# Off-Policy Average Reward Actor-Critic with Deterministic Policy Search

**Naman Saxena** [1] **Subhojyoti Khastigir** [1] **Shishir Kolathaya** [1 2] **Shalabh Bhatnagar** [1]

## Abstract

The average reward criterion is relatively less studied as most existing works in the Reinforcement Learning literature consider the discounted reward criterion. There are few recent works that present on-policy average reward actor-critic algorithms, but average reward off-policy actor-critic is relatively less explored. In this work, we present both on-policy and off-policy deterministic policy gradient theorems for the average reward performance criterion. Using these theorems, we also present an Average Reward Off-Policy Deep Deterministic Policy Gradient (ARO-DDPG) Algorithm. We first show asymptotic convergence analysis using the ODE-based method. Subsequently, we provide a finite time analysis of the resulting stochastic approximation scheme with linear function approximator and obtain an $\epsilon$-optimal stationary policy with a sample complexity of $\Omega(\epsilon^{-2.5})$. We compare the average reward performance of our proposed ARO-DDPG algorithm and observe better empirical performance compared to state-of-the-art on-policy average reward actor-critic algorithms over MuJoCo-based environments.

## 1. Introduction

The reinforcement learning (RL) paradigm has shown significant promise for finding solutions to decision making problems that rely on a reward-based feedback from the environment. Here one is mostly concerned with the long-term reward acquired by the algorithm. In the case of infinite horizon problems, the discounted reward criterion has largely been studied because of its simplicity. Major recent development in the context of RL in continuous state-action spaces has considered the discounted reward criterion (Schulman et al., 2015; 2017; Lillicrap et al., 2016; Haarnoja et al., 2018). However, there are very few works which focus on the average reward performance criterion in the continuous state-action setting (Zhang & Ross, 2021; Ma et al., 2021).

The average reward criterion has started receiving attention in recent times and there are papers that discuss the benefits of using this criterion over the discounted reward (Dewanto & Gallagher, 2021; Naik et al., 2019). One of the reasons being, average reward criteria only considers recurrent states and it happens to be the most selective optimization criterion in recurrent Markov Decision Processes (MDPs) according to n-discount optimality criterion. Please refer Mahadevan (1996) for more details on n-discount optimality criterion. Further, optimization in average reward setting is not dependent on the initial state distribution. Moreover, the discrepancy between the objective function and the evaluation metric, that exists for discounted reward setting, is resolved by opting for the average reward criterion. We encourage the readers to go through Dewanto & Gallagher (2021); Naik et al. (2019) for better understanding of the benefits mentioned.

There are very few algorithms in literature that optimize the average reward and all of them happen to be on-policy algorithms (Zhang & Ross, 2021; Ma et al., 2021). It has been demonstrated several times that on-policy algorithms are less sample efficient than off-policy algorithms (Lillicrap et al., 2016; Haarnoja et al., 2018; Fujimoto et al., 2018) for the discounted reward criterion. In this paper we try to find whether the same is true for the average reward criterion. We try to overcome the research gap in development of off-policy average reward algorithms for continuous state and action spaces by proposing an Average Reward Off-Policy Deep Deterministic Policy Gradient (ARO-DDPG) Algorithm.

The policy evaluation step in the case of the average reward algorithm is equivalent to finding the solution to the Poisson equation (i.e., the Bellman equation for a given policy). Poisson equation, because of its form, does not admit a unique solution but only solutions that are unique up to a constant term. Further, the policy evaluation step in this

[1]Department of Computer Science and Automation, Indian Institute of Science, Bangalore, India [2]Robert Bosch Centre for Cyber-Physical Systems, Indian Institute of Science, Bangalore, India. Correspondence to: Naman Saxena <namansaxena@iisc.ac.in>.

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

case consists of finding not just the Differential Q-value function but also the average reward. Thus, because of the required estimation of two quantities instead of one, the role of the optimization algorithm and the target network increases here. Therefore we implement the proposed ARO-DDPG algorithm by using target network and by carefully selecting the optimization algorithm.

The following are the broad contributions of our paper:

- We provide both on-policy and off-policy deterministic policy gradient theorems for the average reward performance metric.
- We present our Average Reward Off-Policy Deep Deterministic Policy Gradient (ARO-DDPG) algorithm.
- We show a comparison of our algorithm on several environments with other state-of-the-art average reward algorithms in the literature.
- We perform asymptotic convergence analysis using ODE-based method and also provide a finite time analysis of our three timescale stochastic approximation based actor-critic algorithm using a linear function approximator.

Silver et al. (2014); Lillicrap et al. (2016); Xiong et al. (2022), individually address one of the aspects of discounted reward performance criteria for deterministic policies such as policy gradient theorem, implementation of practical algorithm and convergence analysis. In this paper we provide a comprehensive treatment of average reward performance criteria for deterministic policies by covering policy gradient theorem, implementation of practical algorithm and convergence analysis. The rest of the paper is structured as follows: In Section 2, we present the preliminaries on the MDP framework, the basic setting as well as the policy gradient algorithm. Section 3 presents the deterministic policy gradient theorem and our proposed ARO-DDPG algorithm. Section 4 then presents the main theoretical results related to the convergence analysis. Section 5 presents the experimental results. In Section 6, we discuss other related work and Section 7 presents the conclusions. The detailed proofs for the convergence analysis are available in the Appendix.

## 2. Preliminaries

Consider a Markov Decision Process (MDP) $M = \{S, A, R, P, \pi\}$ where $S \subset \mathbb{R}^n$ is the (continuous) state space, $A \subset \mathbb{R}^m$ is the (continuous) action space, $R : S \times A \mapsto \mathbb{R}$ denotes the reward function with $R(s, a)$ being the reward obtained under state $s$ and action $a$. Further, $P(\cdot|s, a)$ denotes the state transition function defined as $P : S \times A \mapsto \mu(\cdot)$, where $\mu : \mathcal{B}(S) \mapsto [0, 1]$ is a probability measure. Deterministic policy $\pi$ is defined as $\pi : S \mapsto A$.In the above, $\mathcal{B}(S)$ represents the Borel sigma algebra on $S$. Stochastic policy $\pi_r$ is defined as $\pi_r : S \mapsto \mu'(\cdot)$, where $\mu' : \mathcal{B}(A) \mapsto [0, 1]$ and $\mathcal{B}(A)$ is the Borel sigma algebra on $A$.

**Assumption 2.1.** The Markov process obtained under any policy $\pi$ is ergodic.

Assumption 2.1 is necessary to ensure existence of steady state distribution of Markov process.

### 2.1. Discounted Reward MDPs

In discounted reward MDPs, discounting is controlled by $\gamma \in (0, 1)$. The following performance metric is optimized with respect to the policy:

$$\eta(\pi) = \mathbb{E}^{\pi}[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)] = \int_S \rho_0(s) V^{\pi}(s)\, ds. \quad (1)$$

Here, $\rho_0$ is the initial state distribution and $V^{\pi}$ is the value function. $V^{\pi}(s)$ denotes the long term reward acquired when starting in the state $s$.

$$V^{\pi}(s_t) = \mathbb{E}^{\pi}\Big[R(s_t, a_t) + \gamma V^{\pi}(s_{t+1})|s_t\Big]. \quad (2)$$

### 2.2. Average reward MDPs

The performance metric in the case of average reward MDPs is the long-run average reward $\rho(\pi)$ defined as follows:

$$\rho(\pi) = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}^{\pi}\Big[\sum_{t=0}^{N-1} R(s_t, a_t)\Big] = \int_S d^{\pi}(s) R^{\pi}(s)\, ds, \quad (3)$$

where $R^{\pi}(s) \triangleq R(s, \pi(s))$. The limit in the first equality in (3) exists because of Assumption 2.1. The quantity $d^{\pi}(s)$ in the second equality in (3) corresponds to the steady state probability of the Markov process being in state $s \in S$ and it exists and is unique given $\pi$ from Assumption 2.1 as well.

$V^{\pi}_{diff}$ is the differential value function corresponding to the policy $\pi$ and is defined in (4). Further, the differential Q-value or action-value function $Q^{\pi}_{diff}$ is defined in (5).

$$V^{\pi}_{diff}(s_t) = \mathbb{E}^{\pi}[\sum_{i=t}^{\infty} R(s_i, a_i) - \rho(\pi)|s_t]. \quad (4)$$

$$Q^{\pi}_{diff}(s_t, a_t) = \mathbb{E}^{\pi}[\sum_{i=t}^{\infty} R(s_i, a_i) - \rho(\pi)|s_t, a_t]. \quad (5)$$

**Lemma 2.2.** *There exists a unique constant $k(= \rho(\pi))$ which satisfies the following equation for differential value function $V_{diff}$ :*

$$V^{\pi}_{diff}(s_t) = \mathbb{E}^{\pi}[R(s_t, a_t) - k + V^{\pi}_{diff}(s_{t+1})|s_t] \quad (6)$$

*Proof.* See Lemma A.9 in the appendix for the proof. □

### 2.3. Policy Gradient Theorem

Unlike in Q-learning where we try to find the optimal Q-value function and then infer the policy from it, the policy gradient theorem (Sutton et al., 1999; Silver et al., 2014; Degris et al., 2012) allows us to directly optimize the performance metric via its gradient with respect to the policy parameters. Q-learning can be visualized to be a value iteration scheme while an algorithm based on the policy gradient theorem can be seen as mimicking policy iteration. Sutton et al. (1999) provided the policy gradient theorem for on-policy optimization of both the discounted reward and the average reward algorithms, see (7)-(8), respectively.

$$\nabla_\theta \eta(\pi) = \int_S \omega^\pi(s) \int_A \nabla_\theta \pi_r(a|s,\theta) Q^{\pi_r}(s,a)\, da\, ds. \tag{7}$$

$$\nabla_\theta \rho(\pi) = \int_S d^\pi(s) \int_A \nabla_\theta \pi_r(a|s,\theta) Q^{\pi_r}_{diff}(s,a)\, da\, ds. \tag{8}$$

In (7) $\omega^\pi$ denotes the long term discounted state visitation probability density which is defined in (9) while $d^\pi(s) = \lim_{t\to\infty} P^\pi_t(s)$ is the steady state probability density on states. $P^\pi$ denotes the transition probability kernel for the Markov chain induced by policy $\pi$ and $P^\pi_t$ is the state distribution at instant $t$ given by (10).

$$\omega^\pi(s) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t P^\pi_t(s). \tag{9}$$

$$P^\pi_t(s) = \int_{S\times S\ldots} \rho_0(s_0) \prod_{k=0}^{t-1} P^\pi(s_{k+1}|s_k)\, ds_0 \ldots\, ds_{t-1}. \tag{10}$$

The policy gradient theorem in Sutton et al. (1999) is only valid for on-policy algorithms. Degris et al. (2012) proposed an approximate off-policy policy gradient theorem for stochastic policies, see (11), where $d^\mu$ stands for the steady state density function corresponding to the policy $\mu$.

$$\nabla_\theta \eta(\pi) \approx \int_S d^\mu(s) \int_A \nabla_\theta \pi_r(a|s,\theta) Q^\pi(s,a)\, da\, ds. \tag{11}$$

Silver et al. (2014) came up with the deterministic policy gradient theorem for discounted reward setting, see (12), which eventually led to the development of very successful Deep Deterministic Policy Gradient (DDPG) (Lillicrap et al., 2016) algorithm and Twin Delayed DDPG (TD3) algorithm (Fujimoto et al., 2018). In the next section we show how we extend the same development for average reward criterion.

$$\nabla_\theta \eta(\pi) = \int_S \omega^\pi(s) \nabla_a Q^\pi(s,a)|_{a=\pi(s)} \nabla_\theta \pi(s,\theta)\, ds. \tag{12}$$

## 3. Proposed Average Reward Algorithm

We now propose the deterministic policy gradient theorem for the average reward criterion. The policy gradient estimator has to be derived separately for both the on-policy and off-policy settings. Obtaining the on-policy deterministic policy gradient estimator is straight forward but dealing with the off-policy gradient estimates involves an approximate gradient (Degris et al., 2012).

### 3.1. On-Policy Policy Gradient Theorem

We cannot directly use the second equality of (3) to derive the policy gradient theorem because of the inability to take the derivative of steady state density function. Therefore one needs to use Lemma 2.2 to obtain the average reward deterministic policy gradient theorem.

**Theorem 3.1.** *The gradient of $\rho(\pi)$ with respect to policy parameter $\theta$ is given as follows:*

$$\nabla_\theta \rho(\pi) = \int_S d^\pi(s) \nabla_a Q^\pi_{diff}(s,a)|_{a=\pi(s)} \nabla_\theta \pi(s,\theta)\, ds. \tag{13}$$

*Proof.* See Theorem A.10 in the appendix for the proof. □

### 3.2. Compatible Function Approximation

The result in this section is mostly inspired from Silver et al. (2014). Recall that $Q^\pi_{diff}(s,a)$ is the 'true' differential $Q$-value of the state-action tuple $(s,a)$ under the parameterized policy $\pi$. Now let $Q^w_{diff}(s,a)$ denote the approximate differential $Q$-value of the $(s,a)$-tuple when function approximation with parameter $w$ is used. Lemma 3.2 says that when the function approximator satisfies a compatibility condition (cf. (14,15)), then the gradient expression in (13) is also satisfied by $Q^w_{diff}$ in place of $Q^\pi_{diff}$.

**Lemma 3.2.** *Assume that the differential Q-value function (5) satisfies the following:*

1. $$\nabla_w \nabla_a Q^w_{diff}(s,a)|_{a=\pi(s)} = \nabla_\theta \pi(s,\theta). \tag{14}$$

2. *Differential Q-value function parameter $w = w^*_\epsilon$ optimizes the following error function:*

$$\zeta(\theta,w) = \frac{1}{2}\int_S d^\pi(s) \| \nabla_a Q^\pi_{diff}(s,a)|_{a=\pi(s)} - \nabla_a Q^w_{diff}(s,a)|_{a=\pi(s)} \|^2\, ds. \tag{15}$$

*Then,*

$$\int_S d^\pi(s) \nabla_a Q^\pi_{diff}(s,a)|_{a=\pi(s)} \nabla_\theta \pi(s,\theta)\, ds = \int_S d^\pi(s) \nabla_a Q^w_{diff}(s,a)|_{a=\pi(s)} \nabla_\theta \pi(s,\theta)\, ds. \tag{16}$$

*Further, in the case when a linear function approximator is used, we obtain*

$$\nabla_a Q^w_{diff}(s,a)|_{a=\pi(s)} = \nabla_\theta \pi(s,\theta)^\intercal w. \tag{17}$$

*Proof.* See Lemma A.11 in the appendix for a proof. □

An important implication of Lemma 3.2 also is that the dimension of the matrix on the left hand side and the right hand side of (14) should be the same. Hence the dimensions of the parameters $\theta$ (used in the parameterized policy) and $w$ (used to approximate the differential Q-value function) are the same. Lemma 3.2 shows that the compatible function approximation theorem has the same form in the average reward setting as the discounted reward setting.

### 3.3. Off-Policy Policy gradient theorem

In order to derive off-policy policy gradient theorem it is not possible to use the direction adopted by Degris et al. (2012) for off-policy stochastic policy gradient theorem for the discounted reward setting. We first mention our proposed approximate off-policy deterministic policy gradient theorem and then explain why some alternatives would not have worked.

**Assumption 3.3.** For the Markov chain obtained from the policy $\pi$, let $K(\cdot|\cdot)$ be the transition kernel and $S^\pi$ the steady state measure. Then there exists $a > 0$ and $\kappa \in (0,1)$ such that

$$D_{TV}(K^t(\cdot|s), S^\pi(\cdot)) \leq a\kappa^t, \forall t, \forall s \in S.$$

Assumption 3.3 states that Markov chain generated by a policy $\pi$ follows uniform ergodicity property. This assumption is necessary to get an upper bound on the total variation norm of steady state probability distribution of two policies. Further this assumption allows for fast mixing of markov chain and i.i.d sampling of transitions from buffer for convergence analysis purpose.

**Theorem 3.4.** *The approximate gradient ($\widehat{\nabla_\theta \rho}(\pi)$) of the average reward $\rho(\pi)$ with respect to the policy parameter $\theta$ is given by the following expression:*

$$\widehat{\nabla_\theta \rho}(\pi) = \int_S d^\mu(s) \nabla_a Q^\pi_{diff}(s,a)|_{a=\pi(s)} \nabla_\theta \pi(s,\theta)\, ds. \tag{18}$$

*Further, the approximation error is $\mathcal{E}(\pi,\mu) = \|\nabla_\theta \rho(\pi) - \widehat{\nabla_\theta \rho}(\pi)\|$, where $\mu$ represents the behaviour policy with parameter $\theta^\mu$ and $\nabla_\theta \rho(\pi)$ is the on-policy policy gradient from Theorem 3.1. $\mathcal{E}$ satisfies*

$$\mathcal{E}(\pi,\mu) \leq Z\|\theta - \theta^\mu\|, \tag{19}$$

*where, $Z = 2^{n+1}C(\lceil \log_\kappa a^{-1} \rceil + 1/\kappa)L_t$ with $L_t$ being the Lipchitz constant for the transition probability density function (Assumption A.1). Constants $a$ and $\kappa$ are from Assumption 3.3, $n$ is the dimension of the state space, and $C = \max_s \|\nabla_a Q^\pi_{diff}(s,a)|_{a=\pi(s)} \nabla_\theta \pi(s,\theta)\|$.*

*Proof.* See Theorem A.12 in the appendix for a proof. □

Theorem 3.4 suggests that the approximation error in the gradient increases as the difference between the target policy $\pi$ and the behaviour policy $\mu$ increases.

### 3.4. Off-Policy Alternatives

In this section we will talk about what alternatives could be thought of in place of what is suggested in Section 3.3 and why those alternatives would not work.

1. One can possibly take inspiration from Degris et al. (2012) and define an objective function, $\rho_{new}(\pi)$, as in (20), which is a naive off-policy version of (3).

$$\rho_{new}(\pi) = \int_S d^\mu(s) R^\pi(s)\, ds. \tag{20}$$

If, however, we take the derivative of $\rho_{new}(\pi)$ defined above, we get the policy update rule as in (21).

$$\nabla_\theta \rho_{new}(\pi) = \int_S d^\mu(s) \nabla_a R(s,a)|_{a=\pi(s)} \nabla_\theta \pi(s,\theta)\, ds. \tag{21}$$

The update rule (21) only considers the reward function and not the transition dynamics of the MDP. In (A.2), the derivative of the objective function includes the differential Q-value function which encapsulates both the information of the reward function and the transition dynamics of the MDP and hence is valid derivative.

2. A lot of work in the off-policy setting relies on importance sampling ratios. Recently a few works devised a method to estimate the steady state probability density ratio of the target and behavior policies (Zhang et al., 2020a;b; Liu et al., 2018; Nachum et al., 2019). The ratio of steady state densities could be used for deterministic policy optimization but there are certain issues which prohibit its usage, see (22).

$$\nabla_\theta \rho(\pi) = \int_S d^\mu(s) \tau(s) \nabla_a Q^\pi_{diff}(s,a)|_{a=\pi(s)} \nabla_\theta \pi(s,\theta)\, ds. \tag{22}$$

Here, $\tau(s)$ is the steady state probability density ratio defined as $d^\pi(s)/d^\mu(s)$. In order to calculate $\tau(s)$ we need information about $(\pi(a|s), \mu(a|s)$ and $P(s'|s,a))$. We need the ratio $\pi(a|s)/\mu(a|s)$ and

for deterministic policies the ratio would be $\delta(a - \pi(s)/\delta(a - \mu(s))$, where $\delta(\cdot)$ is the Dirac-Delta function:

$$\frac{\delta(a-\pi(s))}{\delta(a-\mu(s))} = \begin{cases} 0 & \text{if } a = \mu(s) \\ \infty & \text{if } a = \pi(s) \\ \frac{0}{0} & \text{otherwise.} \end{cases} \tag{23}$$

From (23), it is clear that the ratio $\delta(a - \pi(s)/\delta(a - \mu(s))$ will be undefined for almost all actions $a \in A$. Thus, we cannot use this ratio for deterministic policies. Otherwise, we need $P(s'|s, \pi(a))$ and $P(s'|s, \mu(a))$. It is possible to get the information about $P(s'|s, \mu(a))$ by sampling from the Markov process generated by the policy $\mu$ but obtaining this information about $P(s'|s, \pi(a))$ is impossible as in the off-policy setting data from $\pi$ is assumed to be simply unavailable.

### 3.5. Actor-Critic Update rule

**Assumption 3.5.** $\alpha_t, \beta_t$, and $\gamma_t$ are the step sizes for critic, target estimator, and actor parameter updates respectively.

$$\alpha_t = \frac{C_\alpha}{(1+t)^\sigma} \quad \beta_t = \frac{C_\beta}{(1+t)^u} \quad \gamma_t = \frac{C_\gamma}{(1+t)^v}$$

Here, $C_\alpha, C_\beta, C_\gamma > 0$ and $0 < \sigma < u < v < 1$. $\alpha_t$ is at the fastest timescale, $\beta_t$ is at slower timescale and $\gamma_t$ is at the slowest timescale.

The critic and average reward parameters are estimated using the TD(0) update rule but use target estimators. We are using target estimators to ensure stability of the iterates of the algorithm. Let $\{s_i, a_i, s'_i\}_{i=0}^{n-1}$ denote the batch of sampled data from the replay buffer.

$$\xi_t^j = \frac{1}{2}\sum_{i=0}^{n-1}\Big(R(s_i,a_i) - \overline{\rho_t} - Q_{diff}^{w^j}(s_i,a_i) + \min(Q_{diff}^{\overline{w^1}}, Q_{diff}^{\overline{w^2}})(s'_i, \pi(s'_i, \overline{\theta_t}))\Big)^2 \quad j \in \{1,2\} \tag{24}$$

$$\xi_t^3 = \frac{1}{2}\sum_{i=0}^{n-1}\Big(R(s_i,a_i) - \rho_t - \min(Q_{diff}^{\overline{w^1}}, Q_{diff}^{\overline{w^2}})(s_i,a_i) + \min(Q_{diff}^{\overline{w^1}}, Q_{diff}^{\overline{w^2}})(s'_i, \pi(s'_i, \overline{\theta_t}))\Big)^2 \tag{25}$$

Equations (24) and (25) correspond to the Bellman error for the differential Q-value function approximator and the average reward estimator respectively. Note that we are using the double Q-value function approximator. Here $\bar{\rho}_t$ represents the target estimator for average reward at time t, $Q_{diff}^{\overline{w^i}}$ represents the differential Q-value function parameterized by target differential Q-value parameter $\bar{w}_t^i$ and $\bar{\theta}_t$ represents the target parameter for actor at time t, respectively.

$$w_{t+1}^j = w_t^j - \alpha_t \nabla_{w^j} \xi_t^j \quad j \in \{1,2\} \tag{26}$$

$$\rho_{t+1} = \rho_t - \alpha_t \nabla_p \xi_t^3 \tag{27}$$

Our aim is to find the value of parameters for differential Q-value function and average reward estimator such that the Bellman equation is satisfied. Hence, the Bellman error in (24) is used to update the Q-value function approximator parameters $w_t^j$ using (26) and the Bellman error in (25) is used to update the estimator of the average reward $\rho_t$ using (27). Our approach is motivated from the update rule for the differential Q-value function and the average reward parameters given in Wan et al. (2021b) (equations 3 and 4) and Zhang et al. (2021b)(Algorithm 2).

$$\nu_i = \nabla_a min(Q_{diff}^{w^1}, Q_{diff}^{w^2})(s_i,a)|_{a=\pi(s_i)} \nabla_\theta \pi(s_i, \theta_t) \tag{28}$$

$$\theta_{t+1} = \theta_t + \gamma_t \left(\sum_{i=0}^{n-1} \nu_i\right) \tag{29}$$

Actor update is performed using theorem 3.4. Actor parameter, $\theta_t$, is updated using empirical estimate (28) of the gradient in A.2.

$$\overline{w_{t+1}^j} = \overline{w_t^j} + \beta_t(w_{t+1}^j - \overline{w_t^j}) \quad j \in \{1,2\} \tag{30}$$

$$\overline{\rho_{t+1}} = \overline{\rho_t} + \beta_t(\rho_{t+1} - \overline{\rho_t}) \tag{31}$$

$$\overline{\theta_{t+1}} = \overline{\theta_t} + \beta_t(\theta_{t+1} - \overline{\theta_t}) \tag{32}$$

Equation 30-32 are used to update the target Q-value function approximator $\overline{w_t^j}$, target average reward estimator $\overline{\rho_t}$ and target actor parameter $\overline{\theta_t}$.

## 4. Convergence Analysis

In this section we present the asymptotic convergence analysis and finite time analysis of the on-policy and off-policy average reward actor critic algorithm with linear function approximators. First we mention the assumptions taken to perform the convergence analysis followed by the main results.

**Assumption 4.1.** $\phi^\pi(s)\big( = \phi(s, \pi(s))\big)$ denotes the feature vector of state s and satisfies $\|\phi^\pi(s)\| \leq 1$.

The assumption above is just taken for the sake of convenience.

**Assumption 4.2.** The reward function is uniformly bounded, viz., $|R^\pi(s)| \leq C_r < \infty$.

Assumption 4.2 is required to make sure that the average reward objective function is bounded from above.

**Assumption 4.3.** $Q^w_{diff}(s,a)$ is Lipchitz continuous w.r.t to $a$. Thus, $\forall w \quad \|Q^w_{diff}(s,a_1) - Q^w_{diff}(s,a_2)\| \leq L_a\|a_1 - a_2\|$.

Continuity of approximate Q-value function w.r.t action is enforced using Assumption 4.3. Without the continuity property approximate differential Q-values will not generalize for unseen action values.

**Assumption 4.4.** Parameterised policy $\pi(s,\theta)$ is Lipchitz continuous w.r.t $\theta$. Thus, $\|\pi(s,\theta_1) - \pi(s,\theta_2)\| \leq L_\pi\|\theta_1 - \theta_2\|$.

Assumption 4.4 is a common regularity assumption for convergence of actor. It can be found in Wu et al. (2020), Xiong et al. (2022) and Zou et al. (2019).

**Assumption 4.5.** The state feature mapping ($\phi^\pi(s) = \phi(s,\pi(s))$ defined for a policy $\pi$ with parameter $\theta$ is Lipschitz continuous w.r.t $\theta$. Thus, $\max_s \|\phi^{\pi_1}(s) - \phi^{\pi_2}(s)\| \leq L_\phi\|\theta_1 - \theta_2\|$.

Continuity of state action feature w.r.t action is required to ensure generalisation of Q-values to unseen action values. Using this continuity of state action feature with Assumption 4.4 we can satisfy Assumption 4.5.

### 4.1. Asymptotic Convergence

We prove the asymptotic convergence of the three timescale stochastic approximation on-policy algorithm (Algorithm 4) using ODE-based method (Borkar, 2009; Kushner & Clark, 2012; Lakshminarayanan & Bhatnagar, 2017) in two steps. First we keep the policy parameter $\theta$ fixed and prove the convergence of differential Q-value function parameter $w_t$, average reward estimator $\rho_t$, target differential Q-value function parameter $\bar{w}_t$ and target average reward estimator $\bar{\rho}_t$ in Theorem 4.6 (given below).

**Theorem 4.6.** *In Algorithm 4, let policy parameter $\theta_t$ be kept constant at $\theta$. The critic parameter $w_t$ and the target critic parameter $\bar{w}_t$ converges to $w(\theta)^*$. Also, average reward estimator $\rho_t$ and target average reward estimator $\bar{\rho}_t$ converges to $\rho(\theta)^*$.(Note: The point of convergence $w(\theta)^*$ and $\rho(\theta)^*$ are defined in Theorem A.32.)*

*Proof.* See Theorem A.32 in the appendix for the proof. □

Theorem 4.6 uses the two timescale stochastic approximation stability result from Lakshminarayanan & Bhatnagar (2017). Later, taking inspiration from Bhatnagar & Lakshmanan (2012) and invoking the Theorem 5.3.1 of Kushner & Clark (2012) we prove the convergence of policy parameter $\theta_t$ in Theorem 4.7.

**Theorem 4.7.** *$\Gamma_{C_\theta} : \mathbb{R}^d \to C_\theta$ is a projection operator, where $C_\theta$ is compact convex set and $\hat{\Gamma}_{C_\theta}(\theta)\nabla_\theta\rho(\theta)$ refers to directional derivative of $\Gamma_{C_\theta}(\cdot)$ in the direction $\nabla_\theta\rho(\theta)$ at $\theta$. Let $K = \{\theta \in C_\theta | \hat{\Gamma}_{C_\theta}(\theta)\nabla_\theta\rho(\theta) = 0\}$ and $K^\epsilon = \{\theta' \in C_\theta | \exists\ \theta \in K\ \|\theta' - \theta\| < \epsilon\}$. $\forall \epsilon > 0\ \exists \delta$ such that if $\sup_\pi \|e^\pi\| < \delta$ then $\theta_t$ converges to $K^\epsilon$ as $t \to \infty$ with probability one. $e^\pi$ is the function approximation error defined in Lemma A.33.*

*Proof.* See Theorem A.34 in the appendix for the proof. □

Theorem 4.7 essentially argues that the actor update scheme in Algorithm 4 tracks the ODE $\dot{\theta}(t) = \hat{\Gamma}_{C_\theta}(\theta(t))(\nabla_\theta\rho(\theta(t)) + e^{\pi(t)})$ and converges to an $\epsilon-$neighbourhood of the set $K$. Moreover, when $\sup_\pi \|e^\pi\| \to 0$, the actor update scheme tracks $\dot{\theta}(t) = \hat{\Gamma}_{C_\theta}(\theta(t))(\nabla_\theta\rho(\theta(t)))$ and convergence to the set $K$.

Conclusions of Theorem 4.7 will continue to hold for off policy algorithm (Algorithm 5) by suitably setting the value of l2-regularisation coefficient.

### 4.2. Finite Time Analysis

We perform finite time analysis by finding an upper bound on the expected squared norm of policy gradient ($\min_{0\leq t\leq T} \mathbb{E}\|\nabla_\theta\rho(\theta_t)\|^2$) for both Algorithm 2 and 3. We first identify error in the parameters of the algorithm and define dependency graph of error, as shown in Figure 1 for Algorithm 2. In Figure 1, arrow from one error (source) to the other error (destination) shows that an upper bound on the destination error depend on an upper bound on the source error. Exploiting the dependency graph of errors we finally find an upper bound on the expected squared norm of policy gradient ($\min_{0\leq t\leq T} \mathbb{E}\|\nabla_\theta\rho(\theta_t)\|^2$) in term of time $T$.

