1}^*\|^2 && (160)\\
&= \|\omega_t - \omega_t^*\|^2 + 2\beta_t\langle \omega_t - \omega_t^*, g_t\rangle + 2\langle \omega_t - \omega_t^*, \omega_t^* - \omega_{t+1}^*\rangle && (161)\\
&\quad + 2\beta_t\langle \omega_t^* - \omega_{t+1}^*, g_t\rangle + \|\omega_t^* - \omega_{t+1}^*\|^2 + \beta_t^2\|h_t\|^2 && (162)\\
&\overset{(a)}{\leq} \|\omega_t - \omega_t^*\|^2 + 2\beta_t\langle \omega_t - \omega_t^*, g_t\rangle + 2\langle \omega_t - \omega_t^*, \omega_t^* - \omega_{t+1}^*\rangle && (163)\\
&\quad + 2\|\omega_t^* - \omega_{t+1}^*\|^2 + 2\beta_t^2\|h_t\|^2 && (164)\\
&= \|\omega_t - \omega_t^*\|^2 + 2\beta_t\langle \omega_t - \omega_t^*, \nabla G(\omega_t)\rangle + 2\beta_t\langle \omega_t - \omega_t^*, g_t - \nabla G(\omega_t)\rangle && (165)\\
&\quad + 2\langle \omega_t - \omega_t^*, \omega_t^* - \omega_{t+1}^*\rangle + 2\|\omega_t^* - \omega_{t+1}^*\|^2 + 2\beta_t^2\|h_t\|^2 && (166)\\
&\overset{(b)}{\leq} (1 - 2m\beta_t)\|\omega_t - \omega_t^*\|^2 + 2\beta_t\langle \omega_t - \omega_t^*, g_t - \nabla G(\omega_t)\rangle && (167)\\
&\quad + 2\langle \omega_t - \omega_t^*, \omega_t^* - \omega_{t+1}^*\rangle + 2\|\omega_t^* - \omega_{t+1}^*\|^2 + 2\beta_t^2\|h_t\|^2, && (168)
\end{aligned}$$

where (a) follows from completing the square with the last three terms and the fact that $(a+b)^2 \leq 2a^2 + 2b^2$, and (b) follows from the strong concavity of $G(\omega)$.

Rearranging, dividing by $2m\beta_t$, taking expectations, and summing yields

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}[\|w_t - w_t^*\|^2] \leq &\underbrace{\sum_{t=1}^{T}\frac{1}{2m\beta_t}\mathbb{E}[\|w_t - w_t^*\|^2 - \|w_t - w_t^*\|^2]}_{M_1} + \underbrace{\sum_{t=1}^{T}\frac{1}{m}\mathbb{E}[\langle w_t - w_t^*, \nabla G(w_t) - g_t\rangle]}_{M_2}\\
&+ \underbrace{\sum_{t=1}^{T}\frac{1}{m\beta_t}\mathbb{E}[\langle w_t - w_t^*, w_t^* - w_{t+1}^*\rangle]}_{M_3} + \underbrace{\sum_{t=1}^{T}\frac{1}{m\beta_t}\mathbb{E}[\|w_t^* - w_{t+1}^*\|^2]}_{M_4} + \underbrace{\sum_{t=1}^{T}\frac{\beta_t}{m}\mathbb{E}[\|g_t\|^2]}_{M_5}. \qquad (169)
\end{aligned}$$

As in the proof of Theorem D.1, we first provide intermediate bounds on M_1, M_2, M_3, M_4, M_5, then manipulate the resulting expressions to obtain the desired, final bound on the critic error. With the exception of M_2, the intermediate bounds follow by the same reasoning as their counterparts in Theorem D.1.

