On the Second-Order Convergence of Biased Policy Gradient Algorithms

Siqiao Mu [1] **Diego Klabjan** [2]

Abstract

Since the objective functions of reinforcement learning problems are typically highly nonconvex, it is desirable that policy gradient, the most popular algorithm, escapes saddle points and arrives at second-order stationary points. Existing results only consider vanilla policy gradient algorithms with unbiased gradient estimators, but practical implementations under the infinite-horizon discounted reward setting are biased due to finite-horizon sampling. Moreover, actor-critic methods, whose second-order convergence has not yet been established, are also biased due to the critic approximation of the value function. We provide a novel second-order analysis of biased policy gradient methods, including the vanilla gradient estimator computed from Monte-Carlo sampling of trajectories as well as the double-loop actor-critic algorithm, where in the inner loop the critic improves the approximation of the value function via TD(0) learning. Separately, we also establish the convergence of TD(0) on Markov chains irrespective of initial state distribution.

1. Introduction

In the standard reinforcement learning framework, an agent interacts with an environment according to some policy, which dictates the best actions to take given the state of the environment. The ultimate goal of the agent is to adopt a policy that maximizes some measure of cumulative reward. To efficiently search for the optimal policy, policy gradient methods optimize a policy parameter θ through updates that approximate the gradient of the objective function with respect to θ. These algorithms can be fast and flexible, and under mild assumptions, they share convergence properties with mainstream gradient descent algorithms.

The policy gradient can be estimated by way of the policy gradient theorem, which enables the straightforward application of existing techniques in gradient-based optimization towards understanding PG convergence. The theorem provides a formula for the exact gradient of the objective function as the expectation of the state-action value function over the discounted state-action measure (Sutton et al., 1999). In practice, the gradient is estimated through two common approaches: the "vanilla" approach, where the gradient is approximated via Monte-Carlo sampling, or the "actor-critic" approach, where the policy parameter θ, called the "actor parameter," is updated simultaneously alongside a "critic parameter" w, which parametrizes the state-action value function. The critic parameter is typically updated via bootstrapping temporal difference methods such as TD(0), although the actor updates may also be bootstrapped (Konda & Tsitsiklis, 1999; Bhatnagar et al., 2009).

Although it is well-established that policy gradient converges to first-order stationary points where the norm of the gradient is approximately zero (Sutton et al., 1999; Yuan et al., 2022; Agarwal et al., 2020), it is of interest whether policy gradient algorithms yield second-order stationary points (local maxima) as opposed to saddle points. This is because the function landscape of RL problems can be highly nonconvex and features suboptimal stationary points even in very simple examples (Agarwal et al., 2020; Bhandari & Russo, 2019; Zhang et al., 2020). We can utilize seminal results in nonconvex optimization that show that stochastic gradient descent can leverage randomness to escape saddle points either with added noise (Ge et al., 2015; Jin et al., 2019) or as long as there is a component of noise in the direction of curvature (Daneshmand et al., 2018; Vlaski & Sayed, 2022).

However, unlike stochastic gradient descent, practical implementations of policy gradient are frequently biased. For the discounted infinite-horizon reward objective function, the vanilla policy gradient estimator is biased due to truncation of sampled trajectory horizons (Yuan et al., 2022). In addition, actor-critic algorithms introduce a second source of bias in the critic's approximation of the value function. This therefore requires novel techniques for controlling and

[*]Equal contribution [1]Department of Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, IL [2]Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL. Correspondence to: Siqiao Mu <siqiaomu2026@u.northwestern.edu>, Diego Klabjan <d-klabjan@northwestern.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

bounding the bias. In comparison, existing works on second-order convergence of policy gradient only consider unbiased gradient estimators; for example, (Zhang et al., 2020) require artificially constructing an unbiased gradient estimator via Q-sampling as well as periodic step size enlargement, and (Yang et al., 2021) assume the gradient estimator is unbiased. Both works also only consider the vanilla policy gradient algorithm and not the actor-critic scheme.

In this paper, we address the aforementioned gap by showing that biased policy gradient methods, including both vanilla policy gradient and actor-critic methods, can escape saddle points. Borrowing from the analyses of (Vlaski & Sayed, 2022) regarding unbiased stochastic gradient descent, we tackle the biased policy gradient setting by showing that the components of the bias can be sufficiently controlled to yield second-order convergence guarantees.

Related Work

Escaping Saddle Points. As referenced earlier, escaping saddle points has become a central research topic in nonconvex optimization in the last few years, with natural extensions to the policy gradient setting. In addition to the seminal works showing the second-order convergence of gradient descent and stochastic gradient descent (Ge et al., 2015; Jin et al., 2019; 2017; Daneshmand et al., 2018; Vlaski & Sayed, 2022), an additional body of work has focused on second-order methods that use Hessian information to arrive at second-order stationary points faster. Some examples of reinforcement learning approaches in this direction include (Shen et al., 2019; Wang et al., 2022; Khorasani et al., 2023). However, these works only consider unbiased gradient estimators. For an in-depth comparison of sample complexities, see Appendix A.2.

Global Convergence. Separate from our line of work, there are several "global convergence" results for policy gradient and actor-critic algorithms that ensure convergence to a global optimum for specific policy parametrization or function structure. For instance, global convergence is established for the following settings: tabular or tabular softmax policy parametrization with exact gradients (Agarwal et al., 2020; Bhandari & Russo, 2021), objective functions that satisfy the gradient domination property (Bhandari & Russo, 2019), linear quadratic and nearly linear-quadratic control systems (Yang et al., 2019; Han et al., 2022), and over-parametrized neural networks (Wang et al., 2020; Fu et al., 2020). The premise of "global convergence" also requires some assumption that the proposed policy parametrization can approximate the optimal policy with arbitrary precision, i.e., the optimal policy lies within the policy class. In comparison, our second-order convergence guarantees pertain to finding the best policy parameter θ under a generic policy parametrization, for policy gradient algorithms with noisy and biased updates. For an in-depth discussion of global convergence rates, see Appendix A.1.

Vanilla Policy Gradient. In this work, we refer to policy gradient algorithms that estimate the gradient via Monte-Carlo sampling of trajectories as "vanilla policy gradient." Early formulas include REINFORCE (Williams, 2004) as well as GPOMDP (Baxter & Bartlett, 2001), a version of REINFORCE that enjoys reduced variance (Peters & Schaal, 2008) by employing the "reward-to-go" trick. Our work pertains to the GPOMDP estimator. Both REINFORCE and GPOMDP are unbiased estimators of objectives with deterministic, fixed horizons, but they are biased estimators of the infinite-horizon discounted reward due to truncation (Yuan et al., 2022).

Actor-critic Algorithms. The asymptotic convergence of actor-critic algorithms was first established in (Konda & Tsitsiklis, 1999) for gradient actor updates and bootstrapped critic updates, and in (Bhatnagar et al., 2009) for bootstrapped actor and critic updates. Since then, finite-time convergence has been established for a variety of actor-critic frameworks, although to the best of our knowledge no second-order convergence result exists. In this work we consider double-loop actor-critic algorithms where the critic parameter undergoes TD(0) updates in the inner loop and the actor parameter undergoes gradient updates in the outer loop. Several works have shown first-order convergence of these algorithms with various additional settings; (Yang et al., 2018) consider i.i.d sampling in the actor and the critic, (Qiu et al., 2021) consider i.i.d. policy samples and critic sampling from a stationary Markov chain, and (Xu et al., 2020b) study mini-batch Markovian sampling for controlling the bias error. Separately, (Wang et al., 2020) establish global convergence for double-loop algorithms where the actor and critic functions are compatible overparametrized neural networks. In comparison, we consider linear critic parametrization and arbitrary actor parametrization with Markovian sampling, and we do not require the Markov chain to be stationary.

Recently, two-timescale (Xu et al., 2020a; Wu et al., 2020) and single-timescale (Fu et al., 2020; Olshevsky & Gharesifard, 2023) actor-critic algorithms have shown a slight performance advantage over double-loop algorithms, although existing works still only analyze their first-order or global convergence under specific function parametrizations, while we focus on second-order convergence.

Temporal Difference Algorithms. Actor-critic algorithms typically feature some bootstrapping element; in particular, we consider actor-critic algorithms where the critic updates via TD(0) learning. The finite-time convergence of TD(0) has been established recently under independent and identically distributed sampling (Dalal et al., 2018; Kumar et al., 2019) as well as under Markovian sampling (Liu & Olshevsky, 2020; Bhandari et al., 2018). The latter is more

relevant to the RL setting; however, both (Bhandari et al., 2018) and (Liu & Olshevsky, 2020) assume the Markov chain begins in the stationary distribution, which is unrealistic for practical implementations such as in actor-critic algorithms.

Our Contributions

Our contributions are as follows.

- We provide the first finite-time convergence analysis of vanilla policy gradient with biased gradient estimator to ϵ-second-order stationary points. This results in a sample complexity of $\tilde{O}(\epsilon^{-6.5})$ iterations, where $\tilde{O}(\cdot)$ hides logarithmic dependencies. We note that this is a stronger result than $\tilde{O}(\epsilon^{-9})$ from (Zhang et al., 2020) and a weaker result than $\tilde{O}(\epsilon^{-4.5})$ from (Yang et al., 2021), both of which only analyze unbiased gradient estimators.

- We provide the first finite-time convergence analysis of an actor-critic policy gradient algorithm to ϵ-second-order stationary points. We show that our double-loop actor-critic algorithm, where the critic updates via $TD(0)$ and the actor updates via policy gradient, arrives at an ϵ-second-order stationary point in $\tilde{O}(\epsilon^{-6.5})$ outer loop iterations with $\tilde{O}(\epsilon^{-8})$ inner-loop TD(0) steps. In contrast to existing first-order analyses of actor-critic algorithms, we allow for Markovian sampling in both the actor and the critic.

- Of separate interest, we provide the first finite-time convergence analysis of the classic TD(0) algorithm on nonstationary Markov chains; i.e., we do not assume that the initial state distribution is the stationary distribution of the Markov chain. This allows realistic analyses of the actor-critic setting, where we have no guarantee that after every policy update the new underlying Markov chain is in its stationary distribution. We show that for K constant timesteps $\alpha = \frac{1}{\sqrt{K}}$ and exponential mixing, the algorithm converges at the rate of $O(\frac{1}{\sqrt{K}}) + O(\frac{1}{K})$, as opposed to $O(\frac{1}{\sqrt{K}})$ when starting from the stationary distribution.

The structure of the paper is as follows. In Section 2, we formalize the problem and introduce notation used in the rest of the paper. In Section 3, we establish second-order convergence for biased policy gradient estimators and apply our results to vanilla policy gradient. In Section 4, we present our new analysis of the TD(0) algorithm, which is incorporated to bound the critic approximation bias and show second-order convergence of the actor-critic algorithm. Finally, in Section 5, we summarize our results and discuss the next steps of our work.

2. Problem Formulation

2.1. Markov Decision Process

We define the Markov decision process as a quadruple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P}(s'|s,a)$ is the transition probability from state s to state s' by taking action a, and $\mathcal{R}(s,a)$ is the reward function for performing action a in state s. The agent is trying to learn a stochastic policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ is the space of probability distributions over $\mathcal{A}$, such that $\pi(a|s)$ is the probability that the agent performs action a given state s. As the agent interacts with the environment, it generates a sequence of states, actions, and rewards referred to as a trajectory $\tau = \{s_0, a_0, s_1, a_1, ...\}$. The trajectory is sampled from the probability distribution $p(\cdot|\pi)$, which describes the probability of a trajectory generated by some policy π, where $a_k \sim \pi(\cdot|s_k)$ and $s_{k+1} \sim \mathcal{P}(\cdot|s_k, a_k)$.

We want to learn a policy π that maximizes the expected infinite-horizon discounted reward, J. For policy gradient algorithms, we parameterize the policy π with some parameter $\theta \in \mathbb{R}^M$ so that J is a function of θ to obtain

$$J(\theta) = \mathbb{E}_{s_0 \sim \rho_0(\cdot), \tau \sim p(\cdot|\pi_\theta)}[\sum_{k=0}^{\infty} \gamma^k \mathcal{R}(s_k, a_k)]$$

where $\gamma \in (0,1)$ is the discount factor and the expectation is taken with respect to an initial state distribution $s_0 \sim \rho_0(\cdot)$ and a stochastic policy π_θ under which trajectories are sampled $\tau \sim p(\cdot|\pi_\theta)$. The RL problem is to find an optimal policy parameter θ^* such that $\theta^* = \arg\max_\theta J(\theta)$.

We also define the state value function and state-action value function as $V^\pi(s) = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k \mathcal{R}(s_k, a_k)|s_0 = s]$ and $Q^\pi(s,a) = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k \mathcal{R}(s_k, a_k)|s_0 = s, a_0 = a]$ respectively, such that the objective can be alternatively formulated as $J(\theta) = \mathbb{E}_{s \sim \rho_0}[V^{\pi_\theta}(s)]$. Finally, we reference the discounted state-weighting measure as $d^{\pi_\theta, \rho_0}(s) = \sum_{k=0}^{\infty} \gamma^k \mathbb{E}_{\rho_0}[Pr(s_k = s\,|s_0, \pi_\theta)]$.

2.2. Policy Gradient Algorithm

Algorithm 1 Biased Policy Gradient Algorithm

Input: initial policy parameters θ_0
for $t = 0, 1, 2, ...T-1$ **do**
  Sample a trajectory τ_t of length H under π_{θ_t},

$$\tau_t = \{s_0, a_0, s_1, a_1, ...s_{H-1}, a_{H-1}\}$$

  where $s_0 \sim \rho_0(s)$
  Compute $\hat{G}(\theta_t; \tau_t)$ via **Algorithm 2** or **Algorithm 3**
  Update $\theta_{t+1} = \theta_t + \mu \hat{G}(\theta_t; \tau_t)$
end for

In this work we consider policy gradient algorithms of the form shown in Algorithm 1. We note that these gradient

algorithms can be mini-batched for additional variance reduction, but our analysis does not rely on batching and we omit this notation for simplicity. At each time step $t+1$, a random trajectory τ_t is sampled and $\hat{G}(\theta_t;\tau_t)$, a biased estimator of the true policy gradient $\nabla J(\theta_t)$, is computed. Then the policy parameter θ is updated with step-size μ as follows

$$\theta_{t+1} = \theta_t + \mu\hat{G}(\theta_t;\tau_t). \tag{1}$$

Let $G(\theta_t) = \mathbb{E}_\tau[\hat{G}(\theta_t;\tau)]$, where the expectation is taken with respect to the sampled trajectory τ. Then in order to analyze the convergence of Algorithm 1, we decompose the updates in terms of the noise ξ_{t+1} and bias d_{t+1} to obtain

$$\theta_{t+1} = \theta_t + \mu\nabla J(\theta_t) + \mu\xi_{t+1} + \mu d_{t+1},$$

where $\xi_{t+1} = \hat{G}(\theta_t;\tau_t) - G(\theta_t)$ represents a zero-mean noise term induced from sampling the trajectory of length H and $d_{t+1} = G(\theta_t) - \nabla J(\theta_t)$ represents the bias induced by the gradient estimator algorithm. Specifically, we denote by $\{\mathcal{F}_t\}_{t\geq 0}$ the filtration generated by the random process $\{\theta_t\}_{t\geq 0}$ such that θ_t is measurable with respect to $\mathcal{F}_t$. Then the gradient noise process $\{\xi_t\}_{t\geq 0}$ satisfies

$$\mathbb{E}[\xi_{t+1}|\mathcal{F}_t] = \mathbb{E}[\hat{G}(\theta_t;\tau_t) - G(\theta_t)|\mathcal{F}_t] = 0.$$

Although d_{t+1} might be random or deterministic depending on the gradient estimator algorithm, it is generated by a different process than ξ_{t+1}. In the sequel, our analyses rely on assumptions on ξ_{t+1} and d_{t+1}.

2.3. Second-Order Stationary Points

We define second-order stationary points and approximate second-order stationary points as follows (Jin et al., 2017; Nesterov & Polyak, 2006).

Definition 2.1. For the twice-differentiable function $J(\theta)$, θ is a second-order stationary point if $\|\nabla J(\theta)\| = 0$ and $\lambda_{max}(\nabla^2 J(\theta)) \leq 0$. In addition, if $\nabla^2 J(\theta)$ is χ-Lipschitz, θ is an ϵ-second order stationary point of $J(\theta)$ if $\|\nabla J(\theta)\| \leq \epsilon$ and $\lambda_{max}(\nabla^2 J(\theta)) \leq \sqrt{\chi\epsilon}$.

In line with the *strict-saddle* definition introduced in (Ge et al., 2015) and later used in (Jin et al., 2019; 2017; Daneshmand et al., 2018), we focus on escaping saddle points with at least one strictly positive eigenvalue. We divide the parameter space of the objective function $J(\theta)$ into regions where the gradient is large or small with respect to the step-size μ, which we assume to be a small hyperparameter. We define the following sets

$$\mathcal{G} = \left\{\theta : \|\nabla J(\theta)\|^2 \geq \mu\ell\left(1+\frac{1}{\delta}\right)\right\},$$

$$\mathcal{H} = \{\theta : \theta \in \mathcal{G}^C, \lambda_{max}(\nabla^2 J(\theta)) \geq \omega\},$$

$$\mathcal{M} = \{\theta : \theta \in \mathcal{G}^C, \lambda_{max}(\nabla^2 J(\theta)) < \omega\},$$

where $\ell > 0$ is a parameter depending on the problem and $\delta > 0$, $\omega > 0$ are parameters depending on the desired accuracy of the algorithm that we choose later on. The set $\mathcal{G}^C$ represents approximate first-order stationary points, and within $\mathcal{G}^C$, the region $\mathcal{H}$ represents "strict-saddle" points, whereas the region $\mathcal{M}$ represents approximately second-order stationary points. Specifically, points in $\mathcal{M}$ represent the set of local maxima with respect to first and second-order information.

3. Second-Order Convergence of Vanilla Policy Gradient

In this section, we establish the second-order convergence of biased policy gradient in general and apply it to vanilla policy gradient. We begin with the original policy gradient theorem (Sutton et al., 1999), which states

$$\nabla J(\theta) = \sum_{s\in\mathcal{S}} d^{\pi_\theta,\rho_0}(s)\sum_{a\in\mathcal{A}}\pi_\theta(a|s)\nabla\log\pi_\theta(a|s)Q^{\pi_\theta}(s,a),$$

which is equivalent to the temporal summation

$$\nabla J(\theta) = \mathbb{E}_{\pi_\theta,\rho_0}[\sum_{k=0}^{\infty}\sum_{t=k}^{\infty}\gamma^t\nabla\log\pi_\theta(a_k|s_k)\mathcal{R}(s_t,a_t)].$$

For further details on the connection between these two formulations, see Appendix F.