#### 4.2.1. On-Policy Analysis

**Theorem 4.8.** *The on-policy average reward actor critic algorithm (Algorithm 2) obtains an $\epsilon$-accurate optimal point with sample complexity of $\Omega(\epsilon^{-2.5})$. We obtain*

$$\min_{0\leq t\leq T-1} \mathbb{E}||\nabla_\theta\rho(\theta_t)||^2 = \mathcal{O}\left(\frac{1}{T^{2/5}}\right) + 3C_\pi^4(\tau^2 + \frac{4}{M}C^2_{w^*_\epsilon}),$$
$$\leq \epsilon + \mathcal{O}(1).$$

*Here, $\|\nabla_\theta\pi(s)\| \leq C_\pi$ (Assumption 4.4), $\tau = \max_t \|w^*_t - w^*_{\epsilon,t}\|$, $w^*_\epsilon$ is the optimal critic parameter according to Lemma 3.2. Constant $C_{w^*_\epsilon}$ is defined in Lemma A.28. M is the size of batch of samples used to update parameters.*

*Proof.* See Theorem A.18 in the appendix for the proof. □

We started the analysis with a three timescale stochastic approximation algorithm but later observed that the best sample complexity is achieved when critic parameter and target critic parameters are updated on the same time-scale,

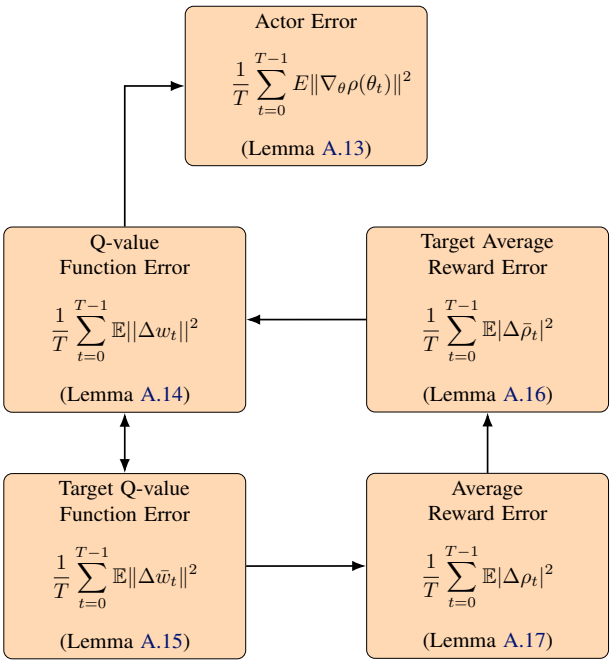


*Figure 1.* Dependency of errors in different types of parameters in Algorithm 2 on one another.

i.e. $u = \sigma$ (Assumption 3.5). The extra terms $3C_\pi^4\tau^2$ and $12C_\pi^4 C_{w_\epsilon^*}^2/M$ exist in the bound established in Theorem 4.8 because of function approximation error and empirical expectation respectively. $3C_\pi^4\tau^2$ can be reduced using high capacity function approximator such as neural network. $12C_\pi^4 C_{w_\epsilon^*}^2/M$ can be made small by increasing the size of the batch $M$ used for empirical expectation. The same error terms are also present in the finite time analysis by Xiong et al.. Let $K_2 = \{\theta \mid \nabla_\theta \rho(\theta) = 0\}$ and $K_2^\epsilon = \{\theta' | \exists\ \theta \in K_2\ \|\theta' - \theta\| < \epsilon\}$. $\forall \epsilon > 0\ \exists \delta$ such that if $|3C_\pi^4(\tau^2 + \frac{4}{M}C_{w_\epsilon^*}^2)| < \delta$ then $\theta_t$ converges to $K_2^\epsilon$ with rate $\mathcal{O}(T^{-2/5})$.

#### 4.2.2. Off-Policy Analysis

**Theorem 4.9.** *The off-policy average reward actor critic algorithm (Algorithm 3) with behavior policy $\mu$ obtains an $\epsilon$-accurate optimal point with sample complexity of $\Omega(\epsilon^{-2.5})$. Here $\theta_\mu$ refers to the behavior policy parameter and $\theta_t$ refers to the target or current policy parameter. We obtain*

$$\begin{aligned}
&\min_{0\le t\le T-1} \mathbb{E}\|\widehat{\nabla_\theta \rho(\theta_t)}\|^2 \\
&= \mathcal{O}\left(\frac{1}{T^{2/5}}\right) + 3C_\pi^4(\tau^2 + \frac{4}{M}C_{w_\epsilon^*}^2) + \mathcal{O}(W_\theta^2) \\
&\le \epsilon + 3C_\pi^4(\tau^2 + \frac{4}{M}C_{w_\epsilon^*}^2) + \mathcal{O}(W_\theta^2)
\end{aligned}$$

*where* $W_\theta := \sup_t \|\theta_\mu - \theta_t\|$.

*Here,* $\|\nabla_\theta \pi(s)\| \le C_\pi$ *(Assumption 4.4),* $\tau = \max_t \|w_t^* - w_{\epsilon,t}^*\|$, $w_\epsilon^*$ *is the optimal critic parameter according to Lemma 3.2. Constant* $C_{w_\epsilon^*}$ *is defined in Lemma A.28. M is the size of batch of samples used to update parameters.*

*Proof.* See Theorem A.20 the appendix for a proof. □

Here also we found that two timescale stochastic approximation algorithm has better sample complexity than three timescale version. We have the same extra term in the bound as established in Theorem 4.8 with an additional term of $\mathcal{O}(W_\theta^2)$. The extra term $\mathcal{O}(W_\theta^2)$ denotes the error induced because of not using the samples from the current policy for performing updates. $W_\theta^2$ will be small when replay buffer is used because replay buffer contains data from policies similar to the current policy. This explains why policy gradient theorem in Theorem 3.4 can be used with replay buffer. Let $K_3 = \{\theta \mid \widehat{\nabla_\theta \rho}(\theta) = 0\}$ and $K_3^\epsilon = \{\theta' | \exists\ \theta \in K_3\ \|\theta' - \theta\| < \epsilon\}$. $\forall \epsilon > 0\ \exists \delta$ such that if $|3C_\pi^4(\tau^2 + \frac{4}{M}C_{w_\epsilon^*}^2) + ZW_\theta^2| < \delta$ then $\theta_t$ converges to $K_3^\epsilon$ with rate $\mathcal{O}(T^{-2/5})$.

## 5. Experimental Results

We conducted experiments on six different environments using the DeepMind control suite (Tassa et al., 2018) and found the performance of ARO-DDPG [1] to be superior than the other algorithms (Figure 2). All the environments selected are infinite horizon tasks. Maximum reward per time step is 1. None of the tasks have a goal reaching nature. We performed all the experiments using 10 different seeds. We show here performance comparisons with two state-of-the-art algorithms: the Average Reward TRPO (ATRPO) (Zhang & Ross, 2021) and the Average Policy Optimization (APO) (Ma et al., 2021) respectively. In general for the average reward performance, not many algorithms are available in the literature. We implemented the ATRPO algorithm using the instructions available in the original paper. We performed hyperparameter tuning and found the original hyper-parameters suggested by the author for ATRPO are the best.

For our proposed algorithm we trained the agent for 1 million time steps and evaluated the agent after every 5,000 time steps in the concerned environment. The length of each episode for the training phase was taken to be 1,000 and for the evaluation phase it was taken to be 10,000. The reason for taking longer episode length for evaluation phase was to compare the long term average reward performance of the algorithms. We also tried using episode length of 10,000 for training phase and found that to be giving poor average

[1] Pytorch implementation of ARO-DDPG could be found at this URL: https://github.com/namansaxena9/ARO-DDPG

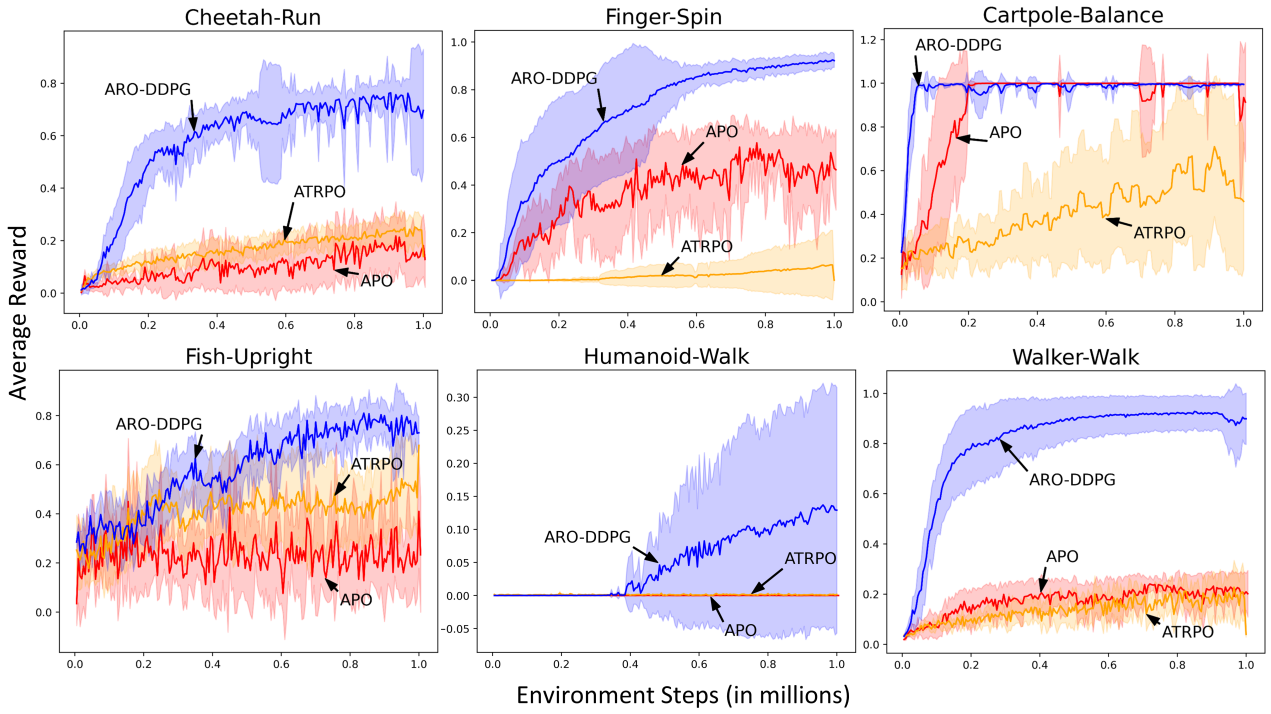


*Figure 2.* Comparison of performance of different average reward algorithms

reward performance. We do not reset the agent if before completing 10,000 steps, it lands in a state from where it is unable to escape of its own. The agent continues to get a reward of zero by default for the remaining length of the episode. That way the cost of failure is high. While training we updated the actor after performing a fixed number of environment steps. We updated the critic neural network with more frequency as compared to the actor neural network. We used target actor and critic networks along with target estimator of the average reward parameter for stability while using bootstrapping updates. We updated the target network using polyak averaging. We tried to enforce multiple timescales in our algorithm by using different update frequency for actor, critic and polyak averaging for target networks. We also borrowed the double Q-network trick from Fujimoto et al. (2018). Complete information regarding the set of hyper-parameters used is provided in the appendix.

## 6. Related Work

Actor-Critic algorithms for average reward performance criterion is much less studied compared to discounted reward performance criterion. One of the earliest works on the average reward criterion is Mahadevan (1996). In this paper, Mahadevan compares the performance of R-learning with that of Q-learning and concludes that fine tuning is required to get better results from R-learning. R-learning is the average reward version of Q-learning. Later in 1999, Sutton et al. derived the policy gradient theorem for both discounted and average reward criteria (Sutton et al., 1999), which formed the bedrock for development of the average reward actor-critic algorithms. The first proof of asymptotic convergence of average reward actor-critic algorithms with function approximation appeared in Konda & Tsitsiklis (2003). A temporal difference learning based off-policy control algorithm has been proposed in Maei et al. (2010). An incremental off-policy search algorithm based on the cross entropy method has been proposed in (Joseph & Bhatnagar, 2018). Further, in Bhatnagar et al. (2007; 2009), incremental update natural policy gradient algorithms for the average reward setting have been proposed in the on-policy setting and asymptotic convergence proofs of the same provided. An off-policy variant of the natural actor-critic algorithm has been proposed in Diddigi et al. (2022).

Recently, Wan et al. presented a Differential Q-learning algorithm and claimed that their algorithm is able to find the exact differential value function without an offset. Further, Wan et al. provided an extension of the options framework from the discounted setting to the average reward setting and demonstrated the performance of the algorithm in the Four-Room domain task. One of the major contributions in off-policy policy evaluation is made by Zhang et al. (2021a). Here Zhang et al. gave a convergent off-policy evaluation scheme inspired from the gradient temporal difference learn-

ing algorithms but involving a primal-dual formulation making the policy evaluation step feasible for a neural network implementation. Zhang et al. (2021b) provided another convergent off-policy evaluation algorithm using target network and $l_2$-regularisation. In our work we use the same policy evaluation update.

Our work in this paper is actually an extension of the work of Silver et al. (2014) from the discounted to the average reward setting. In Xiong et al. (2022), a finite time analysis for deterministic policy gradient algorithm was done for the discounted reward setting. We performed the finite time analysis for the average reward deterministic policy gradient algorithm and in particular obtain the same sample complexity for our algorithm as reported by Wu et al. (2020) for stochastic policies.

## 7. Conclusion and Future Work

In this paper we presented a deterministic policy gradient theorem for both on-policy and off-policy settings considering average reward performance criteria. We then proposed the Average Reward Off-policy Deep Deterministic Policy Gradient(ARO-DDPG) algorithm using neural network and replay buffer for high dimensional MuJoCo based environments. We observed superior performance of ARO-DDPG over existing average reward algorithms (ATRPO and APO). We first showed the asymptotic convergence using ODE-based method. Later we provided finite time analysis for the on-policy and off-policy algorithms based on the proposed policy gradient theorem and obtained the sample complexity of $\Omega(\epsilon^{-2.5})$. Lastly to extend the current line of work, one could try using natural gradient descent based update rule for deterministic policy. Further in the current work we tried optimizing the average reward performance (gain optimality). In the literature, optimizing the differential value function for all the states is mentioned as part of achieving Blackwell optimality. Hence actor-critic algorithms could be designed that not only optimize average reward performance but also differential value function (bias optimality). It would also be interesting to devise similar algorithms for constrained MDPs as with (Bhatnagar, 2010; Bhatnagar & Lakshmanan, 2012; Bhatnagar et al., 2013).

## Acknowledgement

S.Bhatnagar was supported by the J.C. Bose Fellowship, Project No. DFTM/02/3125/M/04/AIR-04 from DRDO under the DIA-RCOE scheme, a project from DST-ICPS, and the RBCCPS, IISc. S Kolathaya is supported by Pratiksha Trust Young Investigator Fellowship and the SERB grant no CRG/2021/008115.

## References

Bertsekas, D. Convergence of discretization procedures in dynamic programming. *IEEE Transactions on Automatic Control*, 20(3):415–419, 1975.

Bhatnagar, S. An actor-critic algorithm with function approximation for discounted cost constrained markov decision processes. *Systems & Control Letters*, 59:760–766, 12 2010.

Bhatnagar, S. and Lakshmanan, K. *Journal of Optimization Theory and Applications*, 153(3):688–708, 2012.

Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M., and Lee, M. Incremental natural actor-critic algorithms. *Advances in neural information processing systems*, 20, 2007.

Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M., and Lee, M. Natural actor–critic algorithms. *Automatica*, 45(11): 2471–2482, 2009.

Bhatnagar, S., Prasad, H., and Prashanth, L. Stochastic recursive algorithms for optimization: Simultaneous perturbation methods. In *Stochastic Recursive Algorithms for Optimization: Simultaneous Perturbation Methods*, volume 434, pp. 320. Springer, 2013.

Borkar, V. S. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

Borkar, V. S. and Meyn, S. P. The o.d.e. method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000. doi: 10.1137/S0363012997331639. URL https://doi.org/10.1137/S0363012997331639.

Chow, C.-S. and Tsitsiklis, J. N. An optimal one-way multigrid algorithm for discrete-time stochastic control. *IEEE transactions on automatic control*, 36(8):898–914, 1991.

Degris, T., White, M., and Sutton, R. S. Linear off-policy actor-critic. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress, 2012. URL http://icml.cc/2012/papers/268.pdf.

Dewanto, V. and Gallagher, M. Examining average and discounted reward optimality criteria in reinforcement learning. *CoRR*, abs/2107.01348, 2021. URL https://arxiv.org/abs/2107.01348.

Diddigi, R. B., Jain, P., Prabuchandran, K., and Bhatnagar, S. Neural network compatible off-policy natural actor-critic algorithm. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–10, 2022. doi: 10.1109/IJCNN55064.2022.9892303.

Dufour, F. and Prieto-Rumeau, T. Approximation of average cost markov decision processes using empirical distributions and concentration inequalities. *Stochastics An International Journal of Probability and Stochastic Processes*, 87(2):273–307, 2015.

Fujimoto, S., van Hoof, H., and Meger, D. Addressing function approximation error in actor-critic methods. In Dy, J. G. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1582–1591. PMLR, 2018. URL http://proceedings.mlr.press/v80/fujimoto18a.html.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Dy, J. G. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1856–1865. PMLR, 2018. URL http://proceedings.mlr.press/v80/haarnoja18b.html.

Joseph, A. G. and Bhatnagar, S. An online prediction algorithm for reinforcement learning with linear function approximation using cross entropy method. *Machine Learning*, 107(8):1385–1429, Sep 2018. ISSN 1573-0565. doi: 10.1007/s10994-018-5727-z. URL https://doi.org/10.1007/s10994-018-5727-z.

Konda, V. R. and Tsitsiklis, J. N. Onactor-critic algorithms. *SIAM journal on Control and Optimization*, 42(4):1143–1166, 2003.

Kushner, H. J. and Clark, D. S. *Stochastic approximation methods for constrained and unconstrained systems*, volume 26. Springer Science & Business Media, 2012.

Lakshminarayanan, C. and Bhatnagar, S. A stability criterion for two timescale stochastic approximation schemes. *Automatica*, 79:108–114, 2017.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. In Bengio, Y. and LeCun, Y. (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1509.02971.

Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 5361–5371, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/dda04f9d634145a9c68d5dfe53b21272-Abstract.html.

Ma, X., Tang, X., Xia, L., Yang, J., and Zhao, Q. Average-reward reinforcement learning with trust region methods. In Zhou, Z. (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pp. 2797–2803. ijcai.org, 2021. doi: 10.24963/ijcai.2021/385. URL https://doi.org/10.24963/ijcai.2021/385.

Maei, H. R., Szepesvári, C., Bhatnagar, S., and Sutton, R. S. Toward off-policy learning control with function approximation. In *International Conference on Machine Learning (ICML)*, 2010.

Mahadevan, S. Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Mach. Learn.*, 22(1-3):159–195, 1996. doi: 10.1023/A:1018064306595. URL https://doi.org/10.1023/A:1018064306595.

Mao, Y. and Song, Y. Perturbation theory and uniform ergodicity for discrete-time markov chains. *arXiv preprint arXiv:2003.06978*, 2020.

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

Nachum, O., Chow, Y., Dai, B., and Li, L. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 2315–2325, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/cf9a242b70f45317ffd281241fa66502-Abstract.html.

Naik, A., Shariff, R., Yasui, N., and Sutton, R. S. Discounted reinforcement learning is not an optimization problem. *CoRR*, abs/1910.02140, 2019. URL http://arxiv.org/abs/1910.02140.

Schulman, J., Levine, S., Moritz, P., Jordan, M., and Abbeel, P. Trust region policy optimization. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pp. 1889–1897. JMLR.org, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. A. Deterministic policy gradient algorithms. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pp. 387–395. JMLR.org, 2014. URL http://proceedings.mlr.press/v32/silver14.html.

Sutton, R. S., McAllester, D. A., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In Solla, S. A., Leen, T. K., and Müller, K. (eds.), *Advances in Neural Information Processing Systems 12, [NIPS Conference, Denver, Colorado, USA, November 29 - December 4, 1999]*, pp. 1057–1063. The MIT Press, 1999. URL http://papers.nips.cc/paper/1713-policy-gradient-methods-for-reinforcement-learning-with-function-approximation.

Tassa, Y., Doron, Y., Muldal, A., Erez, T., Li, Y., de Las Casas, D., Budden, D., Abdolmaleki, A., Merel, J., Lefrancq, A., Lillicrap, T. P., and Riedmiller, M. A. Deepmind control suite. *CoRR*, abs/1801.00690, 2018. URL http://arxiv.org/abs/1801.00690.

Wan, Y., Naik, A., and Sutton, R. Average-reward learning and planning with options. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 22758–22769. Curran Associates, Inc., 2021a. URL https://proceedings.neurips.cc/paper/2021/file/c058f544c737782deacefa532d9add4c-Paper.pdf.

Wan, Y., Naik, A., and Sutton, R. S. Learning and planning in average-reward markov decision processes. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10653–10662. PMLR, 18–24 Jul 2021b. URL https://proceedings.mlr.press/v139/wan21a.html.

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/2020/file/cc9b3c69b56df284846bf2432f1cba90-Paper.pdf.

Xiong, H., Xu, T., Zhao, L., Liang, Y., and Zhang, W. Deterministic policy gradient: Convergence analysis. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022.

Zhang, R., Dai, B., Li, L., and Schuurmans, D. Gendice: Generalized offline estimation of stationary values. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020a. URL https://openreview.net/forum?id=HkxlcnVFwB.

Zhang, S., Liu, B., and Whiteson, S. Gradientdice: Rethinking generalized offline estimation of stationary values. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11194–11203. PMLR, 2020b. URL http://proceedings.mlr.press/v119/zhang20r.html.

Zhang, S., Wan, Y., Sutton, R. S., and Whiteson, S. Average-reward off-policy policy evaluation with function approximation. In *International Conference on Machine Learning*, pp. 12578–12588. PMLR, 2021a.

Zhang, S., Yao, H., and Whiteson, S. Breaking the deadly triad with a target network. In *International Conference on Machine Learning*, pp. 12621–12631. PMLR, 2021b.

Zhang, Y. and Ross, K. W. On-policy deep reinforcement learning for the average-reward criterion. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12535–12545. PMLR, 2021. URL http://proceedings.mlr.press/v139/zhang21q.html.

Zou, S., Xu, T., and Liang, Y. Finite-sample analysis for SARSA with linear function approximation. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14,*

*2019, Vancouver, BC, Canada*, pp. 8665–8675, 2019.
URL https://proceedings.neurips.cc/paper/2019/hash/9f9e8cba3700df6a947a8cf91035ab84-Abstract.html.

# A. Assumptions, Lemmas and Theorems

## A.1. Additional Assumptions

We make the following additional assumptions.

**Assumption A.1.** The transition probability density function for a policy $\pi$ with parameter $\theta$ is Lipschitz continuous w.r.t $\theta$. Thus, $max_{s',s}|P^{\pi_1}(s'|s) - P^{\pi_2}(s'|s)| \leq L_t\|\theta_1 - \theta_2\|$.

The above assumption is a standard assumption in theoretical studies in literature. Reference for those assumptions can be found in Xiong et al. (2022); Bertsekas (1975); Chow & Tsitsiklis (1991) and Dufour & Prieto-Rumeau (2015).

**Assumption A.2.** The reward function for a policy $\pi$ with parameter $\theta$ is Lipschitz continuous w.r.t $\theta$. Thus, $max_s|R^{\pi_1}(s) - R^{\pi_2}(s)| \leq L_r\|\theta_1 - \theta_2\|$.

The above assumption can be satified by using a well defined reward function to ensure Lipchitz continuity of reward function w.r.t action and then evoking Assumption 4.4.

**Assumption A.3.** The initial value of target estimators is bounded. Thus, $\|\bar{w_0}\| \leq C_w$ and $\|\bar{\rho_0}\| \leq (Cr + 2C_w)$.

Assumption A.3 is used to enforce the stability of the iterates of target estimators.

**Assumption A.4.** Let $A(\theta) = \int d^\pi(s)(\phi^\pi(s)(\int P^\pi(s'|s)\phi^\pi(s')\,ds' - \phi^\pi(s))^\intercal - \eta I)\,ds$. $\lambda_{min}$ is the lower bound on the minimum eigenvalue of $A(\theta)$ for all values of $\theta$.

The assumption above is used in Lemma A.21 to prove the Lipchitz continuity of optimal critic parameter $w^*$ for a particular value of policy parameter $\theta$ with respect to $\theta$.

**Assumption A.5.** Let $A'(\theta) = \int d^\pi(s)(\phi^\pi(s)(\int P^\pi(s'|s)\phi^\pi(s')\,ds' - \phi^\pi(s))^\intercal)\,ds$. $\lambda^{all}_{max}$ is the upper bound on maximum eigenvalue of $(A'(\theta) + A'(\theta)^\intercal)/2$ for all values of $\theta$.

Assumption A.5 is used to prove the negative definiteness of the matrix $A_\theta$ (defined in Assumption A.4) in Lemma A.26.

**Assumption A.6.** Let $H_\theta = \int_S d^\pi(s)\nabla_\theta\pi(s,\theta)\nabla_\theta\pi(s,\theta)^\intercal\,ds$. $\lambda^\epsilon_{min} > 0$ is the lower bound on the minimum eigenvalues of $H_\theta$ for all values of $\theta$.

The above assumption is used in Lemma A.28 to make sure $H_\theta$ is invertible and optimal critic parameter $w^*_\epsilon$ according to compatible function approximation lemma (Lemma 3.2) can be obtained. Similar assumption is present in (Xiong et al., 2022).

**Assumption A.7.** Let $A^{\mu'}_{off}(\theta) = \int d^\mu(s)(\phi^\pi(s)(\int P^\pi(s'|s)\phi^\pi(s')\,ds' - \phi^\pi(s))^\intercal)\,ds$. $\chi^{all}_{max}$ is the upper bound on maximum eigenvalue of $(A^{\mu'}_{off}(\theta) + A^{\mu'}_{off}(\theta)^\intercal)/2$ for behaviour policy $\mu$ and all values of $\theta$.

Assumption A.7 is used to prove the negative definiteness of the matrix $A_\theta$ (defined in Lemma A.30) in Lemma A.31.

**Assumption A.8.** $\forall$ s policy $\pi$ is twice continuously differentiable i.e. $\nabla^2_\theta\pi(s)$ exists.

Assumption A.8 can be satisfied by using neural network to parameterize the policy $\pi$.

## A.2. Lemmas and Theorems for Policy Gradient

**Lemma A.9.** *There exists a unique constant $k(= \rho(\pi))$ which satisfies the following equation for differential value function $V_{diff}$ :*

$$V^\pi_{diff}(s_t) = \mathbb{E}^\pi[R(s_t,a_t) - k + V^\pi_{diff}(s_{t+1})|s_t].$$

*Proof.*

$$\begin{aligned}
&V^\pi_{diff}(s_t) = R(s_t,\pi(s_t)) - k + \int_S P^\pi(s_{t+1}|s_t)V^\pi_{diff}(s_{t+1})\,ds_{t+1}\\
&\implies V^\pi_{diff}(s_t) - \int_S P^\pi(s_{t+1}|s_t)V^\pi_{diff}(s_{t+1})\,ds_{t+1} = R(s_t,\pi(s_t)) - k\\
&\implies \sum_{t=0}^{T-1}\Big(V^\pi_{diff}(s_t) - \int_S P^\pi(s_{t+1}|s_t)V^\pi_{diff}(s_{t+1})\,ds_{t+1}\Big) = \sum_{t=0}^{T-1} R(s_t,\pi(s_t)) - kT
\end{aligned}$$

Integrating w.r.t the stationary distribution $d^\pi$ of policy $\pi$ :

$$\sum_{t=0}^{T-1} \int_S d^\pi(s_t) \Big( V^\pi_{diff}(s_t) - \int_S P^\pi(s_{t+1}|s_t) V^\pi_{diff}(s_{t+1})\, ds_{t+1} \Big)\, ds_t$$
$$= \sum_{t=0}^{T-1} \int_S d^\pi(s_t) R(s_t, \pi(s_t))\, ds - kT$$

$$\sum_{t=0}^{T-1} \Big( \int_S d^\pi(s_t) V^\pi_{diff}(s_t)\, ds_t - \int_S d^\pi(s_{t+1}) V^\pi_{diff}(s_{t+1})\, ds_{t+1} \Big)$$
$$= \sum_{t=0}^{T-1} \int_S d^\pi(s_t) R(s_t, \pi(s_t))\, ds_t - kT$$

Note: $\Big( \int_S d^\pi(s_t) V^\pi_{diff}(s_t)\, ds_t - \int_S d^\pi(s_{t+1}) V^\pi_{diff}(s_{t+1})\, ds_{t+1} \Big) = 0.$

$$\implies k = \frac{1}{T} \sum_{t=0}^{T-1} \int_S d^\pi(s_t) R(s_t, \pi(s_t))\, ds_t$$
$$\implies k = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \int_S d^\pi(s_t) R(s_t, \pi(s_t))\, ds_t$$
$$\implies k = \rho(\pi) \quad \text{(using (3))}.$$

□

**Theorem A.10.** *The gradient of $\rho(\pi)$ with respect to the policy parameter $\theta$ is given as follows:*

$$\nabla_\theta \rho(\pi) = \int_S d^\pi(s) \nabla_a Q^\pi_{diff}(s,a)|_{a=\pi(s)} \nabla_\theta \pi(s,\theta)\, ds.$$

*Proof.* Using Lemma 2.2:

$$V^\pi_{diff}(s_t) = R(s_t, \pi(s_t)) - \rho(\pi) + \int_S P^\pi(s_{t+1}|s_t) V^\pi_{diff}(s_{t+1})\, ds_{t+1}$$
$$\implies Q^\pi_{diff}(s_t, \pi(s_t)) = R(s_t, \pi(s_t)) - \rho(\pi) + \int_S P^\pi(s_{t+1}|s_t) Q^\pi_{diff}(s_{t+1}, \pi(s_{t+1}))\, ds_{t+1}$$

Differentiating w.r.t $\theta$, we obtain

$$\nabla_\theta Q^\pi_{diff}(s_t, \pi(s_t)) = \nabla_\theta R(s_t, \pi(s_t)) - \nabla_\theta \rho(\pi)$$
$$+ \nabla_\theta \Big( \int_S P^\pi(s_{t+1}|s_t) Q^\pi_{diff}(s_{t+1}, \pi(s_{t+1}))\, ds_{t+1} \Big)$$
$$= \nabla_a R(s_t, a)|_{a=\pi(s_t)} \nabla_\theta \pi(s_t) - \nabla_\theta \rho(\pi)$$
$$+ \int_S \nabla_a P^\pi(s_{t+1}|s_t, a)|_{a=\pi(s_t)} \nabla_\theta \pi(s_t) Q^\pi_{diff}(s_{t+1}, \pi(s_{t+1}))\, ds_{t+1}$$
$$+ \int_S P^\pi(s_{t+1}|s_t) \nabla_\theta Q^\pi_{diff}(s_{t+1}, \pi(s_{t+1}))\, ds_{t+1}.$$