Bound for M_1: By the same reasoning as for I_1,

$$M_1 \leq \frac{2R_\omega^2}{m\beta_t}. \qquad (170)$$

Bound for M_2: Since ω_t, ω_t^* are deterministic given $\mathcal{F}_{t-1}$, by the law of total expectation and Lemma D.4 we have

$$M_2 = \sum_{t=1}^{T}\frac{1}{m}\mathbb{E}\left[\langle \omega_t - \omega_t^*, g_t^{j_{max}} - \nabla G(\omega_t)\rangle\right]. \qquad (171)$$

Furthermore,

$$\begin{aligned}
|M_2| &\overset{(a)}{\leq} \sum_{t=1}^{T}\frac{1}{m}\mathbb{E}\left[\|\omega_t - \omega_t^*\| \cdot \|g_t^{j_{max}} - \nabla G(\omega_t)\|\right] && (172)\\
&\overset{(b)}{\leq} \sum_{t=1}^{T}\frac{1}{m}\left(\mathbb{E}\left[\|\omega_t - \omega_t^*\|^2\right]\right)^{1/2}\left(\mathbb{E}\left[\|g_t^{j_{max}} - \nabla G(\omega_t)\|^2\right]\right)^{1/2} && (173)\\
&\overset{(c)}{\leq} \left(\frac{1}{m^2}\sum_{t=1}^{T}\mathbb{E}\left[\|\omega_t - \omega_t^*\|^2\right]\right)^{1/2}\left(\sum_{t=1}^{T}\mathbb{E}\left[\|g_t^{j_{max}} - \nabla G(\omega_t)\|^2\right]\right)^{1/2} && (174)\\
&\overset{(d)}{\leq} \left(\frac{1}{m^2}\sum_{t=1}^{T}\mathbb{E}\left[\|\omega_t - \omega_t^*\|^2\right]\right)^{1/2}\left(T\widetilde{\mathcal{O}}\left(G_G^2\max_{t\in[T]}\tau_{mix}^{\theta_t}\frac{\log T_{max}}{T_{max}}\right) + D^2\sum_{t=1}^{T}\mathbb{E}\left[\|\eta_t - \eta_t^*\|^2\right]\right)^{1/2}, && (175)
\end{aligned}$$

where (a) follows by applying the triangle, Jensen's, and Cauchy-Schwarz inequalities, (b) and (c) follow from Hölder's inequality, and (d) results from applying Lemma D.3.

Bound for M_3: Since $\omega^*(\theta)$ is L_ω-Lipschitz in θ by Lemma D.2, we have $\|\omega_t^* - \omega_{t+1}^*\| \leq L_\omega \|\theta_t - \theta_{t+1}\| \leq L_\omega G_H \alpha_t$, where we recall that $\sup_\theta \|\nabla J(\theta)\| \leq G_H$. Thus, by reasoning analogous to I_3,

$$|M_3| \leq \left(\sum_{t=1}^{T} \mathbb{E}\left[\|\omega_t - \omega_t^*\|^2\right]\right)^{1/2} \left(\frac{L_\omega^2 G_H^2}{m^2} \sum_{t=1}^{T} \frac{\alpha_t^2}{\beta_t^2}\right)^{1/2}. \tag{176}$$

Bound for M_4: Similarly,

$$M_4 \leq \frac{L_\omega^2 G_H^2}{m} \sum_{k=1}^{T} \frac{\alpha_t^2}{\beta_t}. \tag{177}$$

Bound for M_5: Finally, by Lemma D.4 and the fact that $|\eta_t| \leq R$, for all t,

$$M_5 \leq \sum_{t=1}^{T} \frac{\beta_t}{m} \left[\widetilde{\mathcal{O}}\left(G_H^2 \tau_{mix}^{\theta_t} \log T_{max}\right) + 8D^2 \log(T_{max}) T_{max} \mathbb{E}\left[(\eta_t - \eta_t^*)^2\right]\right] \tag{178}$$

$$\leq \left[\widetilde{\mathcal{O}}\left(G_H^2 \tau_{mix}^{\theta_t} \log T_{max}\right) + 16D^2 R^2 \log(T_{max}) T_{max}\right] \sum_{k=1}^{T} \frac{\beta_t}{m}. \tag{179}$$

Combining the foregoing and recalling the definitions of $\beta_t, \alpha_t, \alpha_t'$, we have

$$\sum_{t=1}^{T} \mathbb{E}\left[\|\omega_t - \omega_t^*\|^2\right] \leq \frac{2R_\omega}{m}(1+t)^\nu \tag{180}$$

$$+ \left(\frac{1}{m^2} \sum_{t=1}^{T} \mathbb{E}\left[\|\omega_t - \omega_t^*\|^2\right]\right)^{1/2} \left(T\widetilde{\mathcal{O}}\left(G_G^2 \max_{t\in[T]} \tau_{mix}^{\theta_t} \frac{\log T_{max}}{T_{max}}\right) + D^2 \sum_{t=1}^{T} \mathbb{E}\left[\|\eta_t - \eta_t^*\|^2\right]\right)^{1/2} \tag{181}$$