Through Monte-Carlo sampling of trajectories τ of fixed length H, we construct the GPOMDP gradient estimator (Baxter & Bartlett, 2001) as follows

$$\hat{G}^{VPG}(\theta;\tau) = \sum_{h=0}^{H-1}\nabla\log\pi_\theta(a_h|s_h)\sum_{i=h}^{H-1}\gamma^i\mathcal{R}(s_i,a_i).$$

Algorithm 2 Vanilla Policy Gradient Estimator

Input: policy parameter θ_t, trajectory $\tau_t = \{s_0,a_0,s_1,a_1,...s_{H-1},a_{H-1}\}$
Compute gradient estimator from τ_t:

$$\hat{G}^{VPG}(\theta_t;\tau_t) = \sum_{h=0}^{H-1}\nabla\log\pi_\theta(a_h|s_h)\sum_{i=h}^{H-1}\gamma^i\mathcal{R}(s_i,a_i)$$

return $\hat{G}^{VPG}(\theta_t;\tau_t)$

This yields the "vanilla" policy gradient algorithm as outlined in Algorithm 2. We define the truncated or finite-horizon objective $J_H(\theta) = \mathbb{E}_{\pi_\theta,\rho_0}[\sum_{t=0}^{H-1}\gamma^t\mathcal{R}(s_t,a_t)]$. As observed in (Yuan et al., 2022; Zhang et al., 2020; Wu et al., 2022), since we cannot practically sample infinite-horizon trajectories, $\hat{G}^{VPG}(\theta;\tau)$ is a *biased* gradient estimator of $J(\theta)$ and an unbiased gradient estimator of $J_H(\theta)$,

such that

$$\mathbb{E}[\hat{G}^{VPG}(\theta;\tau)] = \nabla J_H(\theta) \neq \nabla J(\theta).$$

3.1. Assumptions

We require the following assumptions for the convergence of biased policy gradient algorithms.

Assumption 3.1. The following conditions hold for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and θ:

1. The rewards are bounded such that there exists $\mathcal{R}_{max} > 0$ with $|\mathcal{R}(s,a)| \leq \mathcal{R}_{max}$.
2. The policy score function $\nabla \log \pi_\theta(a|s)$ exists and its norm is bounded by a constant $G > 0$ such that $\|\nabla \log \pi_\theta(a|s)\| \leq G$.
3. The Jacobian of the score function exists and its norm is bounded by a constant $B > 0$ such that $\|\nabla^2 \log \pi_\theta(a|s)\| < B$, and it is Lipschitz continuous such that for all θ_1, θ_2, we have

$$\|\nabla^2 \log \pi_{\theta_1}(a|s) - \nabla^2 \log \pi_{\theta_2}(a|s)\| \leq \iota \|\theta_1 - \theta_2\|.$$

In Assumption 3.1, we assume that the reward and the policy log gradient are bounded, assumptions first put forward in (Papini et al., 2018) and since widely adopted in many theoretical analyses of policy gradient (Zhu & Gong, 2023; Zhang et al., 2020; Yang et al., 2021). Assumption 3.1 is satisfied by commonly used policy parametrizations, including the softmax policy parametrization $\pi_\theta(a|s) = \frac{e^{h(s,a,\theta)}}{\sum_b e^{h(s,b,\theta)}}$. The action preferences $h(s,a,\theta)$ can be parametrized via deep neural networks or other functions of the feature vectors $\phi(s,a)$. The assumption is also satisfied by Gaussian policies such as $\pi_\theta(a|s) \sim \mathcal{N}(\phi(s)^\intercal \theta, \sigma^2)$ if the parameter θ lies in some bounded set and the actions and the feature vectors $\phi(s)$ are bounded (Zhang et al., 2020). In addition, Assumption 3.1 has several important implications as follows.

Lemma 3.2. *(Lemma 3.2 of (Zhang et al., 2020)) The score function $\nabla \log \pi_\theta(a|s)$ is B-Lipschitz continuous. Moreover, the policy gradient $\nabla J(\theta)$ is Lipschitz continuous such that for all θ_1, θ_2, we have*

$$\|\nabla J(\theta_1) - \nabla J(\theta_2)\| \leq L\|\theta_1 - \theta_2\|$$

where $L = \frac{\mathcal{R}_{max} B}{(1-\gamma)^2} + \frac{(1+\gamma)\mathcal{R}_{max} G^2}{(1-\gamma)^3}$.

Lemma 3.3. *(Lemma 5.4 from (Zhang et al., 2020)) The Hessian matrix of $J(\theta)$ is Lipschitz continuous such that for all θ_1, θ_2, we have*

$$\|\nabla^2 J(\theta_1) - \nabla^2 J(\theta_2)\| \leq \chi \|\theta_1 - \theta_2\|$$

where $\chi = \frac{\mathcal{R}_{max} G B}{(1-\gamma)^2} + \frac{\mathcal{R}_{max} G^3 (1+\gamma)}{(1-\gamma)^3} + \frac{\mathcal{R}_{max} G}{1-\gamma} \cdot \max\{B, \frac{G^2\gamma}{1-\gamma}, \frac{\iota}{G}, \frac{B\gamma}{1-\gamma}, \frac{G^2(1+\gamma)+B(1-\gamma)\gamma}{(1-\gamma)^2}\}$.

Finally, we require assumptions on the noise of the algorithm that allows the iterates to escape saddle points.

Assumption 3.4. The covariance matrix of the noise ξ_{t+1} generated at iterate θ_t, defined as $R_\xi(\theta_t) = \mathbb{E}[\xi_{t+1}\xi_{t+1}^\intercal | \mathcal{F}_t]$, is Lipschitz such that

$$\|R_\xi(\theta_1) - R_\xi(\theta_2)\| \leq \beta_R \|\theta_1 - \theta_2\|^\nu$$

where $0 < \nu \leq 4$ and $\beta_R > 0$.

Assumption 3.5. Let $\nabla^2 J(\theta) = V_\theta \Lambda_\theta V_\theta^T$ be the eigendecomposition of the Hessian matrix at θ where the eigenvalues and eigenvectors are ordered as follows

$$V_\theta = \begin{bmatrix} V_\theta^{>0} & V_\theta^{\leq 0} \end{bmatrix}, \qquad \Lambda_\theta = \begin{bmatrix} \Lambda_\theta^{>0} & 0 \\ 0 & \Lambda_\theta^{\leq 0} \end{bmatrix}$$

where $\Lambda_\theta^{>0} > 0$ and $\Lambda_\theta^{\leq 0} \leq 0$. Then, we assume that there exists $\sigma_l^2 > 0$ such that for any approximate strict-saddle point $\theta_t \in \mathcal{H}$

$$\lambda_{min}((V_{\theta_t}^{>0})^\intercal \mathbb{E}[\xi_{t+1}\xi_{t+1}^\intercal] V_{\theta_t}^{>0}) \geq \sigma_l^2.$$

Assumption 3.4 states that the covariance matrix is Lipschitz, an assumption proposed in (Vlaski & Sayed, 2022; Vlaski et al., 2020). Assumption 3.4 requires that the covariance of noise does not change too much between iterates, which can be ensured by restricting our parameter search space to a compact set, which reflects common practices. Assumption 3.5, or the condition of "correlated negative curvature," states that there must be a component of noise in the direction of negative curvature in order for the noise to help the iterates escape the saddle point. It is necessary for most second-order convergence analyses (Daneshmand et al., 2018; Zhang et al., 2020; Yang et al., 2021). Like (Vlaski & Sayed, 2022), we note that although Assumption 3.5 is a technical requirement for convergence, the condition can be achieved by simply adding isotropic noise at each parameter update iteration like in (Ge et al., 2015; Jin et al., 2019).

3.2. Main Result

We first present the following theorem for general biased policy gradient, which shows that if the gradient estimator and bias are sufficiently bounded, we can conclude second-order convergence. Theorem 3.6 adapts Theorem 3 from (Vlaski & Sayed, 2022) regarding second-order convergence of unbiased stochastic gradient descent. Our approach follows the framework proposed in (Vlaski et al., 2020) for showing second-order convergence of federated learning, a form of biased stochastic gradient descent.

Theorem 3.6. *For the iterates θ_t of Algorithm 1, suppose that Assumptions 3.1-3.5 hold and for $\sigma > 0$, $D > 0$, the following hold*

$$\|\hat{G}(\theta_t;\tau_t)\| \leq \sigma \tag{2}$$

$$\mathbb{E}[\|d_{t+1}\|^4|\mathcal{F}_t] \leq D^4\mu^4 \quad (3)$$

$$\mathbb{E}[\|d_{t+1}\|^2\|\xi_{t+1}\|^2|\mathcal{F}_t] \leq \sigma^2\mathbb{E}[\|d_{t+1}\|^2|\mathcal{F}_t]. \quad (4)$$

Let $\ell = L\sigma^2 - D^2\mu$. Then with probability $1-\delta$, Algorithm 1 yields $\theta_T \in \mathcal{M}$ such that $\|\nabla J(\theta_T)\|^2 \leq \mu\ell(1+\frac{1}{\delta})$ and $\lambda_{max}(\nabla^2 J(\theta_T)) \leq \omega$ in T iterations, where

$$T \leq \frac{4\mathcal{R}_{max}}{\mu^2(1-\gamma)(L\sigma^2+D^2\mu)\delta}\cdot\mathcal{T}$$

$$\mathcal{T} = \frac{\log(2M\frac{\sigma^2}{\sigma_l^2}+1)}{\log(1+2\mu\omega)}.$$

Proof. See Appendix B.

Therefore, for both the vanilla and actor-critic settings, our key challenge is establishing the conditions (2) (3) (4). By applying Theorem 3.6 and choosing μ to be small enough with respect to ϵ, we can arrive at the following theorem establishing convergence of vanilla policy gradient to ϵ-second order stationary points, specifying the required horizon of each sampled trajectory.

Theorem 3.7. *Suppose Assumptions 3.1-3.5 hold and let $\epsilon > 0$. For $\mu < \frac{\epsilon^2\delta}{L\sigma^2+D^2}$ where $D = \frac{G\mathcal{R}_{max}}{1-\gamma}$ and $\sigma = \frac{G\mathcal{R}_{max}}{(1-\gamma)^2}$, we have with probability $1-\delta$ that Algorithm 1 with vanilla policy gradient estimator computed via Algorithm 2 and $H = O(\log(\epsilon^{-2}))$ reaches an ϵ-second order stationary point in $\tilde{O}(\epsilon^{-6.5})$ iterations, where $\tilde{O}(\cdot)$ hides logarithmic dependencies.*

In Theorem 3.7, $O(\cdot)$ hides dependency on γ, and $\tilde{O}(\cdot)$ hides dependencies on L, G, $\mathcal{R}_{max}$, γ, M, σ_l, χ, δ.

The detailed proof of Theorem 3.7 is provided in Appendix C. The proof sketch is as follows. In order to apply Theorem 3.6, we first show that the gradient estimator is uniformly bounded based on our assumptions (Lemma C.1). This also implies that the second and fourth moment of the noise are also bounded. We then establish that the gradient bias due to truncation is deterministically bounded and decays as the trajectory horizon H increases (Lemma C.3). For large enough H, the bias error is proportional to μ, allowing us to bound away its effects.

4. Second-Order Convergence of Actor-Critic Policy Gradient

Now we consider the second-order convergence of actor-critic policy gradient algorithms. Actor-critic methods can reduce the variance of policy gradient methods by separately learning to approximate the state-action value function Q^π. With some algebraic manipulation, the policy gradient can be expressed as follows

$$\nabla J(\theta) = \mathbb{E}_{\pi_\theta,\rho_0}[\sum_{k=0}^{\infty}\gamma^k\nabla\log\pi_\theta(a_k|s_k)Q^{\pi_\theta}(s_k,a_k)].$$

This motivates the construction of the following biased gradient estimator from a trajectory τ of length H

$$\hat{G}^{AC}(\theta;\tau) = \sum_{k=0}^{H}\gamma^k\nabla\log\pi_\theta(a_k|s_k)Q_w(s_k,a_k),$$

where $Q_w(s,a)$ is a function with parameter w that approximates $Q^{\pi_\theta}(s,a)$. Note that $\hat{G}^{AC}(\theta;\tau)$ depends on w, although we omit this dependency in notation. In contrast to vanilla policy gradient, the actor-critic gradient estimator has two sources of bias: bias from the horizon truncation that depends on H, and bias from the critic approximation Q_w. Our following analysis addresses both. Let

$$\begin{aligned}G(\theta) = G_H(\theta) &= \mathbb{E}_\tau[\hat{G}^{AC}(\theta;\tau)]\\ &= \mathbb{E}_\tau[\sum_{k=0}^{H}\gamma^k\nabla\log\pi_\theta(a_k|s_k)Q_w(s_k,a_k)]\end{aligned}$$

represent the expectation of the gradient estimator with respect to the sampled trajectory τ of length H. Let

$$G_\infty(\theta) = \mathbb{E}_\tau[\sum_{k=0}^{\infty}\gamma^k\nabla\log\pi_\theta(a_k|s_k)Q_w(s_k,a_k)]$$

represent the expectation of an infinite-horizon gradient estimator with respect to a sampled trajectory of infinite horizon. As before, we have the following noise-bias decomposition of the policy updates

$$\theta_{t+1} = \theta_t + \mu\nabla J(\theta_t) + \mu\xi_{t+1} + \mu d_{t+1},$$

where $\xi_{t+1} = \hat{G}^{AC}(\theta_t;\tau_t) - G_H(\theta_t)$ once again represents a zero-mean noise term. Then we can further decompose the bias term d_{t+1} as follows

$$\begin{aligned}d_{t+1} &= G_H(\theta_t) - \nabla J(\theta_t)\\ &= G_H(\theta_t) - G_\infty(\theta_t) + G_\infty(\theta_t) - \nabla J(\theta_t)\\ &= p_{t+1} + q_{t+1}\end{aligned}$$

where $p_{t+1} = G_H(\theta_t) - G_\infty(\theta_t)$ represents the bias component due to truncation of the infinite horizon and $q_{t+1} = G_\infty(\theta_t) - \nabla J(\theta_t)$ represents the bias component induced by the critic approximation. In order to apply Theorem 3.6, we need to bound both p_{t+1} and q_{t+1}; although the former can be bounded using the approach of the previous section, the latter requires novel techniques.

The structure of our algorithm is as follows. We consider a double-loop actor-critic algorithm with linear function approximation of Q^{π_θ} and an arbitrary policy parametrization. In the inner loop, the critic parameter w is updated via TD(0) and projected onto a convex set Θ, and in the outer loop, the policy parameter θ is updated via gradient updates. The gradient estimator algorithm is outlined in Algorithm 3.

Algorithm 3 Actor-Critic Gradient Estimator

Input: initial critic parameters w_0, policy parameter θ_t, trajectory $\tau_t = \{s_0, a_0, s_1, a_1, ...s_{H-1}, a_{H-1}\}$
Sample initial state-action pair $s'_0 \sim \rho_0$ and $a'_0 \sim \pi_{\theta_t}(\cdot|s'_0)$
for $k = 0, 1, 2, ...K-1$ **do**
Sample $s'_{k+1} \sim \mathcal{P}(\cdot|s'_k, a'_k)$ and $a'_{k+1} \sim \pi_{\theta_t}(\cdot|s'_{k+1})$
Compute the TD(0) semi-gradient $g_k(w_k)$

$$g_k(w_k) = (\mathcal{R}(s'_k, a'_k) + \gamma Q_{w_k}(s'_{k+1}, a'_{k+1}) - Q_{w_k}(s'_k, a'_k))\nabla Q_{w_k}(s'_k, a'_k)$$

$w_{k+1} = Proj_\Theta[w_k + \alpha_k g_k(w_k)]$
end for
Calculate the averaged parameter $\bar{w}_{K,t} = \frac{1}{K}\sum_{k=0}^{K-1} w_k$
Compute gradient estimator from τ_t:

$$\hat{G}^{AC}(\theta_t; \tau_t) = \sum_{h=0}^{H-1} \gamma^h Q_{\bar{w}_{K,t}}(s_h, a_h)\nabla \log \pi_{\theta_t}(a_h|s_h)$$

return $\hat{G}^{AC}(\theta_t; \tau_t)$

To control the actor-critic bias q_{t+1} and ensure second-order convergence by Theorem 3.6, we require that for each policy parameter iterate, the inner loop converges to the global optima $w^*(\theta_t)$ and that $Q_{w^*}(\theta_t)$ precisely approximates the true value function $Q^{\pi_{\theta_t}}$. The first requirement is satisfied by linear TD, as shown in Section 4.1, and the second is formalized later in Assumption 4.9.

4.1. Convergence of TD(0) on Nonstationary Markov Chains

The inner-loop structure of Algorithm 3 suggests that we should apply existing results regarding the finite-time convergence of TD(0) with Markovian sampling (Bhandari et al., 2018; Liu & Olshevsky, 2020). However, these analyses rely on an additional assumption that the Markov chain begins in the stationary distribution, which is unrealistic for the actor-critic setting since there is no guarantee that after each policy update we can begin at the new stationary distribution with respect to the updated policy. As argued in (Bhandari et al., 2018; Liu & Olshevsky, 2020; Dalal et al., 2018), an exponentially mixing Markov chain approximately arrives at its stationary distribution after a logarithmic number of time steps. However, although this explanation justifies the assumption in a practical sense, it is not conducive for finite-time convergence analysis, since we can never reach the exact stationary distribution after a finite number of time steps.

As it turns out, the general convergence of TD(0) does hold without this initial state distribution assumption after some proof adjustments. In this section we reestablish the core results of (Bhandari et al., 2018) for a nonstationary Markov chain (i.e. a Markov chain that has not reached steady-state). These results, which are also utilized in (Qiu et al., 2021; Liu & Olshevsky, 2020), may be of independent interest.