Note: $\nabla_a \rho(\pi) = \nabla_a \Big( \int_S d^\pi(s) R^\pi(s)\, ds \Big) = 0.$

$$\implies \nabla_\theta Q^\pi_{diff}(s_t, \pi(s_t)) = \nabla_a Q^\pi_{diff}(s_t, a)|_{a=\pi(s_t)} \nabla_\theta \pi(s_t) - \nabla_\theta \rho(\pi) + \int_S P^\pi(s_{t+1}|s_t) \nabla_\theta Q^\pi_{diff}(s_{t+1}, \pi(s_{t+1}))\, ds_{t+1}.$$

Integrating w.r.t stationary distribution $d^\pi(\cdot)$ of policy $\pi$:

$$\int_S d^\pi(s_t) \nabla_\theta Q^\pi_{diff}(s_t, \pi(s_t)) ds_t = \int_S d^\pi(s_t) \nabla_a Q^\pi_{diff}(s_t, a)|_{a=\pi(s_t)} \nabla_\theta \pi(s_t) ds_t - \nabla_\theta \rho(\pi) + \int_S d^\pi(s_t) \int_S P^\pi(s_{t+1}|s_t) \nabla_\theta Q^\pi_{diff}(s_{t+1}, \pi(s_{t+1}))\, ds_{t+1}\, ds_t.$$

Note: $\int_S d^\pi(s) P^\pi(s'|s)\, ds = d^\pi(s')$. Thus,

$$\nabla_\theta \rho(\pi) = \int_S d^\pi(s_t) \nabla_a Q^\pi_{diff}(s_t, a)|_{a=\pi(s_t)} \nabla_\theta \pi(s_t) ds_t + \int_S d^\pi(s_{t+1}) \nabla_\theta Q^\pi_{diff}(s_{t+1}, \pi(s_{t+1}))\, ds_{t+1} - \int_S d^\pi(s_t) \nabla_\theta Q^\pi_{diff}(s_t, \pi(s_t)) ds_t.$$

$$\nabla_\theta \rho(\pi) = \int_S d^\pi(s) \nabla_a Q^\pi_{diff}(s, a)|_{a=\pi(s)} \nabla_\theta \pi(s)\, ds.$$

□

**Lemma A.11.** *Assume that the differential Q-value function (5) satisfies the following:*

1. $$\nabla_w \nabla_a Q^w_{diff}(s, a)|_{a=\pi(s)} = \nabla_\theta \pi(s, \theta).$$

2. *The differential Q-value function parameter $w = w^*_\epsilon$ optimizes the following error function:*

$$\zeta(\theta, w) = \frac{1}{2} \int_S d^\pi(s) \| \nabla_a Q^\pi_{diff}(s, a)|_{a=\pi(s)} - \nabla_a Q^w_{diff}(s, a)|_{a=\pi(s)} \|^2\, ds.$$

*Then,*

$$\int_S d^\pi(s) \nabla_a Q^\pi_{diff}(s, a)|_{a=\pi(s)} \nabla_\theta \pi(s, \theta)\, ds = \int_S d^\pi(s) \nabla_a Q^w_{diff}(s, a)|_{a=\pi(s)} \nabla_\theta \pi(s, \theta)\, ds.$$

*Further,*

$$\nabla_a Q^w_{diff}(s, a) = \nabla_\theta \pi(s, \theta)^\intercal w \quad (\textit{for linear function approximator}).$$

*Proof.* Let $\mathcal{E}(\theta, w, s) = \nabla_a Q^\pi_{diff}(s, a)|_{a=\pi(s)} - \nabla_a Q^w_{diff}(s, a)|_{a=\pi(s)}$,

$$\zeta(\theta, w) = \frac{1}{2} \int_S d^\pi(s) \mathcal{E}(\theta, w, s)^\intercal \mathcal{E}(\theta, w, s)\, ds.$$

Differentiating w.r.t the critic parameter $w$, we obtain:

$$\begin{aligned}\nabla_w \zeta(\theta, w) &= \int_S d^\pi(s) \nabla_w \mathcal{E}(\theta, w, s) \mathcal{E}(\theta, w, s)\, ds \\ &= -\int_S d^\pi(s) \nabla_w \nabla_a Q^w_{diff}(s,a)|_{a=\pi(s)} \Big( \nabla_a Q^\pi_{diff}(s,a)|_{a=\pi(s)} \\ &\quad - \nabla_a Q^w_{diff}(s,a)|_{a=\pi(s)} \Big)\, ds = 0.\end{aligned}$$

Letting $\nabla_w \nabla_a Q^w_{diff}(s,a)|_{a=\pi(s)} = \nabla_\theta \pi(s)$, we obtain

$$\int_S d^\pi(s) \nabla_a Q^\pi_{diff}(s,a)|_{a=\pi(s)} \nabla_\theta \pi(s,\theta)\, ds = \int_S d^\pi(s) \nabla_a Q^w_{diff}(s,a)|_{a=\pi(s)} \nabla_\theta \pi(s,\theta)\, ds.$$

Let us consider the case of linear function approximator with parameter $w$, i.e., $Q^w_{diff}(s, \pi(s)) = \phi^\pi(s, \pi(s))^\intercal w$.

We know from above,

$$\begin{aligned}&\nabla_w \nabla_a Q^w_{diff}(s,a)|_{a=\pi(s)} = \nabla_\theta \pi(s) \\ &\implies \nabla_a \phi^\pi(s,a)|_{a=\pi(s)} = \nabla_\theta \pi(s).\end{aligned} \tag{A.1}$$

Thus,

$$\begin{aligned}&Q^w_{diff}(s, \pi(s)) = \phi^\pi(s, \pi(s))^\intercal w \\ &\implies \nabla_a Q^w_{diff}(s,a)|_{a=\pi(s)} = \nabla_a \phi^\pi(s,a)|^\intercal_{a=\pi(s)} w \\ &\implies \nabla_a Q^w_{diff}(s,a)|_{a=\pi(s)} = \nabla_\theta \pi(s)^\intercal w \ \ \text{(using } (A.1)\text{)}.\end{aligned}$$

□

**Theorem A.12.** *The approximate gradient ($\widehat{\nabla_\theta \rho(\pi)}$) of the average reward $\rho(\pi)$ with respect to the policy parameter $\theta$ is given by the following expression:*

$$\widehat{\nabla_\theta \rho(\pi)} = \int_S d^\mu(s) \nabla_a Q^\pi_{diff}(s,a)|_{a=\pi(s)} \nabla_\theta \pi(s,\theta)\, ds. \tag{A.2}$$

*Further, the approximation error is $\mathcal{E}(\pi, \mu) = \|\nabla_\theta \rho(\pi) - \widehat{\nabla_\theta \rho(\pi)}\|$, where $\mu$ represents the behaviour policy with parameter $\theta^\mu$ and $\nabla_\theta \rho(\pi)$ is the on-policy policy gradient from Theorem 3.1. $\mathcal{E}$ satisfies*

$$\mathcal{E}(\pi, \mu) \leq Z \|\theta - \theta^\mu\|, \tag{A.3}$$

*where, $Z = 2^{n+1} C(\lceil \log_\kappa a^{-1} \rceil + 1/\kappa) L_t$ with $L_t$ being the Lipchitz constant for the transition probability density function (Assumption A.1). Constants $a$ and $\kappa$ are from Assumption 3.3, $n$ is the dimension of the state space, and $C = \max_s \|\nabla_a Q^\pi_{diff}(s,a)|_{a=\pi(s)} \nabla_\theta \pi(s,\theta)\|$.*

*Proof.*

$$\begin{aligned}\mathcal{E}(\pi, \mu) &= \|\nabla_\theta \rho(\pi) - \widehat{\nabla_\theta \rho(\pi)}\| \\ &= \| \int_S d^\pi(s) \nabla_a Q^\pi_{diff}(s,a)|_{a=\pi(s)} \nabla_\theta \pi(s,\theta)\, ds \\ &\quad - \int_S d^\mu(s) \nabla_a Q^\pi_{diff}(s,a)|_{a=\pi(s)} \nabla_\theta \pi(s,\theta)\, ds \| \\ &\leq \int_S |d^\pi(s) - d^\mu(s)| \|\nabla_a Q^\pi_{diff}(s,a)|_{a=\pi(s)} \nabla_\theta \pi(s,\theta)\|\, ds \\ &\leq C \int_S |d^\pi(s) - d^\mu(s)|\, ds.\end{aligned}$$

Here, $C = \max_s \|\nabla_a Q^{\pi}_{diff}(s,a)|_{a=\pi(s)} \nabla_\theta \pi(s,\theta)\|$. Thus,

$$\mathcal{E}(\pi,\mu) \leq C L_d \|\theta - \theta^\mu\| = Z\|\theta - \theta^\mu\| \text{ (using Lemma A.27)}.$$

Here, $Z = 2^{n+1} C(\lceil \log_\kappa a^{-1} \rceil + 1/\kappa) L_t$.

□

### A.3. Finite Time Analysis

Figure 3 shows the relation between different types of error in the on-policy algorithm (Algorithm 2). Here, arrow from value function error to actor error shows that the actor error is dependent on the value function error. Similar relationship follows for the rest of the error types. Figure 3 intuitively explains which Lemma utilizes which Lemma to arrive at the final result (Theorem A.18).

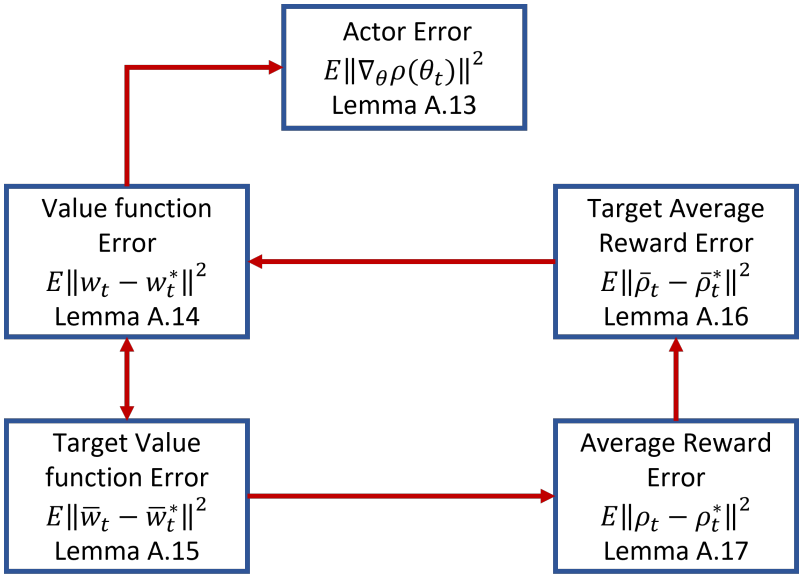


*Figure 3.* Dependency of different types of errors on each other.

#### A.3.1. Main Lemmas and Theorems

**Lemma A.13.** *Let the cumulative error of on-policy actor be $\sum_{t=0}^{T-1} E||\nabla_\theta \rho(\theta_t)||^2$ and cumulative error of critic be $\sum_{t=0}^{T-1} E||\Delta w_t||^2$. $\theta_t$ and $w_t$ are the actor and linear critic parameter at time t.Bound on the cumulative error of on-policy actor is proven using cumulative error of critic as follows:*

$$\frac{1}{T}\sum_{t=0}^{T-1} E||\nabla_\theta \rho(\theta_t)||^2 \leq 2\frac{C_r}{C_\gamma} T^{v-1} + 3C_\pi^4 (\frac{1}{T}\sum_{t=0}^{T-1} E||\Delta w_t||^2) + 3C_\pi^4(\tau^2 + \frac{4}{M} C^2_{w^*_\epsilon}),$$
$$+ \frac{C_\gamma L_J G_\theta^2}{1-v} T^{-v}$$

*Here, $C_r$ is the upper bound on rewards (Assumption 4.2) , $C_\gamma$, $v$ are constants used for step size $\gamma_t$ (Assumption 3.5, $\|\nabla_\theta \pi(s)\| \leq C_\pi$ (Assumption 4.4), $\Delta w_t = w_t - w_t^*, \tau = \max_t \|w_t^* - w^*_{\epsilon,t}\|$, $w^*_\epsilon$ is the optimal critic parameter according to Lemma 3.2. $w_t^*$ is the optimal parameters given by TD(0) algorithm corresponding to policy parameter $\theta_t$. Constant $C_{w^*_\epsilon}$ is defined in Lemma A.28. $L_J$ is the coefficient used in smoothness condition of the non convex function $\rho(\theta)$. Constant $G_\theta$ is defined in Lemma A.22. M is the size of batch of samples used to update parameters.*

*Proof.* By $[-L_J, L_J]$-smoothness of non-convex function we have:

$$E[\rho(\theta_{t+1})] \geq E[\rho(\theta_t)] + E\langle \nabla_\theta \rho(\theta_t), \theta_{t+1} - \theta_t \rangle - \frac{L_J}{2} E\|\theta_{t+1} - \theta_t\|^2. \tag{A.4}$$

Now,

$$h(B_t, w_t, \theta_t) = \frac{1}{M} \sum_i \nabla_a Q^\pi(s_{t,i}, a)|_{a=\pi(s_{t,i})} \nabla_\theta \pi(s_{t,i}).$$

$$\begin{aligned} E\langle \nabla_\theta \rho(\theta_t), \theta_{t+1} - \theta_t \rangle &= \gamma_t E\langle \nabla_\theta \rho(\theta_t), h(B_t, w_t, \theta_t) \rangle \\ &= \gamma_t E\langle \nabla_\theta \rho(\theta_t), h(B_t, w_t, \theta_t) - \nabla_\theta \rho(\theta_t) \rangle + \gamma_t E\|\nabla_\theta \rho(\theta_t)\|^2. \end{aligned} \tag{A.5}$$

From (A.5), we have

$$\begin{aligned} E\langle \nabla_\theta \rho(\theta_t), h(B_t, w_t, \theta_t) - \nabla_\theta \rho(\theta_t) \rangle &\geq -\frac{1}{2} E\|\nabla_\theta \rho(\theta_t)\|^2 - \frac{1}{2} E\|h(B_t, w_t, \theta_t) - \nabla_\theta \rho(\theta_t)\|^2 \\ (\because x^\intercal y \geq -\|x\|^2/2 - \|y\|^2/2). & \end{aligned} \tag{A.6}$$

From (A.6):

$$\begin{aligned} &E\|h(B_t, w_t, \theta_t) - \nabla_\theta \rho(\theta_t)\|^2 \\ &= E\|h(B_t, w_t, \theta_t) - h(B_t, w_t^*, \theta_t) + h(B_t, w_t^*, \theta_t) - h(B_t, w_{\epsilon,t}^*, \theta_t) + h(B_t, w_{\epsilon,t}^*, \theta_t) - \nabla_\theta \rho(\theta_t)\|^2 \\ &\leq 3(E\|h(B_t, w_t, \theta_t) - h(B_t, w_t^*, \theta_t)\|^2 \ \textcircled{1} \\ &\quad + E\|h(B_t, w_t^*, \theta_t) - h(B_t, w_{\epsilon,t}^*, \theta_t)\|^2 \ \textcircled{2} \\ &\quad + E\|h(B_t, w_{\epsilon,t}^*, \theta_t) - \nabla_\theta \rho(\theta_t)\|^2 \ \textcircled{3} \end{aligned} \tag{A.7}$$

From (A.7):

①:

$$\begin{aligned} &E||h(B_t, w_t, \theta_t) - h(B_t, w_t^*, \theta_t)||^2 \\ &= \frac{1}{M} || \sum_{i=0} \nabla_a Q^{w_t}(s_{t,i}, a)|_{a=\pi(s_{t,i})} \nabla_\theta \pi(s_{t,i}) - \sum_{i=0} \nabla_a Q^{w_t^*}(s_{t,i}, a)|_{a=\pi(s_{t,i})} \nabla_\theta \pi(s_{t,i}) ||^2. \end{aligned}$$

Here, by compatible function approximation lemma 3.2: $\nabla_a Q^{w_t^*}(s_i, a)|_{a=\pi(s_i)} = \nabla_\theta \pi(s)^\intercal w$.

$$\begin{aligned} E||h(B_t, w_t, \theta_t) - h(B_t, w_t^*, \theta_t)||^2 &= E|| \frac{1}{M} \sum_{i=0} \nabla_\theta \pi(s_{t,i}) \nabla_\theta \pi(s_{t,i})^\intercal (w_t - w_t^*)||^2 \\ &\leq C_\pi^4 E||w_t - w_t^*||^2. \end{aligned}$$

② is similar as ①:

$$\begin{aligned} E||h(B_t, w_t^*, \theta_t) - h(B_t, w_{\epsilon,t}^*, \theta_t)||^2 &\leq C_\pi^4 E||w_t^* - w_{\epsilon,t}^*||^2 \\ &\leq C_\pi^4 \tau^2. \end{aligned}$$

③ :

- By compatible function approximation lemma 3.2: $\nabla_\theta \rho(\theta_t) = \int_S d(s, \pi(\theta_t)) \nabla_\theta \pi(s) \nabla_\theta \pi(s)^\intercal w^*_{\epsilon,t} \, ds = E[h(B_t, w^*_{\epsilon,t}, \theta_t)]$

- By lemma 4 (Xiong et al., 2022), if $E[\hat{Y}] = \bar{Y}, ||\hat{Y}||, ||\bar{Y}|| \leq C_Y$ then,

$$E||\frac{1}{M}\sum_{i=0}^{M-1} \hat{Y}_i - \bar{Y}|| \leq 4\frac{C_Y^2}{M}.$$

Using above two bullet points:

$$\begin{aligned} E||h(B_t, w^*_{\epsilon,t}, \theta_t) - \nabla_\theta \rho(\theta_t)||^2 &\leq \frac{4}{M}||\nabla_\theta \pi(s) \nabla_\theta \pi(s)^\intercal w^*_{\epsilon,t}||^2 \\ &\leq \frac{4C_\pi^4 C_{w_\epsilon}^2}{M}. \end{aligned}$$

Combining ①,② and ③ and using in (A.7):

$$E||h(B_t, w_t, \theta_t) - \nabla_\theta \rho(\theta_t)||^2 \leq 3C_\pi^4 (E||w_t - w_t^*||^2 + \tau^2 + \frac{4C_{w_\epsilon}^2}{M}). \tag{A.8}$$

Using (A.8) in (A.6):

$$\begin{aligned} E\langle \nabla_\theta \rho(\theta_t), h(B_t, w_t, \theta_t) - \nabla_\theta \rho(\theta_t) \rangle \geq &-\frac{1}{2} E||\nabla_\theta \rho(\theta_t)||^2 \\ &- \frac{3}{2} C_\pi^4 (E||w_t - w_t^*||^2 + \tau^2 + \frac{4C_{w_\epsilon}^2}{M}). \end{aligned} \tag{A.9}$$

Using (A.9) in (A.5):

$$\begin{aligned} E\langle \nabla_\theta \rho(\theta_t), \theta_{t+1} - \theta_t \rangle \geq &\frac{\gamma_t}{2} E||\nabla_\theta \rho(\theta_t)||^2 \\ &- \frac{3\gamma_t}{2} C_\pi^4 (E||w_t - w_t^*||^2 + \tau^2 + \frac{4C_{w_\epsilon}^2}{M}). \end{aligned} \tag{A.10}$$

Using (A.10) in (A.4):

$$\begin{aligned} E[\rho(\theta_{t+1})] - E[\rho(\theta_t)] \geq &\frac{\gamma_t}{2} E||\nabla_\theta \rho(\theta_t)||^2 - \frac{L_J}{2} E||\theta_{t+1} - \theta_t||^2 \\ &- \frac{3\gamma_t}{2} C_\pi^4 (E||w_t - w_t^*||^2 + \tau^2 + \frac{4C_{w_\epsilon}^2}{M}) \end{aligned}$$

$$\begin{aligned} \implies E||\nabla_\theta \rho(\theta_t)||^2 \geq &\frac{2}{\gamma_t} E[\rho(\theta_{t+1})] - E[\rho(\theta_t)] + 3C_\pi^4 (E||w_t - w_t^*||^2) \\ &+ 3C_\pi^4 (\tau^2 + \frac{4C_{w_\epsilon}^2}{M}) + L_J \gamma_t G_\theta^2 \quad \text{(using lemma A.22)} \end{aligned}$$

$$\begin{aligned}\implies \sum_{t=0}^{T-1} E||\nabla_\theta \rho(\theta_t)||^2 \geq \sum_{t=0}^{T-1} \frac{2}{\gamma_t} E[\rho(\theta_{t+1})] - E[\rho(\theta_t)] \ \textcircled{1} \\ + \sum_{t=0}^{T-1} 3C_\pi^4 (E||w_t - w_t^*||^2) \ \textcircled{2} \\ + \sum_{t=0}^{T-1} 3C_\pi^4 (\tau^2 + \frac{4C_{w_\epsilon}^2}{M}) \ \textcircled{3} \\ + \sum_{t=0}^{T-1} L_J \gamma_t G_\theta^2 \ \textcircled{4} \quad \text{(using lemma A.22)}\end{aligned} \tag{A.11}$$

From (A.11)

①:

$$\begin{aligned}\sum_{t=0}^{T-1} \frac{2}{\gamma_t} E[\rho(\theta_{t+1})] - E[\rho(\theta_t)] &= 2\Big(\sum_{t=0}^{T-1} \Big(\frac{1}{\gamma_t} - \frac{1}{\gamma_{t-1}}\Big) E[\rho(\theta_t)] + \frac{E[\rho(\theta_0)]}{\gamma_0} - \frac{E[\rho(\theta_T)]}{\gamma_{T-1}}\Big) \\ &\leq 2\Big(\sum_{t=0}^{T-1} \Big(\frac{1}{\gamma_t} - \frac{1}{\gamma_{t-1}}\Big) E[\rho(\theta_t)] + \frac{E[\rho(\theta_0)]}{\gamma_0}\Big) \\ &\leq 2\Big(\sum_{t=0}^{T-1} \Big(\frac{1}{\gamma_t} - \frac{1}{\gamma_{t-1}}\Big) + \gamma_0\Big) C_r \\ &\leq \frac{2C_r}{\gamma_{T-1}} = \frac{2C_r T^v}{C_\gamma}\end{aligned}$$

②:

$$\sum_{t=0}^{T-1} 3C_\pi^4 (E||w_t - w_t^*||^2) = \sum_{t=0}^{T-1} 3C_\pi^4 (E||\Delta w_t||^2)$$

④:

$$\sum_{t=0}^{T-1} L_J \gamma_t G_\theta^2 \leq L_J G_\theta^2 C_\gamma \frac{T^{1-v}}{1-v} \quad \Big(\because \sum_{t=0}^{T-1} \frac{1}{1+t^v} \leq \int_0^T \frac{1}{t^v}\, dt = \frac{T^{1-v}}{1-v}\Big)$$

Using ①-④ and dividing (A.11) by T:

$$\begin{aligned}\frac{1}{T} \sum_{t=0}^{T-1} E||\nabla_\theta \rho(\theta_t)||^2 \leq 2\frac{C_r}{C_\gamma} T^{v-1} + 3C_\pi^4 (\frac{1}{T} \sum_{t=0}^{T-1} E||\Delta w_t||^2) + 3C_\pi^4 (\tau^2 + \frac{4}{M} C_{w_\epsilon}^2) \\ + \frac{C_\gamma L_J G_\theta^2}{1-v} T^{-v}\end{aligned}$$

□

**Lemma A.14.** *Let the cumulative error of linear critic be $\sum_{t=0}^{T-1} \mathbb{E}||\Delta w_t||^2$, the cumulative error of target linear critic be $\sum_{t=0}^{T-1} \mathbb{E}||\Delta \bar{w}_t||^2$, and the cumulative error of target average reward estimator be $\sum_{t=0}^{T-1} \mathbb{E}||\Delta \bar{\rho}_t||^2$. $w_t$, $\bar{w}_t$ and $\bar{\rho}_t$ are linear critic parameter, target linear critic parameter and target average reward estimator at time t respectively. Bound on the cumulative error of critic is proven using cumulative error of target average reward estimator and target linear critic*

*parameter as follows:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2 \leq 2\left(\sqrt{\frac{2C_w^2}{(\lambda-1)C_\alpha}T^{\sigma-1}+\frac{C_gC_\alpha}{1-\sigma}T^{-\sigma}}+\frac{L_wG_\theta C_\gamma}{(\lambda-1)C_\alpha}\left(\frac{T^{-2(v-\sigma)}}{1-2(v-\sigma)}\right)^{\frac{1}{2}}\right)^2 + \frac{4}{(\lambda-1)^2}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}|\Delta\bar{\rho}_t|^2+\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta\bar{w}_t||^2\right)$$

*Here, $\Delta w_t = w_t - w_t^*$, $\Delta\bar{w}_t = \bar{w}_t - w_t^*$, and $\Delta\bar{\rho}_t = \bar{\rho}_t - \rho_t^*$. $w_t^*$ and $\rho_t^*$ are the optimal parameters given by TD(0) algorithm corresponding to policy parameter $\theta_t$. $C_\alpha$, $C_\gamma$, $\sigma$, $v$ are constants and $\gamma_t, \alpha_t$ are step-sizes defined in Assumption 3.5, $\|w_t\| \leq C_w$ (Algorithm 2, step 8), $C_r$ is the upper bound on rewards (Assumption 4.2), Constant $G_\theta$ is defined in Lemma A.22, $C_g = \frac{L_w^2}{(\lambda-1)}\max_t\frac{\gamma_t^2}{\alpha_t^2}G_\theta^2 + \frac{C_\delta^2}{(\lambda-1)}$, $C_\delta = 2C_r + (4+\eta)C_w$. $\eta$ is the l2-regularisation coefficient from Algorithm 2 and $\eta > \lambda_{max}^{all}$, where $\lambda_{max}^{all}$ is defined in Lemma A.26. $\lambda$ is defined in Lemma A.26. $L_w$ is defined in Lemma A.21.*

*Proof.*

$$\begin{aligned}
w_{t+1} &= w_t + \alpha_t\frac{1}{M}\sum_{i=0}^{M-1}\Big(R^\pi(s_{t,i}) - \bar{\rho}_t + \phi^\pi(s'_{t,i})^\intercal\bar{w}_t - \phi^\pi(s_{t,i})^\intercal w_t\Big)\phi^\pi(s_{t,i}) - \alpha_t\eta w_t \\
\implies w_{t+1} - w_{t+1}^* &= w_t - w_t^* + w_t^* - w_{t+1}^* \quad ① \\
&\quad + \alpha_t\frac{1}{M}\sum_{i=0}^{M-1}\Big(R^\pi(s_{t,i}) - \rho_t^* + \phi^\pi(s'_{t,i})^\intercal\bar{w}_t - \phi^\pi(s_{t,i})^\intercal w_t\Big)\phi^\pi(s_{t,i}) - \alpha_t\eta w_t \quad ② \\
&\quad + \alpha_t\frac{1}{M}\sum_{i=0}^{M-1}\Big(\rho_t^* - \bar{\rho}_t\Big)\phi^\pi(s_{t,i}) \quad ③
\end{aligned} \tag{A.12}$$

From (A.12):
②:

$$\begin{aligned}
&\frac{1}{M}\sum_{i=0}^{M-1}\Big(R^\pi(s_{t,i}) - \rho_t^* + \phi^\pi(s'_{t,i})^\intercal\bar{w}_t - \phi^\pi(s_{t,i})^\intercal w_t\Big)\phi^\pi(s_{t,i}) - \eta w_t \\
&\quad= \frac{1}{M}\sum_{i=0}^{M-1}\Big(R^\pi(s_{t,i}) - \rho_t^* + \phi^\pi(s'_{t,i})^\intercal w_t - \phi^\pi(s_{t,i})^\intercal w_t\Big)\phi^\pi(s_{t,i}) - \eta w_t \\
&\quad\quad + \frac{1}{M}\sum_{i=0}^{M-1}\phi^\pi(s_{t,i})\phi^\pi(s'_{t,i})^\intercal(\bar{w}_t - w_t^*) - \frac{1}{M}\sum_{i=0}^{M-1}\phi^\pi(s_{t,i})\phi^\pi(s'_{t,i})^\intercal(w_t - w_t^*) \\
&\quad= \frac{1}{M}\sum_{i=0}^{M-1}\phi^\pi(s_{t,i})\phi^\pi(s'_{t,i})^\intercal(\bar{w}_t - w_t^*) - \frac{1}{M}\sum_{i=0}^{M-1}\phi^\pi(s_{t,i})\phi^\pi(s'_{t,i})^\intercal(w_t - w_t^*) \\
&\quad\quad + g(B_t, w_t, \theta_t) - \bar{g}(w_t, \theta_t) + \bar{g}(w_t, \theta_t) - \bar{g}(w_t^*, \theta_t)
\end{aligned} \tag{A.13}$$

Let $g(B_t, w_t, \theta_t) := \frac{1}{M}\sum_{i=0}^{M-1}\Big(R^\pi(s_{t,i}) - \rho_t^*\Big)\phi^\pi(s_{t,i}) + \frac{1}{M}\sum_{i=0}^{M-1}\Big(\phi^\pi(s_{t,i})(\phi^\pi(s'_{t,i}) - \phi^\pi(s_{t,i}))^\intercal - \eta I\Big)w_t$

Let $\bar{g}(w_t, \theta_t) := \int d(s, \pi(\theta_t))\phi^\pi(s)\Big(R^\pi(s) - \rho_t^* + \int P^\pi(s'|s)\phi^\pi(s')^\intercal w_t\, ds' - \phi^\pi(s)^\intercal w_t\Big)\, ds$