$$+ \left(\sum_{t=1}^{T} \mathbb{E}\left[\|\omega_t - \omega_t^*\|^2\right]\right)^{1/2} \left(\frac{L_\omega^2 G_H^2}{m^2} \sum_{t=1}^{T} (1+t)^{-2(\sigma-\nu)}\right)^{1/2} \tag{182}$$

$$+ \frac{L_\omega^2 G_H^2}{m} \sum_{k=1}^{T} (1+t)^{\nu-2\sigma} \tag{183}$$

$$+ \widetilde{\mathcal{O}}\left(G_H^2 \max_{t\in[T]} \tau_{mix}^{\theta_t} \log(T_{max}) T_{max}\right) \sum_{t=1}^{T} (1+t)^{-\nu}. \tag{184}$$

Define

$$Z(T) = \sum_{t=1}^{T} \mathbb{E}\left[\|\omega_t - \omega_t^*\|^2\right], \tag{185}$$

$$F(T) = \frac{L_\omega^2 G_H^2}{4m^2} \sum_{t=1}^{T} (1+t)^{-2(\sigma-\nu)}, \tag{186}$$

$$G(T) = \frac{1}{16m}\left[T\widetilde{\mathcal{O}}\left(G_G^2 \max_{t\in[T]} \tau_{mix}^{\theta_t} \frac{\log T_{max}}{T_{max}}\right) + D^2 \sum_{t=1}^{T} \mathbb{E}\left[\|\eta_t - \eta_t^*\|^2\right]\right], \tag{187}$$

$$A(T) = \frac{2R_\omega}{m}(1+t)^\nu + \frac{L_\omega^2 G_H^2}{m} \sum_{k=1}^{T} (1+t)^{\nu-2\sigma} + \widetilde{\mathcal{O}}\left(G_H^2 \max_{t\in[T]} \tau_{mix}^{\theta_t} \log(T_{max}) T_{max}\right) \sum_{t=1}^{T} (1+t)^{-\nu}. \tag{188}$$

The previous inequality is thus the same as

$$Z(T) \leq A(T) + 2\sqrt{Z(T)}\sqrt{F(T)} + 2\sqrt{Z(T)}\sqrt{G(T)}, \tag{189}$$

which is in turn equivalent to

$$\left(\sqrt{Z(T)} - \sqrt{F(T)} - \sqrt{G(T)}\right)^2 \leq A(T) + \left(\sqrt{F(T)} + \sqrt{G(T)}\right)^2. \tag{190}$$

This yields

$$\sqrt{Z(T)} - \sqrt{F(T)} - \sqrt{G(T)} \leq \left(A(T) + \left(\sqrt{F(T)} + \sqrt{G(T)}\right)^2\right)^{1/2} \tag{191}$$

$$\leq \sqrt{A(T)} + \sqrt{F(T)} + \sqrt{G(T)}, \tag{192}$$

whence

$$\sqrt{Z(T)} \leq \sqrt{A(T)} + 2\sqrt{F(T)} + 2\sqrt{G(T)} \tag{193}$$

and thus

$$Z(T) \leq 2A(T) + 2\left(2\sqrt{F(T)} + 2\sqrt{G(T)}\right)^2 \tag{194}$$

$$\leq 2A(T) + 16F(T) + 16G(T). \tag{195}$$

Noticing that $2A(T) + 16F(T) = \mathcal{O}\left(T^{\nu}\right) + \mathcal{O}\left(T^{1+\nu-2\sigma}\right) + \mathcal{O}\left(T^{1-\nu}\right)$ and using the bound $\sum_{t=1}^{T}(1+t)^{-\xi} \leq (1+t)^{1-\xi}/(1-\xi)$, we have

$$\sum_{t=1}^{T}\mathbb{E}\left[\left\|\omega_t - \omega_t^*\right\|^2\right] \leq \frac{1}{m}\left[T\widetilde{\mathcal{O}}\left(G_G^2 \max_{t\in[T]}\tau_{mix}^{\theta_t}\frac{\log T_{max}}{T_{max}}\right) + D^2\sum_{t=1}^{T}\mathbb{E}\left[\left\|\eta_t - \eta_t^*\right\|^2\right]\right] \tag{196}$$

$$+ \mathcal{O}\left(T^{\nu}\right) + \mathcal{O}\left(T^{1+\nu-2\sigma}\right) + \mathcal{O}\left(T^{1-\nu}\right). \tag{197}$$

Dividing by T, combining with Theorem D.1, and absorbing constants into the order notation finishes the proof.