4.1.1. Setup

We consider TD(0) with linear function approximation and a projection step. We define a set of N feature functions $\phi_n : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, $0 < n \leq N$. For each state-action pair (s, a), we define the vector $\phi(s, a) = (\phi_1(s, a), \phi_2(s, a), ..., \phi_N(s, a))^\intercal$ as the vector representing the features of (s, a). Then we denote as $\Phi \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times N}$ the matrix $\Phi = [\phi_1, ...\phi_N]$. Finally, we denote our linear parametrization of Q^π by the parameter $w \in \mathbb{R}^N$ as

$$Q_w(s, a) = w^\intercal \phi(s, a).$$

Let g_k represent the stochastic semi-gradient at time step k such that

$$g_k(w) = \mathcal{R}(s_k, a_k)\phi(s_k, a_k) + (\gamma\phi(s_{k+1}, a_{k+1})^\intercal w - \phi(s_k, a_k)^\intercal w)\phi(s_k, a_k).$$

As is common in the TD convergence literature, we consider TD(0) projected onto a convex set Θ that contains the limit point w^* of the algorithm (Bhandari et al., 2018). Therefore, the TD update can be written as

$$w_{k+1} = Proj_\Theta[w_k + \alpha g_k(w_k)]. \tag{5}$$

4.1.2. Assumptions

The following assumptions and Assumption 3.1 are necessary to show the convergence of TD(0).

Assumption 4.1. For all θ and $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have $\pi_\theta(s, a) > 0$.

Assumption 4.1 is satisfied by popular policy parametrizations such as the softmax policy parametrization.

Assumption 4.2. For any $\pi > 0$, The Markov chain defined by $P(s, a, s', a') = \mathcal{P}(s'|s, a)\pi(a'|s')$ is ergodic.

Combined with Assumption 4.1, Assumption 4.2 is satisfied if there is a positive probability of transitioning between any two state-action pairs in a finite number of steps. Assumption 4.2 is a common assumption in the TD and actor-critic literature (Bhandari et al., 2018; Qiu et al., 2021; Liu & Olshevsky, 2020; Wu et al., 2020; Xu et al., 2020a;b) that has a number of implications; the Markov chain is irreducible and aperiodic, it has a unique stationary distribution $\eta_\pi(s, a)$, and $\eta_\pi(s, a) \neq 0$ for all (s, a). In addition, the Markov chain mixes at a uniform geometric rate, i.e., there exists

$m > 0, r \in (0,1)$, such that for $t \in \mathbb{N}$ we have

$$\sup_{\substack{s\in\mathcal{S}\\a\in\mathcal{A}}} d_{TV}(\mathbb{P}(s_t = \cdot, a_t = \cdot | s_0 = s, a_0 = a), \eta_\pi) \leq m r^t \tag{6}$$

and $\tau^{\text{mix}}(\epsilon) = \min\{t \in \mathbb{N} \,|\, mr^t \leq \epsilon\}$ is the mixing time.

Assumption 4.3. The number of states and actions is finite.

As discussed in (Bhandari et al., 2018), Assumption 4.3 can be relaxed to consider countably infinite state-action pairs.

Assumption 4.4. The feature matrix Φ has full column rank, i.e., the feature vectors $\{\phi_1, ...\phi_N\}$ are linearly independent. In addition, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, $\|\phi(s,a)\|^2 \leq 1$.

Assumption 4.5. There exists $R > 0$ such that $diam(\Theta) \leq R$, where $diam$ is the diameter.

Assumptions 4.4 and 4.5 are also common in the literature (Bhandari et al., 2018; Liu & Olshevsky, 2020; Tsitsiklis & Van Roy, 1997; Qiu et al., 2021). See Section 8.2 of (Bhandari et al., 2018) and Proposition 3 of (Qiu et al., 2021) for additional details on defining R.

Finally, let A_θ denote the positive definite matrix $\mathbb{E}_{(s,a)\sim\eta_\theta,(s',a')\sim P(s,a,\cdot)}[\phi(s,a)(\phi(s,a) - \gamma\phi(s',a'))^\intercal]$. Based on Assumptions 4.2 and 4.4, we can conclude that A_θ is positive definite (Tsitsiklis & Van Roy, 1997). We further assume the smallest eigenvalue of A_θ is uniformly bounded away from zero in the following assumption, which is also utilized in (Qiu et al., 2021).

Assumption 4.6. There exists a lower bound $\varsigma > 0$, such that for all $\theta \in \mathbb{R}^M$ we have $\lambda_{min}(A_\theta + A_\theta^\intercal) \geq \varsigma$.

4.1.3. Finite-Time Convergence of TD(0)

We first require the following theorem from (Tsitsiklis & Van Roy, 1997), which establishes the existence and uniqueness of the solution to the projected Bellman equation as the limit point of the TD(0) algorithm.

Theorem 4.7. *(Tsitsiklis & Van Roy, 1997) Denote by $\mathbb{T}^\pi$ the Bellman operator under policy π such that for a value function $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ we have*

$$(\mathbb{T}^\pi Q)(s,a) = \mathcal{R}(s,a) + \gamma \sum_{\substack{s'\in\mathcal{S}\\a'\in\mathcal{A}}} \mathcal{P}(s'|s,a)\pi(a'|s')Q(s',a').$$

Then given Assumptions 3.1, 4.1-4.5, the limit point w^* of the TD(0) algorithm with linear function approximation exists, and it is the unique solution to the projected Bellman equation

$$\Phi w = Proj_\Phi(\mathbb{T}^\pi \Phi w)$$

where $Proj_\Phi$ is the projection operator onto the subspace $\{\Phi x | x \in \mathbb{R}^N\}$ spanned by the feature vectors ϕ_n.

Now we share our main result regarding the convergence of TD(0). Theorem 4.8 establishes the convergence rate of the Projected TD(0) algorithm after T constant time steps on a nonstationary Markov chain.

Theorem 4.8. *Suppose Assumptions 3.1, 4.1-4.5 hold and $\bar{w}_K = \frac{1}{K}\sum_{k=0}^{K-1} w_k$ is generated by K steps of the Projected TD(0) algorithm with $w^* \in \Theta$ and $\alpha = \frac{1}{\sqrt{K}}$. Then*

$$\mathbb{E}[\|Q_{w^*} - Q_{\bar{w}_K}\|_{\eta_\pi}^2] \leq \frac{\|w^* - w_0\|^2 + F^2(17 + 12\tau^{mix}(\frac{1}{\sqrt{K}}))}{2(1-\gamma)\sqrt{K}} + \frac{10F^2 m}{(1-r)(1-\gamma)K},$$

where $F = \mathcal{R}_{max} + 2R$ and

$$\|Q_{w^*} - Q_{\bar{w}_K}\|_{\eta_\pi}^2 = \sum_{\substack{s\in\mathcal{S}\\a\in\mathcal{A}}} \eta_\pi(s,a)(Q_{w^*}(s,a) - Q_{\bar{w}_K}(s,a))^2.$$

Proof. See Appendix D.

We can compare this result with Theorem 3 from (Bhandari et al., 2018) which shows

$$\mathbb{E}[\|Q_{w^*} - Q_{\bar{w}_K}\|_{\eta_\pi}^2] \leq \frac{\|w^* - w_0\|^2 + F^2(9 + 12\tau^{\text{mix}}(\frac{1}{\sqrt{K}}))}{2(1-\gamma)\sqrt{K}} \tag{7}$$

for TD(0) on a stationary Markov chain. We note that the nonstationary result involves slightly different constants and an additional term that decays at the rate of $O(1/K)$.

4.2. Main Result

Returning to the actor-critic setting, the following assumption combined with the existence and uniqueness theorem (Theorem 4.7) implies for the limit point of the TD(0) algorithm w^*, the resulting function $Q_{w^*} = {w^*}^T\phi(s,a)$ approximates the true state-action value function Q^{π_θ} with arbitrarily close precision.

Assumption 4.9. The value function lies in the linear function class such that

$$\inf_{w\in\Theta} \mathbb{E}_{(s,a)\sim\eta_{\pi_\theta}}[(\mathbb{T}^{\pi_\theta}(w^\intercal\phi(s,a)) - w^\intercal\phi(s,a))^2] = 0.$$

Assumption 4.9 is a linear realizability assumption that states that the value function can be sufficiently represented by a linear model of the feature vectors. This can be satisfied via an appropriate choice of feature vectors, such as radial basis functions, Fourier basis functions, or neural networks (Ji et al., 2019). Other actor-critic works (Xu et al., 2020a; Kumar et al., 2019; Fu et al., 2020) also require similar assumptions.

By characterizing the convergence of TD(0), we show in the following lemma that the norm of the bias term q_{t+1}

decays with respect to the number of inner-loop iterations. Lemma 4.10 features diminishing step sizes because it requires a stronger fourth moment bound showing the direct convergence of $\bar{w}_K$ to w^*, as opposed to the weaker result obtained in Theorem 4.8. The details of the proof are in Appendix E.1.

Lemma 4.10. *For $K = O(\frac{\log^2(\mu^{-4})}{\mu^4})$ as $\mu \to 0$, for which $O(\cdot)$ does not hide dependencies on other constants, and*

$$D_q = G\left(\frac{192F^2R^2}{\varsigma^2\log^2(r^{-1})} + O(\frac{1}{\log\mu^{-4}})\right)^{-1/4}$$

for which $O(\cdot)$ hides dependencies on F, r, ς, m, the expected function approximation bias is bounded such that

$$\mathbb{E}[\|q_{t+1}\|^4|\mathcal{F}_t] \leq G^4\mathbb{E}[\|\bar{w}_{K,t} - w^*\|^4|\mathcal{F}_t] \leq D_q^4\mu^4.$$

Finally, by combining Theorem 3.6 with Lemma 4.10 and bounds on the truncation bias p_{t+1} (Lemma E.2) and noise ξ_{t+1} (Lemma E.1), we arrive at Theorem 4.11.

Theorem 4.11. *Suppose Assumptions 3.1-3.5, 4.1-4.6, 4.9 hold and let $\epsilon > 0$. For $\mu < \frac{\epsilon^2\delta}{L\sigma^2+D^2}$ where $\sigma = \frac{GR}{1-\gamma}$, $D = 2(D_p^4 + D_q^4)^{1/4}$ and $D_p = \frac{GR}{1-\gamma}$, we have with probability $1-\delta$ that Algorithm 1 with actor-critic policy gradient estimator computed via Algorithm 3 with $H = O(\log(\epsilon^{-2}))$ and $K = \tilde{O}(\epsilon^{-8})$ reaches an ϵ-second order stationary point in $\tilde{O}(\epsilon^{-6.5})$ iterations.*

Proof. See Appendix E.

In Theorem 4.11, $O(\cdot)$ hides dependency on γ and $\tilde{O}(\cdot)$ hides dependencies on L, G, R, γ, M, σ_l, χ, δ, F, r, m, ς.

5. Conclusion

In this work, we provide a novel analysis on the convergence of biased policy gradient methods to second-order stationary points. Our work applies to general policy parametrization and Markovian sampling. We also show the convergence of TD(0) on nonstationary Markov chains, which pertains to realistic actor-critic implementations.

Future directions may involve extending this work to two-timescale or single-timescale actor-critic algorithms, which may provide some performance improvement. In addition, instead of assuming Assumption 4.9, we may want to show second-order convergence of actor-critic algorithms with respect to some irremoveable approximation error ϵ_{app} representing the imperfect critic approximation, similar to several first-order analyses (Wu et al., 2020; Qiu et al., 2021).

Impact Statement

This paper advances the field of reinforcement learning by characterizing the solutions achieved by policy gradient algorithms. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. Optimality and approximation with policy gradient methods in markov decision processes. In Abernethy, J. and Agarwal, S. (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 64–66. PMLR, 09–12 Jul 2020.

Alfano, C., Yuan, R., and Rebeschini, P. A novel framework for policy mirror descent with general parameterization and linear convergence. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Baxter, J. and Bartlett, P. L. Infinite-horizon policy-gradient estimation. *J. Artif. Int. Res.*, 15(1):319–350, Nov 2001. ISSN 1076-9757.

Bhandari, J. and Russo, D. Global optimality guarantees for policy gradient methods. *CoRR*, abs/1906.01786, 2019.

Bhandari, J. and Russo, D. On the linear convergence of policy gradient methods for finite mdps. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2386–2394. PMLR, 13–15 Apr 2021.

Bhandari, J., Russo, D., and Singal, R. A finite time analysis of temporal difference learning with linear function approximation. In Bubeck, S., Perchet, V., and Rigollet, P. (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1691–1692. PMLR, 06–09 Jul 2018.

Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M., and Lee, M. Natural actor–critic algorithms. *Automatica*, 45(11): 2471–2482, 2009. ISSN 0005-1098. doi: https://doi.org/10.1016/j.automatica.2009.07.008.

Cayci, S., He, N., and Srikant, R. Finite-time analysis of entropy-regularized neural natural actor-critic algorithm, 2022.

Dalal, G., Szorenyi, B., Thoppe, G., and Mannor, S. Finite sample analyses for td(0) with function approximation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Daneshmand, H., Kohler, J., Lucchi, A., and Hofmann, T. Escaping saddles with stochastic gradients. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1155–1164. PMLR, 10–15 Jul 2018.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Fu, Z., Yang, Z., and Wang, Z. Single-timescale actor-critic provably finds globally optimal policy. *CoRR*, abs/2008.00483, 2020.

Ge, R., Huang, F., Jin, C., and Yuan, Y. Escaping from saddle points - online stochastic gradient for tensor decomposition. *CoRR*, abs/1503.02101, 2015.

Han, Y., Razaviyayn, M., and Xu, R. Policy gradient finds global optimum of nearly linear-quadratic control systems. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022.

Ji, Z., Telgarsky, M., and Xian, R. Neural tangent kernels, transportation mappings, and universal approximation. *CoRR*, abs/1910.06956, 2019.

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. How to escape saddle points efficiently. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1724–1732. PMLR, 06–11 Aug 2017.

Jin, C., Netrapalli, P., Ge, R., Kakade, S. M., and Jordan, M. I. Stochastic gradient descent escapes saddle points efficiently. *CoRR*, abs/1902.04811, 2019.

Khorasani, S., Salehkaleybar, S., Kiyavash, N., He, N., and Grossglauser, M. Efficiently escaping saddle points for non-convex policy optimization, 2023.

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.

Kumar, H., Koppel, A., and Ribeiro, A. On the sample complexity of actor-critic method for reinforcement learning with function approximation. *CoRR*, abs/1910.08412, 2019.

Liu, R. and Olshevsky, A. Temporal difference learning as gradient splitting. *CoRR*, abs/2010.14657, 2020.

Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. On the global convergence rates of softmax policy gradient methods. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6820–6829. PMLR, 13–18 Jul 2020.

Nesterov, Y. and Polyak, B. Cubic regularization of newton method and its global performance. *Math. Program.*, 108: 177–205, 08 2006. doi: 10.1007/s10107-006-0706-8.

Olshevsky, A. and Gharesifard, B. A small gain analysis of single timescale actor critic. *SIAM Journal on Control and Optimization*, 61(2):980–1007, 2023. doi: 10.1137/22M1483335.

Papini, M., Binaghi, D., Canonaco, G., Pirotta, M., and Restelli, M. Stochastic variance-reduced policy gradient. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4026–4035. PMLR, 10–15 Jul 2018.

Peters, J. and Schaal, S. Reinforcement learning of motor skills with policy gradients. *Neural Networks*, 21(4):682–697, 2008. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2008.02.003. Robotics and Neuroscience.

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, 2021. doi: 10.1109/JSAIT.2021.3078754.

Shen, Z., Ribeiro, A., Hassani, H., Qian, H., and Mi, C. Hessian aided policy gradient. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5729–5738. PMLR, 09–15 Jun 2019.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In Solla, S., Leen, T., and Müller, K. (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

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. doi: 10.1109/9.580874.

Vlaski, S. and Sayed, A. H. Second-order guarantees of stochastic gradient descent in nonconvex optimization. *IEEE Transactions on Automatic Control*, 67(12):6489–6504, 2022. doi: 10.1109/TAC.2021.3131963.

Vlaski, S., Rizk, E., and Sayed, A. H. Second-order guarantees in federated learning. In *2020 54th Asilomar Conference on Signals, Systems, and Computers*, pp. 915–922, 2020. doi: 10.1109/IEEECONF51394.2020.9443421.

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*, 2020. URL https://openreview.net/forum?id=BJgQfkSYDS.

Wang, P., Wang, H., and Zheng, N. Stochastic cubic-regularized policy gradient method. *Knowledge-Based Systems*, 255:109687, 2022. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2022.109687.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 2004.

Wu, S., Shi, L., Wang, J., and Tian, G. Understanding policy gradient algorithms: A sensitivity-based approach. 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. 24131–24149. PMLR, 17–23 Jul 2022.

Wu, Y. F., ZHANG, W., Xu, P., and Gu, Q. A finite-time analysis of two time-scale actor-critic methods. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17617–17628. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/cc9b3c69b56df284846bf2432f1cba90-Paper.pdf.

Xu, T., Wang, Z., and Liang, Y. Improving sample complexity bounds for (natural) actor-critic algorithms. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 4358–4369. Curran Associates, Inc., 2020a.

Xu, T., Wang, Z., and Liang, Y. Improving sample complexity bounds for actor-critic algorithms. *CoRR*, abs/2004.12956, 2020b.

Yang, L., Zheng, Q., and Pan, G. Sample complexity of policy gradient finding second-order stationary points. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):10630–10638, May 2021.

Yang, Z., Zhang, K., Hong, M., and Başar, T. A finite sample analysis of the actor-critic algorithm. In *2018 IEEE Conference on Decision and Control (CDC)*, pp. 2759–2764, 2018. doi: 10.1109/CDC.2018.8619440.

Yang, Z., Chen, Y., Hong, M., and Wang, Z. Provably global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Yuan, R., Gower, R. M., and Lazaric, A. A general sample complexity analysis of vanilla policy gradient. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 3332–3380. PMLR, 28–30 Mar 2022.

Yuan, R., Du, S. S., Gower, R. M., Lazaric, A., and Xiao, L. Linear convergence of natural policy gradient methods with log-linear policies. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=-z9hdsyUwVQ.

Zhang, K., Koppel, A., Zhu, H., and Başar, T. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6):3586–3612, 2020. doi: 10.1137/19M1288012.

Zhu, Y. and Gong, X. Distributed policy gradient with heterogeneous computations for federated reinforcement learning. In *2023 57th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6, 2023. doi: 10.1109/CISS56502.2023.10089771.

A. Additional Discussion of Related Work

A.1. Global Convergence

We first present some of the many global convergence results available in terms of ϵ-optimality. In reinforcement learning, global convergence guarantees for policy gradient algorithms tend to arise from some underlying structure of the optimization problem, due to specific policy parametrization or algorithm.

For first-order algorithms, the following papers achieve global optimality.

- In (Agarwal et al., 2020), tabular parametrization with exact gradients, $O(\epsilon^{-2})$ iterations
- In (Bhandari & Russo, 2019), objective functions that satisfy a gradient dominance property with exact gradients, $O(\epsilon^{-2})$ iterations
- In (Mei et al., 2020), tabular softmax parametrization with exact gradients, $O(\epsilon^{-1})$ iterations
- In (Wang et al., 2020), neural policy gradient with extremely wide shallow neural networks, $O(\epsilon^{-2})$ iterations

More powerful global convergence results are achieved for quasi-second-order methods like natural policy gradient and mirror descent policy gradient. These results can go beyond the tabular setting.