Using (A.13) in (A.12):

$$\begin{aligned}
w_{t+1} - w^*_{t+1} =& w_t - w^*_t + w^*_t - w^*_{t+1} + \\
&+ \alpha_t \frac{1}{M} \sum_{i=0}^{M-1} (\rho^*_t - \bar{\rho}_t) \phi^\pi(s_{t,i}) \\
&+ \alpha_t \frac{1}{M} \sum_{i=0}^{M-1} \phi^\pi(s_{t,i}) \phi^\pi(s'_{t,i})^\intercal (\bar{w}_t - w^*_t) \\
&\alpha_t \frac{1}{M} \sum_{i=0}^{M-1} \phi^\pi(s_{t,i}) \phi^\pi(s'_{t,i})^\intercal (w^*_t - w_t) \\
&+ \alpha_t (g(B_t, w_t, \theta_t) - \bar{g}(w_t, \theta_t)) \\
&+ \alpha_t (\bar{g}(w_t, \theta_t) - \bar{g}(w^*_t, \theta_t))
\end{aligned}$$

$$\begin{aligned}
\text{Let,} \quad f(B_t, w_t, \theta_t) :=& \frac{1}{M} \sum_{i=0}^{M-1} (\rho^*_t - \bar{\rho}_t) \phi^\pi(s_{t,i}) \\
&+ \frac{1}{M} \sum_{i=0}^{M-1} \phi^\pi(s_{t,i}) \phi^\pi(s'_{t,i})^\intercal (\bar{w}_t - w^*_t) \\
&+ \frac{1}{M} \sum_{i=0}^{M-1} \phi^\pi(s_{t,i}) \phi^\pi(s'_{t,i})^\intercal (w^*_t - w_t) \\
&+ (g(B_t, w_t, \theta_t) - \bar{g}(w_t, \theta_t)) \\
&+ (\bar{g}(w_t, \theta_t) - \bar{g}(w^*_t, \theta_t))
\end{aligned}$$

$$\begin{aligned}
||w_{t+1} - w^*_{t+1}||^2 &= \|\Gamma_{C_w}\big(w_t + \alpha_t f(B_t, w_t, \theta_t)\big) - w^*_{t+1}\| \\
&\leq \|w_t + \alpha_t f(B_t, w_t, \theta_t)) - w^*_{t+1}\| \quad (\because \text{Projection is a non-expansive operator}) \\
&\leq ||(w_t - w^*_t) + (w^*_t - w^*_{t+1}) + \alpha_t f(B_t, w_t, \theta_t)||^2 \\
&\leq ||w_t - w^*_t||^2 + ||w^*_t - w^*_{t+1}||^2 \\
&\quad + \alpha_t^2 ||f(B_t, w_t, \theta_t)||^2 \\
&\quad + 2\langle \Delta w_t, w^*_t - w^*_{t+1} \rangle + 2\alpha_t \langle \Delta w_t, f(B_t, w_t, \theta_t) \rangle \\
&\quad + 2\alpha_t \langle w^*_t - w^*_{t+1}, f(B_t, w_t, \theta_t) \rangle
\end{aligned}$$

$$
\begin{aligned}
\mathbb{E}||w_{t+1} - w^*_{t+1}||^2 \leq\ & \mathbb{E}||\Delta w_t||^2 + 2\mathbb{E}||w^*_t - w^*_{t+1}||^2 \\
& + 2\alpha_t^2 \mathbb{E}||f(B_t, w_t, \theta_t)||^2 \\
& + 2\mathbb{E}\langle \Delta w_t, w^*_t - w^*_{t+1}\rangle \\
& + 2\alpha_t \mathbb{E}\langle \Delta w_t, f(B_t, w_t, \theta_t)\rangle \\
=\ & \mathbb{E}||\Delta w_t||^2 + 2\mathbb{E}||w^*_t - w^*_{t+1}||^2 \ \textcircled{1} \\
& + 2\alpha_t^2 \mathbb{E}||f(B_t, w_t, \theta_t)||^2 \ \textcircled{2} \\
& + 2\mathbb{E}\langle \Delta w_t, w^*_t - w^*_{t+1}\rangle \ \textcircled{3} \\
& + 2\alpha_t \mathbb{E}\langle \Delta w_t, \frac{1}{M}\sum_{i=0}^{M-1} (\bar{\rho}_t - \rho^*_t)\phi^\pi(s_{t,i})\rangle \ \textcircled{4} \\
& + 2\alpha_t \mathbb{E}\langle \Delta w_t, \frac{1}{M}\sum_{i=0}^{M-1} \phi^\pi(s_{t,i})\phi^\pi(s'_{t,i})^\intercal(\bar{w}_t - w^*_t)\rangle \ \textcircled{5} \\
& + 2\alpha_t \mathbb{E}\langle \Delta w_t, \frac{1}{M}\sum_{i=0}^{M-1} \phi^\pi(s_{t,i})\phi^\pi(s'_{t,i})^\intercal(w^*_t - w_t)\rangle \ \textcircled{6} \\
& + 2\alpha_t \mathbb{E}\langle \Delta w_t, g(B_t, w_t, \theta_t) - \bar{g}(w_t, \theta_t)\rangle \ \textcircled{7} \\
& + 2\alpha_t \mathbb{E}\langle \Delta w_t, \bar{g}(w_t, \theta_t) - \bar{g}(w^*_t, \theta_t)\rangle \ \textcircled{8}
\end{aligned} \tag{A.14}
$$

From (A.14):
①:

$$
\begin{aligned}
\mathbb{E}||w^*_t - w^*_{t+1}||^2 &\leq L_w^2 \mathbb{E}||\theta_{t+1} - \theta_t||^2 \quad \text{(using lemma A.21)} \\
&\leq L_w^2 \gamma_t^2 G_\theta^2 \quad \text{(using lemma A.22)}
\end{aligned}
$$

②:

$$
\begin{aligned}
&\mathbb{E}||f(B_t, w_t, \theta_t)||^2 \\
&= \mathbb{E}||\frac{1}{M}\sum_{i=0}^{M-1} (R^\pi(s_{t,i}) - \bar{\rho}_t + \phi^\pi(s'_{t,i})^\intercal \bar{w}_t - \phi^\pi(s_{t,i})^\intercal w_t)\phi^\pi(s_{t,i}) - \eta w_t||^2 \\
&\leq \mathbb{E}\big(||\frac{1}{M}\sum_{i=0}^{M-1} (R^\pi(s_{t,i}) - \bar{\rho}_t + \phi^\pi(s'_{t,i})^\intercal \bar{w}_t - \phi^\pi(s_{t,i})^\intercal w_t)\phi^\pi(s_{t,i})|| + \eta||w_t||\big)^2
\end{aligned}
$$

Here,

$$
\begin{aligned}
||\phi^\pi(s)|| &< 1 \quad \text{(Assumption 4.1)} \\
|R^\pi(s)| &\leq C_r \quad \text{(Assumption 4.2)} \\
||w_t|| &\leq C_w \quad \text{(Algorithm 2, step 8)} \\
|\rho_t| &\leq C_r + 2C_w \quad \text{(lemma A.23)} \\
||\bar{w}_t|| &\leq C_w \quad \text{(lemma A.24)} \\
||\bar{\rho}_t|| &\leq C_r + 2C_w \quad \text{(lemma A.25)}
\end{aligned}
$$

$$
\begin{aligned}
&\leq \mathbb{E}\Big(\frac{1}{M}\sum_{i=0}^{M-1} ||(R^\pi(s_{t,i}) - \bar{\rho}_t + \phi^\pi(s'_{t,i})^\intercal \bar{w}_t - \phi^\pi(s_{t,i})^\intercal w_t)\phi^\pi(s_{t,i})|| + \eta||w_t||\Big)^2 \\
&\leq \mathbb{E}(C_r + C_r + 2C_w + 2C_w + \eta C_w)^2 \\
&\leq \mathbb{E}(C_\delta)^2 \quad (C_\delta = 2C_r + (4+\eta)C_w) \\
&\leq C_\delta^2
\end{aligned}
$$

③:

$$\begin{aligned}\mathbb{E}\langle \Delta w_t, w_t^* - w_{t+1}^* \rangle &\leq \mathbb{E}||\Delta w_t||\,||w_t^* - w_{t+1}^*|| \\ &\leq L_w \mathbb{E}||\Delta w_t||\,||\theta_{t+1} - \theta_t|| \quad \text{(using Lemma A.21)}\end{aligned}$$

④:

$$\begin{aligned}\mathbb{E}[\langle \Delta w_t, \frac{1}{M}\sum_{i=0}^{M-1}(\rho_t^* - \bar{\rho}_t)\phi^\pi(s_{t,i})\rangle] &= \mathbb{E}[\frac{1}{M}\sum_{i=0}^{M-1}\langle \Delta w_t, \phi^\pi(s_{t,i})\rangle(\rho_t^* - \bar{\rho}_t)] \\ &\leq \mathbb{E}[\frac{1}{M}\sum_{i=0}^{M-1}||\Delta w_t||||\phi^\pi(s_{t,i})||||(\rho_t^* - \bar{\rho}_t)|] \\ &\leq \mathbb{E}||\Delta w_t|||\rho_t^* - \bar{\rho}_t| \\ &= \mathbb{E}||\Delta w_t|||\Delta\bar{\rho}_t|\end{aligned}$$

⑤:

$$\begin{aligned}\mathbb{E}\langle \Delta w_t, \frac{1}{M}\sum_{i=0}^{M-1}\phi^\pi(s_{t,i})\phi^\pi(s'_{t,i})^\intercal(\bar{w}_t - w_t^*)\rangle &\leq \mathbb{E}||\Delta w_t||\,||\frac{1}{M}\sum_{i=0}^{M-1}\phi^\pi(s_{t,i})\phi^\pi(s'_{t,i})^\intercal(\bar{w}_t - w_t^*)|| \\ &\leq \mathbb{E}||\Delta w_t||\,||\bar{w}_t - w_t^*|| \\ &\leq \mathbb{E}||\Delta w_t||\,||\Delta\bar{w}_t||\end{aligned}$$

⑥:

$$\begin{aligned}\mathbb{E}\langle \Delta w_t, \frac{1}{M}\sum_{i=0}^{M-1}\phi^\pi(s_{t,i})\phi^\pi(s'_{t,i})^\intercal(w_t^* - w_t)\rangle &\leq \mathbb{E}||\Delta w_t||\,||\frac{1}{M}\sum_{i=0}^{M-1}\phi^\pi(s_{t,i})\phi^\pi(s'_{t,i})^\intercal(w_t^* - w_t)|| \\ &\leq \mathbb{E}||\Delta w_t||\,||w_t^* - w_t|| \\ &\leq \mathbb{E}||\Delta w_t||^2\end{aligned}$$

⑦:

$$\begin{aligned}&\mathbb{E}[\langle \Delta w_t, g(B_t, w_t, \theta_t) - \bar{g}(w_t, \theta_t)\rangle] = \mathbb{E}[\langle \Delta w_t, \mathbb{E}[g(B_t, w_t, \theta_t) - \bar{g}(w_t, \theta_t)|\Delta w_t]\rangle] \\ &\text{Note:}\quad \mathbb{E}[g(B_t, w_t, \theta_t) - \bar{g}(w_t, \theta_t)] = 0 \\ &\text{Hence,}\quad \mathbb{E}[\langle \Delta w_t, g(B_t, w_t, \theta_t) - \bar{g}(w_t, \theta_t)\rangle] = 0\end{aligned}$$

⑧:

$$\begin{aligned}&\mathbb{E}[\langle \Delta w_t, \bar{g}(w_t, \theta_t) - \bar{g}(w_t^*, \theta_t)\rangle] \\ &\qquad A(\theta_t) = \int_S d^\pi(s, \theta_t)\big(\phi^\pi(s)(\mathbb{E}[\phi^\pi(s')] - \phi^\pi(s))^\intercal - \eta I\big)ds \\ &\qquad b(\theta_t) = \int_S d^\pi(s, \theta_t) r^\pi(s)\phi^\pi(s)ds \\ &\bar{g}(w_t, \theta_t) - \bar{g}(w_t^*, \theta_t) = b(\theta_t) + A(\theta_t)w_t - b(\theta_t) - A(\theta_t)w_t^* \\ &\qquad = A(\theta_t)(w_t - w_t^*) \\ &\text{Now,}\quad \mathbb{E}[\langle \Delta w_t, \bar{g}(w_t, \theta_t) - \bar{g}(w_t^*, \theta_t)\rangle] = \mathbb{E}[\langle \Delta w_t, A(\theta_t)\Delta w_t\rangle] \\ &\qquad = \mathbb{E}[\Delta w_t^\intercal A(\theta_t)\Delta w_t] \\ &\qquad \leq -\lambda\mathbb{E}||\Delta w_t||^2 \qquad \text{(Lemma A.26)}\end{aligned}$$

Combining ① - ⑧ into (A.14):

$$\begin{aligned}\mathbb{E}||w_{t+1} - w^*_{t+1}||^2 \leq (1 - 2(\lambda-1)\alpha_t)\mathbb{E}||\Delta w_t||^2 + 2L_w^2\gamma_t^2 G_\theta^2 + 2\alpha_t^2 C_\delta^2 \\ + 2L_w\mathbb{E}||\Delta w_t||||\theta_{t+1} - \theta_t|| + 2\alpha_t\mathbb{E}||\Delta w_t|||\Delta\bar{\rho}_t| \\ + 2\alpha_t\mathbb{E}||\Delta w_t||||\Delta\bar{w}_t||\end{aligned}$$

$$\begin{aligned}\implies 2(\lambda-1)\alpha_t\mathbb{E}||\Delta w_t||^2 \leq \mathbb{E}[||\Delta w_t||^2] - \mathbb{E}||\Delta w_{t+1}||^2 + 2L_w^2\gamma_t^2 G_\theta^2 + 2\alpha_t^2 C_\delta^2 \\ + 2L_w\gamma_t G_\theta\mathbb{E}||\Delta w_t|| + 2\alpha_t\mathbb{E}||\Delta w_t|||\Delta\bar{\rho}_t| \\ + 2\alpha_t\mathbb{E}||\Delta w_t||||\Delta\bar{w}_t||\end{aligned}$$

$$\begin{aligned}\implies \mathbb{E}||\Delta w_t||^2 \leq \frac{1}{2(\lambda-1)\alpha_t}(\mathbb{E}||\Delta w_t||^2 - \mathbb{E}||w_{t+1}||^2) \\ + \Big(\frac{L_w^2\gamma_t^2}{(\lambda-1)\alpha_t}G_\theta^2 + \frac{\alpha_t}{(\lambda-1)}C_\delta^2\Big) \\ + \frac{L_w}{(\lambda-1)}\frac{\gamma_t}{\alpha_t}G_\theta\mathbb{E}||\Delta w_t|| \\ + \frac{\mathbb{E}||\Delta w_t|||\Delta\bar{\rho}_t|}{(\lambda-1)} \\ + \frac{\mathbb{E}||\Delta w_t||||\Delta\bar{w}_t||}{(\lambda-1)}\end{aligned}$$

$$\begin{aligned}\implies \sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2 \leq \sum_{t=0}^{T-1}\frac{1}{2(\lambda-1)\alpha_t}(\mathbb{E}||\Delta w_t||^2 - \mathbb{E}||\Delta w_{t+1}||^2) \ \textcircled{1} \\ + \sum_{t=0}^{T-1}\Big(\frac{L_w}{(\lambda-1)}\frac{\gamma_t^2}{\alpha_t}G_\theta^2 + \frac{\alpha_t}{(\lambda-1)}C_\delta^2\Big) \ \textcircled{2} \\ + \sum_{t=0}^{T-1}\frac{L_w}{(\lambda-1)}\frac{\gamma_t}{\alpha_t}G_\theta\mathbb{E}||\Delta w_t|| \ \textcircled{3} \\ + \sum_{t=0}^{T-1}\frac{\mathbb{E}||\Delta w_t||\big(|\Delta\bar{\rho}_t| + ||\Delta\bar{w}_t||\big)}{(\lambda-1)} \ \textcircled{4}\end{aligned} \tag{A.15}$$

From (A.15):
①:

$$\begin{aligned}\frac{1}{2(\lambda-1)}\sum_{t=0}^{T-1}(\mathbb{E}||\Delta w_t||^2 - \mathbb{E}||\Delta w_{t+1}||^2)\frac{1}{\alpha_t} &= \frac{1}{2(\lambda-1)}\Big(\sum_{t=1}^{T-1}\Big(\frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}}\Big)\mathbb{E}||\Delta w_t||^2 \\ &\quad + \frac{1}{\alpha_0}\mathbb{E}||\Delta w_0||^2 - \frac{1}{\alpha_{T-1}}\mathbb{E}||\Delta w_T||^2\Big) \\ &\leq \frac{1}{2(\lambda-1)}\left(\sum_{t=1}^{T-1}\Big(\frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}}\Big) + \frac{1}{\alpha_0}\right)4C_w^2 \\ &\leq \frac{4C_w^2}{2(\lambda-1)\alpha_{T-1}} = \frac{C_w^2}{(\lambda-1)C_\alpha}T^\sigma \quad (\because \alpha_t = \frac{C_\alpha}{(1+t)^\alpha})\end{aligned}$$

②:

$$\begin{aligned}\sum_{t=0}^{T-1}\Big(\frac{L_w^2}{(\lambda-1)}\frac{\gamma_t^2}{\alpha_t}G_\theta^2+\frac{\alpha_t}{(\lambda-1)}C_\delta^2\Big)&=\sum_{t=0}^{T-1}\Big(\frac{L_w^2}{(\lambda-1)}\frac{\gamma_t^2}{\alpha_t^2}G_\theta^2+\frac{C_\delta^2}{(\lambda-1)}\Big)\alpha_t\\&\leq\sum_{t=0}^{T-1}\Big(\frac{L_w^2}{(\lambda-1)}\max_t\frac{\gamma_t^2}{\alpha_t^2}G_\theta^2+\frac{C_\delta^2}{(\lambda-1)}\Big)\alpha_t\\&=\sum_{t=0}^{T-1}C_g\alpha_t=\frac{C_gC_\alpha}{1-\sigma}T^{1-\sigma}\quad\Big(C_g=\frac{L_w^2}{(\lambda-1)}\max_t\frac{\gamma_t^2}{\alpha_t^2}G_\theta^2+\frac{C_\delta^2}{(\lambda-1)}\Big)\end{aligned}$$

③:

$$\begin{aligned}\sum_{t=0}^{T-1}\frac{L_w}{(\lambda-1)}\frac{\gamma_t}{\alpha_t}G_\theta\mathbb{E}||\Delta w_t||&=\frac{L_w}{(\lambda-1)}G_\theta\sum_{t=0}^{T-1}\frac{\gamma_t}{\alpha_t}\mathbb{E}||\Delta w_t||\\&\leq\frac{L_w}{(\lambda-1)}G_\theta\Big(\sum_{t=0}^{T-1}\Big(\frac{\gamma_t}{\alpha_t}\Big)^2\Big)^{\frac{1}{2}}\Big(\sum_{t=0}^{T-1}(\mathbb{E}||\Delta w_t||)^2\Big)^{\frac{1}{2}}\\&\text{(Using Cauchy Schwartz inequality)}\\&\leq\frac{L_w}{(\lambda-1)}G_\theta\Big(\sum_{t=0}^{T-1}\Big(\frac{\gamma_t}{\alpha_t}\Big)^2\Big)^{\frac{1}{2}}\Big(\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2\Big)^{\frac{1}{2}}\\&\text{(Using Jensen's inequality)}\\&\leq\frac{L_wG_\theta C_\gamma}{(\lambda-1)C_\alpha}\Big(\frac{T^{1-2(v-\sigma)}}{1-2(v-\sigma)}\Big)^{\frac{1}{2}}\Big(\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2\Big)^{\frac{1}{2}}\end{aligned}$$

④:

$$\begin{aligned}\frac{1}{(\lambda-1)}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t|||\Delta\bar{\rho}_t|&\leq\frac{1}{(\lambda-1)}\Big(\sum_{t=0}^{T-1}(\mathbb{E}||\Delta w_t||)^2\Big)^{\frac{1}{2}}\Big(\sum_{t=0}^{T-1}\big(\mathbb{E}(|\Delta\bar{\rho}_t|+||\Delta\bar{w}_t||)\big)^2\Big)^{\frac{1}{2}}\\&\leq\frac{1}{(\lambda-1)}\Big(\sum_{t=0}^{T-1}(\mathbb{E}||\Delta w_t||)^2\Big)^{\frac{1}{2}}\Big(\sum_{t=0}^{T-1}\mathbb{E}(|\Delta\bar{\rho}_t|+||\Delta\bar{w}_t||)^2\Big)^{\frac{1}{2}}\\&\leq\frac{1}{(\lambda-1)}\Big(\sum_{t=0}^{T-1}(\mathbb{E}||\Delta w_t||)^2\Big)^{\frac{1}{2}}\Big(2\sum_{t=0}^{T-1}\mathbb{E}(|\Delta\bar{\rho}_t|^2+||\Delta\bar{w}_t||^2)\Big)^{\frac{1}{2}}\end{aligned}$$

Combining ① - ⑤ into (A.15) and dividing by T:

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2\leq&\frac{2C_w^2}{(\lambda-1)C_\alpha}T^{\sigma-1}+\frac{C_gC^\alpha}{1-\sigma}T^{-\sigma}\\&+\frac{L_wG_\theta C_\gamma}{(\lambda-1)C_\alpha}\Big(\frac{T^{-2(v-\sigma)}}{1-2(v-\sigma)}\Big)^{\frac{1}{2}}\Big(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2\Big)^{\frac{1}{2}}\\&+\frac{2^{1/2}}{(\lambda-1)}\Big(\frac{1}{T}\sum_{t=0}^{T-1}(\mathbb{E}||\Delta w_t||)^2\Big)^{\frac{1}{2}}\Big(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big(|\Delta\bar{\rho}_t|^2+||\Delta\bar{w}_t||^2\big)\Big)^{\frac{1}{2}}\end{aligned}$$

Let,

$$M(T) = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2$$
$$N(T) = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}|\Delta\bar{\rho}_t|^2 + \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta\bar{w}_t||^2$$

$$M(T) \leq K_1 + K_2\sqrt{M(T)} + K_3\sqrt{M(T)}\sqrt{N(T)}$$
$$K_1 := \frac{2C_w^2}{(\lambda-1)C_\alpha}T^{\sigma-1} + \frac{C_gC_\alpha}{1-\sigma}T^{-\sigma}$$
$$K_2 := \frac{L_wG_\theta C_\gamma}{(\lambda-1)C_\alpha}\left(\frac{T^{-2(v-\sigma)}}{1-2(v-\sigma)}\right)^{\frac{1}{2}}$$
$$K_3 := \frac{2^{1/2}}{\lambda-1}$$

$$M(T) - 2\frac{K_2}{2}\sqrt{M(T)} - 2\frac{K_3}{2}\sqrt{M(T)}\sqrt{N(T)} + 2\frac{K_2}{2}\frac{K_3}{2}\sqrt{N(T)}$$
$$+\left(\frac{K_2}{2}\right)^2 + \left(\frac{K_3}{2}\sqrt{N(T)}\right)^2 \leq K_1 + \left(\frac{K_2}{2}\right)^2 + \left(\frac{K_3}{2}\sqrt{N(T)}\right)^2 + 2\frac{K_2}{2}\frac{K_3}{2}\sqrt{N(T)}$$
$$\implies \left(\sqrt{M(T)} - \frac{K_2}{2} - \frac{K_3}{2}\sqrt{N(T)}\right)^2 \leq K_1 + \left(\frac{K_2}{2} + \frac{K_3}{2}\sqrt{N(T)}\right)^2$$
$$\implies \sqrt{M(T)} - \frac{K_2}{2} - \frac{K_3}{2}\sqrt{N(T)} \leq \sqrt{K_1} + \frac{K_2}{2} + \frac{K_3}{2}\sqrt{N(T)}$$
$$\implies \sqrt{M(T)} \leq \sqrt{K_1} + K_2 + K_3\sqrt{N(T)}$$
$$\implies M(T) \leq 2(\sqrt{K_1} + K_2)^2 + 2K_3^2N(T)$$

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2 \leq 2\left(\sqrt{\frac{2C_w^2}{(\lambda-1)C_\alpha}T^{\sigma-1} + \frac{C_gC_\alpha}{1-\sigma}T^{-\sigma}} + \frac{L_wG_\theta C_\gamma}{(\lambda-1)C_\alpha}\left(\frac{T^{-2(v-\sigma)}}{1-2(v-\sigma)}\right)^{\frac{1}{2}}\right)^2$$
$$+ \frac{4}{(\lambda-1)^2}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}|\Delta\bar{\rho}_t|^2 + \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta\bar{w}_t||^2\right)$$

□

**Lemma A.15.** *Let the cumulative error of target linear critic be $\sum_{t=0}^{T-1}\mathbb{E}||\Delta\bar{w}_t||^2$ and cumulative error of linear critic be $\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2$. $\bar{w}_t$ and $w_t$ are target linear critic parameter and linear critic parameter at time t respectively. Bound on the cumulative error of target linear critic parameter is proven using cumulative error of linear critic parameter as follows:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\Delta\bar{w}_t\|^2 \leq 2\left(\sqrt{\frac{2C_w^2T^{u-1}}{C_\beta} + \frac{C_{gt}C_\beta T^{-u}}{1-u}}\right.$$
$$\left. + L_pG_\theta C_\gamma\left(\frac{T^{-v}}{(1-2v)^{1/2}} + \frac{T^{-(v-u)}}{C_\beta(1-2(v-u))^{1/2}}\right)\right)^2$$
$$+ \frac{2}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_{t+1}||^2$$

*Here, $\Delta w_t = w_t - w_t^*$, $\Delta \bar{w}_t = \bar{w}_t - w_t^*$. $\bar{w}_t^*$ and $w_t^*$ are the optimal parameters given by TD(0) algorithm corresponding to policy parameter $\theta_t$. $C_\beta$, $C_\gamma$, $u$, and $v$ are constants defined in Assumption 3.5, $\|w_t\| \leq C_w$ (Algorithm 2, step 8), $C_r$ is the upper bound on rewards (Assumption 4.2), Constant $G_\theta$ is defined in Lemma A.22. $C_{gt} = L_p^2 G_\theta^2 \max_t(\gamma_t^2/\beta_t^2) + 4C_w^2$. $L_p$ is Lipchitz constant defined in Lemma A.29.*

*Proof.*

$$
\begin{aligned}
\bar{w}_{t+1} =& \bar{w}_t + \beta_t(w_{t+1} - \bar{w}_t) \\
\implies \bar{w}_{t+1} - w_{t+1}^* =& \bar{w}_t - w_t^* + w_t^* - w_{t+1}^* + \beta_t(w_{t+1} - \bar{w}_t) \\
\implies \|\Delta\bar{w}_{t+1}\|^2 =& \|\Delta\bar{w}_t + w_t^* - w_{t+1}^* + \beta_t(w_{t+1} - \bar{w}_t)\|^2 \\
\leq& \|\Delta\bar{w}_t\|^2 + 2\|w_t^* - w_{t+1}^*\|^2 + 2\|\beta_t(w_{t+1} - \bar{w}_t)\|^2 \\
&+ 2\beta\langle\Delta\bar{w}_t, w_{t+1} - \bar{w}_t\rangle + 2\langle\Delta\bar{w}_t, w_t^* - w_{t+1}^*\rangle \\
=& \|\Delta\bar{w}_t\|^2 + 2\|w_t^* - w_{t+1}^*\|^2 + 2\|\beta_t(w_{t+1} - \bar{w}_t)\|^2 \\
&+ 2\beta_t\langle\Delta\bar{w}_t, \Delta w_{t+1} - \Delta\bar{w}_t\rangle + 2\beta_t\langle\Delta\bar{w}_t, w_{t+1}^* - w_t^*\rangle \\
&+ 2\langle\Delta\bar{w}_t, w_t^* - w_{t+1}^*\rangle \\
\implies \|\Delta\bar{w}_{t+1}\|^2 \leq& (1 - 2\beta_t)\|\Delta\bar{w}_t\|^2 + 2\|w_t^* - w_{t+1}^*\|^2 + 2\|\beta_t(w_{t+1} - \bar{w}_t)\|^2 \\
&+ 2\beta_t\langle\Delta\bar{w}_t, \Delta w_{t+1}\rangle + 2\beta_t\langle\Delta\bar{w}_t, w_{t+1}^* - w_t^*\rangle \\
&+ 2\langle\Delta\bar{w}_t, w_t^* - w_{t+1}^*\rangle
\end{aligned}
$$

$$
\begin{aligned}
\implies \|\Delta\bar{w}_t\|^2 =& \frac{1}{2\beta_t}\big(\|\Delta\bar{w}_t\|^2 - \|\Delta\bar{w}_{t+1}\|^2\big) + \big(\frac{1}{\beta_t}\|w_t^* - w_{t+1}^*\|^2 + \beta_t\|(w_{t+1} - \bar{w}_t)\|^2\big) \\
&+ \langle\Delta\bar{w}_t, \Delta w_{t+1}\rangle + \langle\Delta\bar{w}_t, w_{t+1}^* - w_t^*\rangle + \frac{1}{\beta_t}\langle\Delta\bar{w}_t, w_t^* - w_{t+1}^*\rangle
\end{aligned}
$$

$$
\begin{aligned}
\implies \sum_{t=0}^{T-1}\mathbb{E}\|\Delta\bar{w}_t\|^2 =& \sum_{t=0}^{T-1}\frac{1}{2\beta_t}\big(\mathbb{E}\|\Delta\bar{w}_t\|^2 - \mathbb{E}\|\Delta\bar{w}_{t+1}\|^2\big) \ \textcircled{1} \\
&+ \sum_{t=0}^{T-1}\big(\frac{1}{\beta_t}\mathbb{E}\|w_t^* - w_{t+1}^*\|^2 + \beta_t\mathbb{E}\|(w_{t+1} - \bar{w}_t)\|^2\big) \ \textcircled{2} \\
&+ \sum_{t=0}^{T-1}\mathbb{E}\langle\Delta\bar{w}_t, \Delta w_{t+1}\rangle \ \textcircled{3} \\
&+ \sum_{t=0}^{T-1}\mathbb{E}\langle\Delta\bar{w}_t, w_{t+1}^* - w_t^*\rangle \ \textcircled{4} \\
&+ \sum_{t=0}^{T-1}\frac{1}{\beta_t}\mathbb{E}\langle\Delta\bar{w}_t, w_t^* - w_{t+1}^*\rangle \ \textcircled{5}
\end{aligned} \tag{A.16}
$$

From A.16:

①:

$$
\begin{aligned}
&\sum_{t=0}^{T-2} \frac{1}{2\beta_t}\big(\mathbb{E}\|\Delta\bar{w}_t\|^2 - \mathbb{E}\|\Delta\bar{w}_{t+1}\|^2\big) \\
&= \frac{1}{2}\Big(\sum_{t=0}^{T-1}\big(\frac{1}{\beta_{t+1}} - \frac{1}{\beta_t}\big)\mathbb{E}\|\Delta\bar{w}_{t+1}\|^2 + \frac{1}{\beta_0}\mathbb{E}\|\Delta\bar{w}_0\|^2 - \frac{1}{\beta_{T-2}}\mathbb{E}\|\Delta\bar{w}_T\|^2\Big) \\
&\leq \frac{1}{2}\Big(\sum_{t=0}^{T-1}\big(\frac{1}{\beta_{t+1}} - \frac{1}{\beta_t}\big)\mathbb{E}\|\Delta\bar{w}_{t+1}\|^2 + \frac{1}{\beta_0}\mathbb{E}\|\Delta\bar{w}_0\|^2\Big) \\
&\leq \frac{1}{2}\Big(\sum_{t=0}^{T-2}\big(\frac{1}{\beta_{t+1}} - \frac{1}{\beta_t}\big) + \frac{1}{\beta_0}\Big)4C_w^2 \quad \text{(Using Lemma A.25)} \\
&= \frac{2C_w^2}{\beta_{T-1}} = \frac{2C_w^2 T^u}{C_\beta}
\end{aligned}
$$

②:

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\big(\frac{1}{\beta_t}\mathbb{E}\|w_t^* - w_{t+1}^*\|^2 + \beta_t\mathbb{E}\|(w_{t+1} - \bar{w}_t)\|^2\big) \\
&\leq \sum_{t=0}^{T-1}\big(\frac{L_p^2}{\beta_t}\mathbb{E}\|\theta_t - \theta_{t+1}\|^2 + \beta_t\mathbb{E}\|(w_{t+1} - \bar{w}_t)\|^2\big) \quad \text{(Using Lemma A.29 )} \\
&\leq \sum_{t=0}^{T-1}\big(L_p^2 G_\theta^2\frac{\gamma_t^2}{\beta_t} + 4\beta_t C_w^2\big) \quad \text{(Using Lemma A.22,A.23, and A.26 )} \\
&\leq \sum_{t=0}^{T-1}\big(L_p^2 G_\theta^2\max_t\frac{\gamma_t^2}{\beta_t^2} + 4C_w^2\big)\beta_t \\
&= \sum_{t=0}^{T-1} C_{gt}\beta_t = \frac{C_{gt}C_\beta T^{1-u}}{1-u} \quad (C_{gt} = L_p^2 G_\theta^2\max_t\frac{\gamma_t^2}{\beta_t^2} + 4C_w^2)
\end{aligned}
$$

③:

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}\langle\Delta\bar{w}_t, \Delta w_{t+1}\rangle \\
&\leq \Big(\sum_{t=0}^{T-1}\mathbb{E}\|\Delta\bar{w}_t\|^2\Big)^{1/2}\Big(\sum_{t=0}^{T-1}\mathbb{E}\|\Delta w_{t+1}\|^2\Big)^{1/2} \quad \text{(Using Cauchy-Schwarz inequality)}
\end{aligned}
$$

④:

$$
\begin{aligned}
&\sum_{t=0}^{T-1} \mathbb{E}\langle \Delta\bar{w}_t, w^*_{t+1} - w^*_t \rangle \\
&\leq \sum_{t=0}^{T-1} \mathbb{E}\|\Delta\bar{w}_t\| \|w^*_{t+1} - w^*_t\| \\
&\leq \sum_{t=0}^{T-1} L_p G_\theta \gamma_t \mathbb{E}\|\Delta\bar{w}_t\| \quad \text{(Using Lemma A.22,A.29)} \\
&\leq L_p G_\theta \Big( \sum_{t=0}^{T-1} \gamma_t^2 \Big)^{1/2} \Big( \sum_{t=0}^{T-1} \mathbb{E}\|\Delta\bar{w}_t\|^2 \Big)^{1/2} \quad \text{(Using Cauchy-Schwarz inequality)} \\
&\leq L_p G_\theta C_\gamma \frac{T^{-v}}{(1-2v)^2} T^{1/2} \Big( \sum_{t=0}^{T-1} \mathbb{E}\|\Delta\bar{w}_t\|^2 \Big)^{1/2}
\end{aligned}
$$

⑤:

$$
\begin{aligned}
&\sum_{t=0}^{T-1} \frac{1}{\beta_t} \mathbb{E}\langle \Delta\bar{w}_t, w^*_{t+1} - w^*_t \rangle \\
&\leq \sum_{t=0}^{T-1} \frac{1}{\beta_t} \mathbb{E}\|\Delta\bar{w}_t\| \|w^*_{t+1} - w^*_t\| \\
&\leq \sum_{t=0}^{T-1} L_p G_\theta \frac{\gamma_t}{\beta_t} \mathbb{E}\|\Delta\bar{w}_t\| \quad \text{(Using Lemma A.22,A.29)} \\
&\leq L_p G_\theta \Big( \sum_{t=0}^{T-1} \frac{\gamma_t^2}{\beta_t^2} \Big)^{1/2} \Big( \sum_{t=0}^{T-1} \mathbb{E}\|\Delta\bar{w}_t\|^2 \Big)^{1/2} \quad \text{(Using Cauchy-Schwarz inequality)} \\
&\leq \frac{L_p G_\theta C_\gamma}{C_\beta} \frac{T^{-(v-u)}}{(1-2(v-u))^2} T^{1/2} \Big( \sum_{t=0}^{T-1} \mathbb{E}\|\Delta\bar{w}_t\|^2 \Big)^{1/2}
\end{aligned}
$$

Combining ①-⑤ into A.16:

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\Delta\bar{w}_t\|^2 =& \frac{2C_w^2 T^{u-1}}{C_\beta} + \frac{C_{gt} C_\beta T^{-u}}{1-u} \\
&+ \Big( \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\Delta\bar{w}_t\|^2 \Big)^{1/2} \Big( \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\Delta w_{t+1}\|^2 \Big)^{1/2} \\
&+ L_p G_\theta C_\gamma \frac{T^{-v}}{(1-2v)^2} \Big( \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\Delta\bar{w}_t\|^2 \Big)^{1/2} \\
&+ \frac{L_p G_\theta C_\gamma}{C_\beta} \frac{T^{-(v-u)}}{(1-2(v-u))^2} \Big( \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\Delta\bar{w}_t\|^2 \Big)^{1/2}
\end{aligned} \tag{A.17}
$$

$$M(T) = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\Delta\bar{w}_t\|^2$$
$$N(T) = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\Delta w_{t+1}\|^2$$
$$M(T) \leq K_1 + K_2\sqrt{M(T)} + K_3\sqrt{M(T)N(T)}$$

Here,

$$K_1 = \frac{2C_w^2T^{u-1}}{C_\beta} + \frac{C_{gt}C_\beta T^{1-u}}{1-u}$$
$$K_2 = L_pG_\theta C_\gamma\frac{T^{-v}}{(1-2v)^2} + \frac{L_pG_\theta C_\gamma}{C_\beta}\frac{T^{-(v-u)}}{(1-2(v-u))^2}$$
$$K_3 = 1$$

From Lemma A.14, we know that

$$M(T) \leq 2(\sqrt{K_1} + K_2)^2 + 2K_3^2N(T)$$

Hence,

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\Delta\bar{w}_t\|^2 \leq& 2\Bigg(\sqrt{\frac{2C_w^2T^{u-1}}{C_\beta} + \frac{C_{gt}C_\beta T^{1-u}}{1-u}} \\ &+ L_pG_\theta C_\gamma\frac{T^{-v}}{(1-2v)^2} + \frac{L_pG_\theta C_\gamma}{C_\beta}\frac{T^{-(v-u)}}{(1-2(v-u))^2}\Bigg)^2 \\ &+ \frac{2}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\Delta w_{t+1}\|^2\end{aligned} \tag{A.18}$$

□

**Lemma A.16.** *Let the cumulative error of target average reward estimator be $\sum_{t=0}^{T-1}\mathbb{E}||\Delta\bar{\rho}_t||^2$ and cumulative error of average reward estimator be $\sum_{t=0}^{T-1}\mathbb{E}||\Delta\rho_t||^2$. $\bar{\rho}_t$ and $\rho_t$ are target average reward estimator and average reward estimator at time t respectively. Bound on the cumulative error of target average reward estimator is proven using cumulative error of average reward estimator as follows:*

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\Delta\bar{\rho}_t\|^2 \leq 2\Bigg(&\sqrt{\frac{2(C_r+2C_w)^2T^{u-1}}{C_\beta} + \frac{C_{st}C_\beta T^{-u}}{1-u}} \\ &+ L_pG_\theta C_\gamma\Big(\frac{T^{-v}}{(1-2v)^{1/2}} + \frac{T^{-(v-u)}}{C_\beta(1-2(v-u))^{1/2}}\Big)\Bigg)^2 \\ &+ \frac{2}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta\rho_{t+1}||^2\end{aligned}$$

*Here, $\Delta\rho_t = \rho_t - \rho_t^*$, $\Delta\bar{\rho}_t = \bar{\rho}_t - \rho_t^*$. $\bar{\rho}_t^*$ and $\rho_t^*$ are the optimal parameters given by TD(0) algorithm corresponding to policy parameter $\theta_t$. $C_\beta$, $C_\gamma$, $u$, and $v$ are constants defined in Assumption 3.5, $\|w_t\| \leq C_w$ (Algorithm 2, step 8), $C_r$ is the upper bound on rewards (Assumption 4.2), Constant $G_\theta$ is defined in Lemma A.22. $C_{st} = L_p^2G_\theta^2\max_t(\gamma_t^2/\beta_t^2) + 4(C_r+2C_w)^2$. $L_p$ is Lipchitz constant defined in Lemma A.29.*

*Proof.*

$$
\begin{aligned}
\bar{\rho}_{t+1} =& \bar{\rho}_t + \beta_t(\rho_{t+1} - \bar{\rho}_t) \\
\implies \bar{\rho}_{t+1} - \rho^*_{t+1} =& \bar{\rho}_t - \rho^*_t + \rho^*_t - \rho^*_{t+1} + \beta_t(\rho_{t+1} - \bar{\rho}_t) \\
\implies \|\Delta\bar{\rho}_{t+1}\|^2 =& \|\Delta\bar{\rho}_t + \rho^*_t - \rho^*_{t+1} + \beta_t(\rho_{t+1} - \bar{\rho}_t)\|^2 \\
\leq& \|\Delta\bar{\rho}_t\|^2 + 2\|\rho^*_t - \rho^*_{t+1}\|^2 + 2\|\beta_t(\rho_{t+1} - \bar{\rho}_t)\|^2 \\
&+ 2\beta\langle\Delta\bar{\rho}_t, \rho_{t+1} - \bar{\rho}_t\rangle + 2\langle\Delta\bar{\rho}_t, \rho^*_t - \rho^*_{t+1}\rangle \\
=& \|\Delta\bar{\rho}_t\|^2 + 2\|\rho^*_t - \rho^*_{t+1}\|^2 + 2\|\beta_t(\rho_{t+1} - \bar{\rho}_t)\|^2 \\
&+ 2\beta_t\langle\Delta\bar{\rho}_t, \Delta\rho_{t+1} - \Delta\bar{\rho}_t\rangle + 2\beta_t\langle\Delta\bar{\rho}_t, \rho^*_{t+1} - \rho^*_t\rangle \\
&+ 2\langle\Delta\bar{\rho}_t, \rho^*_t - \rho^*_{t+1}\rangle \\
\implies \|\Delta\bar{\rho}_{t+1}\|^2 \leq& (1 - 2\beta_t)\|\Delta\bar{\rho}_t\|^2 + 2\|\rho^*_t - \rho^*_{t+1}\|^2 + 2\|\beta_t(\rho_{t+1} - \bar{\rho}_t)\|^2 \\
&+ 2\beta_t\langle\Delta\bar{\rho}_t, \Delta\rho_{t+1}\rangle + 2\beta_t\langle\Delta\bar{\rho}_t, \rho^*_{t+1} - \rho^*_t\rangle \\
&+ 2\langle\Delta\bar{\rho}_t, \rho^*_t - \rho^*_{t+1}\rangle
\end{aligned}
$$

$$
\begin{aligned}
\implies \|\Delta\bar{\rho}_t\|^2 =& \frac{1}{2\beta_t}\big(\|\Delta\bar{\rho}_t\|^2 - \|\Delta\bar{\rho}_{t+1}\|^2\big) + \big(\frac{1}{\beta_t}\|\rho^*_t - \rho^*_{t+1}\|^2 + \beta_t\|(\rho_{t+1} - \bar{\rho}_t)\|^2\big) \\
&+ \langle\Delta\bar{\rho}_t, \Delta\rho_{t+1}\rangle + \langle\Delta\bar{\rho}_t, \rho^*_{t+1} - \rho^*_t\rangle + \frac{1}{\beta_t}\langle\Delta\bar{\rho}_t, \rho^*_t - \rho^*_{t+1}\rangle
\end{aligned}
$$

$$
\begin{aligned}
\implies \sum_{t=0}^{T-1}\mathbb{E}\|\Delta\bar{\rho}_t\|^2 =& \sum_{t=0}^{T-1}\frac{1}{2\beta_t}\big(\mathbb{E}\|\Delta\bar{\rho}_t\|^2 - \mathbb{E}\|\Delta\bar{\rho}_{t+1}\|^2\big) \quad ① \\
&+ \sum_{t=0}^{T-1}\big(\frac{1}{\beta_t}\mathbb{E}\|\rho^*_t - \rho^*_{t+1}\|^2 + \beta_t\mathbb{E}\|(\rho_{t+1} - \bar{\rho}_t)\|^2\big) \quad ② \\
&+ \sum_{t=0}^{T-1}\mathbb{E}\langle\Delta\bar{\rho}_t, \Delta\rho_{t+1}\rangle \quad ③ \\
&+ \sum_{t=0}^{T-1}\mathbb{E}\langle\Delta\bar{\rho}_t, \rho^*_{t+1} - \rho^*_t\rangle \quad ④ \\
&+ \sum_{t=0}^{T-1}\frac{1}{\beta_t}\mathbb{E}\langle\Delta\bar{\rho}_t, \rho^*_t - \rho^*_{t+1}\rangle \quad ⑤
\end{aligned} \tag{A.19}
$$

From A.19:

①:

$$
\begin{aligned}
&\sum_{t=0}^{T-2}\frac{1}{2\beta_t}\big(\mathbb{E}\|\Delta\bar{\rho}_t\|^2 - \mathbb{E}\|\Delta\bar{\rho}_{t+1}\|^2\big) \\
&= \frac{1}{2}\Big(\sum_{t=0}^{T-1}\big(\frac{1}{\beta_{t+1}} - \frac{1}{\beta_t}\big)\mathbb{E}\|\Delta\bar{\rho}_{t+1}\|^2 + \frac{1}{\beta_0}\mathbb{E}\|\Delta\bar{\rho}_0\|^2 - \frac{1}{\beta_{T-2}}\mathbb{E}\|\Delta\bar{\rho}_T\|^2\Big) \\
&\leq \frac{1}{2}\Big(\sum_{t=0}^{T-1}\big(\frac{1}{\beta_{t+1}} - \frac{1}{\beta_t}\big)\mathbb{E}\|\Delta\bar{\rho}_{t+1}\|^2 + \frac{1}{\beta_0}\mathbb{E}\|\Delta\bar{\rho}_0\|^2\Big) \\
&\leq \frac{1}{2}\Big(\sum_{t=0}^{T-2}\big(\frac{1}{\beta_{t+1}} - \frac{1}{\beta_t}\big) + \frac{1}{\beta_0}\Big)4(C_r + 2C_w)^2 \quad \text{(Using Lemma A.25)} \\
&= \frac{2(C_r + 2C_w)^2}{\beta_{T-1}} = \frac{2(C_r + 2C_w)^2 T^u}{C_\beta}
\end{aligned}
$$

②:

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\big(\frac{1}{\beta_t}\mathbb{E}\|\rho_t^*-\rho_{t+1}^*\|^2+\beta_t\mathbb{E}\|(\rho_{t+1}-\bar{\rho}_t)\|^2\big)\\
&\leq\sum_{t=0}^{T-1}\big(\frac{L_p^2}{\beta_t}\mathbb{E}\|\theta_t-\theta_{t+1}\|^2+\beta_t\mathbb{E}\|(\rho_{t+1}-\bar{\rho}_t)\|^2\big)\quad\text{(Using Lemma A.29 )}\\
&\leq\sum_{t=0}^{T-1}\big(L_p^2G_\theta^2\frac{\gamma_t^2}{\beta_t}+\beta_t 4(C_r+2C_w)^2\big)\quad\text{(Using Lemma A.22,A.23, and A.26 )}\\
&\leq\sum_{t=0}^{T-1}\big(L_p^2G_\theta^2\max_t\frac{\gamma_t^2}{\beta_t^2}+4(C_r+2C_w)^2\big)\beta_t\\
&=\sum_{t=0}^{T-1}C_{st}\beta_t=\frac{C_{st}C_\beta T^{1-u}}{1-u}\quad(C_{st}=L_p^2G_\theta^2\max_t\frac{\gamma_t^2}{\beta_t^2}+4(C_r+2C_w)^2)
\end{aligned}
$$

③:

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}\langle\Delta\bar{\rho}_t,\Delta\rho_{t+1}\rangle\\
&\leq\Big(\sum_{t=0}^{T-1}\mathbb{E}\|\Delta\bar{\rho}_t\|^2\Big)^{1/2}\Big(\sum_{t=0}^{T-1}\mathbb{E}\|\Delta\rho_{t+1}\|^2\Big)^{1/2}\quad\text{(Using Cauchy-Schwarz inequality)}
\end{aligned}
$$

④:

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}\langle\Delta\bar{\rho}_t,\rho_{t+1}^*-\rho_t^*\rangle\\
&\leq\sum_{t=0}^{T-1}\mathbb{E}\|\Delta\bar{\rho}_t\|\|\rho_{t+1}^*-\rho_t^*\|\\
&\leq\sum_{t=0}^{T-1}L_pG_\theta\gamma_t\mathbb{E}\|\Delta\bar{\rho}_t\|\quad\text{(Using Lemma A.22,A.29)}\\
&\leq L_pG_\theta\Big(\sum_{t=0}^{T-1}\gamma_t^2\Big)^{1/2}\Big(\sum_{t=0}^{T-1}\mathbb{E}\|\Delta\bar{\rho}_t\|^2\Big)^{1/2}\quad\text{(Using Cauchy-Schwarz inequality)}\\
&\leq L_pG_\theta C_\gamma\frac{T^{-v}}{(1-2v)^2}T^{1/2}\Big(\sum_{t=0}^{T-1}\mathbb{E}\|\Delta\bar{\rho}_t\|^2\Big)^{1/2}
\end{aligned}
$$

⑤:

$$\begin{aligned}
&\sum_{t=0}^{T-1}\frac{1}{\beta_t}\mathbb{E}\langle\Delta\bar{\rho}_t,\rho^*_{t+1}-\rho^*_t\rangle\\
&\leq\sum_{t=0}^{T-1}\frac{1}{\beta_t}\mathbb{E}\|\Delta\bar{\rho}_t\|\|\rho^*_{t+1}-\rho^*_t\|\\
&\leq\sum_{t=0}^{T-1}L_pG_\theta\frac{\gamma_t}{\beta_t}\mathbb{E}\|\Delta\bar{\rho}_t\| \quad \text{(Using Lemma A.22,A.29)}\\
&\leq L_pG_\theta\Big(\sum_{t=0}^{T-1}\frac{\gamma_t^2}{\beta_t^2}\Big)^{1/2}\Big(\sum_{t=0}^{T-1}\mathbb{E}\|\Delta\bar{\rho}_t\|^2\Big)^{1/2} \quad \text{(Using Cauchy-Schwarz inequality)}\\
&\leq\frac{L_pG_\theta C_\gamma}{C_\beta}\frac{T^{-(v-u)}}{(1-2(v-u))^2}T^{1/2}\Big(\sum_{t=0}^{T-1}\mathbb{E}\|\Delta\bar{\rho}_t\|^2\Big)^{1/2}
\end{aligned}$$

Combining ①-⑤ into A.19:

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\Delta\bar{\rho}_t\|^2 =&\frac{2(C_r+2C_w)^2T^{u-1}}{C_\beta}+\frac{C_{st}C_\beta T^{-u}}{1-u}\\
&+\Big(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\Delta\bar{\rho}_t\|^2\Big)^{1/2}\Big(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\Delta\rho_{t+1}\|^2\Big)^{1/2}\\
&+L_pG_\theta C_\gamma\frac{T^{-v}}{(1-2v)^2}\Big(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\Delta\bar{\rho}_t\|^2\Big)^{1/2}\\
&+\frac{L_pG_\theta C_\gamma}{C_\beta}\frac{T^{-(v-u)}}{(1-2(v-u))^2}\Big(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\Delta\bar{\rho}_t\|^2\Big)^{1/2}
\end{aligned}\tag{A.20}$$

$$\begin{aligned}
M(T)&=\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\Delta\bar{\rho}_t\|^2\\
N(T)&=\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\Delta\rho_{t+1}\|^2\\
M(T)&\leq K_1+K_2\sqrt{M(T)}+K_3\sqrt{M(T)N(T)}
\end{aligned}$$

Here,

$$\begin{aligned}
K_1&=\frac{2(C_r+2C_w)^2T^{u-1}}{C_\beta}+\frac{C_{st}C_\beta T^{1-u}}{1-u}\\
K_2&=L_pG_\theta C_\gamma\frac{T^{-v}}{(1-2v)^2}+\frac{L_pG_\theta C_\gamma}{C_\beta}\frac{T^{-(v-u)}}{(1-2(v-u))^2}\\
K_3&=1
\end{aligned}$$

From Lemma A.14, we know that

$$M(T)\leq 2(\sqrt{K_1}+K_2)^2+2K_3^2N(T)$$

Hence,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\Delta\bar{\rho}_t\|^2 \leq 2\Bigg(\sqrt{\frac{2(C_r+2C_w)^2T^{u-1}}{C_\beta}+\frac{C_{st}C_\beta T^{1-u}}{1-u}} + L_pG_\theta C_\gamma\frac{T^{-v}}{(1-2v)^2}+\frac{L_pG_\theta C_\gamma}{C_\beta}\frac{T^{-(v-u)}}{(1-2(v-u))^2}\Bigg)^2 + \frac{2}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\Delta\rho_{t+1}\|^2 \tag{A.21}$$

□

**Lemma A.17.** *Let the cumulative error of average reward estimator be $\sum_{t=0}^{T-1}\mathbb{E}||\Delta\rho_t||^2$ and cumulative error of target linear critic be $\sum_{t=0}^{T-1}\mathbb{E}||\Delta\bar{w}_t||^2$. $\bar{w}_t$ and $\rho_t$ are the target linear critic parameter and average reward estimator at time t respectively. Bound on the cumulative error of average reward estimator is proven using cumulative error of target critic as follows:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}|\Delta\rho_t|^2 \leq 2\Bigg(\sqrt{\frac{2(C_r+2C_w)^2}{C_\alpha}T^{\sigma-1}+\frac{C_sC_\alpha}{1-\sigma}T^{-\sigma}}+\frac{L_pG_\theta C_\gamma}{C_\alpha}\Big(\frac{T^{-2(v-\sigma)}}{1-2(v-\sigma)}\Big)^{1/2}\Bigg)^2 + 8\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta\bar{w}_t||^2$$

*Here, $\Delta\rho_t=\rho_t-\rho_t^*$, $\Delta\bar{w}_t=\bar{w}_t-w_t^*$. $w_t^*$ and $\rho_t^*$ are the optimal parameters given by TD(0) algorithm corresponding to policy parameter $\theta_t$. $C_\alpha$, $\sigma$ are constants and $\gamma_t,\alpha_t$ are step-sizes defined in Assumption 3.5, $\|w_t\|\leq C_w$ (Algorithm 2, step 8), $C_r$ is the upper bound on rewards (Assumption 4.2), Constant $G_\theta$ is defined in Lemma A.22. $C_s=L_p^2G_\theta^2\max_t\frac{\gamma_t^2}{\alpha_t^2}+4(C_r+2C_w)^2$. $L_p$ is Lipchitz constant defined in Lemma A.29.*

*Proof.*

$$
\begin{aligned}
\rho_{t+1} &= \rho_t + \alpha_t \frac{1}{M} \sum_{i=0}^{M-1} \Big( R^\pi(s_{t,i}) - \rho_t + \phi^\pi(s'_{t,i})^\intercal \bar{w}_t - \phi^\pi(s_{t,i})^\intercal \bar{w}_t \Big) \\
\rho_{t+1} - \rho^*_{t+1} &= \rho_t - \rho^*_t + \rho^*_t - \rho^*_{t+1} \\
&\quad + \alpha_t \frac{1}{M} \sum_{i=0}^{M-1} \Big( R^\pi(s_{t,i}) - \rho_t + \phi^\pi(s'_{t,i})^\intercal \bar{w}_t - \phi^\pi(s_{t,i})^\intercal \bar{w}_t \Big) \\
&= \rho_t - \rho^*_t + \rho^*_t - \rho^*_{t+1} \\
&\quad + \alpha_t \frac{1}{M} \sum_{i=0}^{M-1} \Big( R^\pi(s_{t,i}) - \rho^*_t + \phi^\pi(s'_{t,i})^\intercal \bar{w}_t - \phi^\pi(s_{t,i})^\intercal \bar{w}_t \Big) \\
&\quad + \alpha_t(\rho^*_t - \rho_t) \\
\rho_{t+1} - \rho^*_{t+1} &= \rho_t - \rho^*_t + \rho^*_t - \rho^*_{t+1} \\
&\quad + \alpha_t(\rho^*_t - \rho_t) \\
&\quad + \alpha_t \Big( \frac{1}{M} \sum_{i=0}^{M-1} (\phi^\pi(s'_{t,i}) - \phi^\pi(s_{t,i}))^\intercal (\bar{w}_t - w^*_t) \Big) \\
&\quad + \alpha_t \Big( \frac{1}{M} \sum_{i=0}^{M-1} (R^\pi(s_{t,i}) - \rho^*_t + \phi^\pi(s'_{t,i})^\intercal w^*_t - \phi^\pi(s_{t,i})^\intercal w^*_t) \Big) \\
&= \rho_t - \rho^*_t + \rho^*_t - \rho^*_{t+1} \\
&\quad + \alpha_t(\rho^*_t - \rho_t) \\
&\quad + \alpha_t \Big( \frac{1}{M} \sum_{i=0}^{M-1} (\phi^\pi(s'_{t,i}) - \phi^\pi(s_{t,i}))^\intercal (\bar{w}_t - w^*_t) \Big) \\
&\quad + \alpha_t \bar{l}(w^*_t, \theta_t) \\
&= \rho_t - \rho^*_t + \rho^*_t - \rho^*_{t+1} + \alpha_t l(B_t, \rho_t, w^*_t, \theta_t)
\end{aligned}
$$

Here,

$$
\begin{aligned}
\bar{l}(w_t, \theta_t) &:= \int_S d^\pi\big(s, \pi(\theta_t)\big) \big( R^\pi(s) - \rho(\pi(\theta_t)) + \mathbb{E}[\phi^\pi(s')^\intercal w_t] - \phi^\pi(s)^\intercal w_t \big)\, ds \\
l(B_t, \rho_t, w_t, \theta_t) &:= (\rho^*_t - \rho_t) \\
&\quad + \Big( \frac{1}{M} \sum_{i=0}^{M-1} (\phi^\pi(s'_{t,i}) - \phi^\pi(s_{t,i}))^\intercal (\bar{w}_t - w_t) \Big) \\
&\quad + \bar{l}(w_t, \theta_t)
\end{aligned}
$$

$$
\begin{aligned}
||\Delta\rho_{t+1}||^2 &= ||\Delta\rho_t + \rho^*_t + \alpha_t l(B_t, w^*_t, \rho_t, \theta_t)||^2 \\
&= ||\Delta\rho_t||^2 + ||\rho^*_t - \rho^*_{t+1}||^2 + \alpha_t^2 ||l(B_t, w^*_t, \rho_t, \theta_t)||^2 \\
&\quad + 2\langle \Delta\rho_t, \rho^*_t - \rho^*_{t+1} \rangle \\
&\quad + 2\alpha_t \langle \Delta\rho_t, l(B_t, w^*_t, \rho_t, \theta_t) \rangle \\
&\quad + 2\alpha_t \langle \rho^*_t - \rho^*_{t+1}, l(B_t, \rho_t, w^*_t, \theta_t) \rangle \\
&\leq ||\Delta\rho_t||^2 + 2||\rho^*_t - \rho^*_{t+1}||^2 + 2\alpha_t^2 ||l(B_t, w^*_t, \rho_t, \theta_t)||^2 \\
&\quad + 2\langle \Delta\rho_t, \rho^*_t - \rho^*_{t+1} \rangle \\
&\quad + 2\alpha_t \langle \Delta\rho_t, l(B_t, w^*_t, \rho_t, \theta_t) \rangle
\end{aligned}
$$

$$\begin{aligned}\mathbb{E}||\Delta\rho_{t+1}||^2 \leq \mathbb{E}||\Delta\rho_t||^2 &+ 2\mathbb{E}||\rho_t^* - \rho_{t+1}^*||^2 \ \textcircled{1} \\ &+ 2\alpha_t^2\mathbb{E}||l(B_t, w_t^*, \rho_t, \theta_t)||^2 \ \textcircled{2} \\ &+ 2\mathbb{E}\langle\Delta\rho_t, \rho_t^* - \rho_{t+1}^*\rangle \ \textcircled{3} \\ &+ 2\alpha_t\mathbb{E}\langle\Delta\rho_t, -\Delta\rho_t\rangle \ \textcircled{4} \\ &+ 2\alpha_t\mathbb{E}\langle\Delta\rho_t, \frac{1}{M}\sum_{i=0}^{M-1}(\phi^\pi(s'_{t,i}) - \phi^\pi(s_{t,i}))^\intercal(\bar{w}_t - w_t^*)\rangle \ \textcircled{5} \\ &+ 2\alpha_t\mathbb{E}\langle\Delta\rho_t, \bar{l}(w_t^*, \theta_t)\rangle \ \textcircled{6}\end{aligned} \tag{A.22}$$

From (A.22):
①:

$$\begin{aligned}\mathbb{E}||\rho_t^* - \rho_{t+1}^*||^2 &\leq L_p^2\mathbb{E}||\theta_{t+1} - \theta_t||^2 \text{(Lemma A.29)} \\ &\leq L_p^2\gamma_t^2 G_\theta^2 \quad \text{(Using Lemma A.22)}\end{aligned}$$

②:

$$\begin{aligned}\mathbb{E}||l(B_t, \rho_t, \bar{w}_t, \theta_t)||^2 &= \mathbb{E}||\frac{1}{M}\sum_{i=0}^{M-1}\big(R^\pi(s_{t,i}) - \rho_t + \big(\phi^\pi(s'_{t,i}) - \phi^\pi(s_{t,i})\big)^\intercal\bar{w}_t\big)||^2 \\ &\leq \mathbb{E}\Big(\frac{1}{M}\sum_{i=0}^{M-1}(C_r + C_r + 2C_w + 2C_w)\Big)^2 \\ &= 4(C_r + 2C_w)^2\end{aligned}$$

③:

$$\begin{aligned}\mathbb{E}\langle\Delta\rho_t, \rho_t^* - \rho_{t+1}^*\rangle &\leq \mathbb{E}||\Delta\rho_t||\,|\rho_t^* - \rho_{t+1}^*| \\ &\leq L_p\mathbb{E}|\Delta\rho_t|\,||\theta_{t+1} - \theta_t|| \\ &\leq L_p\gamma_t G_\theta\mathbb{E}|\Delta\rho_t| \quad \text{(Using Lemma A.22}\end{aligned}$$

④:

$$\mathbb{E}\langle\Delta\rho_t, -\Delta\rho_t\rangle = -\mathbb{E}|\Delta\rho_t|^2$$

⑤:

$$\begin{aligned}&\mathbb{E}\langle\Delta\rho_t, \frac{1}{M}\sum_{i=0}^{M-1}\big(\phi^\pi(s'_{t,i})^\intercal - \phi^\pi(s_{t,i})^\intercal\big)(\bar{w}_t - w_t^*)\rangle \\ &\leq \mathbb{E}\Big[\frac{1}{M}\sum_{i=0}^{M-1}||\phi^\pi(s'_{t,i}) - \phi^\pi(s_{t,i})||\,||\bar{w}_t - w_t^*||\,|\Delta\rho_t|\Big] \\ &\leq 2\mathbb{E}|\Delta\rho_t|\|\Delta\bar{w}_t\|\end{aligned}$$