□

E. Proof of Theorem 4.8

Proof. From the statement of Theorems 4.6 and 4.7, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla J(\theta_t)\right\|^2\right] \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \mathcal{O}\left(\frac{1}{T}\sum_{t=1}^{T}\mathcal{E}(t)\right) + \widetilde{\mathcal{O}}\left(\sqrt{\max_{t\in[T]}\tau_{mix}^{\theta_t}\frac{\log T_{\max}}{T_{\max}}}\right) + \mathcal{O}\left(\mathcal{E}_{app}\right), \tag{198}$$

and

$$\frac{1}{T}\sum_{t=1}^{T}\mathcal{E}(t) \leq \mathcal{O}\left(T^{\nu-1}\right) + \mathcal{O}\left(T^{-2(\sigma-\nu)}\right) + \widetilde{\mathcal{O}}\left(\max_{t\in[T]}\tau_{mix}^{\theta_t}\log T_{\max}\right)\mathcal{O}\left(T^{-\nu}\right) + \widetilde{\mathcal{O}}\left(\sqrt{\max_{t\in[T]}\tau_{mix}^{\theta_t}\frac{\log T_{\max}}{T_{\max}}}\right). \tag{199}$$

utilizing the upper bound in (199) into the right hand side of (198), we get

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla J(\theta_t)\right\|^2\right] \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \mathcal{O}\left(T^{\nu-1}\right) + \mathcal{O}\left(T^{-2(\sigma-\nu)}\right) + \widetilde{\mathcal{O}}\left(\max_{t\in[T]}\tau_{mix}^{\theta_t}\log T_{\max}\right)\mathcal{O}\left(T^{-\nu}\right)$$

$$+ \widetilde{\mathcal{O}}\left(\sqrt{\max_{t\in[T]}\tau_{mix}^{\theta_t}\frac{\log T_{\max}}{T_{\max}}}\right) + \mathcal{O}\left(\mathcal{E}_{app}\right). \tag{200}$$

For the selection $\nu = 0.5$ and $\sigma = 0.75$ (which satisfies the constraint that $0 < \nu < \sigma < 1$), we obtain

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right] \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \widetilde{\mathcal{O}}\left(\max_{t\in[T]}\tau_{mix}^{\theta_t}\log T_{\max}\right)\mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \widetilde{\mathcal{O}}\left(\sqrt{\max_{t\in[T]}\tau_{mix}^{\theta_t}\frac{\log T_{\max}}{T_{\max}}}\right) + \mathcal{O}\left(\mathcal{E}_{app}\right). \tag{201}$$

Therefore, after further simplification, we can write

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right] \leq \widetilde{\mathcal{O}}\left(\max_{t\in[T]}\tau_{mix}^{\theta_t}\log T_{\max}\right)\mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \widetilde{\mathcal{O}}\left(\sqrt{\max_{t\in[T]}\tau_{mix}^{\theta_t}\frac{\log T_{\max}}{T_{\max}}}\right) + \mathcal{O}\left(\mathcal{E}_{app}\right). \tag{202}$$

completes the proof.

□

F. Hyperparametrs for the Experiments

We list all the hyperparameters in Table 2 here.

Table 2. This table compares the hyperparameters and performance between the four experiments, each run for five trials. From the table, we see that given the same learning rates, environment, and the number of samples, MAC and Vanilla AC converge to the same reward value.

Method	Learning Rate			Grid Size	$T_{\max}$	Samples Processed	Limiting Mean Reward	Limiting Policy Gradient Norm
	Actor	Critic	Reward Estimator					
MAC	.01	.01	.01	6×6	8	$3 \cdot 10^6$	0.4	0
Vanilla AC	.01	.01	.01	6×6	3	$3 \cdot 10^6$	0.4	0
MAC	.005	.005	.005	10×10	16	$4 \cdot 10^6$	0.5	0
Vanilla AC	.005	.005	.005	10×10	4	$4 \cdot 10^6$	0.5	0