- In (Agarwal et al., 2020), natural policy gradient with tabular softmax parametrization with exact gradients, $O(\epsilon^{-1})$ iterations
- In (Xu et al., 2020b), natural actor-critic with function approximation where the feature vectors vary in each iteration, $O(\epsilon^{-4})$ outer-loop iterations
- In (Cayci et al., 2022), natural actor-critic policy gradient where the actor and critic are parametrized with extremely wide neural networks, $O(\epsilon^{-2})$ iterations
- In (Yuan et al., 2023), natural policy gradient for log-linear policies with compatible function approximation, linear convergence in terms of outer loop iterations
- In (Alfano et al., 2023), policy mirror descent with general parametrization, linear convergence in terms of outer loop iterations

In comparison, we achieve an ϵ-second-order stationary point in $\tilde{O}(\epsilon^{-6.5})$ iterations. Our work focuses on first-order algorithms and general policy parametrization which are simpler to implement and more widely used in practice. We also do not use oracles or exact gradients in our analysis. Finally, our work also has no dependence on the distribution mismatch coefficient, which is widely used in global convergence results and roughly quantifies how well the initial state distribution matches the optimal state distribution.

A.2. Second-Order Convergence

Our work inherits the sample complexity of $\tilde{O}(\epsilon^{-6.5})$ from (Vlaski & Sayed, 2022), which is weaker than the best sample complexity of $\tilde{O}(\epsilon^{-4})$ obtained by (Jin et al., 2019), both of which analyze the second-order convergence of vanilla stochastic gradient descent. Faster second-order convergence is available for algorithms with exact gradients, such as perturbed gradient descent which converges in $\tilde{O}(\epsilon^{-2})$ iterations (Jin et al., 2017). However, exact gradient computations are intractable for realistic policy gradient implementations. Second-order or quasi-second-order algorithms that utilize Hessian information can also converge faster. SPIDER-SFO from (Fang et al., 2018) obtains a sample complexity of $\tilde{O}(\epsilon^{-3})$ via a negative curvature search method. Similar techniques are extended to the policy gradient setting in (Khorasani et al., 2023) with complexity $\tilde{O}(\epsilon^{-3})$ and (Wang et al., 2022) with complexity $\tilde{O}(\epsilon^{-3.5})$. All of these aforementioned works only deal with unbiased gradient estimators. To the best of our knowledge only (Vlaski et al., 2020) show second-order convergence with biased gradient estimators in the form of federated learning, requiring $\tilde{O}(\epsilon^{-6.5})$ global iterations. This matches our sample complexity, as expected.

B. Proof of Theorem 3.6

B.1. Key Lemmas and Proof Sketch

Our approach for proving that Algorithm 1 arrives at an ϵ-second order stationary point relies on bounding the bias and noise of the gradient estimator and applying the techniques developed in (Vlaski & Sayed, 2022). Broadly speaking, (Vlaski & Sayed, 2022) show second-order convergence for unbiased stochastic gradient descent by first showing the iterates on the second-order Taylor expansion of the objective function escape saddle points, and then showing that the iterates and those on its Taylor approximation are sufficiently close. Therefore, if we also show that the policy gradient iterates are close to the iterates on the Taylor expansion, we can conveniently apply the convergence results of (Vlaski & Sayed, 2022) to our setting.

We begin with establishing fourth moment bounds on the noise term ξ_{t+1} by the following lemma.

Lemma B.1. *Suppose for some random variable X we have $\mathbb{E}[X] = \mu$ and $\|X\| \leq \sigma$. Then*

$$\mathbb{E}[\|X - \mu\|^2] \leq \sigma^2,$$

$$\mathbb{E}[\|X - \mu\|^4] \leq 4\sigma^2.$$

Proof.

$$\mathbb{E}[\|X - \mu\|^2] = \mathbb{E}[\|X\|^2] - \mu^2 \leq \mathbb{E}[\|X\|^2] \leq \sigma^2$$

$$\|X - \mu\|^4 \leq \|X - \mu\|^2 \cdot 4\sigma^2$$

$$\mathbb{E}[\|X - \mu\|^4] \leq 4\sigma^4$$

□

So by Lemma B.1, we have

$$\mathbb{E}[\|\xi_{t+1}\|^2 | \mathcal{F}_t] \leq \sigma^2, \tag{8}$$

$$\mathbb{E}[\|\xi_{t+1}\|^4 | \mathcal{F}_t] \leq 4\sigma^4. \tag{9}$$

In addition, by Jensen's inequality, the fourth-moment bound

$$\mathbb{E}[\|d_{t+1}\|^4 | \mathcal{F}_t] \leq D^4\mu^4$$

also implies the following second-moment bound

$$\mathbb{E}[\|d_{t+1}\|^2 | \mathcal{F}_t] \leq D^2\mu^2. \tag{10}$$

To proceed with the proof, we first show that in the large gradient regime, we observe a large ascent in function value, whereas around local maxima, the possible descent is bounded. Then we construct a pair of coupled sequences $\{\theta_{i+j}\}$ and $\{\theta'_{i+j}\}$, where $\{\theta_{i+j}\}$ represents the gradient iterates on the original objective function and $\{\theta'_{i+j}\}$ represents gradient ascent iterates on the second-order Taylor approximation of the function centered at θ_i with the same noise term. Through moment bounds, we show that the difference between the coupled sequences is sufficiently small. These results allow us to leverage Theorems 2 and 3 in (Vlaski & Sayed, 2022).

The following lemma establishes that for small enough step sizes, we have sufficient ascent starting in the large gradient regime $\theta_i \in \mathcal{G}$, and descent is bounded starting near a local maxima $\theta_i \in \mathcal{M}$.

Lemma B.2. *For $\mu < \frac{1}{L}$, we have after one iteration of Algorithm 1,*

$$\mathbb{E}[J(\theta_{i+1}) | \theta_i \in \mathcal{G}] \geq \mathbb{E}[J(\theta_i) | \theta_i \in \mathcal{G}] + \frac{\mu^2(L\sigma^2 + D^2\mu)}{2\delta}$$

$$\mathbb{E}[J(\theta_{i+1}) | \theta_i \in \mathcal{M}] \geq \mathbb{E}[J(\theta_i) | \theta_i \in \mathcal{M}] - \frac{\mu^2(L\sigma^2 + D^2\mu)}{2}.$$

Proof of Lemma B.2: see Appendix B.3.

Beginning from $\theta_i \in \mathcal{H}$, we define $\{\theta'_{i+j}\}$ as the gradient ascent iterates on the second-order Taylor approximation of $J(\theta)$ plus the noise term ξ_{i+j+1} from the original sequence. Denote the Taylor expansion around θ_i as $\hat{J}$ as follows

$$\hat{J}(\theta) = J(\theta_i) + \nabla J(\theta_i)^T(\theta - \theta_i) + \frac{1}{2}(\theta - \theta_i)^\intercal \nabla^2 J(\theta_i)(\theta - \theta_i)$$

$$\nabla \hat{J}(\theta) = \nabla J(\theta_i) + \nabla^2 J(\theta_i)(\theta - \theta_i).$$

So we have $\{\theta'_{i+j}\}$ and our original sequence iterates $\{\theta_i\}$ defined as follows

$$\theta'_{i+j+1} = \theta'_{i+j} + \mu \nabla J(\theta_i) + \mu \nabla^2 J(\theta_i)(\theta'_{i+j} - \theta_i) + \mu \xi_{i+j+1}$$

$$\theta_{i+j+1} = \theta_{i+j} + \mu \nabla J(\theta_{i+j}) + \mu \xi_{i+j+1} + \mu d_{i+j+1}.$$

Then we can conclude in the following lemma that in the vicinity of a saddle point, the distance between θ_i and θ_{i+j+1} is bounded, and the distance between θ_i and θ'_i is bounded.

Lemma B.3. *For $\{\theta_i\}$ and $\{\theta'_i\}$ defined above, and $j \leq \frac{C}{\mu}$, where C is a constant independent of μ, we have*

$$\mathbb{E}[\|\theta_i - \theta_{i+j+1}\|^2 | \theta_i \in \mathcal{H}] \leq O(\mu) \tag{11}$$

$$\mathbb{E}[\|\theta_i - \theta_{i+j+1}\|^4 | \theta_i \in \mathcal{H}] \leq O(\mu^2) \tag{12}$$

$$\mathbb{E}[\|\theta'_{i+j+1} - \theta_{i+j+1}\|^2 | \theta_i \in \mathcal{H}] \leq O(\mu^2). \tag{13}$$

Corollary B.4. *From the results of Lemma B.3, we can conclude*

$$\mathbb{E}[\|\theta_i - \theta_{i+j+1}\|^3 | \theta_i \in \mathcal{H}] \leq O(\mu^{3/2}) \tag{14}$$

$$\mathbb{E}[\|\theta_i - \theta'_{i+j+1}\|^2 | \theta_i \in \mathcal{H}] \leq O(\mu) \tag{15}$$

$$\mathbb{E}[\|\theta_i - \theta'_{i+j+1}\|^3 | \theta_i \in \mathcal{H}] \leq O(\mu^{3/2}). \tag{16}$$

The first inequality follows from Jensen's inequality, and the second and third follow from the bounds on $\|\theta_i - \theta_{i+j+1}\|$ and $\|\theta'_{i+j+1} - \theta_{i+j+1}\|$.

Proof of Lemma B.3. See Appendix B.4.

B.2. Proof of Theorem 3.6

Proof. For the sequences $\{\theta_i\}$ and $\{\theta'_i\}$ defined above, suppose the moment bounds in Lemma B.3 hold. Then from Corollary 1 in (Vlaski & Sayed, 2022), beginning at $\theta_i \in \mathcal{H}$ for the finite horizon $j \leq \frac{C}{\mu}$ we have

$$\mathbb{E}[J(\theta_{i+j}) | \theta_i \in \mathcal{H}] \geq \mathbb{E}[J(\theta'_{i+j}) | \theta_i \in \mathcal{H}] - O(\mu^{3/2}),$$

which basically states that the function values on $\{\theta_i\}$ stay close to the function values on $\{\theta'_i\}$. This allows us to conclude that sufficient ascent occurs on the Taylor approximation as well as the original function by way of Theorem 2 from (Vlaski & Sayed, 2022). Beginning at a strict saddle point $\theta_i \in \mathcal{H}$, gradient ascent iterates on the short-term model for $\mathcal{T}$ iterations after i with

$$\mathcal{T} = \frac{\log(2M\frac{\sigma^2}{\sigma_l^2} + 1)}{\log(1 + 2\mu\omega)} \leq O(\frac{1}{\mu\omega})$$

guarantees

$$\mathbb{E}[J(\theta'_{i+\mathcal{T}}) | \theta_i \in \mathcal{H}] \geq \mathbb{E}[J(\theta_i) | \theta_i \in \mathcal{H}] + \frac{\mu}{2} M\sigma^2 - o(\mu).$$

Combined with the bounds on the iterates from Lemma B.3, this implies

$$\mathbb{E}[J(\theta_{i+\mathcal{T}}) | \theta_i \in \mathcal{H}] \geq \mathbb{E}[J(\theta_i) | \theta_i \in \mathcal{H}] + \frac{\mu}{2} M\sigma^2 - o(\mu).$$

This result, in combination with Lemma B.2, and the observation that $|J(\theta)| \leq \frac{\mathcal{R}_{\max}}{1-\gamma}$ for all θ allows us to apply Theorem 3 from (Vlaski & Sayed, 2022), yielding our final result. □

B.3. Proof of Lemma B.2

Proof. Our iterates are

$$\theta_{i+1} = \theta_i + \mu\nabla J(\theta_i) + \mu\xi_{i+1} + \mu d_{i+1}.$$

Because J is Lipschitz smooth by Lemma 3.2, we have

$$\begin{aligned}
J(\theta_{i+1}) \geq& J(\theta_i) + \nabla J(\theta_i)^T(\theta_{i+1} - \theta_i) - \frac{L}{2}\|\theta_{i+1} - \theta_i\|^2 \\
\geq& J(\theta_i) + \mu\nabla J(\theta_i)^T(\nabla J(\theta_i) + \xi_{i+1} + d_{i+1}) - \frac{L\mu^2}{2}\|\nabla J(\theta_i) + \xi_{i+1} + d_{i+1}\|^2 \\
\geq& J(\theta_i) + \mu\|\nabla J(\theta_i)\|^2 + \mu\nabla J(\theta_i)^T\xi_{i+1} + \mu\nabla J(\theta_i)^T d_{i+1} \\
&- \frac{L\mu^2}{2}(\|\nabla J(\theta_i) + d_{i+1}\|^2 + \|\xi_{i+1}\|^2 + 2(\nabla J(\theta_i) + d_{i+1})^T\xi_{i+1}).
\end{aligned}$$

We can take expectation with respect to the filtration $\mathcal{F}_i$ on either side to remove the cross terms with the noise term ξ_{i+1}, and then we have by (8)

$$\begin{aligned}
\mathbb{E}[J(\theta_{i+1})|\mathcal{F}_i] \geq& J(\theta_i) + \mu\|\nabla J(\theta_i)\|^2 + \mu\mathbb{E}[\nabla J(\theta_i)^T d_{i+1}|\mathcal{F}_i] - \frac{L\mu^2}{2}\mathbb{E}[\|\nabla J(\theta_i) + d_{i+1}\|^2|\mathcal{F}_i] - \frac{L\mu^2}{2}\mathbb{E}[\|\xi_{i+1}\|^2|\mathcal{F}_i] \\
\geq& J(\theta_i) + \mu\|\nabla J(\theta_i)\|^2 + \mu\mathbb{E}[\nabla J(\theta_i)^T d_{i+1}|\mathcal{F}_i] - \frac{L\mu^2}{2}\mathbb{E}[\|\nabla J(\theta_i) + d_{i+1}\|^2|\mathcal{F}_i] - \frac{L\mu^2\sigma^2}{2}.
\end{aligned}$$

We assume that $\mu < \frac{1}{L}$ to obtain

$$\mathbb{E}[J(\theta_{i+1})|\mathcal{F}_i] \geq J(\theta_i) + \mu\|\nabla J(\theta_i)\|^2 + \mu\mathbb{E}[\nabla J(\theta_i)^T d_{i+1}|\mathcal{F}_i] - \frac{\mu}{2}\mathbb{E}[\|\nabla J(\theta_i) + d_{i+1}\|^2|\mathcal{F}_i] - \frac{L\mu^2\sigma^2}{2}.$$

Then we use the fact that $\|a+b\|^2 = \|a\|^2 + 2a^Tb + \|b\|^2$ and (10) to obtain

$$\begin{aligned}
\mathbb{E}[J(\theta_{i+1})|\mathcal{F}_i] \geq& J(\theta_i) + \mu\|\nabla J(\theta_i)\|^2 + \mu\mathbb{E}[\nabla J(\theta_i)^T d_{i+1}|\mathcal{F}_i] - \frac{\mu}{2}\|\nabla J(\theta_i)\|^2 - \frac{\mu}{2}\mathbb{E}[\|d_{i+1}\|^2|\mathcal{F}_i] \\
&- \mu\mathbb{E}[\nabla J(\theta_i)^T d_{i+1}|\mathcal{F}_i] - \frac{L\mu^2\sigma^2}{2} \\
=& J(\theta_i) + \frac{\mu}{2}\|\nabla J(\theta_i)\|^2 - \frac{\mu}{2}\mathbb{E}[\|d_{i+1}\|^2|\mathcal{F}_i] - \frac{L\mu^2\sigma^2}{2} \\
\geq& J(\theta_i) + \frac{\mu}{2}\|\nabla J(\theta_i)\|^2 - \frac{D^2\mu^3}{2} - \frac{L\mu^2\sigma^2}{2}.
\end{aligned}$$

Now we want to apply the law of total expectation and condition on where w_i is located in the parameter space. We first condition on $\theta_i \in \mathcal{G}$, where we have $\|\nabla J(\theta_i)\|^2 > \mu(L\sigma^2 + D^2\mu)(1 + \frac{1}{\delta})$ to arrive at

$$\mathbb{E}[J(\theta_{i+1})|\theta_i \in \mathcal{G}] \geq \mathbb{E}[J(\theta_i)|\theta_i \in \mathcal{G}] + \frac{\mu}{2}\mu(L\sigma^2 + D^2\mu)(1 + \frac{1}{\delta}) - \frac{L\mu^2\sigma^2}{2} - \frac{D^2\mu^3}{2}$$

$$\mathbb{E}[J(\theta_{i+1})|\theta_i \in \mathcal{G}] \geq \mathbb{E}[J(\theta_i)|\theta_i \in \mathcal{G}] + \frac{\mu^2(L\sigma^2 + D^2\mu)}{2\delta}.$$

If we instead condition on $\theta_i \in \mathcal{M}$, we have that

$$\mathbb{E}[J(\theta_{i+1})|\theta_i \in \mathcal{M}] \geq \mathbb{E}[J(\theta_i)|\theta_i \in \mathcal{M}] - \frac{\mu^2(L\sigma^2 + D^2\mu)}{2}.$$

□

B.4. Proof of Lemma B.3

Before we proceed with the proof of Lemma B.3, we require a preliminary lemma from (Vlaski & Sayed, 2022) that will help us show that our product does not blow up for small μ.

Lemma B.5. *For $C, \mu, L > 0$ and $k \in \mathbb{Z}_+$ with $\mu < \frac{1}{L}$*

$$\lim_{\mu \to 0} \left(\frac{(1+\mu L)^k + O(\mu^2)}{(1-\mu L)^{k-1}} \right)^{C/\mu} = O(1).$$

Proof of Lemma B.5. See (Vlaski & Sayed, 2022).