⑥:

$$\mathbb{E}\langle\Delta\rho_t, \bar{l}(w_t^*, \theta_t)\rangle = 0 \quad (\because \bar{l}(w_t^*, \theta_t) = 0)$$

Combining ①-⑦ into (A.22):

$$\begin{aligned}\mathbb{E}||\Delta\rho_{t+1}||^2 \leq (1 - 2\alpha_t)\mathbb{E}||\Delta\rho_t||^2 &+ L_p^2\gamma_t^2 G_\theta^2 \\ &+ 8\alpha_t^2(C_r + 2C_w)^2 + L_p\gamma_t G_\theta\mathbb{E}|\Delta\rho_t| \\ &+ 4\alpha_t\mathbb{E}|\Delta\rho_t|\|\Delta\bar{w}_t\|\end{aligned}$$

$$
\begin{aligned}
\implies \sum_{t=0}^{T-1} \mathbb{E}||\Delta\rho_t||^2 \leq \sum_{t=0}^{T-1} \frac{1}{2\alpha_t}\Big(\mathbb{E}||\Delta\rho_t||^2 - \mathbb{E}||\Delta\rho_{t+1}||^2\Big) \;\; \textcircled{1} \\
+ \sum_{t=0}^{T-1} \Big(\frac{L_p^2\gamma_t^2}{\alpha_t} G_\theta^2 + 4\alpha_t(C_r + 2C_w)^2\Big) \;\; \textcircled{2} \\
+ \sum_{t=0}^{T-1} \Big(L_p G_\theta \frac{\gamma_t}{\alpha_t}\Big)\mathbb{E}|\Delta\rho_t| \;\; \textcircled{3} \\
+ \sum_{t=0}^{T-1} 2\mathbb{E}||\Delta\bar{w}_t||\,|\Delta\rho_t| \;\; \textcircled{4}
\end{aligned} \tag{A.23}
$$

From (A.23):

①:

$$
\begin{aligned}
\frac{1}{2}\sum_{t=0}^{T-1} \frac{1}{\alpha_t}(\mathbb{E}||\Delta\rho_t||^2 - \mathbb{E}||\Delta\rho_{t+1}||^2) &= \frac{1}{2}\left(\sum_{t=0}^{T-1}\Big(\frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}}\Big)\mathbb{E}|\Delta\rho_t|^2 + \frac{1}{\alpha_0}\mathbb{E}|\Delta\rho_0|^2 - \frac{1}{\alpha_{T-1}}\mathbb{E}|\Delta\rho_t|^2\right) \\
&\leq \frac{1}{2}\left(\sum_{t=0}^{T-1}\Big(\frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}}\Big) + \frac{1}{\alpha_0}\right)4(C_r + 2C_w)^2 \\
&\leq \frac{2(C_r + 2C_w)^2}{C_\alpha} T^\sigma
\end{aligned}
$$

②:

$$
\begin{aligned}
\sum_{t=0}^{T-1}\Big(L_p^2 G_\theta^2 \frac{\gamma_t^2}{\alpha_t} + 4\alpha_t(C_r + 2C_w)^2\Big) &\leq \sum_{t=0}^{T-1}\Big(L_p^2 G_\theta^2 \max_t \frac{\gamma_t^2}{\alpha_t^2} + 4(C_r + 2C_w)^2\Big)\alpha_t \\
&\leq \sum_{t=0}^{T-1} C_s\alpha_t \quad (C_s = L_p^2 G_\theta^2 \max_t \frac{\gamma_t^2}{\alpha_t^2} + 4(C_r + 2C_w)^2) \\
&\leq \frac{C_s C_\alpha}{1-\sigma} T^{1-\sigma}
\end{aligned}
$$

③:

$$
\begin{aligned}
\sum_{t=0}^{T-1}\Big(L_p G_\theta \frac{\gamma_t}{\alpha_t}\Big)\mathbb{E}||\Delta\rho_t|| &= \sum_{t=0}^{T-1} L_p G_\theta \frac{\gamma_t}{\alpha_t}\mathbb{E}||\Delta\rho_t|| \\
&\leq L_p G_\theta \left(\sum_{t=0}^{T-1}\Big(\frac{\gamma_t}{\alpha_t}\Big)^2\right)^{1/2}\Big(\sum_{t=0}^{T-1}\mathbb{E}|\Delta\rho_t|^2\Big)^{1/2} \\
&\leq \frac{L_p G_\theta C_\gamma}{C_\alpha}\Big(\frac{T^{1-2(v-\sigma)}}{1-2(v-\sigma)}\Big)^{1/2}\Big(\sum_{t=0}^{T-1}\mathbb{E}|\Delta\rho_t|^2\Big)^{1/2} \\
&\text{(using cauchy schwarz inequality)}
\end{aligned}
$$

④:

$$
2\sum_{t-0}^{T-1}\mathbb{E}||\Delta\bar{w}_t||\,|\Delta\rho_t| \leq 2(\sum_{t=0}^{T-1}\mathbb{E}||\Delta\bar{w}_t||^2)^{1/2}(\sum_{t=0}^{T-1}\mathbb{E}|\Delta\rho_t|^2)^{1/2}
$$

(using cauchy schwarz inequality)

Combining ①-④ into (A.23)

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta\rho_t||^2 \leq \frac{2(C_r+2C_w)^2T^{\sigma-1}}{C_\alpha}+\frac{C_sC_\alpha T^{-\sigma}}{1-\sigma} + \frac{L_pG_\theta C_\gamma}{C_\alpha}\Big(\frac{T^{-2(v-\sigma)}}{1-2(v-\sigma)}\Big)^{1/2}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}|\Delta\rho_t|^2\right)^{1/2} + 2\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta\bar{w}_t||^2\right)^{1/2}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}|\Delta\rho_t|^2\right)^{1/2}$$

$$M(T) = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta\rho_t||^2$$
$$N(T) = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta\bar{w}_t||^2$$
$$M(T) \leq K_1 + K_2\sqrt{M(T)} + K_3\sqrt{M(T)}\sqrt{N(T)}$$

Here,

$$K_1 = \frac{2(C_r+2C_w)^2T^{\sigma-1}}{C_\alpha}+\frac{C_sC_\alpha T^{-\sigma}}{1-\sigma}$$
$$K_2 = \frac{L_pG_\theta C_\gamma}{C_\alpha}\Big(\frac{T^{-2(v-\sigma)}}{1-2(v-\sigma)}\Big)^{1/2}$$
$$K_3 = 2$$

From Lemma A.14, we know that

$$M(T) \leq 2(\sqrt{K_1}+K_2)^2 + 2K_3^2N(T)$$

Hence,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}|\Delta\rho_t|^2 \leq 2\left(\sqrt{\frac{2(C_r+2C_w)^2}{C_\alpha}T^{\sigma-1}+\frac{C_sC_\alpha}{1-\sigma}T^{-\sigma}}+\frac{L_pG_\theta C_\gamma}{C_\alpha}\Big(\frac{T^{-2(v-\sigma)}}{1-2(v-\sigma)}\Big)^{1/2}\right)^2 + 8\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta\bar{w}_t||^2$$

□

**Theorem A.18.** *The on-policy average reward actor critic algorithm (Algorithm 2) obtains an $\epsilon$-accurate optimal point with sample complexity of $\Omega(\epsilon^{-2.5})$. We obtain*

$$\min_{0\leq t\leq T-1}\mathbb{E}||\nabla_\theta\rho(\theta_t)||^2 = \mathcal{O}\left(\frac{1}{T^{2/5}}\right)+3C_\pi^4(\tau^2+\frac{4}{M}C_{w_\epsilon^*}^2),$$
$$\leq \epsilon + \mathcal{O}(1).$$

*Here, $\|\nabla_\theta\pi(s)\| \leq C_\pi$ (Assumption 4.4), $\tau = \max_t\|w_t^* - w_{\epsilon,t}^*\|$, $w_\epsilon^*$ is the optimal critic parameter according to Lemma 3.2. Constant $C_{w_\epsilon^*}$ is defined in Lemma A.28. M is the size of batch of samples used to update parameters.*

*Proof.* From Lemma A.14 we have:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2 \leq 2\left(\sqrt{\frac{2C_w^2}{(\lambda-1)C_\alpha}T^{\sigma-1}+\frac{C_gC_\alpha}{1-\sigma}T^{-\sigma}}+\frac{L_wG_\theta C_\gamma}{(\lambda-1)C_\alpha}\left(\frac{T^{-2(v-\sigma)}}{1-2(v-\sigma)}\right)^{\frac{1}{2}}\right)^2 + \frac{4}{(\lambda-1)^2}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}|\Delta\bar{\rho}_t|^2+\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta\bar{w}_t||^2\right)$$

Using Lemma A.15 and Lemma A.16:

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2 \leq & 2\left(\sqrt{\frac{2C_w^2}{(\lambda-1)C_\alpha}T^{\sigma-1}+\frac{C_gC_\alpha}{1-\sigma}T^{-\sigma}}+\frac{L_wG_\theta C_\gamma}{(\lambda-1)C_\alpha}\left(\frac{T^{-2(v-\sigma)}}{1-2(v-\sigma)}\right)^{\frac{1}{2}}\right)^2 \\ & + \frac{4}{(\lambda-1)^2}\left(2\left(\sqrt{\frac{2C_w^2T^{u-1}}{C_\beta}+\frac{C_{gt}C_\beta T^{-u}}{1-u}}\right.\right. \\ & \left.\left. + L_pG_\theta C_\gamma\left(\frac{T^{-v}}{(1-2v)^{1/2}}+\frac{T^{-(v-u)}}{C_\beta(1-2(v-u))^{1/2}}\right)\right)^2\right) \\ & + \frac{4}{(\lambda-1)^2}\left(2\left(\sqrt{\frac{2(C_r+2C_w)^2T^{u-1}}{C_\beta}+\frac{C_{st}C_\beta T^{-u}}{1-u}}\right.\right. \\ & \left.\left. + L_pG_\theta C_\gamma\left(\frac{T^{-v}}{(1-2v)^{1/2}}+\frac{T^{-(v-u)}}{C_\beta(1-2(v-u))^{1/2}}\right)\right)^2\right) \\ & + \frac{2}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta\rho_{t+1}||^2 \\ & + \frac{2}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_{t+1}||^2\end{aligned}$$

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2 \leq & 2\left(\sqrt{\frac{2C_w^2}{(\lambda-1)C_\alpha}T^{\sigma-1}+\frac{C_gC_\alpha}{1-\sigma}T^{-\sigma}}+\frac{L_wG_\theta C_\gamma}{(\lambda-1)C_\alpha}\left(\frac{T^{-2(v-\sigma)}}{1-2(v-\sigma)}\right)^{\frac{1}{2}}\right)^2 \\ & + \frac{4}{(\lambda-1)^2}\left(2\left(\sqrt{\frac{2C_w^2T^{u-1}}{C_\beta}+\frac{C_{gt}C_\beta T^{-u}}{1-u}}\right.\right. \\ & \left.\left. + L_pG_\theta C_\gamma\left(\frac{T^{-v}}{(1-2v)^{1/2}}+\frac{T^{-(v-u)}}{C_\beta(1-2(v-u))^{1/2}}\right)\right)^2\right) \\ & + \frac{4}{(\lambda-1)^2}\left(2\left(\sqrt{\frac{2(C_r+2C_w)^2T^{u-1}}{C_\beta}+\frac{C_{st}C_\beta T^{-u}}{1-u}}\right.\right. \\ & \left.\left. + L_pG_\theta C_\gamma\left(\frac{T^{-v}}{(1-2v)^{1/2}}+\frac{T^{-(v-u)}}{C_\beta(1-2(v-u))^{1/2}}\right)\right)^2\right) \\ & + \frac{8}{(\lambda-1)^2T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta\rho_t||^2+\frac{8}{(\lambda-1)^2T}(\mathbb{E}||\Delta\rho_T||^2-\mathbb{E}||\Delta\rho_0||^2) \\ & + \frac{8}{(\lambda-1)^2T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2+\frac{8}{(\lambda-1)^2T}(\mathbb{E}||\Delta w_T||^2-\mathbb{E}||\Delta w_0||^2)\end{aligned}$$

Using Lemma A.17 for $\sum_{t=0}^{T-1}\mathbb{E}||\Delta\rho_t||^2$:

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2 \leq & 2\Bigg(\sqrt{\frac{2C_w^2}{(\lambda-1)C_\alpha}T^{\sigma-1}+\frac{C_gC_\alpha}{1-\sigma}T^{-\sigma}}+\frac{L_wG_\theta C_\gamma}{(\lambda-1)C_\alpha}\Big(\frac{T^{-2(v-\sigma)}}{1-2(v-\sigma)}\Big)^{\frac{1}{2}}\Bigg)^2 \\ & + \frac{4}{(\lambda-1)^2}\Bigg(2\Bigg(\sqrt{\frac{2C_w^2T^{u-1}}{C_\beta}+\frac{C_{gt}C_\beta T^{-u}}{1-u}} \\ & + L_pG_\theta C_\gamma\Big(\frac{T^{-v}}{(1-2v)^{1/2}}+\frac{T^{-(v-u)}}{C_\beta(1-2(v-u))^{1/2}}\Big)\Bigg)^2\Bigg) \\ & + \frac{4}{(\lambda-1)^2}\Bigg(2\Bigg(\sqrt{\frac{2(C_r+2C_w)^2T^{u-1}}{C_\beta}+\frac{C_{st}C_\beta T^{-u}}{1-u}} \\ & + L_pG_\theta C_\gamma\Big(\frac{T^{-v}}{(1-2v)^{1/2}}+\frac{T^{-(v-u)}}{C_\beta(1-2(v-u))^{1/2}}\Big)\Bigg)^2\Bigg) \\ & + \frac{16}{(\lambda-1)^2}\Bigg(\sqrt{\frac{2(C_r+2C_w)^2}{C_\alpha}T^{\sigma-1}+\frac{C_sC_\alpha}{1-\sigma}T^{-\sigma}}+\frac{L_pG_\theta C_\gamma}{C_\alpha}\Big(\frac{T^{-2(v-\sigma)}}{1-2(v-\sigma)}\Big)^{1/2}\Bigg)^2 \\ & + \frac{64}{(\lambda-1)^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta\bar{w}_t||^2+\frac{8}{(\lambda-1)^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2 \\ & + \frac{8}{(\lambda-1)^2T}(\mathbb{E}||\Delta\rho_T||^2-\mathbb{E}||\Delta\rho_0||^2)+\frac{8}{(\lambda-1)^2T}(\mathbb{E}||\Delta w_T||^2-\mathbb{E}||\Delta w_0||^2)\end{aligned}$$

Using Lemma A.15 for $\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2$:

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2 \leq & 2\Bigg(\sqrt{\frac{2C_w^2}{(\lambda-1)C_\alpha}T^{\sigma-1}+\frac{C_gC_\alpha}{1-\sigma}T^{-\sigma}}+\frac{L_wG_\theta C_\gamma}{(\lambda-1)C_\alpha}\Big(\frac{T^{-2(v-\sigma)}}{1-2(v-\sigma)}\Big)^{\frac{1}{2}}\Bigg)^2 \\ & + \frac{136}{(\lambda-1)^2}\Bigg(\sqrt{\frac{2C_w^2T^{u-1}}{C_\beta}+\frac{C_{gt}C_\beta T^{-u}}{1-u}} \\ & + L_pG_\theta C_\gamma\Big(\frac{T^{-v}}{(1-2v)^{1/2}}+\frac{T^{-(v-u)}}{C_\beta(1-2(v-u))^{1/2}}\Big)\Bigg)^2 \\ & + \frac{8}{(\lambda-1)^2}\Bigg(\sqrt{\frac{2(C_r+2C_w)^2T^{u-1}}{C_\beta}+\frac{C_{st}C_\beta T^{-u}}{1-u}} \\ & + L_pG_\theta C_\gamma\Big(\frac{T^{-v}}{(1-2v)^{1/2}}+\frac{T^{-(v-u)}}{C_\beta(1-2(v-u))^{1/2}}\Big)\Bigg)^2 \\ & + \frac{16}{(\lambda-1)^2}\Bigg(\sqrt{\frac{2(C_r+2C_w)^2}{C_\alpha}T^{\sigma-1}+\frac{C_sC_\alpha}{1-\sigma}T^{-\sigma}}+\frac{L_pG_\theta C_\gamma}{C_\alpha}\Big(\frac{T^{-2(v-\sigma)}}{1-2(v-\sigma)}\Big)^{1/2}\Bigg)^2 \\ & + \frac{136}{(\lambda-1)^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2 \\ & + \frac{8}{(\lambda-1)^2T}(\mathbb{E}||\Delta\rho_T||^2-\mathbb{E}||\Delta\rho_0||^2)+\frac{136}{(\lambda-1)^2T}(\mathbb{E}||\Delta w_T||^2-\mathbb{E}||\Delta w_0||^2)\end{aligned}$$

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2 \\
\leq&\frac{2(\lambda-1)^2}{(\lambda-1)^2-136}\left(\sqrt{\frac{2C_w^2}{(\lambda-1)C_\alpha}T^{\sigma-1}+\frac{C_gC_\alpha}{1-\sigma}T^{-\sigma}}+\frac{L_wG_\theta C_\gamma}{(\lambda-1)C_\alpha}\left(\frac{T^{-2(v-\sigma)}}{1-2(v-\sigma)}\right)^{\frac{1}{2}}\right)^2 \\
&+\frac{136}{(\lambda-1)^2-136}\left(\sqrt{\frac{2C_w^2T^{u-1}}{C_\beta}+\frac{C_{gt}C_\beta T^{-u}}{1-u}}\right. \\
&\left.+L_pG_\theta C_\gamma\Big(\frac{T^{-v}}{(1-2v)^{1/2}}+\frac{T^{-(v-u)}}{C_\beta(1-2(v-u))^{1/2}}\Big)\right)^2 \\
&+\frac{8}{(\lambda-1)^2-136}\left(\sqrt{\frac{2(C_r+2C_w)^2T^{u-1}}{C_\beta}+\frac{C_{st}C_\beta T^{-u}}{1-u}}\right. \\
&\left.+L_pG_\theta C_\gamma\Big(\frac{T^{-v}}{(1-2v)^{1/2}}+\frac{T^{-(v-u)}}{C_\beta(1-2(v-u))^{1/2}}\Big)\right)^2 \\
&+\frac{16}{(\lambda-1)^2-136}\left(\sqrt{\frac{2(C_r+2C_w)^2}{C_\alpha}T^{\sigma-1}+\frac{C_sC_\alpha}{1-\sigma}T^{-\sigma}}+\frac{L_pG_\theta C_\gamma}{C_\alpha}\Big(\frac{T^{-2(v-\sigma)}}{1-2(v-\sigma)}\Big)^{1/2}\right)^2 \\
&+\frac{8}{(\lambda-1)^2-136}\frac{1}{T}(\mathbb{E}||\Delta\rho_T||^2-\mathbb{E}||\Delta\rho_0||^2)+\frac{136}{(\lambda-1)^2-136}\frac{1}{T}(\mathbb{E}||\Delta w_T||^2-\mathbb{E}||\Delta w_0||^2)
\end{aligned}
$$

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2 \leq&\mathcal{O}\Big(\frac{1}{T^{1-\sigma}}\Big)+\mathcal{O}\Big(\frac{1}{T^{\sigma}}\Big)+\mathcal{O}\Big(\frac{1}{T^{2(v-\sigma)}}\Big)+\mathcal{O}\Big(\frac{1}{T^{1-u}}\Big)+\mathcal{O}\Big(\frac{1}{T^{u}}\Big)+\mathcal{O}\Big(\frac{1}{T^{2v}}\Big) \\
&+\mathcal{O}\Big(\frac{1}{T^{2(v-u)}}\Big)+\frac{8}{(\lambda-1)^2-136}\frac{1}{T}(\mathbb{E}||\Delta\rho_T||^2-\mathbb{E}||\Delta\rho_0||^2) \\
&+\frac{136}{(\lambda-1)^2-136}\frac{1}{T}(\mathbb{E}||\Delta w_T||^2-\mathbb{E}||\Delta w_0||^2)
\end{aligned}
\tag{A.24}
$$

From Lemma A.13 we have :

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}E||\nabla_\theta\rho(\theta_t)||^2 \leq 2\frac{C_r}{C_\gamma}T^{v-1}+3C_\pi^4(\frac{1}{T}\sum_{t=0}^{T-1}E||\Delta w_t||^2)+3C_\pi^4(\tau^2+\frac{4}{M}C_{w_\epsilon^*}^2), \\
+\frac{C_\gamma L_J G_\theta^2}{1-v}T^{-v}
\end{aligned}
$$

Using A.24:

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\nabla_\theta\rho(\theta_t)||^2 \leq&\mathcal{O}\Big(\frac{1}{T^{1-\sigma}}\Big)+\mathcal{O}\Big(\frac{1}{T^{\sigma}}\Big)+\mathcal{O}\Big(\frac{1}{T^{2(v-\sigma)}}\Big)+\mathcal{O}\Big(\frac{1}{T^{1-u}}\Big)+\mathcal{O}\Big(\frac{1}{T^{u}}\Big)+\mathcal{O}\Big(\frac{1}{T^{2v}}\Big) \\
&+\mathcal{O}\Big(\frac{1}{T^{2(v-u)}}\Big)+\mathcal{O}\Big(\frac{1}{T^{1-v}}\Big)+\mathcal{O}\Big(\frac{1}{T^{v}}\Big)+3C_\pi^4(\tau^2+\frac{4}{M}C_{w_\epsilon^*}^2) \\
&+\underbrace{\frac{8(\mathbb{E}||\Delta\rho_T||^2-\mathbb{E}||\Delta\rho_0||^2)}{(\lambda-1)^2-136}\frac{1}{T}}_{\text{(I)}}+\underbrace{\frac{136(\mathbb{E}||\Delta w_T||^2-\mathbb{E}||\Delta w_0||^2)}{(\lambda-1)^2-136}\frac{1}{T}}_{\text{(II)}}
\end{aligned}
$$

$\|\Delta\rho_t\|$ is bounded because of Lemma A.23 and because $\rho_t^*(=\rho(\theta_t))$ is bounded. $\|\Delta w_t\|$ is bounded because of projection operator $\Gamma_{C_w}$ and Lemma A.21. Hence, we have, Ⓘ and Ⓘ Ⓘ are $\mathcal{O}\Big(\frac{1}{T}\Big)$ terms.

By setting $\sigma = 2/5, u = 2/5$, and $v = 3/5$, we obtain the following bound :

$$\min_{0\le t\le T-1} \mathbb{E}||\nabla_\theta \rho(\theta_t)||^2 \le \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}||\nabla_\theta \rho(\theta_t)||^2 \le \mathcal{O}\Big(\frac{1}{T^{2/5}}\Big) + 3C_\pi^4(\tau^2 + \frac{4}{M}C_{w_\epsilon^*}^2)$$
$$\le \epsilon + \mathcal{O}(1)$$

The sample complexity of the on-policy algorithm (Algorithm 2) is $\Omega(\epsilon^{-2.5})$.

□

**Lemma A.19.** *Let the cumulative error of off-policy actor be $\sum_{t=0}^{T-1} E||\widehat{\nabla_\theta \rho}(\theta_t)||^2$ and cumulative error of critic be $\sum_{t=0}^{T-1} E||\Delta w_t||^2$. $\theta_t$ and $w_t$ are the actor and linear critic parameter at time t. $\theta^\mu$ is the policy parameter for behavior policy $\mu$. Bound on the cumulative error of off-policy actor with behaviour policy $\mu$ is proven using cumulative error of critic as:*

$$\frac{1}{T}\sum_{t=0}^{T-1} E||\widehat{\nabla_\theta \rho}(\theta_t)||^2 \le 4\frac{C_r}{C_\gamma}T^{v-1} + 6C_\pi^4(\frac{1}{T}\sum_{t=0}^{T-1} E||\Delta w_t||^2) + 6C_\pi^4(\tau^2 + \frac{4}{M}C_{w_\epsilon^*}^2)$$
$$+ 2\frac{C_\gamma L_J G_\theta^2}{1-v}T^{-v} + \frac{Z}{T}\sum_{t=0}^{T-1} E||\theta^\mu - \theta_t||^2$$

*Here, $C_r$ is the upper bound on rewards (Assumption 4.2) , $C_\gamma$, $v$ are constants used for step size $\gamma_t$ (Assumption 3.5, $\|\nabla_\theta \pi(s)\| \le C_\pi$ (Assumption 4.4), $\Delta w_t = w_t - w_t^*, \tau = \max_t \|w_t^* - w_{\epsilon,t}^*\|$, $w_\epsilon^*$ is the optimal critic parameter according to Lemma3.2. $w_t^*$ is the optimal parameters given by TD(0) algorithm corresponding to policy parameter $\theta_t$. Constant $C_{w_\epsilon^*}$ is defined in Lemma A.28. $L_J$ is the coefficient used in smoothness condition of the non convex function $\rho(\theta)$. Constant $G_\theta$ is defined in Lemma A.22. M is the size of batch of samples used to update parameters. $Z = 2^{n+1}C(\lceil \log_\kappa a^{-1} \rceil + 1/\kappa)L_t$ with $L_t$ being the Lipchitz constant for the transition probability density function (Assumption A.1). Constants a and $\kappa$ are from Assumption 3.3, n is the dimension of state space, and $C = \max_s \|\nabla_a Q_{diff}^\pi(s,a)|_{a=\pi(s)} \nabla_\theta \pi(s,\theta)\|$.*

*Proof.*

$$\frac{1}{T}\sum_{t=0}^{T-1} E||\widehat{\nabla_\theta \rho}(\theta_t)||^2 = \frac{1}{T}\sum_{t=0}^{T-1} E||\nabla_\theta \rho(\theta_t) + \widehat{\nabla_\theta \rho}(\theta_t) - \nabla_\theta \rho(\theta_t)||^2$$
$$\le \frac{1}{T}\sum_{t=0}^{T-1} E||\nabla_\theta \rho(\theta_t)||^2 + \frac{1}{T}\sum_{t=0}^{T-1} E||\widehat{\nabla_\theta \rho}(\theta_t) - \nabla_\theta \rho(\theta_t)||^2$$

Using Theorem 3.4 and Lemma A.13:

$$\frac{1}{T}\sum_{t=0}^{T-1} E||\widehat{\nabla_\theta \rho}(\theta_t)||^2 \le 4\frac{C_r}{C_\gamma}T^{v-1} + 6C_\pi^4(\frac{1}{T}\sum_{t=0}^{T-1} E||\Delta w_t||^2) + 6C_\pi^4(\tau^2 + \frac{4}{M}C_{w_\epsilon^*}^2)$$
$$+ 2\frac{C_\gamma L_J G_\theta^2}{1-v}T^{-v} + \frac{Z}{T}\sum_{t=0}^{T-1} E||\theta^\mu - \theta_t||^2$$

□

**Theorem A.20.** *The off-policy average reward actor critic algorithm (Algorithm 3) with behavior policy $\mu$ obtains an $\epsilon$-accurate optimal point with sample complexity of $\Omega(\epsilon^{-2.5})$. Here $\theta_\mu$ refers to the behavior policy parameter and $\theta_t$ refers to the target or current policy parameter. We obtain*

$$\begin{aligned}\min_{0\leq t\leq T-1}\mathbb{E}\|\widehat{\nabla_\theta\rho}(\theta_t)\|^2 &= \mathcal{O}\left(\frac{1}{T^{0.4}}\right)+3C_\pi^4(\tau^2+\frac{4}{M}C_{w_\epsilon^*}^2)+\mathcal{O}(W_\theta^2)\\ &\leq \epsilon+3C_\pi^4(\tau^2+\frac{4}{M}C_{w_\epsilon^*}^2)+\mathcal{O}(W_\theta^2)\\ \text{where } W_\theta &:= \sup_t\|\theta_\mu-\theta_t\|.\end{aligned}$$

*Here, $\|\nabla_\theta\pi(s)\|\leq C_\pi$ (Assumption 4.4), $\tau=\max_t\|w_t^*-w_{\epsilon,t}^*\|$, $w_\epsilon^*$ is the optimal critic parameter according to Lemma 3.2. Constant $C_{w_\epsilon^*}$ is defined in Lemma A.28. M is the size of batch of samples used to update parameters.*

*Proof.* Lemma A.14 and Lemma A.17 will hold in the case of off-policy update. Lemma A.14 will require Lemma A.30 instead of Lemma A.21.

Using Lemma A.14 and Lemma A.17 and using the procedure followed in Theorem 4.8 to obtain asymptotic notations, we have:

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2 \leq& \mathcal{O}\left(\frac{1}{T^{1-\sigma}}\right)+\mathcal{O}\left(\frac{1}{T^{\sigma}}\right)+\mathcal{O}\left(\frac{1}{T^{2(v-\sigma)}}\right)+\mathcal{O}\left(\frac{1}{T^{1-u}}\right)+\mathcal{O}\left(\frac{1}{T^{u}}\right)+\mathcal{O}\left(\frac{1}{T^{2v}}\right)\\ &+\mathcal{O}\left(\frac{1}{T^{2(v-u)}}\right)+\frac{8}{(\lambda-1)^2-136}\frac{1}{T}(\mathbb{E}||\Delta\rho_T||^2-\mathbb{E}||\Delta\rho_0||^2)\\ &+\frac{136}{(\lambda-1)^2-136}\frac{1}{T}(\mathbb{E}||\Delta w_T||^2-\mathbb{E}||\Delta w_0||^2)\end{aligned}\tag{A.25}$$

Using Lemma A.19 and (A.25):

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\widehat{\nabla_\theta\rho}(\theta_t)||^2 \leq& \mathcal{O}\left(\frac{1}{T^{1-\sigma}}\right)+\mathcal{O}\left(\frac{1}{T^{\sigma}}\right)+\mathcal{O}\left(\frac{1}{T^{2(v-\sigma)}}\right)+\mathcal{O}\left(\frac{1}{T^{1-u}}\right)+\mathcal{O}\left(\frac{1}{T^{u}}\right)+\mathcal{O}\left(\frac{1}{T^{2v}}\right)\\ &+\mathcal{O}\left(\frac{1}{T^{2(v-u)}}\right)+\mathcal{O}\left(\frac{1}{T^{1-v}}\right)+\mathcal{O}\left(\frac{1}{T^{v}}\right)+3C_\pi^4(\tau^2+\frac{4}{M}C_{w_\epsilon^*}^2)\\ &+\underbrace{\frac{8(\mathbb{E}||\Delta\rho_T||^2-\mathbb{E}||\Delta\rho_0||^2)}{(\lambda-1)^2-136}\frac{1}{T}}_{\text{(I)}}+\underbrace{\frac{136(\mathbb{E}||\Delta w_T||^2-\mathbb{E}||\Delta w_0||^2)}{(\lambda-1)^2-136}\frac{1}{T}}_{\text{(II)}}\\ &+\frac{Z}{T}\sum_{t=0}^{T-1}E||\theta^\mu-\theta_t||^2\end{aligned}$$

We have, (I) and (II) are $\mathcal{O}\left(\frac{1}{T}\right)$ terms as discussed in Theorem 4.8. By setting $\sigma=2/5, u=2/5$ and $v=2/5$, we obtain:

$$\begin{aligned}\min_{0\leq t\leq T-1}E||\widehat{\nabla_\theta\rho}(\theta_t)||^2 &= \mathcal{O}\left(\frac{1}{T^{0.4}}\right)+3C_\pi^4(\tau^2+\frac{4}{M}C_{w_\epsilon^*}^2)+\frac{Z}{T}\sum_{t=0}^{T-1}E||\theta^\mu-\theta_t||^2\\ &= \mathcal{O}\left(\frac{1}{T^{0.4}}\right)+3C_\pi^4(\tau^2+\frac{4}{M}C_{w_\epsilon^*}^2)+\mathcal{O}(W_\theta^2).\end{aligned}$$

Further,

$$\mathcal{O}\left(\frac{1}{T^{0.4}}\right) \leq \epsilon.$$

Hence, the sample complexity of off-policy average reward actor-critic algorithm is $\Omega(\epsilon^{-2.5})$. □

### A.3.2. AUXILIARY LEMMAS

**Lemma A.21.** *The optimal critic parameter $w(\theta_t)^*$ as a function of actor parameter $\theta_t$ is Lipchitz continuous with constant $L_w$. Note: $w_t^* := w(\theta_t)^*$.*

$$||w_t^* - w_{t+1}^*|| \leq L_w ||\theta_{t+1} - \theta_t||$$

*Proof.* $\eta$ is the l2-regularisation coefficient from Algorithm 2 and $\eta > \lambda_{max}^{all}$, where $\lambda_{max}^{all}$ is defined in Lemma A.26. Because of carefully setting the value of $\eta$, $A(\theta_t)$ is negative definite. Thus, for on-policy TD(0) with l2-regularization and target estimators, the following condition holds true for optimal critic parameter $w_t^*$:

$$E[(R^\pi(s) - \rho_t^*)\phi^\pi(s) + (\phi^\pi(s)(E[\phi^\pi(s')] - \phi^\pi(s))^\intercal - \eta I)w_t^*] = 0$$

$$\begin{aligned} b(\theta_t) &:= E[(R^\pi(s) - \rho_t^*)\phi^\pi(s)] \\ A(\theta_t) &:= E[(\phi^\pi(s)(E[\phi^\pi(s')] - \phi^\pi(s))^\intercal - \eta I)] \end{aligned}$$

$$\therefore b(\theta_t) + A(\theta_t)w_t^* = 0 \implies w_t^* = -A(\theta_t)^{-1}b(\theta_t)$$

$$\begin{aligned} ||w_t^* - w_{t+1}^*|| &= ||A(\theta_t)^{-1}b(\theta_t) - A(\theta_{t+1})^{-1}b(\theta_{t+1})|| \\ &\leq ||A(\theta_t)^{-1}b(\theta_t) - A(\theta_{t+1})^{-1}b(\theta_t) + A(\theta_{t+1})^{-1}b(\theta_t) - A(\theta_{t+1})^{-1}b(\theta_{t+1})|| \\ &\leq ||A(\theta_t)^{-1} - A(\theta_{t+1})^{-1}||\ ||b(\theta_t)||\ \textcircled{1} \\ &\quad + ||A(\theta_{t+1})^{-1}||\ ||b(\theta_t) - b(\theta_{t+1})||\textcircled{2} \end{aligned} \tag{A.26}$$

From (A.26):

①:

$$\begin{aligned} ||A(\theta_t)^{-1} - A(\theta_{t+1})^{-1}|| &= ||A(\theta_t)^{-1}A(\theta_{t+1})A(\theta_{t+1})^{-1} - A(\theta_t)^{-1}A(\theta_t)^A(\theta_{t+1})^{-1}|| \\ &\leq ||A(\theta_t)^{-1}||\ ||A(\theta_t) - A(\theta_{t+1})||\ ||A(\theta_{t+1})^{-1}|| \end{aligned} \tag{A.27}$$

From (A.27):

Here, $\pi'$ and $\pi$ represents the policy with parameter $\theta_{t+1}$ and $\theta_t$ respectively.

$$\begin{aligned} ||A(\theta_t) - A(\theta_{t+1})|| &\leq || \int d^{\pi'}(s)(\phi^{\pi'}(s)(\int P^{\pi'}(s'|s)\phi^{\pi'}(s')\,ds' - \phi^{\pi'}(s))^\intercal - \eta I)\,ds \\ &\quad - \int d^{\pi}(s)(\phi^{\pi}(s)(\int P^{\pi}(s'|s)\phi^{\pi}(s')\,ds' - \phi^{\pi}(s))^\intercal - \eta I)\,ds|| \\ &\leq || \int d^{\pi'}(s)(\phi^{\pi'}(s)(\int P^{\pi'}(s'|s)\phi^{\pi'}(s')\,ds')^\intercal)\,ds \\ &\quad - \int d^{\pi}(s)(\phi^{\pi}(s)(\int P^{\pi}(s'|s)\phi^{\pi}(s')\,ds')^\intercal)\,ds||\ \textcircled{1} \\ &\leq || \int d^{\pi}(s)(\phi^{\pi}(s)(\phi^{\pi}(s))^\intercal)\,ds - \int d^{\pi'}(s)(\phi^{\pi'}(s)(\phi^{\pi'}(s))^\intercal)\,ds||\ \textcircled{2} \end{aligned} \tag{A.28}$$

From (A.28):

①:

$$
\begin{aligned}
&||\int d^{\pi'}(s)(\phi^{\pi'}(s)(\int P^{\pi'}(s'|s)\phi^{\pi'}(s')\,ds')^\intercal)\,ds - \int d^{\pi}(s)(\phi^{\pi}(s)(\int P^{\pi}(s'|s)\phi^{\pi}(s')\,ds')^\intercal)\,ds||\\
&\leq ||\int (d^{\pi'}(s) - d^{\pi}(s))\phi^{\pi'}(s)(\int P^{\pi'}(s'|s)\phi^{\pi'}(s')\,ds')^\intercal\,ds||\\
&\quad + ||\int d^{\pi}(s)(\phi^{\pi'}(s) - \phi^{\pi}(s))(\int P^{\pi'}(s'|s)\phi^{\pi'}(s')\,ds')^\intercal\,ds||\\
&\quad + ||\int d^{\pi}(s)\phi^{\pi}(s)(\int (P^{\pi'}(s'|s) - P^{\pi}(s'|s))\phi^{\pi'}(s')\,ds')^\intercal\,ds||\\
&\quad + ||\int d^{\pi}(s)\phi^{\pi}(s)(\int P^{\pi}(s'|s)(\phi^{\pi'}(s') - \phi^{\pi}(s'))\,ds'^\intercal)\,ds\\
&\leq L_d||\theta_{t+1} - \theta_t|| \qquad \text{(Lemma A.27)}\\
&+ L_\phi||\theta_{t+1} - \theta_t|| \qquad \text{(Assumption 4.5)}\\
&+ L_t||\theta_{t+1} - \theta_t|| \qquad \text{(Assumption A.1)}\\
&+ L_\phi||\theta_{t+1} - \theta_t|| \qquad \text{(Assumption 4.5)}
\end{aligned}
$$

$$
\begin{aligned}
&||\int d^{\pi'}(s)(\phi^{\pi'}(s)(\int P^{\pi'}(s'|s)\phi^{\pi'}(s')\,ds')^\intercal)\,ds - \int d^{\pi}(s)(\phi^{\pi}(s)(\int P^{\pi}(s'|s)\phi^{\pi}(s')\,ds')^\intercal)\,ds||\\
&\leq (L_d + L_t + 2L_\phi)||\theta_{t+1} - \theta_t||
\end{aligned}
\tag{A.29}
$$

From (A.28):

②:

$$
\begin{aligned}
&||\int d^{\pi}(s)(\phi^{\pi}(s)(\phi^{\pi}(s))^\intercal)\,ds - \int d^{\pi'}(s)(\phi^{\pi'}(s)(\phi^{\pi'}(s))^\intercal)\,ds||\\
&\qquad \leq ||\int (d^{\pi}(s) - d^{\pi'}(s))\phi^{\pi}(s)(\phi^{\pi}(s))^\intercal\,ds||\\
&\qquad\quad + ||\int d^{\pi'}(s)(\phi^{\pi}(s) - \phi^{\pi'}(s))(\phi^{\pi}(s))^\intercal\,ds||\\
&\qquad\quad + ||\int d^{\pi'}(s)\phi^{\pi'}(s)(\phi^{\pi}(s) - \phi^{\pi'}(s))^\intercal\,ds||\\
&\qquad \leq (L_d + 2L_\phi)||\theta_{t+1} - \theta_t||
\end{aligned}
\tag{A.30}
$$

Using (A.29) and (A.30) in (A.28)

$$
||A(\theta_t) - A(\theta_{t+1})|| \leq (2L_d + 4L_\phi + L_t)||\theta_{t+1} - \theta_t|| \tag{A.31}
$$

From (A.26):

②:

$$
\begin{aligned}
||b(\theta_t) - b(\theta_{t+1})|| &= || \int d^{\pi'}(s)((R^{\pi'}(s) - \rho^*_{t+1})\phi^{\pi'}(s)\, ds - \int d^{\pi}(s)(R^{\pi}(s) - \rho^*_{t})\phi^{\pi}(s)\, ds|| \\
&\leq || \int d^{\pi'}(s)(R^{\pi'}(s)\phi^{\pi'}(s)\, ds - \int d^{\pi}(s)R^{\pi}(s)\phi^{\pi}(s)\, ds|| \\
&+ || \int d^{\pi'}(s)\rho^*_{t+1}\phi^{\pi'}(s)\, ds - \int d^{\pi}(s)\rho^*_{t}\phi^{\pi}(s)\, ds|| \\
&\leq || \int (d^{\pi'}(s) - d^{\pi}(s))R^{\pi'}(s)\phi^{\pi'}(s)\, ds|| \\
&+ || \int d^{\pi}(s)(R^{\pi'}(s) - R^{\pi}(s))\phi^{\pi'}(s)\, ds|| \\
&+ || \int d^{\pi}(s)R^{\pi}(s)(\phi^{\pi'}(s) - \phi^{\pi}(s)\, ds|| \\
&+ || \int (d^{\pi'}(s) - d^{\pi}(s))\rho^*_{t+1}\phi^{\pi'}(s)\, ds|| \\
&+ || \int d^{\pi}(s)(\rho^*_{t+1} - \rho^*_{t})\phi^{\pi'}(s)\, ds|| \\
&+ || \int d^{\pi}(s)\rho^*_{t}(\phi^{\pi'}(s) - \phi^{\pi}(s)\, ds|| \\
&\leq C_r L_d||\theta_{t+1} - \theta_t|| \qquad (\text{ Assumption 4.2, Lemma A.27}) \\
&+ L_r||\theta_{t+1} - \theta_t|| \qquad (\text{ Assumption A.2}) \\
&+ C_r L_\phi||\theta_{t+1} - \theta_t|| \qquad (\text{Assumption 4.2, Assumption 4.5}) \\
&+ C_r L_d||\theta_{t+1} - \theta_t|| \qquad (\text{ Assumption 4.2, Lemma A.27}) \\
&+ L_p||\theta_{t+1} - \theta_t|| \qquad (\text{Lemma A.29}) \\
&+ C_r L_\phi||\theta_{t+1} - \theta_t|| \qquad (\text{Assumption 4.2, Assumption 4.5})
\end{aligned}
$$

$$
\implies ||b(\theta_t) - b(\theta_{t+1})|| \leq (2L_d C_r + 2C_r L_\phi + L_r + L_p)||\theta_{t+1} - \theta_t|| \tag{A.32}
$$

Using (A.27), (A.31) and (A.32) in (A.26):

$$
\begin{aligned}
||w^*_t - w^*_{t+1}|| &\leq ||A(\theta_t)^{-1} - A(\theta_{t+1})^{-1}||\ ||b(\theta_t)|| + ||A(\theta_{t+1})^{-1}||\ ||b(\theta_t) - b(\theta_{t+1})|| \\
&\leq ||A(\theta_t)^{-1}||\ ||A(\theta_t) - A(\theta_{t+1})||\ ||A(\theta_{t+1})^{-1}||\ ||b(\theta_t)|| \\
&\quad + ||A(\theta_{t+1})^{-1}||\ ||b(\theta_t) - b(\theta_{t+1})|| \\
&\leq (2L_d + 4L_\phi + L_t)||A(\theta_t)^{-1}||\ ||A(\theta_{t+1})^{-1}||\ ||b(\theta_t)||\ ||\theta_{t+1} - \theta_t|| \\
&\quad + (2L_d C_r + 2C_r L_\phi + L_r + L_p)||A(\theta_{t+1})^{-1}||\ ||\theta_{t+1} - \theta_t||
\end{aligned}
$$

Note:

- $||b(\theta_t)|| = || \int d^{\pi}(s)(\phi^{\pi}(s)(\phi^{\pi}(s))^\intercal)\, ds|| \leq C_r$ (Using Assumption 4.2)
- From Assumption A.4, $\lambda_{min}$ is the lower bound on eigen values of $A(\theta)$ for all $\theta$.

$$
\begin{aligned}
\therefore ||w^*_t - w^*_{t+1}|| &\leq \frac{C_r(2L_d + 4L_\phi + L_t)}{\lambda^2_{min}}||\theta_{t+1} - \theta_t|| \\
&\quad + \frac{(2L_d C_r + 2C_r L_\phi + L_r + L_p)}{\lambda_{min}}||\theta_{t+1} - \theta_t|| \\
&\leq L_w||\theta_{t+1} - \theta_t||
\end{aligned}
$$

where,

$$L_w = \frac{C_r(2L_d + 4L_\phi + L_t)}{\lambda_{min}^2} + \frac{(2L_d C_r + 2C_r L_\phi + L_r + L_p)}{\lambda_{min}}$$

□

**Lemma A.22.** *$Q^w_{diff}$ is the approximate differential Q-value function parameterized by $w$. Then there exist a constant $G_\theta$, independent of policy parameter $\theta$, such that:*

$$||\frac{1}{M}\sum_{i=0}^{M-1} \nabla_a Q^w_{diff}(s,a)|_{a=\pi(s)} \nabla_\theta \pi(s)|| \leq G_\theta$$

*Proof.*

$$\begin{aligned} &||Q^w_{diff}(s,a_1) - Q^w_{diff}(s,a_2)|| \leq L_a||a_1 - a_2|| \ \text{(Assumption 4.3)} \\ &\implies ||\nabla_a Q^w_{diff}(s,a)|| \leq L_a \\ &\implies ||\nabla_a Q^w_{diff}(s,a)|_{a=\pi(s)}|| \leq L_a \end{aligned} \tag{A.33}$$

$$\begin{aligned} &||\pi(s,\theta_1) - \pi(s,\theta_2)|| \leq L_\pi||\theta_1 - \theta_2|| \ \text{(Assumption 4.4)} \\ &\implies ||\nabla_\theta \pi(s)|| \leq L_\pi \end{aligned} \tag{A.34}$$

Using (A.33) and (A.34):

$$\begin{aligned} &\|\frac{1}{M}\sum_{i=0}^{M-1} \nabla_a Q^w_{diff}(s,a)|_{a=\pi(s)} \nabla_\theta \pi(s) \nabla_\theta \pi(s)|| \\ &\leq \frac{1}{M}\sum_{i=0}^{M-1} ||\nabla_a Q^w_{diff}(s,a)|_{a=\pi(s)} \nabla_\theta \pi(s) \nabla_\theta \pi(s)|| \\ &\leq L_a L_\pi = G_\theta \end{aligned}$$

□

**Lemma A.23.** *The average reward estimate $\rho_t$ is bounded.*

$$\forall t > 0 \ \ |\rho_t| \leq C_r + 2C_w$$

*Here, $C_w$ is the upper bound on critic parameter $w_t$ (Algorithm 2, step 8), $C_r$ is the upper bound on rewards (Assumption 4.2).*

*Proof.*

$$|\rho_0| \leq C_r + 2C_w \quad \text{(Assumption A.3)}$$

For t = 1:

$$\begin{aligned} \rho_1 &= \rho_0 + \alpha_0\Big(\frac{1}{M}\sum_{i=0}^{M-1} R_\pi(s_{0,i}) + \phi^\pi(s'_{0,i})^\intercal \bar{w}_0 - \phi^\pi(s_{0,i})^\intercal \bar{w}_0 - \rho_0\Big) \\ &= (1-\alpha_0)\rho_0 + \alpha_0\Big(\frac{1}{M}\sum_{i=0}^{M-1} R^\pi(s_{0,i}) + \phi^\pi(s'_{0,i})^\intercal \bar{w}_0 - \phi^\pi(s_{0,i})^\intercal \bar{w}_0\Big) \end{aligned}$$

$$
\begin{aligned}
|\rho_1| &\leq (1-\alpha_0)|\rho_0| + \alpha_0||\Big(\frac{1}{M}\sum_{i=0}^{M-1} R^\pi(s_{0,i}) + \phi^\pi(s'_{0,i})^\intercal \bar{w}_0 - \phi^\pi(s_{0,i})^\intercal \bar{w}_0\Big)|| \\
&\leq (1-\alpha_0)|\rho_0| + \alpha_0\Big(\frac{1}{M}\sum_{i=0}^{M-1} |R^\pi(s_{0,i})| + ||\phi^\pi(s'_{0,i})||\ ||\bar{w}_0|| + ||\phi^\pi(s_{0,i})||\ ||\bar{w}_0||\Big) \\
&\leq (1-\alpha_0)(C_r + 2C_w) + (\alpha_0)(C_r + 2C_w) = (C_r + 2C_w) \quad \text{(Assumption A.3)}
\end{aligned}
$$

Therefore the bound hold for t = 1.
Let the bound hold for t = k. We will prove that the bound will also hold for k+1

$$
\begin{aligned}
\rho_{k+1} &= \rho_k + \alpha_k\Big(\frac{1}{M}\sum_{i=0}^{M-1} R_\pi(s_{k,i}) + \phi^\pi(s'_{k,i})^\intercal \bar{w}_k - \phi^\pi(s_{k,i})^\intercal \bar{w}_k - \rho_k\Big) \\
&= (1-\alpha_k)\rho_k + \alpha_k\Big(\frac{1}{M}\sum_{i=0}^{M-1} R^\pi(s_{k,i}) + \phi^\pi(s'_{k,i})^\intercal \bar{w}_k - \phi^\pi(s_{k,i})^\intercal \bar{w}_k\Big)
\end{aligned}
$$

$$
\begin{aligned}
|\rho_{k+1}| &\leq (1-\alpha_k)|\rho_k| + \alpha_k||\Big(\frac{1}{M}\sum_{i=0}^{M-1} R^\pi(s_{k,i}) + \phi^\pi(s'_{k,i})^\intercal \bar{w}_k - \phi^\pi(s_{k,i})^\intercal \bar{w}_k\Big)|| \\
&\leq (1-\alpha_k)|\rho_k| + \alpha_k\Big(\frac{1}{M}\sum_{i=0}^{M-1} |R^\pi(s_{k,i})| + ||\phi^\pi(s'_{k,i})||\ ||\bar{w}_k|| + ||\phi^\pi(s_{k,i})||\ ||\bar{w}_k||\Big) \\
&\leq (1-\alpha_k)(C_r + 2C_w) + (\alpha_k)(C_r + 2C_w) = (C_r + 2C_w)
\end{aligned}
$$

The bound hold for t = k+1 as well. Hence by the principle of mathematical induction :

$$\forall t > 0 \ \ |\rho_t| \leq C_r + 2C_w$$

□

**Lemma A.24.** *The norm of target critic estimator $\bar{w}_t$ is bounded*

$$\forall t > 0 \ \ ||\bar{w}_t|| \leq C_w$$

*Here, $C_w$ is the upper bound on critic parameter $w_t$ (Algorithm 2, step 8).*

*Proof.* For t=1:

$$
\begin{aligned}
&\bar{w}_1 = (1-\beta_0)\bar{w}_0 + \beta_0 w_1 \\
&||\bar{w}_1|| \leq (1-\beta_0)||\bar{w}_0|| + \beta_0||w_1|| \\
&||\bar{w}_1|| \leq (1-\beta_0)C_w + \beta_0 C_w \quad \text{(Assumption A.3)} \\
&||\bar{w}_1|| \leq C_w
\end{aligned}
$$

The bound hold for t=1.
Let the bound hold for t = k. We will prove that the bound will also hold for k+1

$$\bar{w}_{k+1} = (1-\beta_k)\bar{w}_k + \beta_k w_{k+1}$$
$$||\bar{w}_{k+1}|| \leq (1-\beta_k)||\bar{w}_k|| + \beta_k||w_{k+1}||$$
$$||\bar{w}_{k+1}|| \leq (1-\beta_k)C_w + \beta_k C_w \quad \text{(Assumption A.3)}$$
$$||\bar{w}_{k+1}|| \leq C_w$$

The bound hold for t = k+1 as well. Hence by the principle of mathematical induction :

$$\forall t > 0 \;\; ||\bar{w}_t|| \leq C_w$$

□

**Lemma A.25.** *The norm of target average reward estimator $\bar{\rho}_t$ is bounded*

$$\forall t > 0 \;\; ||\bar{\rho}_t|| \leq C_r + 2C_w$$

*Here, $C_w$ is the upper bound on critic parameter $w_t$ (Algorithm 2, step 8), $C_r$ is the upper bound on rewards (Assumption 4.2).*

*Proof.* For t=1:

$$\bar{\rho}_1 = (1-\beta_0)\bar{\rho}_0 + \beta_0\rho_1$$
$$||\bar{\rho}_1|| \leq (1-\beta_0)||\bar{\rho}_0|| + \beta_0||\rho_1||$$
$$||\bar{\rho}_1|| \leq (1-\beta_0)(C_r + 2C_w) + \beta_0(C_r + 2C_w) \quad \text{(Assumption A.3)}$$
$$||\bar{\rho}_1|| \leq C_r + 2C_w$$

The bound hold for t=1.
Let the bound hold for t = k. We will prove that the bound will also hold for k+1

$$\bar{\rho}_{k+1} = (1-\beta_k)\bar{\rho}_k + \beta_k\rho_{k+1}$$
$$||\bar{\rho}_{k+1}|| \leq (1-\beta_k)||\bar{\rho}_k|| + \beta_k||\rho_{k+1}||$$
$$||\bar{\rho}_{k+1}|| \leq (1-\beta_k)(C_r + 2C_w) + \beta_k(C_r + 2C_w) \quad \text{(Assumption A.3)}$$
$$||\bar{\rho}_{k+1}|| \leq C_r + 2C_w$$

The bound hold for t = k+1 as well. Hence by the principle of mathematical induction :

$$\forall t > 0 \;\; ||\bar{\rho}_t|| \leq C_r + 2C_w$$

□

**Lemma A.26.** *The $A(\theta)$ matrix defined below is negative definite for all values of $\theta$ ($\theta$ is the policy parameter).*

$$A(\theta) = \int d^\pi(s)(\phi^\pi(s)(\int P^\pi(s'|s)\phi^\pi(s')\,ds' - \phi^\pi(s))^\intercal - \eta I)\,ds$$

$$\forall x \quad x^\intercal A(\theta)x \leq -\lambda||x||^2, \quad \lambda > 0$$

*$\eta$ is the l2-regularisation coefficient from Algorithm 2 and $\eta > \lambda^{all}_{max}$, where $\lambda^{all}_{max}$ is defined in the proof below.*

*Proof.* Let:

$$A'(\theta) = \int d^{\pi}(s)(\phi^{\pi}(s)(\int P^{\pi}(s'|s)\phi^{\pi}(s')\,ds' - \phi^{\pi}(s))^{\intercal})\,ds = A(\theta) + \eta I \tag{A.35}$$

Here, $\eta$ is the l2-regularization coefficient from Algorithm 2.

$$x^{\intercal}A'(\theta)x = x^{\intercal}\Big(\frac{A'(\theta)^{\intercal} + A'(\theta)}{2}\Big)x \leq \lambda_{max}(\theta)||x||^2$$

Here, $\Big(\frac{A'(\theta)^{\intercal} + A'(\theta)}{2}\Big)$ is a symmetric matrix and $\lambda_{max}(\theta)$ is the maximum eigen value of the $\Big(\frac{A'(\theta)^{\intercal} + A'(\theta)}{2}\Big)$. Using $\lambda_{max}^{all}$ from Assumption A.5:

$$\begin{aligned}
&\implies x^{\intercal}A'(\theta)x \leq \lambda_{max}^{all}||x||^2 \\
&x^{\intercal}(A'(\theta) - \eta I)x \leq (\lambda_{max}^{all} - \eta)||x||^2 \\
&x^{\intercal}A(\theta)x \leq (\lambda_{max}^{all} - \eta)||x||^2 \quad \text{(using A.45)}
\end{aligned}$$

Here, if we take $\eta > \lambda_{max}^{all}$ then we can set $\lambda = \eta - \lambda_{max}^{all}$.

$$\implies \forall x \quad x^{\intercal}A(\theta)x \leq -\lambda||x||^2, \quad \lambda > 0$$

□

**Lemma A.27.** *Let $\theta_1$ and $\theta_2$ be the policy parameter for $\pi_1$ and $\pi_2$ respectively. $d^{\pi_1}(\cdot)$ and $d^{\pi_2}(\cdot)$ be the stationary state distribution for $\pi_1$ and $\pi_2$ respectively. Here, $D_{TV}$ denotes the total variation distance between two probability distribution function. We have:*

$$\int |d^{\pi_1}(s) - d^{\pi_2}(s)|\,ds = 2D_{TV}(d^{\pi_1}, d^{\pi_2}) \leq L_d||\theta_1 - \theta_2||$$

*Here, $L_d = 2^{n+1}(\lceil \log_{\kappa} a^{-1} \rceil + 1/\kappa)L_t$. $L_t$ is the Lipchitz constant for the transition probability density function (Assumption A.1). Constants $a$ and $\kappa$ are from Assumption 3.3, $n$ is the dimension of state space.*

*Proof.*

$$\int |d^{\pi_1}(s) - d^{\pi_2}(s)|\,ds = 2D_{TV}(d^{\pi_1}, d^{\pi_2}) = 2D_{TV}(\mu_1, \mu_2)$$

Let $\mu_1$ and $\mu_2$ be the stationary state probability measure for $\pi_1$ and $\pi_2$ respectively. Then we have :

$$\begin{aligned}
d\mu_1 &= d^{\pi_1}(s)\,ds \\
d\mu_2 &= d^{\pi_2}(s)\,ds
\end{aligned}$$

Using the result of Theorem 3.1 of Mitrophanov (2005):

$$2D_{TV}(\mu_1, \mu_2) \leq 2\Big(\lceil \log_{\kappa} a^{-1} \rceil + \frac{1}{\kappa}\Big)\|K_1 - K_2\|_{TV} \tag{A.36}$$

where $K_1$ and $K_2$ are probability transition kernel for markov chain induced by policy $\pi_1$ and $\pi_2$.