Proof. First we want to show (11), restated below

$$\mathbb{E}[\|\theta_i - \theta_{i+j+1}\|^2 | \theta_i \in \mathcal{H}] \leq O(\mu).$$

We have by (8)

$$\begin{aligned}
\|\theta_i - \theta_{i+j+1}\|^2 &= \|\theta_i - \theta_{i+j} - \mu \nabla J(\theta_{i+j}) - \mu \xi_{i+j+1} - \mu d_{i+j+1}\|^2 \\
\mathbb{E}[\|\theta_i - \theta_{i+j+1}\|^2 | \mathcal{F}_{i+j}] &= \mathbb{E}[\|\theta_i - \theta_{i+j} - \mu \nabla J(\theta_{i+j}) - \mu d_{i+j+1}\|^2 | \mathcal{F}_{i+j}] + \mu^2 \mathbb{E}[\|\xi_{i+j+1}\|^2 | \mathcal{F}_{i+j}] \\
&\leq \mathbb{E}[\|\theta_i - \theta_{i+j} - \mu \nabla J(\theta_{i+j}) - \mu d_{i+j+1}\|^2 | \mathcal{F}_{i+j}] + \mu^2 \sigma^2 \\
&= \mathbb{E}[\|\theta_i - \theta_{i+j} - \mu \nabla J(\theta_{i+j}) + \mu \nabla J(\theta_i) - \mu \nabla J(\theta_i) - \mu d_{i+j+1}\|^2 | \mathcal{F}_{i+j}] + \mu^2 \sigma^2.
\end{aligned}$$

By Jensen's inequality, we have for $0 < \alpha < 1$,

$$\|a+b\|^2 \leq \frac{1}{\alpha}\|a\|^2 + \frac{1}{1-\alpha}\|b\|^2.$$

So we have

$$\mathbb{E}[\|\theta_i - \theta_{i+j+1}\|^2 | \mathcal{F}_{i+j}] \leq \frac{1}{1-\mu L}\|\theta_i - \theta_{i+j} - \mu \nabla J(\theta_{i+j}) + \mu \nabla J(\theta_i)\|^2 + \frac{\mu^2}{\mu L}\mathbb{E}[\|\nabla J(\theta_i) + d_{i+j+1}\|^2 | \mathcal{F}_{i+j}] + \mu^2 \sigma^2. \quad (17)$$

We consider the first term on the right hand side of (17) and expand it to obtain

$$\begin{aligned}
&\|\theta_i - \theta_{i+j} - \mu \nabla J(\theta_{i+j}) + \mu \nabla J(\theta_i)\|^2 \\
&\leq \|\theta_i - \theta_{i+j}\|^2 + 2\mu \|\theta_i - \theta_{i+j}\| \cdot \|\nabla J(\theta_{i+j}) - \nabla J(\theta_i)\| + \mu^2 \|\nabla J(\theta_{i+j}) - \nabla J(\theta_i)\|^2.
\end{aligned}$$

By Lipschitz smoothness, we have

$$\begin{aligned}
\|\theta_i - \theta_{i+j} - \mu \nabla J(\theta_{i+j}) + \mu \nabla J(\theta_i)\|^2 &\leq \|\theta_i - \theta_{i+j}\|^2 + 2\mu L \|\theta_i - \theta_{i+j}\| \cdot \|\theta_{i+j} - \theta_i\| + \mu^2 L^2 \|\theta_{i+j} - \theta_i\|^2 \\
&= (1 + 2\mu L + \mu^2 L^2)\|\theta_i - \theta_{i+j}\|^2 \\
&= (1+\mu L)^2 \|\theta_i - \theta_{i+j}\|^2.
\end{aligned}$$

We plug this into our original expression (17) and use (10) to obtain

$$\begin{aligned}
\mathbb{E}[\|\theta_i - \theta_{i+j+1}\|^2 | \mathcal{F}_{i+j}] &\leq \frac{(1+\mu L)^2}{1-\mu L}\|\theta_i - \theta_{i+j}\|^2 + \frac{\mu}{L}\mathbb{E}[\|\nabla J(\theta_i) + d_{i+j+1}\|^2 | \mathcal{F}_{i+j}] + \mu^2 \sigma^2 \\
&\leq \frac{(1+\mu L)^2}{1-\mu L}\|\theta_i - \theta_{i+j}\|^2 + \frac{2\mu}{L}\mathbb{E}[\|d_{i+j+1}\|^2 | \mathcal{F}_{i+j}] + \frac{2\mu}{L}\|\nabla J(\theta_i)\|^2 + \mu^2 \sigma^2 \\
&\leq \frac{(1+\mu L)^2}{1-\mu L}\|\theta_i - \theta_{i+j}\|^2 + \frac{2D^2\mu^3}{L} + \frac{2\mu}{L}\|\nabla J(\theta_i)\|^2 + \mu^2 \sigma^2.
\end{aligned}$$

Now we want to condition on $\theta_i \in \mathcal{H}$ to obtain

$$\begin{aligned}
\mathbb{E}[\|\theta_i - \theta_{i+j+1}\|^2 | \theta_i \in \mathcal{H}] &\leq \mathbb{E}[\frac{(1+\mu L)^2}{1-\mu L}\|\theta_i - \theta_{i+j}\|^2 | \theta_i \in \mathcal{H}] + \frac{2D^2\mu^3}{L} + \frac{2\mu^2}{L} \cdot \mu(L\sigma^2 + D^2\mu)(1 + \frac{1}{\delta}) + \mu^2 \sigma^2 \\
&\leq \frac{(1+\mu L)^2}{1-\mu L}\mathbb{E}[\|\theta_i - \theta_{i+j}\|^2 | \theta_i \in \mathcal{H}] + O(\mu^2).
\end{aligned}$$

Then we can evaluate this recursive formula starting at $j = 0$, since $\mathbb{E}[\|\theta_i - \theta_i\|^2] = 0$, to arrive at

$$\begin{aligned}
\mathbb{E}[\|\theta_i - \theta_{i+j+1}\|^2 | \theta_i \in \mathcal{H}] &\leq O(\mu^2) \sum_{n=0}^{j-1} \Big(\frac{(1+\mu L)^2}{1-\mu L} \Big)^n \\
&\leq O(\mu^2) \frac{1 - (\frac{(1+\mu L)^2}{1-\mu L})^j}{1 - \frac{(1+\mu L)^2}{1-\mu L}} \\
&= O(\mu^2) \frac{(1-\mu L)((\frac{(1+\mu L)^2}{1-\mu L})^j - 1)}{1 + 2\mu L + \mu^2 L^2 - 1 + \mu L} \\
&= O(\mu) \frac{(1-\mu L)((\frac{1+2\mu L+\mu^2 L^2}{1-\mu L})^j - 1)}{3L + \mu L^2} \\
&\leq O(\mu) \frac{(\frac{(1+\mu L)^2}{1-\mu L})^j}{3L} \leq O(\mu) \frac{(\frac{(1+\mu L)^2}{1-\mu L})^{\frac{C}{\mu}}}{3L}.
\end{aligned}$$

By Lemma B.5, this gives us

$$\mathbb{E}[\|\theta_i - \theta_{i+j+1}\|^2 | \theta_i \in \mathcal{H}] \leq O(\mu).$$

Now we want to show the fourth moment bound (12), restated below

$$\mathbb{E}[\|\theta_i - \theta_{i+j+1}\|^4 | \theta_i \in \mathcal{H}] \leq O(\mu^2).$$

We use the inequality $\|a+b\|^4 \leq \|a\|^4 + 3\|b\|^4 + 8\|a\|^2\|b\|^2 + 4\|a\|^2(a^T b)$ to expand the expression as follows

$$\begin{aligned}
\|\theta_i - \theta_{i+j+1}\|^4 =& \|\theta_i - \theta_{i+j} - \mu \nabla J(\theta_{i+j}) - \mu \xi_{i+j+1} - \mu d_{i+j+1}\|^4 \\
\leq& \|\theta_i - \theta_{i+j} - \mu \nabla J(\theta_{i+j}) - \mu d_{i+j+1}\|^4 + 3\mu^4 \|\xi_{i+j+1}\|^4 \\
&+ 8\mu^2 \|\theta_i - \theta_{i+j} - \mu \nabla J(\theta_{i+j}) - \mu d_{i+j+1}\|^2 \cdot \|\xi_{i+j+1}\|^2 \\
&+ 4\|\theta_i - \theta_{i+j} - \mu \nabla J(\theta_{i+j}) - \mu d_{i+j+1}\|^2 (\theta_i - \theta_{i+j} - \mu \nabla J(\theta_{i+j}) - \mu d_{i+j+1})^T \xi_{i+j+1}.
\end{aligned} \tag{18}$$

We first consider the first term on the right hand side of (18) and decompose it via Jensen's inequality:

$$\begin{aligned}
\|\theta_i - \theta_{i+j} - \mu \nabla J(\theta_{i+j}) - \mu d_{i+j+1}\|^4 \leq& \frac{1}{(1-\mu L)^3} \|\theta_i - \theta_{i+j} - \mu \nabla J(\theta_{i+j}) + \mu \nabla J(\theta_i)\|^4 \\
&+ \frac{\mu}{L^3} \|\nabla J(\theta_i) + d_{i+j+1}\|^4 \\
\leq& \frac{(1+\mu L)^4}{(1-\mu L)^3} \|\theta_i - \theta_{i+j}\|^4 + \frac{8\mu}{L^3} \|\nabla J(\theta_i)\|^4 + \frac{8\mu}{L^3} \|d_{i+j+1}\|^4.
\end{aligned} \tag{19}$$

We then consider the third term on the right hand side of (18). From the analysis above, we have

$$\|\theta_i - \theta_{i+j} - \mu \nabla J(\theta_{i+j}) - \mu d_{i+j+1}\|^2 \leq \frac{(1+\mu L)^2}{1-\mu L} \|\theta_i - \theta_{i+j}\|^2 + \frac{2\mu}{L} \|\nabla J(\theta_i)\|^2 + \frac{2\mu}{L} \|d_{i+j+1}\|^2. \tag{20}$$

Now we can plug (19) and (20) into (18) to obtain

$$\begin{aligned}
\|\theta_i - \theta_{i+j+1}\|^4 \leq& \frac{(1+\mu L)^4}{(1-\mu L)^3} \|\theta_i - \theta_{i+j}\|^4 + \frac{8\mu}{L^3} \|\nabla J(\theta_i)\|^4 + \frac{8\mu}{L^3} \|d_{i+j+1}\|^4 + 3\mu^4 \|\xi_{i+j+1}\|^4 \\
&+ 8\mu^2 \|\xi_{i+j+1}\|^2 \left(\frac{(1+\mu L)^2}{1-\mu L} \|\theta_i - \theta_{i+j}\|^2 + \frac{2\mu}{L} \|\nabla J(\theta_i)\|^2 + \frac{2\mu}{L} \|d_{i+j+1}\|^2 \right) \\
&+ 4\|\theta_i - \theta_{i+j} - \mu \nabla J(\theta_{i+j}) - \mu d_{i+j+1}\|^2 (\theta_i - \theta_{i+j} - \mu \nabla J(\theta_{i+j}) - \mu d_{i+j+1})^T \xi_{i+j+1}
\end{aligned}$$

When we take the expectation on both sides, the cross term with ξ_{i+j+1} disappears, and we have by (8), (9), (3) and (4)

$$\begin{aligned}\mathbb{E}[\|\theta_i-\theta_{i+j+1}\|^4|\mathcal{F}_{i+j}] \leq& \frac{(1+\mu L)^4}{(1-\mu L)^3}\|\theta_i-\theta_{i+j}\|^4+\frac{8\mu}{L^3}\|\nabla J(\theta_i)\|^4+\frac{8\mu}{L^3}\mathbb{E}[\|d_{i+j+1}\|^4|\mathcal{F}_{i+j}]+12\mu^4\sigma^4\\ &+\frac{8\mu^2\sigma^2(1+\mu L)^2}{1-\mu L}\|\theta_i-\theta_{i+j}\|^2+\frac{16\mu^3\sigma^2}{L}\|\nabla J(\theta_i)\|^2\\ &+\frac{16\mu^3}{L}\mathbb{E}[\|\xi_{i+j+1}\|^2\cdot\|d_{i+j+1}\|^2|\mathcal{F}_{i+j}]\end{aligned}$$

$$\begin{aligned}\mathbb{E}[\|\theta_i-\theta_{i+j+1}\|^4|\mathcal{F}_{i+j}] \leq& \frac{(1+\mu L)^4}{(1-\mu L)^3}\|\theta_i-\theta_{i+j}\|^4+\frac{8\mu}{L^3}\|\nabla J(\theta_i)\|^4+\frac{8D^4\mu^5}{L^3}+12\mu^4\sigma^4\\ &+\frac{8\mu^2\sigma^2(1+\mu L)^2}{1-\mu L}\|\theta_i-\theta_{i+j}\|^2+\frac{16\mu^3\sigma^2}{L}\|\nabla J(\theta_i)\|^2+\frac{16\sigma^2D^2\mu^5}{L}\end{aligned}$$

Now we take expectation conditioned on $\theta_i\in\mathcal{H}$, allowing us to use the bound (11) derived before on $\|\theta_i-\theta_{i+j}\|^2$ to obtain

$$\begin{aligned}\mathbb{E}[\|\theta_i-\theta_{i+j+1}\|^4|\theta_i\in\mathcal{H}] \leq& \frac{(1+\mu L)^4}{(1-\mu L)^3}\mathbb{E}[\|\theta_i-\theta_{i+j}\|^4|\theta_i\in\mathcal{H}]+\frac{8\mu}{L^3}\mathbb{E}[\|\nabla J(\theta_i)\|^4|\theta_i\in\mathcal{H}]\\ &+\frac{8\mu^2\sigma^2(1+\mu L)^2}{1-\mu L}\mathbb{E}[\|\theta_i-\theta_{i+j}\|^2|\theta_i\in\mathcal{H}]+\frac{16\mu^3\sigma^2}{L}\mathbb{E}[\|\nabla J(\theta_i)\|^2|\theta_i\in\mathcal{H}]+O(\mu^4)\\ \leq& \frac{(1+\mu L)^4}{(1-\mu L)^3}\mathbb{E}[\|\theta_i-\theta_{i+j}\|^4|\theta_i\in\mathcal{H}]+O(\mu^3)\end{aligned}$$

Then we can evaluate this recursive expression as follows

$$\begin{aligned}\mathbb{E}[\|\theta_i-\theta_{i+j+1}\|^4|\theta_i\in\mathcal{H}] &\leq O(\mu^3)\sum_{n=0}^{j-1}(\frac{(1+\mu L)^4}{(1-\mu L)^3})^n\\ &= O(\mu^3)\frac{1-(\frac{(1+\mu L)^4}{(1-\mu L)^3})^j}{1-\frac{(1+\mu L)^4}{(1-\mu L)^3}}\\ &= O(\mu^3)\frac{((\frac{(1+\mu L)^4}{(1-\mu L)^3})^j-1)(1-\mu L)^3}{(1+\mu L)^4-(1-\mu L)^3}\leq O(\mu^2)\frac{(\frac{(1+\mu L)^4}{(1-\mu L)^3})^j}{7L+3\mu L^2+5\mu^2L^3+\mu^3L^4}\end{aligned}$$

By Lemma B.5, this gives us

$$\mathbb{E}[\|\theta_i-\theta_{i+j+1}\|^4|\theta_i\in\mathcal{H}]\leq O(\mu^2).$$

Finally, we want to bound (13), restated below

$$\mathbb{E}[\|\theta_{i+j}-\theta'_{i+j}\|^2|\theta_i\in\mathcal{H}]\leq O(\mu).$$

First we expand the expression as follows using the definition of θ and θ'

$$\begin{aligned}\|\theta_{i+j+1}-\theta'_{i+j+1}\|^2&=\|\theta_{i+j}-\theta'_{i+j}+\mu\nabla J(\theta_{i+j})+\mu d_{i+j+1}-\mu\nabla J(\theta_i)-\mu\nabla^2J(\theta_i)(\theta'_{i+j}-\theta_i)\|^2\\ &=\|(I+\mu\nabla^2J(\theta_i))(\theta_{i+j}-\theta'_{i+j})+\mu\nabla^2J(\theta_i)(\theta_i-\theta_{i+j})+\mu\nabla J(\theta_{i+j})-\mu\nabla J(\theta_i)+\mu d_{i+j+1}\|^2\end{aligned}$$

Define $H_{i+j}=\int_0^1\nabla^2J((1-t)\theta_{i+j}+t\theta_i)dt$, then we can plug this into the expression and expand via Jensens's inequality to obtain

$$\begin{aligned}\|\theta_{i+j+1}-\theta'_{i+j+1}\|^2&=\|(I+\mu\nabla^2J(\theta_i))(\theta_{i+j}-\theta'_{i+j})+\mu(\nabla^2J(\theta_i)-H_{i+j})(\theta_i-\theta_{i+j})+\mu d_{i+j+1}\|^2\\ &\leq\frac{1}{1-\mu L}\|(I+\mu\nabla^2J(\theta_i))(\theta_{i+j}-\theta'_{i+j})\|^2+\frac{\mu}{L}\|(\nabla^2J(\theta_i)-H_{i+j})(\theta_i-\theta_{i+j})+d_{i+j+1}\|^2\\ &\leq\frac{1}{1-\mu L}\|(I+\mu\nabla^2J(\theta_i))(\theta_{i+j}-\theta'_{i+j})\|^2+\frac{2\mu}{L}\|(\nabla^2J(\theta_i)-H_{i+j})(\theta_i-\theta_{i+j})\|^2+\frac{2\mu}{L}\|d_{i+j+1}\|^2\end{aligned}$$

As observed in (Vlaski & Sayed, 2022), we have

$$\begin{aligned}\|\nabla^2 J(\theta_i) - H_{i+j}\| &= \|\nabla^2 J(\theta_i) - \int_0^1 \nabla^2 J((1-t)\theta_{i+j} + t\theta_i)dt\| \\ &= \|\int_0^1 \nabla^2 J(\theta_i) - \nabla^2 J((1-t)\theta_{i+j} + t\theta_i)dt\| \\ &\leq \int_0^1 \|\nabla^2 J(\theta_i) - \nabla^2 J((1-t)\theta_{i+j} + t\theta_i)\|dt \\ &\leq \chi \int_0^1 \|(1-t)\theta_i - (1-t)\theta_{i+j}\|dt \leq \frac{\chi}{2}\|\theta_i - \theta_{i+j}\|,\end{aligned} \tag{21}$$

which implies

$$\|(\nabla^2 J(\theta_i) - H_{i+j})(\theta_i - \theta_{i+j})\|^2 \leq \frac{\chi}{2}\|\theta_i - \theta_{i+j}\|^4. \tag{22}$$

We can plug (22) back into (21) and take expectation of both sides, conditioned on $\theta_i \in \mathcal{H}$. When we apply the fourth moment bound from Lemma B.3, we obtain

$$\mathbb{E}[\|\theta_{i+j+1} - \theta'_{i+j+1}\|^2 | \theta_i \in \mathcal{H}] \leq \frac{(1+\mu L)^2}{(1-\mu L)}\mathbb{E}[\|\theta_{i+j} - \theta'_{i+j}\|^2 | \theta_i \in \mathcal{H}] + O(\mu^3).$$

This is the same recursion as in the proof of (11), so again from Lemma B.5 we have

$$\mathbb{E}[\|\theta_{i+j+1} - \theta'_{i+j+1}\|^2 | \theta_i \in \mathcal{H}] \leq O(\mu^2).$$

□

C. Noise and Bias Bounds for Vanilla Policy Gradient

To apply Theorem 3.6, we first show in Lemma C.1 that the gradient estimator and the second and fourth moment of the noise are bounded. Then we show that the deterministic bias term d_{t+1} is bounded via Lemma C.3. This allows us to directly conclude the results of Theorem 3.7.

Lemma C.1. *The gradient noise process $\{\xi_t\}_{t\geq 0}$ satisfies*

$$\mathbb{E}[\xi_{t+1}|\mathcal{F}_t] = \mathbb{E}[\hat{G}^{VPG}(\theta_t;\tau_t) - \nabla J_H(\theta_t)|\mathcal{F}_t] = 0.$$

In addition, let $\sigma = \frac{G\mathcal{R}_{max}}{(1-\gamma)^2}$. Then we have the following bounds

$$\begin{aligned}\|\hat{G}^{VPG}(\theta_t;\tau_t)\| &\leq \sigma, \\ \mathbb{E}[\|\xi_{t+1}\|^2|\mathcal{F}_t] &\leq \sigma^2, \\ \mathbb{E}[\|\xi_{t+1}\|^4|\mathcal{F}_t] &\leq 4\sigma^4.\end{aligned}$$

Proof.

$$\begin{aligned}\|\hat{G}^{VPG}(\theta;\tau)\| &= \|\sum_{h=0}^{H-1} \nabla_\theta \log \pi_\theta(a_h|s_h) \sum_{t=h}^{H-1} \gamma^t \mathcal{R}(s_t,a_t)\| \\ &\leq \sum_{h=0}^{H-1} \|\nabla_\theta \log \pi_\theta(a_h|s_h) \sum_{t=h}^{H-1} \gamma^t \mathcal{R}(s_t,a_t)\| \\ &\leq \sum_{h=0}^{H-1} \|\nabla_\theta \log \pi_\theta(a_h|s_h)\|\gamma^h \sum_{t=h}^{H-1} \gamma^{t-h}\mathcal{R}_{max} \\ &\leq \sum_{h=0}^{H-1} \|\nabla_\theta \log \pi_\theta(a_h|s_h)\|\gamma^h \frac{\mathcal{R}_{max}}{1-\gamma} \\ &\leq \frac{G\mathcal{R}_{max}}{(1-\gamma)^2}\end{aligned}$$

Then the rest of the bounds follow from Lemma B.1. Thanks to reviewer feedback, we note that the bound on the noise variance $\mathbb{E}[\|\xi_{t+1}\|^2]$ can be tightened by a factor of $\frac{1}{1-\gamma}$ as shown in Lemma 4.2 of (Yuan et al., 2022). □

Before we can prove Lemma C.3, we require the following lemma from (Yuan et al., 2022).

Lemma C.2. *(Lemma 4.5 from (Yuan et al., 2022)) For $D = \frac{G\mathcal{R}_{max}}{1-\gamma}$, we have that the bias term d_{i+1} is bounded such that*

$$\|d_{i+1}\| = \|\nabla J(\theta_i) - \nabla J_H(\theta_i)\| \leq D(\frac{1}{1-\gamma} + H)^{1/2}\gamma^H.$$

Now we can proceed with the proof of Lemma C.3.

Lemma C.3. *For $H = \frac{1}{\log\frac{1}{\gamma}} \cdot O(\log(\frac{1}{\mu}))$ where $\mu \to 0$, we have that the gradient bias is deterministically bounded as follows*

$$\|d_{t+1}\| = \|\nabla J(\theta_t) - \nabla J_H(\theta_t)\| \leq D(\frac{1}{1-\gamma} + H)^{1/2}\gamma^H \leq D\mu$$

where $D = \frac{G\mathcal{R}_{max}}{1-\gamma}$.

Proof. We have the bound on the bias in terms of H from Lemma C.2. We want to choose H large enough so that

$$D(\frac{1}{1-\gamma} + H)^{1/2}\gamma^H \leq D\mu$$

$$(\frac{1}{1-\gamma} + H)^{1/2}\gamma^H \leq \mu.$$

We begin by finding the approximate solution to the following equation using asymptotic expansion

$$(\frac{1}{1-\gamma} + H)^{1/2}\gamma^H = \mu$$

$$\frac{1}{2}\log(\frac{1}{1-\gamma} + H) + H\log\gamma = \log\mu$$

$$\frac{1}{2}\log(\frac{1}{1-\gamma} + H) - H\log\frac{1}{\gamma} = -\log\frac{1}{\mu}$$

$$H\log\frac{1}{\gamma} - \frac{1}{2}\log(\frac{1}{1-\gamma} + H) = \log\frac{1}{\mu}.$$

Now we use the method of dominant balance, treating μ as a small parameter. We have that $\log\frac{1}{\mu}$ and $H\log\frac{1}{\gamma}$ must balance each other out, so

$$H \sim \frac{\log\frac{1}{\mu}}{\log\frac{1}{\gamma}} = \frac{\log\mu}{\log\gamma}.$$

We consider the next term in our asymptotic expansion of H, assuming that it is much smaller than the first term

$$H \sim \frac{\log\mu}{\log\gamma} + x_1.$$

We substitute this into the inequality to obtain

$$(\frac{\log\mu}{\log\gamma} + x_1)\log\frac{1}{\gamma} - \frac{1}{2}\log(\frac{1}{1-\gamma} + \frac{\log\mu}{\log\gamma} + x_1) = \log\frac{1}{\mu}$$

$$x_1\log\frac{1}{\gamma} - \frac{1}{2}\log(\frac{1}{1-\gamma} + \frac{\log\mu}{\log\gamma} + x_1) = 0$$

$$x_1 = \frac{\log(\frac{1}{1-\gamma} + \frac{\log\mu}{\log\gamma} + x_1)}{2\log\frac{1}{\gamma}}.$$

Since we assume x_1 is much smaller than $\frac{\log \mu}{\log \gamma}$ we have

$$x_1 \sim \frac{\log(\frac{1}{1-\gamma} + \frac{\log \mu}{\log \gamma})}{2 \log \frac{1}{\gamma}}.$$

Our asymptotic expansion is $H = \frac{\log \mu}{\log \gamma} + O(\log(\frac{\log \mu}{\log \gamma}))$. So we can pick $H = O(\frac{\log \mu}{\log \gamma})$ to achieve our inequality. □

D. Proof of Theorem 4.8

D.1. Key Lemmas and Proof Sketch

In Theorem 4.8, we establish the convergence of $Q_{\bar{w}_K}$ to Q_{w^*} under constant time steps. Our approach will mirror that of (Bhandari et al., 2018) by establishing a recurrence relation for the iterates and then bounding the bias induced by Markovian sampling. The key challenge is characterizing the distance between the initial state distribution and the stationary distribution in terms of the mixing rate.

We define $\bar{g}(w)$ as the expectation of of the semigradient $g_t(w)$ with respect to the stationary distribution of the Markov chain as follows

$$\bar{g}_(w) = \mathbb{E}_{\eta_\pi}[g_t(w)] = \sum_{s,s',a,a'} \eta_\pi(s,a)\mathcal{P}(s,a,s',a')(\mathcal{R}(s,a) + \gamma\phi(s',a')^\intercal w - \phi(s,a)^\intercal w)\phi(s,a).$$

We also define ζ_t to represent the bias from Markovian sampling as follows

$$\zeta_t(w) = (g_t(w) - \bar{g}(w))^\intercal (w - w^*).$$

First, The following lemma, which is Lemma 6 and 10 from (Bhandari et al., 2018), uniformly bounds the norm of the semi-gradient and Markov bias term ζ_t.

Lemma D.1. *Let $F = \mathcal{R}_{max} + 2R$. Then $R \leq \frac{F}{2}$ and for all $t \geq 0$ we have*

$$\|g_t(w)\|_2 \leq \mathcal{R}_{\max} + 2\|w\|_2 \leq F$$

In addition, for all $w \in \Theta$, the gradient bias is bounded such that

$$|\zeta_t(w)| \leq 2F^2$$

$$|\zeta_t(w) - \zeta_t(w')| \leq 6F\|w - w'\|_2.$$

Then we obtain the following lemma for general nonstationary Markov chains, which differs from Lemma 9 in (Bhandari et al., 2018) by a factor of 2.

Lemma D.2. *Consider two random variables X and Y such that*

$$X \to s_t \to s_{t+\tau} \to Y$$

forms a Markov chain for some fixed $t \geq 0$ and $\tau > 0$. Assume the Markov chain mixes at a uniform geometric rate as in Assumption 4.2. Let X' and Y' denote independent copies drawn from the marginal distributions of X and Y, so $\mathbb{P})X' = \cdot, Y' = \cdot) = \mathbb{P}(X = \cdot) \otimes \mathbb{P}(Y = \cdot)$. Then, for any bounded function v,

$$|\mathbb{E}[v(X,Y)] - \mathbb{E}[v(X',Y')]| \leq 4\|v\|_\infty(mr^\tau),$$

where $\|v\|_\infty = \sup_x |f(x)|$.

Proof. See Appendix D.3.

Now in the following key lemma, we apply Lemma D.2 to bound $\zeta_t(w_t)$ with respect to exponential mixing. Although we follow the proof of Lemma 11 in (Bhandari et al., 2018), it is not sufficient to directly carry the factor of 2 over from Lemma D.2 because we need to account for the fact that the marginal distribution of each observation O_t is now time-dependent and not equal to the stationary distribution.

Lemma D.3. *Consider a non-increasing step-size sequence $\alpha_0 \geq \alpha_1 \geq ... \geq \alpha_T$. Let $\tau_0 = \tau^{mix}(\alpha_T)$. Fix any $t \leq T$ and set $t^* = \max\{0, t - \tau_0\}$. Then,*

$$\mathbb{E}[\zeta_t(w_t)] \leq F^2(8 + 6\tau_0)\alpha_{t^*} + 10F^2 m r^t.$$

Proof. See Appendix D.4.

Finally, we have the following lemma from (Bhandari et al., 2018) that establishes a recursion for the distance between the iterates and the limit point w^*.

Lemma D.4. *With probability 1, for every $t \in \mathbb{N}_0$,*

$$\|w^* - w_{t+1}\|^2 \leq \|w^* - w_t\|^2 - 2\alpha_t(1-\gamma)\|Q_{w^*} - Q_{w_t}\|_{\eta_\pi}^2 + 2\alpha_t\zeta_t(w_t) + \alpha_t^2 F^2.$$

Proof. See the proof of Lemma 8 in (Bhandari et al., 2018).

These key lemmas allow us to proceed with the proof of Theorem 4.8.

D.2. Proof of Theorem 4.8

Proof. Rearranging the terms of the inequality in Lemma D.4 and summing from $t = 0$ to $K - 1$, we arrive at

$$2\alpha_0(1-\gamma)\sum_{t=0}^{K-1}\mathbb{E}[\|Q_{w^*} - Q_{w_t}\|_{\eta_\pi}^2] \leq \|w^* - w_0\|_2^2 + F^2 + 2\alpha_0\sum_{t=0}^{K-1}\mathbb{E}[\zeta_t(w_t)].$$

Then we can use the bound on $\zeta_t(w_t)$ from Lemma D.3 and the fact that our step-sizes are constant to obtain

$$\begin{aligned}
2\alpha_0(1-\gamma)\sum_{t=0}^{K-1}\mathbb{E}[\|Q_{w^*} - Q_{w_t}\|_{\eta_\pi}^2] &\leq \|w^* - w_0\|_2^2 + F^2 + 2\alpha_0\sum_{t=0}^{K-1}F^2(8+6\tau_0)\alpha_0 + 2\alpha_0\sum_{t=0}^{K-1}10F^2 m r^t \\
&\leq \|w^* - w_0\|_2^2 + F^2 + 2\alpha_0^2 K F^2(8+6\tau_0) + \frac{20F^2 m\alpha_0}{1-r}.
\end{aligned}$$

Now we can divide both sides by $2\alpha_0(1-\gamma)$ and substitute $\alpha_0 = \frac{1}{\sqrt{K}}$ to obtain

$$\begin{aligned}
\sum_{t=0}^{K-1}\mathbb{E}[\|Q_{w^*} - Q_{w_t}\|_{\eta_\pi}^2] &\leq \frac{\|w^* - w_0\|_2^2 + F^2}{2\alpha_0(1-\gamma)} + \frac{\alpha_0 K F^2(8+6\tau_0)}{1-\gamma} + \frac{10F^2 m}{(1-r)(1-\gamma)} \\
&= \frac{\sqrt{K}(\|w^* - w_0\|_2^2 + F^2)}{2(1-\gamma)} + \frac{\sqrt{K}F^2(8+6\tau_0)}{1-\gamma} + \frac{10F^2 m}{(1-r)(1-\gamma)} \\
&= \frac{\sqrt{K}(\|w^* - w_0\|_2^2 + 17F^2 + 12F^2\tau_0)}{2(1-\gamma)} + \frac{10F^2 m}{(1-r)(1-\gamma)}.
\end{aligned}$$

Finally, we divide both sides by K and use Jensen's inequality to obtain our final result

$$\mathbb{E}[\|Q_{w^*} - Q_{\bar{w}_K}\|_{\eta_\pi}^2] \leq \frac{1}{K}\sum_{t=0}^{K-1}\mathbb{E}[\|Q_{w^*} - Q_{w_t}\|_{\eta_\pi}^2] \leq \frac{\|w^* - w_0\|_2^2 + F^2(17+12\tau_0)}{2(1-\gamma)\sqrt{K}} + \frac{10F^2 m}{(1-r)(1-\gamma)K}.$$

□

D.3. Proof of Lemma D.2

We first require the following auxiliary lemma.

Lemma D.5. *For any Markov chain $\{s_t\}$ with stationary distribution η and finite state space $\mathcal{S}$,*

$$d_{TV}(\eta, \mathbb{P}(s_{t+\tau} = \cdot)) \leq \sup_{s \in S} d_{TV}(\eta, \mathbb{P}(s_{t+\tau} = \cdot | s_0 = s)).$$

Proof. It follows from proof by induction that for a general convex function f, if $f(x_n) \geq f(x_i)$ for all $x_i \in \{x_1, ...x_n\}$ then $f(x_n) \geq f(x)$ for $x = \sum_{i=1}^n \alpha_i x_n$ where $\sum_{i=1}^n \alpha_i = 1$. In other words, $f(x_n)$ is greater than $f(x)$ for any x that is a convex combination of the other $\{x_1, ...x_n\}$.

Now let $f(x) = \frac{1}{2} \sum_{s_i \in \mathcal{S}} |x^\intercal P^{t+\tau}(s_i) - \eta(s_i)|$, which is a convex function, and let $e_i \in \mathbb{R}^S$ represent the unit vector that is 1 at index i and 0 everywhere else. Then we have

$$d_{TV}(\eta, \mathbb{P}(s_{t+\tau} = \cdot)) = f(\rho_0)$$

$$d_{TV}(\eta, \mathbb{P}(s_{t+\tau} = \cdot | s_0 = s_i)) = f(e_i).$$

Since any initial state distribution ρ_0 will be a convex combination of the e_i unit vectors, we have shown the result. □

Now we can proceed with the proof of Lemma D.2.

Proof. Let $h = \frac{v}{2\|v\|_\infty}$ denote the function v rescaled to take values in $[-1/2, 1/2]$. Then we can follow the steps of Lemma 9 in (Bhandari et al., 2018) to arrive at

$$|\mathbb{E}[h(X,Y)] - \mathbb{E}[h(X',Y')]| \leq \sum_{s \in \mathcal{S}} \mathbb{P}(s_t = s) d_{TV}(\mathbb{P}(s_{t+\tau} = \cdot | s_t = s), \mathbb{P}(s_{t+\tau} = \cdot). \tag{23}$$

We can bound $d_{TV}(\mathbb{P}(s_{t+\tau} = \cdot | s_t = s), \mathbb{P}(s_{t+\tau} = \cdot))$ as follows

$$d_{TV}(\mathbb{P}(s_{t+\tau} = \cdot | s_t = s), \mathbb{P}(s_{t+\tau} = \cdot)) \leq d_{TV}(\mathbb{P}(s_{t+\tau} = \cdot | s_t = s), \eta_\pi) + d_{TV}(\eta_\pi, \mathbb{P}(s_{t+\tau} = \cdot)) \tag{24}$$

because total variation distance is a norm and obeys the triangle inequality. The first term on the right hand side of (24) is bounded by exponential mixing

$$d_{TV}(\mathbb{P}(s_{t+\tau} = \cdot | s_t = s), \eta_\pi) \leq m r^\tau.$$

The second term can be bound as follows by Lemma D.5

$$d_{TV}(\eta_\pi, \mathbb{P}(s_{t+\tau} = \cdot)) \leq \sup_{s \in S} d_{TV}(\eta_\pi, \mathbb{P}(s_{t+\tau} = \cdot | s_0 = s)) \leq m r^{t+\tau}.$$

Returning to (24), we have

$$d_{TV}(\mathbb{P}(s_{t+\tau} = \cdot | s_t = s), \mathbb{P}(s_{t+\tau} = \cdot)) \leq m r^\tau + m r^{t+\tau} \leq 2 m r^\tau,$$

and applying these inequalities to (23), we have

$$|\mathbb{E}[v(X,Y)] - \mathbb{E}[v(X',Y')]| \leq 4 \|v\|_\infty m r^\tau.$$

□

D.4. Proof of Lemma D.3

First we require the following auxiliary lemmas.

Lemma D.6. *Let P and Q represent two different probability distributions. For some bounded function f, and random variable X, we have*

$$|\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)]| \leq 2 \|f\|_\infty d_{TV}(P, Q)$$

Proof. This can be shown with the definition

$$d_{TV}(P, Q) = \sup_{v : \|v\|_\infty \leq \frac{1}{2}} |\int v dP - \int v dQ|$$

□

Lemma D.7. *Let $O'' = (s, a, s', a')$ represent the observation tuple of consecutive state-action pairs where (s, a) is drawn from the stationary distribution of the Markov chain η_π, and let $O_t = (s_t, a_t, s_{t+1}, a_{t+1})$ represent the observation tuple drawn at time t from the Markov chain. Then*

$$d_{TV}(\mathbb{P}(O'' = \cdot), \mathbb{P}(O_t = \cdot)) = d_{TV}(\eta_\pi, \mathbb{P}(s_t = \cdot, a_t = \cdot)) \leq mr^t$$

Now we can proceed with the proof of Lemma D.3, which mirrors the framework of Lemma 11 in (Bhandari et al., 2018). The main difference in our proofs is in Step 2.

Proof. Step 1: Relate $\zeta_t(w_t)$ and $\zeta_t(w_{t-\tau})$

We apply Lemma D.1 to bound ζ_t as follows

$$|\zeta_t(w_t) - \zeta_t(w_{t-\tau})| \leq 6F\|w - w_{t-\tau}\| \leq 6F^2 \sum_{i=t-\tau}^{t-1} \alpha_i$$

$$\zeta_t(w_t) \leq \zeta_t(w_{t-\tau}) + 6F^2 \sum_{i=t-\tau}^{t-1} \alpha_i \tag{25}$$

Step 2: Bound $\mathbb{E}[\zeta_t(w_{t-\tau})]$ using Lemma D.2 and exponential mixing

We denote by $O_t = (s_t, a_t, s_{t+1}, a_{t+1})$ the observation tuple at each time step, and we overload the notation of g_t and ζ_t to make clear their dependence on O_t as follows

$$g_t(w, O_t) = (r(s_t, a_t) + \gamma\phi(s_{t+1}, a_{t+1})^\intercal w - \phi(s_t, a_t)^\intercal w)\phi(s_t, a_t)$$

$$\zeta(w, O_t) = (g_t(w, O_t) - \bar{g}(w))^\intercal (w - w^*)$$

To apply Lemma D.2, we consider random variables $w'_{t-\tau}$ and O'_t drawn independently from their marginal distributions $\mathbb{P}(w_{t-\tau} = \cdot)$ and $\mathbb{P}(O_t = \cdot)$ respectively, such that their joint distribution is defined as follows

$$\mathbb{P}(w'_{t+\tau} = \cdot, O'_t = \cdot) = \mathbb{P}(w_{t-\tau} = \cdot) \otimes \mathbb{P}(O_t = \cdot)$$

Note that typically the random variables $w_{t-\tau}$ and O_t are not independent. Now we want to bound $\mathbb{E}[\zeta(w'_{t-\tau}, O'_t)]$. We can use the law of total expectation as follows

$$\mathbb{E}[\zeta(w'_{t-\tau}, O'_t)] = \mathbb{E}[\mathbb{E}[\zeta(w'_{t-\tau}, O'_t)|w'_{t-\tau}]]$$

Now we consider the conditional expectation term $\mathbb{E}[\zeta(w'_{t-\tau}, O'_t)|w'_{t-\tau}]$. Since $w'_{t-\tau}$ and O'_t are independent, we have

$$\begin{aligned}
\mathbb{E}[\zeta(w'_{t-\tau}, O'_t)|w'_{t-\tau}] &= \mathbb{E}_{O'_t \sim \mathbb{P}(O_t = \cdot)}[\zeta(w'_{t-\tau}, O'_t)|w'_{t-\tau}] \\
&= \mathbb{E}_{O'_t \sim \mathbb{P}(O_t = \cdot)}[(g_t(w'_{t-\tau}, O'_t) - \bar{g}(w'_{t-\tau}))^\intercal (w'_{t-\tau} - w^*)|w'_{t-\tau}] \\
&= (\mathbb{E}_{O'_t \sim \mathbb{P}(O_t = \cdot)}[g_t(w'_{t-\tau}, O'_t)] - \bar{g}(w'_{t-\tau}))^\intercal (w'_{t-\tau} - w^*) \\
&= (\mathbb{E}_{O'_t \sim \mathbb{P}(O_t = \cdot)}[g_t(w'_{t-\tau}, O'_t)] - \mathbb{E}_{O'' \sim \eta_\pi}[g(w'_{t-\tau}, O'')])^\intercal (w'_{t-\tau} - w^*) \\
&\leq \|\mathbb{E}_{O'_t \sim \mathbb{P}(O_t = \cdot)}[g_t(w'_{t-\tau}, O'_t)] - \mathbb{E}_{O'' \sim \eta_\pi}[g(w'_{t-\tau}, O'')]\| \cdot \|w'_{t-\tau} - w^*\| \\
&\leq \|\mathbb{E}_{O'_t \sim \mathbb{P}(O_t = \cdot)}[g_t(w'_{t-\tau}, O'_t)] - \mathbb{E}_{O'' \sim \eta_\pi}[g(w'_{t-\tau}, O'')]\| \cdot 2R,
\end{aligned}$$

where the last inequality is due to the fact that $\|w\| \leq R$. Here O'_t is drawn from the time-dependent marginal distribution $\mathbb{P}(O_t = \cdot)$ whereas O'' is drawn from the stationary distribution of the Markov chain. Then by Lemma D.6 and the fact that $\|g_t\| \leq F$ and $R \leq \frac{F}{2}$, we can bound this difference in expectation by the total variation distance between the two distributions as follows

$$\mathbb{E}[\zeta(w'_{t-\tau}, O'_t)|w'_{t-\tau}] \leq 2F^2 d_{TV}(\mathbb{P}(O'' = \cdot), \mathbb{P}(O_t = \cdot)).$$

Then by Lemma D.5 and Lemma D.7 we have

$$\mathbb{E}[\zeta(w'_{t-\tau}, O'_t)|w'_{t-\tau}] \leq 2F^2 mr^t.$$

Now we can apply Lemma D.2 to bound $\mathbb{E}[\zeta_t(w_{t-\tau}), O_t)]$. By Lemma D.1, we have $|\zeta(w, O_t)| \leq 2F^2$. Since $\theta_{t-\tau} \to s_{t-\tau} \to s_t \to O_t$ forms a Markov chain, applying Lemma D.2 yields

$$|\mathbb{E}[\zeta(w_{t-\tau}, O_t)] - \mathbb{E}[\zeta(w'_{t-\tau}, O'_t)]| \leq 8F^2 m r^\tau$$
$$|\mathbb{E}[\zeta(w_{t-\tau}, O_t)]| \leq 8F^2 m r^\tau + 2F^2 m r^t. \tag{26}$$

Step 3: Combine terms.

We combine (25) and (26) to arrive at

$$\mathbb{E}[\zeta_t(w_t)] \leq 8F^2 m r^\tau + 2F^2 m r^t + 6F^2 \tau \alpha_{t-\tau}.$$

Let $\tau_0 = \tau^{\text{mix}}(\alpha_T)$. For $t \leq \tau_0$, pick $\tau = t$, to arrive at the bound

$$\mathbb{E}[\zeta_t(w_t)] \leq 8F^2 m r^t + 2F^2 m r^t + 6F^2 t \alpha_0 \leq 10F^2 m r^t + 6F^2 \tau_0 \alpha_0.$$

Now for $T \geq t > \tau_0$ we pick $\tau = \tau_0$ to get the bound

$$\mathbb{E}[\zeta_t(w_t)] \leq 8F^2 \alpha_T + 2F^2 m r^t + 6F^2 \tau_0 \alpha_{t-\tau_0} \leq F^2(8 + 6\tau_0)\alpha_{t-\tau_0} + 2F^2 m r^t.$$

□

E. Noise and Bias Bounds for Actor-Critic Policy Gradient

Our general approach for actor-critic policy gradient is similar to that of vanilla policy gradient. Once again, we bound the noise and bias — bounding ξ_t in Lemma E.1, p_t in Lemma E.2, and d_t in Lemma E.1.

We note that for vanilla policy gradient, the noise term ξ_{t+1} is random due to the sampled trajectory τ_t whereas the bias term d_{t+1} is deterministic conditioned on $\mathcal{F}_t$. In contrast, for the actor-critic algorithm, ξ_{t+1} depends on two random variables, τ_t and the averaged critic parameter $\bar{w}_{K,t}$, whereas the bias term depends only on $\bar{w}_{K,t}$. Moreover, τ_t and $\bar{w}_{K,t}$ are generated from independently sampled trajectories and are therefore independent. We quantify these observations in the following lemma.

Lemma E.1. *The gradient noise process $\{\xi_t\}_{t\geq 0}$ satisfies*

$$\mathbb{E}[\xi_{t+1}|\mathcal{F}_t] = 0$$

In addition, let $\sigma = \frac{GR}{1-\gamma}$. Then we have the following bounds

$$\|\hat{G}(\theta_t; \tau_t)\| \leq \sigma,$$
$$\mathbb{E}[\|\xi_{t+1}\|^2|\mathcal{F}_t] \leq \sigma^2$$
$$\mathbb{E}[\|\xi_{t+1}\|^4|\mathcal{F}_t] \leq 4\sigma^4$$
$$\mathbb{E}[\|d_{t+1}\|^2\|\xi_{t+1}\|^2|\mathcal{F}_t] \leq \sigma^2 \mathbb{E}[\|d_{t+1}\|^2|\mathcal{F}_t]$$

Proof. As discussed earlier, ξ_{t+1} depends on two random independent variables: τ_t and $\bar{w}_{K,t}$. We can overload notation and denote the gradient estimator as $\hat{G}(\theta_t; \tau_t; \bar{w}_{K,t})$. We therefore obtain

$$\begin{aligned}
\mathbb{E}[\xi_{t+1}|\mathcal{F}_t] &= \mathbb{E}[\hat{G}(\theta_t; \tau_t; \bar{w}_{K,t}) - \mathbb{E}_\tau[\hat{G}(\theta_t; \tau_t; \bar{w}_{K,t})]|\mathcal{F}_t] \\
&= \mathbb{E}[\mathbb{E}[\hat{G}(\theta_t; \tau_t; \bar{w}_{K,t}) - \mathbb{E}_\tau[\hat{G}(\theta_t; \tau_t; \bar{w}_{K,t})]|\bar{w}_{K,t}]|\mathcal{F}_t] \\
&= \mathbb{E}[\mathbb{E}_\tau[\hat{G}(\theta_t; \tau_t; \bar{w}_{K,t})] - \mathbb{E}_\tau[\hat{G}(\theta_t; \tau_t; \bar{w}_{K,t})]|\mathcal{F}_t] = 0
\end{aligned}$$

Now we want to show the bound on $\hat{G}(\theta_t; \tau_t)$.

$$\begin{aligned}
\|\hat{G}(\theta_t; \tau_t; \bar{w}_{K,t})\| &= \|\sum_{j=0}^{H} \gamma^j Q_{\bar{w}_{K,t}}(s_j, a_j) \nabla \log \pi_{\theta_t}(a_j|s_j)\| \\
&= \|\sum_{j=0}^{H} \gamma^j \bar{w}_{K,t}^\intercal \phi(s_j, a_j) \nabla \log \pi_{\theta_t}(a_j|s_j)\| \leq \frac{RG}{1-\gamma}
\end{aligned}$$

By Lemma B.1

$$\begin{aligned}\mathbb{E}[\|\xi_{t+1}\|^2||\bar{w}_{K,t}] &= \mathbb{E}_\tau[\|\xi_{t+1}\|^2] \leq \sigma^2\\ \mathbb{E}[\|\xi_{t+1}\|^2|\mathcal{F}_t] &= \mathbb{E}[\mathbb{E}[\|\xi_{t+1}\|^2|\bar{w}_{K,t}]|\mathcal{F}_t] \leq \sigma^2\\ \mathbb{E}[\|\xi_{t+1}\|^4|\mathcal{F}_t] &\leq 4\sigma^2\end{aligned}$$

Finally we have

$$\begin{aligned}\mathbb{E}[\|d_{t+1}\|^2\|\xi_{t+1}\|^2|\mathcal{F}_t] &= \mathbb{E}[\mathbb{E}[\|d_{t+1}\|^2\|\xi_{t+1}\|^2|\bar{w}_{K,t}]|\mathcal{F}_t]\\ &= \mathbb{E}[\|d_{t+1}\|^2\mathbb{E}[\|\xi_{t+1}\|^2|\bar{w}_{K,t}]|\mathcal{F}_t]\\ &\leq \mathbb{E}[\|d_{t+1}\|^2|\mathcal{F}_t]\sigma^2\end{aligned}$$

□

In the following lemma, we bound the bias in the actor-critic gradient estimator due to truncation of the infinite horizon. The approach is similar to the proof of Lemma C.2.

Lemma E.2. *For $D_p = \frac{GR}{(1-\gamma)}$ and $H \geq \frac{\log \mu}{\log \gamma}$, the truncation bias is deterministically bounded such that*

$$\|p_{t+1}\| \leq D_p\mu$$

$$\|p_{t+1}\|^4 \leq D_p^4\mu^4.$$

Proof.

$$\begin{aligned}\|p_{t+1}\| &= \|G_H(\theta_t) - G_\infty(\theta_t)\|\\ &= \|\mathbb{E}[\sum_{j=0}^{H}\gamma^j Q_{\bar{w}_{K,t}}(s_j,a_j)\nabla\log\pi_{\theta_t}(a_j|s_j)] - \mathbb{E}[\sum_{j=0}^{\infty}\gamma^j Q_{\bar{w}_{K,t}}(s_j,a_j)\nabla\log\pi_{\theta_t}(a_j|s_j)]\|\\ &= \|\mathbb{E}[\sum_{j=H}^{\infty}\gamma^j Q_{\bar{w}_{K,t}}(s_j,a_j)\nabla\log\pi_{\theta_t}(a_j|s_j)]\|\\ &= \|\mathbb{E}[\sum_{j=H}^{\infty}\gamma^j \bar{w}_{K,t}^\intercal\phi(s_j,a_j)\nabla\log\pi_{\theta_t}(a_j|s_j)]\|\\ &\leq RG\sum_{j=H}^{\infty}\gamma^j \leq \frac{GR}{(1-\gamma)}\gamma^H\end{aligned}$$

□

E.1. Proof of Lemma 4.10

Although Theorem 4.8 establishes the convergence of TD(0) in terms of $Q_{\bar{w}_K}$ under constant step sizes, we actually require a stronger convergence result showing the direct convergence of $\bar{w}_K$ to w^* in order to bound the critic approximation bias. In the following proofs, we use core results proved in Appendix D to prove an alternate fourth-moment bound under diminishing step sizes. These convergence results are formalized in Lemma E.3.

Lemma E.3. *Suppose $\bar{w}_K$ is generated by K steps of the Projected TD(0) algorithm with $w^* \in \Theta$ and diminishing step-size $\alpha_t = \frac{1}{(t+1)\varsigma}$. Then*

$$\mathbb{E}[\|w^* - \bar{w}_K\|^4] \leq \frac{\log^2 K}{K}\cdot\left(\frac{192F^2R^2}{\varsigma^2\log^2(r^{-1})} + O(\frac{1}{\log K}) + O(\frac{1}{\log^2 K})\right)$$

Proof: See Appendix E.2.1

Lemma E.4 follows from Lemma E.3.

Lemma E.4. *For $K = O(\frac{\log^2(\mu^{-4})}{\mu^4})$ as $\mu \to 0$ iterations of Algorithm 3 and*

$$D = \left(\frac{192F^2R^2}{\varsigma^2 \log^2(r^{-1})} + O(\frac{1}{\log \mu^{-4}})\right)^{-1/4}$$

we have

$$\mathbb{E}[\|w^* - \bar{w}_K\|^4] \leq D^4\mu^4$$

Proof. See Appendix E.2.2.

Now we can proceed with the proof of Lemma 4.10.

Proof. We can characterize the bias by the quality of the critic approximation achieved by the TD(0) algorithm. We can "roll up" the temporal summation as follows

$$\begin{aligned}
q_{t+1} &= G_\infty - \nabla J(\theta_t) \\
&= \mathbb{E}_{\pi,\rho_0}[\sum_{k=0}^{\infty} \gamma^k Q_{\bar{w}_{K,t}}(s_k, a_k) \nabla \log \pi_\theta(a_k|s_k)] - \nabla J(\theta) \\
&= \mathbb{E}_{\pi,\rho_0}[\sum_{k=0}^{\infty} \gamma^k Q_{\bar{w}_{K,t}}(s_k, a_k) \nabla \log \pi_\theta(a_k|s_k)] - \mathbb{E}_{\pi,\rho_0}[\sum_{k=0}^{\infty} \gamma^k \nabla \log \pi(a_k|s_k) Q^\pi_\gamma(s_k, a_k)] \\
&= \mathbb{E}_{\pi,\rho_0}[\sum_{k=0}^{\infty} \gamma^k \nabla \log \pi_\theta(a_k|s_k)(Q_{\bar{w}_{K,t}}(s_k, a_k) - Q^\pi_\gamma(s_k, a_k))] \\
&= \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \sum_{k=0}^{\infty} \mathbb{P}(s_k = s|s_0)\pi(a|s)\gamma^k \nabla \log \pi_\theta(a|s)(Q_{\bar{w}_{K,t}}(s, a) - Q^\pi_\gamma(s, a)) \\
&= \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} d^\pi_\gamma(s)\pi(a|s) \nabla \log \pi_\theta(a|s)(Q_{\bar{w}_{K,t}}(s, a) - Q^\pi_\gamma(s, a)),
\end{aligned}$$

where $d^\pi_\gamma(s) = \sum_{k=0}^{\infty} \gamma^k \mathbb{P}(s_k = s|s_0)$ denotes the discounted state visitation measure. Since we have

$$\sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} d^\pi_\gamma(s)\pi(a|s) = \sum_{s\in\mathcal{S}} d^\pi_\gamma(s) = \frac{1}{1-\gamma},$$

then by Jensen's inequality, we can obtain the following bound

$$\begin{aligned}
\|q_{t+1}\|^4 &\leq (1-\gamma) \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} d^\pi_\gamma(s)\pi(a|s) \|\nabla \log \pi_\theta(a|s)(Q_{\bar{w}_{K,t}}(s, a) - Q^\pi_\gamma(s, a))\|^4 \\
&\leq (1-\gamma) \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} d^\pi_\gamma(s)\pi(a|s) G^4 \|Q_{\bar{w}_{K,t}}(s, a) - Q^\pi_\gamma(s, a)\|^4 \\
&= (1-\gamma) \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} d^\pi_\gamma(s)\pi(a|s) G^4 \|\bar{w}^\intercal_{K,t}\phi(s, a) - w^{*\intercal}\phi(s, a)\|^4 \\
&\leq (1-\gamma) \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} d^\pi_\gamma(s)\pi(a|s) G^4 \|\phi(s, a)\|^4 \cdot \|\bar{w}_{K,t} - w^*\|^4 \\
&\leq (1-\gamma) \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} d^\pi_\gamma(s)\pi(a|s) G^4 \cdot \|\bar{w}_{K,t} - w^*\|^4 \\
&\leq G^4 \cdot \|\bar{w}_{K,t} - w^*\|^4.
\end{aligned}$$

Now we take expectation on both sides with respect to the TD(0) algorithm that occurs between timestep t and $t+1$ to obtain

$$\mathbb{E}[\|q_{t+1}\|^4] \leq G^4 \cdot \mathbb{E}[\|\bar{w}_{K,t} - w^*\|^4].$$

Finally, we can apply Lemma E.4 to achieve the final result. □

E.2. Proof of Key Lemmas

E.2.1. Proof of Lemma E.3

We first require the following auxiliary lemma.

Lemma E.5. *Recall the definition of* $A_\theta = \mathbb{E}_{(s,a)\sim\eta_\theta,(s',a')\sim P(s,a,\cdot)}[\phi(s,a)(\phi(s,a)-\gamma\phi(s',a'))^\intercal]$. *Then*

$$(w^*-w_t)^\intercal \bar{g}(w_t) \geq \frac{1}{2}\lambda_{min}(A_\theta + A_\theta^T)\|w^*-w_t\|^2 \geq \frac{1}{2}\varsigma\|w_t-w^*\|^2$$

where ς *is defined in Assumption 4.6.*

Proof.

$$\begin{aligned}
(w^*-w_t)^\intercal \bar{g}(w_t) &= (w^*-w_t)^\intercal(\bar{g}(w_t)-\bar{g}(w^*)) \\
&= (w^*-w_t)^\intercal \mathbb{E}[(\gamma\phi(s',a')^\intercal w_t - \phi(s,a)^\intercal w_t - \gamma\phi(s',a')^\intercal w^* - \phi(s,a)^\intercal w^*)\phi(s,a)] \\
&= (w^*-w_t)^\intercal \mathbb{E}[(\gamma\phi(s',a')-\phi(s,a))^\intercal (w_t-w^*)\phi(s,a)] \\
&= (w^*-w_t)^\intercal A_\theta (w^*-w_t) \\
&= \frac{1}{2}(w^*-w_t)^\intercal (A_\theta + A_\theta^T)(w^*-w_t) \\
&\geq \frac{1}{2}\lambda_{min}(A_\theta + A_\theta^T)\|w^*-w_t\|^2
\end{aligned}$$

□

Now we can proceed with the proof of Lemma E.3.

Proof.

$$\begin{aligned}
\|w^*-w_{t+1}\|^2 &= \|w^* - Proj_\Theta(w_t+\alpha_t g_t(w_t))\|^2 \\
&= \|Proj_\Theta(w^*) - Proj_\Theta(w_t+\alpha_t g_t(w_t))\|^2 \\
&\leq \|w^*-w_t-\alpha_t g_t(w_t)\|^2 \\
&= \|w^*-w_t\|^2 - 2\alpha_t g_t(w_t)^\intercal(w^*-w_t) + \alpha_t^2\|g_t(w_t)\|^2 \\
&= \|w^*-w_t\|^2 - 2\alpha_t \bar{g}(w_t)^\intercal(w^*-w_t) + 2\alpha_t\zeta_t(w_t) + \alpha_t^2 F^2
\end{aligned}$$

So we have the following fourth moment bound (since both sides of the inequality are positive):

$$\begin{aligned}
\|w^*-w_{t+1}\|^4 \leq& \left[\|w^*-w_t\|^2 - 2\alpha_t\bar{g}(w_t)^\intercal(w^*-w_t) + 2\alpha_t\zeta_t(w_t) + \alpha_t^2F^2\right]^2 \\
=& \|w^*-w_t\|^4 + 4\alpha_t^2(\bar{g}(w_t)^\intercal(w^*-w_t))^2 + 4\alpha_t^2(\zeta_t(w_t))^2 + \alpha_t^4F^4 \\
&- 4\alpha_t\|w^*-w_t\|^2(\bar{g}(w_t)^\intercal(w^*-w_t)) + 4\alpha_t\|w^*-w_t\|^2\zeta_t(w_t) + 2\alpha_t^2F^2\|w^*-w_t\|^2 \\
&- 8\alpha_t^2(\bar{g}(w_t)^\intercal(w^*-w_t))\zeta_t(w_t) - 4\alpha_t^3F^2(\bar{g}(w_t)^\intercal(w^*-w_t)) + 4\alpha_t^3F^2\zeta_t(w_t)
\end{aligned}$$

By Lemma E.5,

$$\begin{aligned}
\|w^*-w_{t+1}\|^4 \leq& \|w^*-w_t\|^4 + 4\alpha_t^2(\bar{g}(w_t)^\intercal(w^*-w_t))^2 + 4\alpha_t^2(\zeta_t(w_t))^2 + \alpha_t^4F^4 \\
&- 2\alpha_t\|w^*-w_t\|^2\varsigma\|w^*-w_t\|^2 + 4\alpha_t\|w^*-w_t\|^2\zeta_t(w_t) + 2\alpha_t^2F^2\|w^*-w_t\|^2 \\
&+ 4\alpha_t^2\varsigma\|w^*-w_t\|^2|\zeta_t(w_t)| - 2\alpha_t^3F^2\varsigma\|w^*-w_t\|^2 + 4\alpha_t^3F^2\zeta_t(w_t) \\
\leq& (1-2\alpha_t\varsigma)\|w^*-w_t\|^4 + 4\alpha_t^2(\bar{g}(w_t)^\intercal(w^*-w_t))^2 + 4\alpha_t^2(\zeta_t(w_t))^2 + \alpha_t^4F^4 \\
&+ (4\alpha_t + 4\alpha_t^2\varsigma)\|w^*-w_t\|^2|\zeta_t(w_t)| + (2\alpha_t^2F^2 - 2\alpha_t^3F^2\varsigma)\|w^*-w_t\|^2 + 4\alpha_t^3F^2\zeta_t(w_t).
\end{aligned}$$

We can utilize the bounds $\|w\| \leq R$, $|\zeta_t(w)| \leq 2F^2$, $\|g_t(w)\| \leq F$ from Lemma D.1 to arrive at

$$\begin{aligned}
\|w^*-w_{t+1}\|^4 \leq& (1-2\alpha_t\varsigma)\|w^*-w_t\|^4 + 16\alpha_t^2F^2R^2 + 16\alpha_t^2F^4 + \alpha_t^4F^4 \\
&+ (4\alpha_t + 4\alpha_t^2\varsigma)\|w^*-w_t\|^2|\zeta_t(w_t)| + (2\alpha_t^2F^2 - 2\alpha_t^3F^2\varsigma)4R^2 + 8\alpha_t^3F^4 \\
\leq& (1-2\alpha_t\varsigma)\|w^*-w_t\|^4 + (8\alpha_t + 8\alpha_t^2\varsigma)R^2|\zeta_t(w_t)| + 24\alpha_t^2F^2R^2 + 16\alpha_t^2F^4 \\
&+ \alpha_t^4F^4 - 8\alpha_t^3F^2\varsigma R^2 + 8\alpha_t^3F^4.
\end{aligned}$$

After some rearrangement of terms, we have

$$\begin{aligned}\alpha_t\varsigma\|w^*-w_t\|^4 \leq & (1-\alpha_t\varsigma)\|w^*-w_t\|^4 - \|w^*-w_{t+1}\|^4 + (8\alpha_t+8\alpha_t^2\varsigma)R^2|\zeta_t(w_t)| \\ & + 24\alpha_t^2F^2R^2 + 16\alpha_t^2F^4 + \alpha_t^4F^4 - 8\alpha_t^3F^2\varsigma R^2 + 8\alpha_t^3F^4 \\ \mathbb{E}[\|w^*-w_t\|^4] \leq & (\frac{1}{\alpha_t\varsigma}-1)\mathbb{E}[\|w^*-w_t\|^4] - \frac{1}{\alpha_t\varsigma}\mathbb{E}[\|w^*-w_{t+1}\|^4] + (\frac{8}{\varsigma}+8\alpha_t)R^2\mathbb{E}[|\zeta_t(w_t)|] \\ & + \frac{24\alpha_tF^2R^2}{\varsigma} + \frac{16\alpha_tF^4}{\varsigma} + \frac{\alpha_t^3F^4}{\varsigma} - 8\alpha_t^2F^2R^2 + \frac{8\alpha_t^2F^4}{\varsigma}.\end{aligned}$$

Let $\alpha_t = \frac{1}{(t+1)\varsigma}$, then we have

$$\begin{aligned}\mathbb{E}[\|w^*-w_t\|^4] \leq & t\mathbb{E}[\|w^*-w_t\|^4] - (t+1)\mathbb{E}[\|w^*-w_{t+1}\|^4] + (\frac{8}{\varsigma}+\frac{8}{(t+1)\varsigma})R^2\mathbb{E}[|\zeta_t(w_t)|] \\ & + \frac{24F^2R^2}{\varsigma^2(t+1)} + \frac{16F^4}{\varsigma^2(t+1)} + \frac{F^4}{\varsigma^4(t+1)^3} - \frac{8F^2R^2}{(t+1)^2\varsigma^2} + \frac{8F^4}{\varsigma^3(t+1)^2}.\end{aligned}$$

Then we sum on either side from 0 to $K-1$ and divide by K, using the facts that $\sum_{t=1}^K \frac{1}{t^2} \leq \frac{\pi^2}{6}$ and $\sum_{t=1}^K \frac{1}{t} = \log(K)+O(1)$ to conclude

$$\begin{aligned}\frac{1}{K}\sum_{t=0}^{K-1}\mathbb{E}[\|w^*-w_t\|^4] \leq & \frac{1}{K}\sum_{t=0}^{K-1}[t\mathbb{E}[\|w^*-w_t\|^4] - (t+1)\mathbb{E}[\|w^*-w_{t+1}\|^4]] + \frac{1}{K}\sum_{t=0}^{K-1}(\frac{8}{\varsigma}+\frac{8}{(t+1)\varsigma})R^2\mathbb{E}[|\zeta_t(w_t)|] \\ & + \frac{1}{K}\sum_{t=0}^{K-1}[\frac{24F^2R^2}{\varsigma^2(t+1)} + \frac{16F^4}{\varsigma^2(t+1)} + \frac{F^4}{\varsigma^4(t+1)^2} + \frac{8F^4}{\varsigma^3(t+1)^2}] \\ \leq & \frac{\|w_1-w^*\|^4}{K} + \frac{1}{K}\sum_{t=0}^{K-1}(\frac{8}{\varsigma}+\frac{8}{(t+1)\varsigma})R^2\mathbb{E}[|\zeta_t(w_t)|] + \frac{1}{K}\sum_{t=0}^{K-1}[\frac{24F^2R^2}{\varsigma^2(t+1)} + \frac{16F^4}{\varsigma^2(t+1)}] \\ & + \frac{4F^4\pi^2}{3\varsigma^3K} + \frac{F^4\pi^2}{6\varsigma^4K} \\ \leq & \frac{\|w_1-w^*\|^4}{K} + \frac{1}{K}\sum_{t=0}^{K-1}\frac{16R^2}{\varsigma}\mathbb{E}[|\zeta_t(w_t)|] + (\frac{24F^2R^2+16F^4}{\varsigma^2})\frac{\log K+O(1)}{K} \\ & + \frac{4F^4\pi^2}{3\varsigma^3K} + \frac{F^4\pi^2}{6\varsigma^4K}.\end{aligned}$$

To bound the summation on the right hand side, we can apply the results of Lemma D.3 for $t \leq K-1$, with

$\tau_0 = \tau^{\text{mix}}(\alpha_{T-1}) \leq \frac{\log(m\varsigma T)}{\log(r^{-1})}$. Then we have

$$\begin{aligned}
\sum_{t=0}^{K-1} \mathbb{E}[|\zeta_t(w_t)|] &\leq F^2(8+6\tau_0)\sum_{t=0}^{K-1}\alpha_{t^*} + 10F^2 m\sum_{t=0}^{K-1} r^t \\
&= F^2(8+6\tau_0)\sum_{t=0}^{\tau_0}\alpha_0 + F^2(8+6\tau_0)\sum_{t=\tau_0+1}^{K-1}\alpha_t + 10F^2 m\sum_{t=0}^{K-1} r^t \\
&\leq F^2(8+6\tau_0)\sum_{t=0}^{\tau_0}\alpha_0 + F^2(8+6\tau_0)\sum_{t=\tau_0+1}^{K-1}\alpha_t + \frac{10F^2 m}{1-r} \\
&= F^2(8+6\tau_0)\tau_0\alpha_0 + F^2(8+6\tau_0)\sum_{t=\tau_0+1}^{K-1}\alpha_t + \frac{10F^2 m}{1-r} \\
&\leq \frac{8F^2}{\varsigma}\frac{\log(m\varsigma K)}{\log(r^{-1})} + \frac{6F^2}{\varsigma}\frac{\log^2(m\varsigma K)}{\log^2(r^{-1})} + \frac{F^2}{\varsigma}(8+6\frac{\log(m\varsigma K)}{\log(r^{-1})})(\log K + O(1)) + \frac{10F^2 m}{1-r} \\
&\leq (\frac{6F^2}{\varsigma\log^2(r^{-1})} + \frac{6F^2}{\varsigma\log(r^{-1})})\log^2 K + O(\log(K)) + O(1) \\
&\leq \frac{12F^2}{\varsigma\log^2(r^{-1})}\log^2 K + O(\log(K)) + O(1).
\end{aligned}$$

So we have for the original expression

$$\frac{1}{K}\sum_{t=0}^{K-1}\mathbb{E}[\|w^* - w_t\|^4] \leq (\frac{192F^2R^2}{\varsigma^2\log^2(r^{-1})} + O(\frac{1}{\log K}) + O(\frac{1}{\log^2 K}))\cdot\frac{\log^2 K}{K},$$

and by Jensen's inequality, we obtain

$$\mathbb{E}[\|w^* - \bar{w}_K\|^4] \leq \frac{1}{K}\sum_{t=0}^{K-1}\mathbb{E}[\|w^* - w_t\|^4] \leq (\frac{192F^2R^2}{\varsigma^2\log^2(r^{-1})} + O(\frac{1}{\log K}) + O(\frac{1}{\log^2 K}))\cdot\frac{\log^2 K}{K}.$$

□

E.2.2. PROOF OF LEMMA E.4

Proof. We establish a bound on $\mathbb{E}[\|w^* - \bar{w}_K\|^4]$ via Lemma E.3. Let $\epsilon = \mu^2$. Then we want to find K large enough such that

$$\log(K) \leq \epsilon\sqrt{K}$$

We look for the asymptotic solution to

$$\log(K) = \epsilon\sqrt{K}$$

in the form $K = \frac{k_1}{\epsilon^2}$. Plugging this in, we get

$$\log(k_1) + \log(\frac{1}{\epsilon^2}) = \sqrt{k_1}$$

By the method of dominant balance, we have

$$\begin{aligned}
\sqrt{k_1} &\approx \log(\frac{1}{\epsilon^2}) \\
k_1 &\approx \log^2(\frac{1}{\epsilon^2})
\end{aligned}$$

So we have $K = O(\frac{\log^2(\frac{1}{\epsilon^2})}{\epsilon^2})$.

F. Policy Gradient Theorem

The original policy gradient theorem derived in (Sutton et al., 1999) addresses the gradient of the value function, which is a slightly different objective that we denote as $\tilde{J}(\theta)$ as follows

$$\tilde{J}(\theta) = V^{\pi_\theta}(s_0) = \mathbb{E}_{\pi_\theta}[\sum_{k=0}^{\infty} \gamma^k \mathcal{R}(s_k, a_k)|s_0]$$

Then from (Sutton et al., 1999) we have the original policy gradient theorem:

$$\nabla \tilde{J}(\theta) = \sum_{s\in\mathcal{S}} d^{\pi}(s) \sum_{a\in\mathcal{A}} \nabla \pi_\theta(s,a) Q^{\pi_\theta}(s,a)$$

Where

$$d^{\pi}(s) = \sum_{k=0}^{\infty} \gamma^k Pr(s_k = s|s_0, \pi)$$

We consider instead the expectation of the value function over an initial state distribution:

$$J(\theta) = \mathbb{E}_{\rho_0}[V^{\pi_\theta}(s_0)]$$

Then

$$\nabla J(\theta) = \sum_{s\in\mathcal{S}} d^{\pi_\theta,\rho_0}(s) \sum_{a\in\mathcal{A}} \nabla \pi_\theta(a|s) Q^{\pi}(s,a)$$

Where

$$d^{\pi_\theta,\rho_0}(s) = \sum_{k=0}^{\infty} \gamma^k \mathbb{E}_{\rho_0}[Pr(s_k = s|s_0, \pi_\theta)]$$

When we implement the "log-likelihood trick", we have

$$\nabla J(\theta) = \sum_{s} d^{\pi_\theta,\rho_0} \sum_{a} \pi_\theta(a|s) \nabla \log \pi_\theta(a|s) Q^{\pi_\theta}(s,a)$$

We can "unroll" this result as in (Wu et al., 2022) to acquire the temporal formulation:

$$= \mathbb{E}_{\pi,\rho_0}[\sum_{k=0}^{\infty} \gamma^k \nabla \log \pi(a_k|s_k) Q^{\pi}(s_k, a_k)]$$

To derive the GPOMDP estimator from this result, we use the definition of Q^π

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t)|s_0 = s, a_0 = a] = \mathbb{E}_{\pi}[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_{t+k}, a_{t+k})|s_k = s, a_k = a]$$

$$\begin{aligned}
\nabla J(\theta) &= \mathbb{E}_{\pi,\rho_0}[\sum_{k=0}^{\infty} \gamma^k \nabla \log \pi(a_k|s_k) \mathbb{E}_{\pi}[\sum_{t=0}^{\infty} \gamma^t R(s'_{t+k}, a'_{t+k})|s'_k = s_k, a'_k = a_k]] \\
&= \mathbb{E}_{\pi,\rho_0}[\sum_{k=0}^{\infty} \mathbb{E}_{\pi}[\sum_{t=0}^{\infty} \gamma^{k+t} \nabla \log \pi(a_k|s_k) \mathcal{R}(s'_{t+k}, a'_{t+k})|s'_k = s_k, a'_k = a_k]] \\
&= \mathbb{E}_{\pi,\rho_0}[\sum_{k=0}^{\infty} \sum_{t=0}^{\infty} \gamma^{k+t} \nabla \log \pi(a_k|s_k) \mathcal{R}(s_{t+k}, a_{t+k})] \\
&= \mathbb{E}_{\pi,\rho_0}[\sum_{k=0}^{\infty} \sum_{t=k}^{\infty} \gamma^{t} \nabla \log \pi(a_k|s_k) \mathcal{R}(s_t, a_t)]
\end{aligned}$$

And so we end with an unbiased estimator of the policy gradient

$$\nabla J(\theta) = \mathbb{E}_{\tau\sim p(\cdot|\theta)}[\nabla \log p_\theta(\tau) \mathcal{R}(\tau)].$$