From (A.36):

$$||K_1 - K_2|| \leq \sup_{\|g\|_{TV}=1} || \int g(ds)(K_1(\cdot|s) - K_2(\cdot|s))||_{TV}$$

$$
\begin{aligned}
||\int g(ds)(K_1(\cdot|s) - K_2(\cdot|s))||_{TV} &\leq \sup_{|f|\leq 1} |\iint f(s')(K_1 - K_2)(ds'|s)g(ds)| \\
&\leq \sup_{|f|\leq 1} |\iint f(s')(P^{\pi'}(s'|s) - P^{\pi}(s'|s))(s'|s)g(ds)ds'| \\
&\leq \sup_{|f|\leq 1} \iint |f(s')|\ |(P^{\pi'}(s'|s) - P^{\pi}(s'|s)|g(ds)ds' \\
&\leq L_t||\theta_1 - \theta_2|| \int g(ds) \int ds' \\
&\leq 2^m L_t||\theta_1 - \theta_2|| \\
\implies ||K_1 - K_2|| &\leq 2^m L_t||\theta_1 - \theta_2|| \qquad \text{(A.37)}
\end{aligned}
$$

From (A.36) and (A.37):

$$
\begin{aligned}
\int |d^{\pi'}(s) - d^{\pi}(s)|\, ds = 2D_{TV}(d^{\pi'}, d^{\pi'}) &\leq 2^{n+1}(\lceil \log_\kappa a^{-1} \rceil + \frac{1}{\kappa})L_t||\theta_1 - \theta_2|| \\
&\leq L_d||\theta_1 - \theta_2||
\end{aligned}
$$

□

**Lemma A.28.** *The optimal critic parameter $w_\epsilon^*$ according to compatible function approximation Lemma (3.2) is bounded by constant $C_{w_\epsilon^*}$.*

$$||w_\epsilon^*|| \leq C_{w_\epsilon^*}$$

*Proof.* From Lemma3.2:

$$
\begin{aligned}
\nabla_\theta \rho(\pi) &= \int_S d^\pi(s) \nabla_a Q^\pi_{diff}(s,a)|_{a=\pi(s)} \nabla_\theta \pi(s,\theta)\, ds \\
&= \int_S d^\pi(s) \nabla_a Q^w_{diff}(s,a)|_{a=\pi(s)} \nabla_\theta \pi(s,\theta)\, ds \\
&= \int_S d^\pi(s) \nabla_\theta \pi(s,\theta) \nabla_\theta \pi(s,\theta)^\intercal w_\epsilon^*\, ds \\
&= E[\nabla_\theta \pi(s,\theta) \nabla_\theta \pi(s,\theta)^\intercal] w_\epsilon^*
\end{aligned}
$$

Here,

$$H_\theta = E[\nabla_\theta \pi(s,\theta) \nabla_\theta \pi(s,\theta)^\intercal]$$

$$
\begin{aligned}
\nabla_\theta \rho(\pi) &= H_\theta w_\epsilon^* \\
\implies w_\epsilon^* &= H_\theta^{-1} \nabla_\theta \rho(\pi) \\
\implies ||w_\epsilon^*|| &\leq ||H_\theta^{-1}||\ ||\nabla_\theta \rho(\pi)||
\end{aligned}
$$

By using Assumption A.6, the lower bound on minimum eigenvalue of $H_\theta$ for all $\theta$ is $\lambda^\epsilon_{min}$ and using Assumption 4.3 and 4.4 :

$$||w_\epsilon^*|| \leq \frac{L_a L_\pi}{\lambda^\epsilon_{min}} = C_{w_\epsilon^*}$$

□

**Lemma A.29.** *The average reward performance metric, defined in (3), $\rho(\pi)(\rho(\theta))$ is Lipchitz continuous wrt to the policy (actor) parameter $\theta$.*

$$||\rho(\theta_1) - \rho(\theta_2)|| \leq L_p||\theta_1 - \theta_2||$$

*Proof.* Let $\theta_1$ and $\theta_2$ be the policy parameters of policy $\pi'$ and $\pi$.

$$\begin{aligned}
||\rho(\theta_1) - \rho(\theta_2)|| &= ||\rho(\pi') - \rho(\pi)|| \\
&= || \int_S d^{\pi'}(s) R^{\pi'}(s)\, ds - \int_S d^{\pi}(s) R^{\pi}(s)\, ds|| \\
&\leq || \int_S (d^{\pi'}(s) - d^{\pi}(s)) R^{\pi'}(s)\, ds|| + || \int_S d^{\pi}(s)(R^{\pi'}(s) - R^{\pi}(s))\, ds|| \\
&\leq L_d||\theta_1 - \theta_2|| \quad \text{(Lemma A.27)} \\
&\quad + L_r||\theta_1 - \theta_2|| \quad \text{(Assumption A.2)} \\
&\leq (L_d + L_r)||\theta_1 - \theta_2|| = L_p||\theta_1 - \theta_2|| \quad (L_d + L_r = L_p)
\end{aligned}$$

□

**Lemma A.30.** *The optimal critic parameter $w(\theta_t)^*$ as a function of actor parameter $\theta_t$ is Lipchitz continuous with constant $L_v$ for off-policy case. Note: $w_t^* = w(\theta_t)^*$. $\mu$ is the behaviour policy.*

$$||w_t^* - w_{t+1}^*|| \leq L_v||\theta_{t+1} - \theta_t||$$

*Proof.* $\eta$ is the l2-regularisation coefficient from Algorithm 3 and $\eta > \chi_{max}^{all}$, where $\chi_{max}^{all}$ is defined in Lemma A.31. Because of carefully setting the value of $\eta$, $A(\theta_t)$ is negative definite. Thus, for on-policy TD(0) with l2-regularization and target estimators, the following condition holds true for optimal critic parameter $w_t^*$:

$$E[(R^{\mu}(s) - \rho_t^*)\phi^{\pi}(s) + (\phi^{\pi}(s)(E[\phi^{\pi}(s')] - \phi^{\pi}(s))^{\intercal} - \eta I)w_t^*] = 0$$

$$\begin{aligned}
b(\theta_t) &:= E[(R^{\mu}(s) - \rho_t^*)\phi^{\pi}(s)] \\
A(\theta_t) &:= E[(\phi^{\pi}(s)(E[\phi^{\pi}(s')] - \phi^{\pi}(s))^{\intercal} - \eta I)]
\end{aligned}$$

$$\therefore b(\theta_t) + A(\theta_t)w_t^* = 0 \implies w_t^* = -A(\theta_t)^{-1}b(\theta_t)$$

Expectation above is with respect to stationary state distribution $d^{\mu}(\cdot)$ of policy $\mu$. Please note the abuse of notation here, $A(\theta_t)$ is actually same as $A_{off}^{\mu}(\theta_t)$ of Lemma A.31.

$$\begin{aligned}
||w_t^* - w_{t+1}^*|| &= ||A(\theta_t)^{-1}b(\theta_t) - A(\theta_{t+1})^{-1}b(\theta_{t+1})|| \\
&\leq ||A(\theta_t)^{-1}b(\theta_t) - A(\theta_{t+1})^{-1}b(\theta_t) + A(\theta_{t+1})^{-1}b(\theta_t) - A(\theta_{t+1})^{-1}b(\theta_{t+1})|| \\
&\leq ||A(\theta_t)^{-1} - A(\theta_{t+1})^{-1}||\, ||b(\theta_t)|| \;\; ① \\
&\quad + ||A(\theta_{t+1})^{-1}||\, ||b(\theta_t) - b(\theta_{t+1})|| \;\; ②
\end{aligned} \tag{A.38}$$

From (A.38):

①:

$$\begin{aligned}
||A(\theta_t)^{-1} - A(\theta_{t+1})^{-1}|| &= ||A(\theta_t)^{-1}A(\theta_{t+1})A(\theta_{t+1})^{-1} - A(\theta_t)^{-1}A(\theta_t)^{A}(\theta_{t+1})^{-1}|| \\
&\leq ||A(\theta_t)^{-1}||\, ||A(\theta_t) - A(\theta_{t+1})||\, ||A(\theta_{t+1})^{-1}||
\end{aligned} \tag{A.39}$$

From (A.39):

Here, $\pi'$ and $\pi$ represents the policy with parameter $\theta_{t+1}$ and $\theta_t$ respectively and $\mu$ be the behaviour policy .

$$\begin{aligned}
||A(\theta_t) - A(\theta_{t+1})|| &\leq || \int d^\mu(s)(\phi^{\pi'}(s)(\int P^\mu(s'|s)\phi^{\pi'}(s')\,ds' - \phi^{\pi'}(s))^\intercal - \eta I)\,ds \\
&\quad - \int d^\mu(s)(\phi^{\pi}(s)(\int P^\mu(s'|s)\phi^{\pi}(s')\,ds' - \phi^{\pi}(s))^\intercal - \eta I)\,ds|| \\
&\leq || \int d^\mu(s)(\phi^{\pi'}(s)(\int P^\mu(s'|s)\phi^{\pi'}(s')\,ds')^\intercal)\,ds \\
&\quad - \int d^\mu(s)(\phi^{\pi}(s)(\int P^\mu(s'|s)\phi^{\pi}(s')\,ds')^\intercal)\,ds|| \quad ① \\
&\leq || \int d^\mu(s)(\phi^{\pi}(s)(\phi^{\pi}(s))^\intercal)\,ds - \int d^\mu(s)(\phi^{\pi'}(s)(\phi^{\pi'}(s))^\intercal)\,ds|| \quad ②
\end{aligned} \tag{A.40}$$

From (A.40):

①:

$$\begin{aligned}
&|| \int d^\mu(s)(\phi^{\pi'}(s)(\int P^\mu(s'|s)\phi^{\pi'}(s')\,ds')^\intercal)\,ds - \int d^\mu(s)(\phi^{\pi}(s)(\int P^\mu(s'|s)\phi^{\pi}(s')\,ds')^\intercal)\,ds|| \\
&\leq || \int (d^\mu(s) - d^\mu(s))\phi^{\pi'}(s)(\int P^\mu(s'|s)\phi^{\pi'}(s')\,ds')^\intercal\,ds|| \\
&\quad + || \int d^\mu(s)(\phi^{\pi'}(s) - \phi^{\pi}(s))(\int P^\mu(s'|s)\phi^{\pi'}(s')\,ds')^\intercal\,ds|| \\
&\quad + || \int d^\mu(s)\phi^{\pi}(s)(\int (P^\mu(s'|s) - P^\mu(s'|s))\phi^{\pi'}(s')\,ds')^\intercal\,ds|| \\
&\quad + || \int d^\mu(s)\phi^{\pi}(s)(\int P^\mu(s'|s)(\phi^{\pi'}(s') - \phi^{\pi}(s'))\,ds'^\intercal)\,ds \\
&\leq L_\phi||\theta_{t+1} - \theta_t|| \qquad \text{(Assumption 4.5)} \\
&+ L_\phi||\theta_{t+1} - \theta_t|| \qquad \text{(Assumption 4.5)}
\end{aligned}$$

$$\begin{aligned}
&|| \int d^\mu(s)(\phi^{\pi'}(s)(\int P^\mu(s'|s)\phi^{\pi'}(s')\,ds')^\intercal)\,ds - \int d^\mu(s)(\phi^{\pi}(s)(\int P^\mu(s'|s)\phi^{\pi}(s')\,ds')^\intercal)\,ds|| \\
&\leq 2L_\phi||\theta_{t+1} - \theta_t||
\end{aligned} \tag{A.41}$$

From (A.40):

②:

$$\begin{aligned}
&|| \int d^\mu(s)(\phi^{\pi}(s)(\phi^{\pi}(s))^\intercal)\,ds - \int d^\mu(s)(\phi^{\pi'}(s)(\phi^{\pi'}(s))^\intercal)\,ds|| \\
&\leq || \int d^\mu(s)(\phi^{\pi}(s) - \phi^{\pi'}(s))(\phi^{\pi}(s))^\intercal\,ds|| \\
&\quad + || \int d^\mu(s)\phi^{\pi'}(s)(\phi^{\pi}(s) - \phi^{\pi'}(s))^\intercal\,ds|| \\
&\leq 2L_\phi||\theta_{t+1} - \theta_t||
\end{aligned} \tag{A.42}$$

Using (A.41) and (A.42) in (A.40)

$$||A(\theta_t) - A(\theta_{t+1})|| \leq (4L_\phi + L_t)||\theta_{t+1} - \theta_t|| \tag{A.43}$$

From (A.38):

②:

$$
\begin{aligned}
||b(\theta_t) - b(\theta_{t+1})|| &= || \int d^\mu(s)((R^\mu(s) - \rho^*_{t+1})\phi^{\pi'}(s)\, ds - \int d^\mu(s)(R^\mu(s) - \rho^*_t)\phi^\pi(s)\, ds|| \\
&\leq || \int d^\mu(s)(R^\mu(s)\phi^{\pi'}(s)\, ds - \int d^\mu(s)R^\mu(s)\phi^\pi(s)\, ds|| \\
&+ || \int d^\mu(s)\rho^*_{t+1}\phi^{\pi'}(s)\, ds - \int d^\mu(s)\rho^*_t\phi^\pi(s)\, ds|| \\
&\leq || \int d^\mu(s)(R^\mu(s) - R^\mu(s))\phi^{\pi'}(s)\, ds|| \\
&+ || \int d^\mu(s)R^\mu(s)(\phi^{\pi'}(s) - \phi^\pi(s)\, ds|| \\
&+ || \int d^\mu(s)(\rho^*_{t+1} - \rho^*_t)\phi^{\pi'}(s)\, ds|| \\
&+ || \int d^\mu(s)\rho^*_t(\phi^{\pi'}(s) - \phi^\pi(s)\, ds|| \\
&\leq C_r L_\phi ||\theta_{t+1} - \theta_t|| \qquad \text{(Assumption 4.2)} \\
&+ L_p ||\theta_{t+1} - \theta_t|| \qquad \text{(Lemma A.29)} \\
&+ C_r L_\phi ||\theta_{t+1} - \theta_t|| \qquad \text{(Assumption 4.2)}
\end{aligned}
$$

$$
\implies ||b(\theta_t) - b(\theta_{t+1})|| \leq (2C_r L_\phi + L_p)||\theta_{t+1} - \theta_t|| \tag{A.44}
$$

Using (A.39), (A.43) and (A.44) in (A.38):

$$
\begin{aligned}
||w^*_t - w^*_{t+1}|| &\leq ||A(\theta_t)^{-1} - A(\theta_{t+1})^{-1}||\, ||b(\theta_t)|| + ||A(\theta_{t+1})^{-1}||\, ||b(\theta_t) - b(\theta_{t+1})|| \\
&\leq ||A(\theta_t)^{-1}||\, ||A(\theta_t) - A(\theta_{t+1})||\, ||A(\theta_{t+1})^{-1}||\, ||b(\theta_t)|| \\
&\quad + ||A(\theta_{t+1})^{-1}||\, ||b(\theta_t) - b(\theta_{t+1})|| \\
&\leq 4L_\phi ||A(\theta_t)^{-1}||\, ||A(\theta_{t+1})^{-1}||\, ||b(\theta_t)||\, ||\theta_{t+1} - \theta_t|| \\
&\quad + (2C_r L_\phi + L_p)||A(\theta_{t+1})^{-1}||\, ||\theta_{t+1} - \theta_t||
\end{aligned}
$$

Note:

- $||b(\theta_t)|| = || \int d^\mu(s)(\phi^\pi(s)(\phi^\pi(s))^\intercal)\, ds|| \leq C_r$ (Assumption 4.2)
- Let $\lambda_{min}$ is the lower bound on eigen values of $A(\theta)$ for all $\theta$.

$$
\begin{aligned}
\therefore ||w^*_t - w^*_{t+1}|| &\leq \frac{C_r(4L_\phi)}{\lambda^2_{min}}||\theta_{t+1} - \theta_t|| \\
&\quad + \frac{(2C_r L_\phi + L_p)}{\lambda_{min}}||\theta_{t+1} - \theta_t|| \\
&\leq L_v ||\theta_{t+1} - \theta_t||
\end{aligned}
$$

where,

$$
L_v = \frac{4C_r L_\phi}{\lambda^2_{min}} + \frac{C_r L_\phi}{\lambda_{min}}
$$

□

**Lemma A.31.** *The $A^{\mu}_{off}(\theta)$ matrix defined below is negative definite for all values of $\theta$ ($\theta$ is the policy parameter). $\theta^{\mu}$ is the policy parameter for behaviour policy $\mu$.*

$$A^{\mu}_{off}(\theta) := \int d^{\mu}(s)(\phi^{\pi}(s)(\int P^{\pi}(s'|s)\phi^{\pi}(s')\,ds' - \phi^{\pi}(s))^{\intercal} - \eta I)\,ds$$

$$\forall x \quad x^{\intercal}A^{\mu}_{off}(\theta)x \leq -\lambda||x||^2, \quad \lambda > 0$$

*$\eta$ is the l2-regularisation coefficient from Algorithm 3 and $\eta > \chi^{all}_{max}$, where $\chi^{all}_{max}$ is defined in the proof below.*

*Proof.* Let:

$$A^{\mu'}_{off}(\theta) = \int d^{\mu}(s)(\phi^{\pi}(s)(\int P^{\pi}(s'|s)\phi^{\pi}(s')\,ds' - \phi^{\pi}(s))^{\intercal})\,ds = A^{\mu}_{off}(\theta) + \eta I \tag{A.45}$$

Here, $\eta$ is the l2-regularization coefficient from Algorithm 2.

$$x^{\intercal}A^{\mu'}_{off}(\theta)x = x^{\intercal}\Big(\frac{A^{\mu'}_{off}(\theta)^{\intercal} + A^{\mu'}_{off}(\theta)}{2}\Big)x \leq \chi_{max}(\theta)||x||^2$$

Here, $\Big(\frac{A^{\mu'}_{off}(\theta)^{\intercal} + A^{\mu'}_{off}(\theta)}{2}\Big)$ is a symmetric matrix and $\chi_{max}(\theta)$ is the maximum eigenvalue of the $\Big(\frac{A^{\mu'}_{off}(\theta)^{\intercal} + A^{\mu'}_{off}(\theta)}{2}\Big)$. Using $\chi^{all}_{max}$ from Assumption A.7:

$$\begin{aligned}
&\implies x^{\intercal}A^{\mu'}_{off}(\theta)x \leq \chi^{all}_{max}||x||^2 \\
&x^{\intercal}(A^{\mu'}_{off}(\theta) - \eta I)x \leq (\chi^{all}_{max} - \eta)||x||^2 \\
&x^{\intercal}A^{\mu}_{off}(\theta)x \leq (\chi^{all}_{max} - \eta)||x||^2 \quad \text{(using A.45)}
\end{aligned}$$

Here, if we take $\eta > \chi^{all}_{max}$ then we can set $\lambda = \eta - \chi^{all}_{max}$.

$$\implies \forall x \quad x^{\intercal}A^{\mu}_{off}(\theta)x \leq -\lambda||x||^2, \quad \lambda > 0$$

□

### A.4. Asymptotic Convergence Analysis

**Theorem A.32.** *In Algorithm 4, let policy parameter $\theta_t$ be kept constant at $\theta$. The critic parameter $w_t$ and the target critic parameter $\bar{w}_t$ converges to $w(\theta)^*$. Also, average reward estimator $\rho_t$ and target average reward estimator $\bar{\rho}_t$ converges to $\rho(\theta)^*$.*

*Proof.* For simplicity of proof we are assuming the batch size M to be 1. Critic parameter $w_t \in \mathbb{R}^k$, $\phi^{\pi}(s) \in \mathbb{R}^k$ and $\rho_t$ is a scalar. Let the update rules used for critic parameter and average reward estimator be as follows:

$$\begin{aligned}
w_{t+1} &= w_t + \alpha_t\Big(R^{\pi}(s_t) - \bar{\rho}_t + \phi^{\pi}(s'_t)^{\intercal}\bar{w}_t - \phi^{\pi}(s_t)^{\intercal}w_t\Big)\phi^{\pi}(s_t) - \alpha_t\eta w_t \\
\rho_{t+1} &= \rho_t + \alpha_t\Big(R^{\pi}(s_t) - \rho_t + \phi^{\pi}(s'_t)^{\intercal}\bar{w}_t - \phi^{\pi}(s_t)^{\intercal}\bar{w}_t\Big) \\
\overline{w}_{t+1} &= \overline{w}_t + \beta_t(w_{t+1} - \overline{w}_t) \\
\overline{\rho}_{t+1} &= \overline{\rho}_t + \beta_t(\rho_{t+1} - \overline{\rho}_t)
\end{aligned} \tag{A.46}$$

Let us define $z_t$ as $[w_t \quad \rho_t]^\intercal$ and $\bar{z}_t$ as $[\bar{w}_t \quad \bar{\rho}_t]^\intercal$. $\mathbf{0}$ is a vector in $\mathbb{R}^k$ and $I_0$ is an identity matrix in $\mathbb{R}^{(k+1)\times(k+1)}$ with $I_0[k][k] = 0$ (assuming indexing starts from 0).

$$\begin{aligned}\begin{bmatrix} w_{t+1} \\ \rho_{t+1} \end{bmatrix} &= \begin{bmatrix} w_t \\ \rho_t \end{bmatrix} + \alpha_t \Bigg( R^\pi(s_t) \begin{bmatrix} \phi^\pi(s_t) \\ 1 \end{bmatrix} + \begin{bmatrix} \phi^\pi(s_t)\phi^\pi(s'_t)^\intercal & -\phi^\pi(s_t) \\ \phi^\pi(s'_t)^\intercal - \phi^\pi(s_t)^\intercal & 0 \end{bmatrix} \begin{bmatrix} \bar{w}_t \\ \bar{\rho}_t \end{bmatrix} \\ &\quad - \begin{bmatrix} \phi^\pi(s_t)\phi^\pi(s_t)^\intercal & \mathbf{0} \\ \mathbf{0}^\intercal & 1 \end{bmatrix} \begin{bmatrix} w_t \\ \rho_t \end{bmatrix} - \eta I_0 \begin{bmatrix} w_t \\ \rho_t \end{bmatrix} \Bigg) \\ \begin{bmatrix} \bar{w}_{t+1} \\ \bar{\rho}_{t+1} \end{bmatrix} &= \begin{bmatrix} \bar{w}_t \\ \bar{\rho}_t \end{bmatrix} + \beta_t \Big( \begin{bmatrix} w_{t+1} \\ \rho_{t+1} \end{bmatrix} - \begin{bmatrix} \bar{w}_t \\ \bar{\rho}_t \end{bmatrix} \Big)\end{aligned} \tag{A.47}$$

Here, $R^\pi(s_t) \begin{bmatrix} \phi^\pi(s_t) \\ 1 \end{bmatrix} = R^\pi_\phi(s_t)$, $A_\phi(s_t, s'_t) = \begin{bmatrix} \phi^\pi(s_t)\phi^\pi(s'_t)^\intercal & -\phi^\pi(s_t) \\ \phi^\pi(s'_t)^\intercal - \phi^\pi(s_t)^\intercal & 0 \end{bmatrix}$ and $B_\phi(s_t) = \begin{bmatrix} \phi^\pi(s_t)\phi^\pi(s_t)^\intercal & \mathbf{0} \\ \mathbf{0}^\intercal & 1 \end{bmatrix}$

$$\begin{aligned} z_{t+1} &= z_t + \alpha_t \big( R^\pi_\phi(s_t) + A_\phi(s_t, s'_t)\bar{z}_t - (B_\phi(s_t) + \eta I_0) z_t \big) \\ \bar{z}_{t+1} &= \bar{z}_t + \beta_t (z_{t+1} - \bar{z}_t) \end{aligned} \tag{A.48}$$

Now, we will use the extension of stability criteria for iterates given Borkar & Meyn (2000) to two timescale stochastic approximation scheme (Lakshminarayanan & Bhatnagar, 2017) to show the boundedness of the critic parameter and average reward estimator together. Let us write A.48 in the standard form of stochastic approximation scheme.

$$z_{t+1} = z_t + \alpha_t \big( h(z_t, \bar{z}_t) + \mathcal{M}^1_{t+1} \big)$$

Let, $\bar{R}^\pi_\phi = \int_S d^\pi(s_t) R^\pi_\phi(s_t)\, ds_t$, $\bar{A}_\phi = \int_S d^\pi(s_t) \int_S P^\pi(s'_t|s_t) A_\phi(s_t, s'_t)\, ds'_t\, ds_t$, $\bar{B}_\phi = \int_S d^\pi(s_t) B_\phi(s_t)\, s_t$

Here,

$$\begin{aligned} h(z_t, \bar{z}_t) &= \int_S d^\pi(s_t) \big( R^\pi_\phi(s_t) + A_\phi(s_t, s'_t)\bar{z}_t - (B_\phi(s_t) + \eta I_0) z_t \big)\, ds_t \\ &= \bar{R}^\pi_\phi + \bar{A}_\phi \bar{z}_t - (\bar{B}_\phi + \eta I_0) z_t \\ \mathcal{M}^1_{t+1} &= R^\pi_\phi(s_t) + A_\phi(s_t, s'_t)\bar{z}_t - (B_\phi(s_t) + \eta I_0) z_t - h(z_t, \bar{z}_t) \end{aligned}$$

$$\bar{z}_{t+1} = \bar{z}_t + \beta_t \big( g(z_t, \bar{z}_t) + \mathcal{M}^2_{t+1} + \epsilon(n) \big)$$

Here,

$$\begin{aligned} g(z_t, \bar{z}_t) &= \lambda(\bar{z}_t) - \bar{z}_t \\ \mathcal{M}^2_{t+1} &= 0 \\ \lambda(\bar{z}_t) &= (B_\phi + \eta I_0)^{-1} (R^\pi_\phi + A_\phi \bar{z}_t) \\ \epsilon(n) &= z_{t+1} - \lambda(\bar{z}_t) \end{aligned}$$

$\lambda(\bar{z})$ is the unique globally asymptotically stable equilibrium point of the ODE $\dot{z} = h(z(t), \bar{z})$. $\lambda$ used here has no relation to usage of $\lambda$ in any other section of the paper. Using Lemma 1 of Chapter 6 of (Borkar, 2009), we have $\|z_{t+1} - \lambda(\bar{z}_t)\| \to 0$. Hence $\epsilon(n) = o(1)$. Therefore we can use the conclusion of (Lakshminarayanan & Bhatnagar, 2017).

We will now satisfy condition A1 till condition A5 of (Lakshminarayanan & Bhatnagar, 2017) to prove the boundedness of the critic parameter:

**Condition A1**:

$$\begin{aligned}\|h(z_1,\bar{z}_1)-h(z_2,\bar{z}_2)\| &= \|\bar{A}_\phi(\bar{z}_1-\bar{z}_2)-(\bar{B}_\phi+\eta I_0)(z_1-z_2)\| \\ &\leq \|\bar{A}_\phi\|\|\bar{z}_1-\bar{z}_2\|+\|\bar{B}_\phi+\eta I_0\|\|z_1-z_2\| \\ &\leq \max(\|\bar{A}_\phi\|,\|\bar{B}_\phi+\eta I_0\|)(\|\bar{z}_1-\bar{z}_2\|+\|z_1-z_2\|) \\ &= L_h(\|\bar{z}_1-\bar{z}_2\|+\|z_1-z_2\|)\ \ (L_h=\max(\|\bar{A}_\phi\|,\|\bar{B}_\phi+\eta I_0\|))\end{aligned} \tag{A.49}$$

Therefore, $h(z,\bar{z})$ is Lipchitz continuous with constant $L_h$.

$$\begin{aligned}\|g(z_1,\bar{z}_1)-g(z_2,\bar{z}_2)\| &= \|((\bar{B}_\phi+\eta I_0)A_\phi-I)(\bar{z}_1-\bar{z}_2)\| \\ &\leq \|((\bar{B}_\phi+\eta I_0)A_\phi-I)\|\|\bar{z}_1-\bar{z}_2\| \\ &= L_g\|\bar{z}_1-\bar{z}_2\|\ \ (L_g=\|((\bar{B}_\phi+\eta I_0)A_\phi-I)\|)\end{aligned} \tag{A.50}$$

Therefore, $g(z,\bar{z})$ is Lipchitz continuous with constant $L_g$.

Using A.49 and A.50, condition A1 is satisfied.

**Condition A2:**
Let us define an increasing sequence of $\sigma-$fields $\{\mathcal{F}_t\}$ as $\{z_m,\bar{z}_m,\mathcal{M}^1_m,\mathcal{M}^2_m,m\leq t\}$.

$$\begin{aligned}E[\mathcal{M}^1_{t+1}|\mathcal{F}_t] &= E[R^\pi_\phi(s_t)+A_\phi(s_t,s'_t)\bar{z}_t-(B_\phi(s_t)+\eta I_0)z_t-h(z_t,\bar{z}_t)|\mathcal{F}_t] \\ &= \int_S d^\pi(s_t)\big(R^\pi_\phi(s_t)+A_\phi(s_t,s'_t)\bar{z}_t-(B_\phi(s_t)+\eta I_0)z_t\big)\,ds_t-h(z_t,\bar{z}_t) \\ &= 0\end{aligned}$$

$$E[\mathcal{M}^2_{t+1}|\mathcal{F}_t]=0$$

Hence, $\{\mathcal{M}^1_t\}$ and $\{\mathcal{M}^2_t\}$ are martingale difference sequence.

$$\begin{aligned}\|\mathcal{M}^1_{t+1}\|^2 &= \|(R^\pi_\phi(s_t)-\bar{R}_\phi)+(A_\phi(s_t,s'_t)-\bar{A}_\phi)\bar{z}_t-(B_\phi(s_t)-\bar{B}_\phi(s_t))z_t\|^2 \\ &\leq 3(\|R^\pi_\phi(s_t)-\bar{R}_\phi\|^2+\|(A_\phi(s_t,s'_t)-\bar{A}_\phi)\|^2\|\bar{z}_t\|^2+\|B_\phi(s_t)-\bar{B}_\phi(s_t))\|^2\|z_t\|^2) \\ &\leq K_1(1+\|z_t\|^2+\|\bar{z}_t\|^2)\end{aligned}$$

Here, $K_1 = 6\max(\|R^\pi_\phi(s_t)\|,\|(A_\phi(s_t,s'_t)\|,\|B_\phi(s_t)\|)$ and it follows from Assumption 4.1 and 4.2. We have, $E[\|\mathcal{M}^1_{t+1}\|^2||\mathcal{F}_t]\leq K_1(1+\|z_t\|^2+\|\bar{z}_t\|^2)$ and $E[\|\mathcal{M}^2_{t+1}\|^2||\mathcal{F}_t]\leq K_2(1+\|z_t\|^2+\|\bar{z}_t\|^2)$. $K_2$ can be any positive constant. Hence condition A2 is satisfied.

**Condition A3:**

We have, $\sum_t\alpha_t=\sum_t\frac{C_\alpha}{(1+t)^\sigma}=\infty$, $\sum_t\beta_t=\sum_t\frac{C_\beta}{(1+t)^u}=\infty$ and $\sum_t(\alpha_t^2+\beta_t^2)=\sum_t\big((\frac{C_\alpha}{(1+t)^\sigma})^2+(\frac{C_\beta}{(1+t)^u})^2\big)<\infty$. We can carefully set the value of $\sigma$ and $u$ to satisfy the conditions on step sizes. Further if $\sigma<u$ then $\beta_t=o(\alpha_t)$.

**Condition A4:**

$$\begin{aligned}h_c(z,\bar{z}) &:= \frac{h(cz,c\bar{z})}{c} \\ h_c(z,\bar{z}) &= \frac{\bar{R}^\pi_\phi+c\bar{A}_\phi\bar{z}_t-c(\bar{B}_\phi+\eta I_0)z_t}{c} \\ \lim_{c\to\infty}h_c(z,\bar{z}) &= \lim_{c\to\infty}\frac{\bar{R}^\pi_\phi+c\bar{A}_\phi\bar{z}_t-c(\bar{B}_\phi+\eta I_0)z_t}{c} \\ &= \bar{A}_\phi\bar{z}_t-(\bar{B}_\phi+\eta I_0)z_t\end{aligned}$$

Let us define $h_\infty(z_t, \bar{z}_t) := \bar{A}_\phi \bar{z}_t - (\bar{B}_\phi + \eta I_0) z_t$.The ODE $\dot{z}(t) := h_\infty(z(t), \bar{z})$ has a unique globally asymptotically stable equilibrium point $\lambda_\infty(\bar{z}) = (\bar{B}_\phi + \eta I_0)^{-1} \bar{A}_\phi \bar{z}$ if $(\bar{B}_\phi + \eta I_0)$ is positive definite matrix. Let $C_\phi = \int_S d^\pi(s_t) \phi^\pi(s_t) \phi^\pi(s_t)^\intercal \, ds_t$.

$$\bar{B}_\phi + \eta I_0 = \begin{bmatrix} C_\phi + \eta I & \mathbf{0} \\ \mathbf{0}^\intercal & 1 \end{bmatrix}$$
$$\begin{bmatrix} w^\intercal & \rho \end{bmatrix} \begin{bmatrix} C_\phi + \eta I & \mathbf{0} \\ \mathbf{0}^\intercal & 1 \end{bmatrix} \begin{bmatrix} w \\ \rho \end{bmatrix} = w^\intercal (C_\phi + \eta I) w + \rho^2$$

If $\eta$ is strictly greater than negative of the minimum eigenvalue of $C_\phi$ then,

$$\begin{aligned} \forall \begin{bmatrix} w \\ p \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} w^\intercal & \rho \end{bmatrix} \begin{bmatrix} C_\phi + \eta I & \mathbf{0} \\ \mathbf{0}^\intercal & 1 \end{bmatrix} \begin{bmatrix} w \\ \rho \end{bmatrix} > 0 \\ \forall \begin{bmatrix} w \\ p \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} w^\intercal & \rho \end{bmatrix} \begin{bmatrix} \bar{B}_\phi + \eta I_0 \end{bmatrix} \begin{bmatrix} w \\ \rho \end{bmatrix} > 0 \end{aligned} \tag{A.51}$$

Hence, for $\eta + \lambda_{min}(C_\phi) > 0$, $\bar{B}_\phi + \eta I_0$ is positive definite matrix. Therefore, the ODE $\dot{z}(t) := h_\infty(z(t), \bar{z})$ has a unique globally asymptotically stable equilibrium point $\lambda_\infty(\bar{z})$ and $\lambda_\infty(0) = 0$. Condition A4 is satisfied.

**Condition A5:**

$$\begin{aligned} g_c(\bar{z}) &:= \frac{g(c\lambda_\infty(\bar{z}), c\bar{z})}{c} \\ g_c(\bar{z}) &= \frac{(\bar{B}_\phi + \eta I_0)^{-1}(R^\pi_\phi + cA_\phi \bar{z}) - c\bar{z}}{c} \\ \lim_{c\to\infty} g_c(\bar{z}) &= \lim_{c\to\infty} \frac{(\bar{B}_\phi + \eta I_0)^{-1}(R^\pi_\phi + cA_\phi \bar{z}) - c\bar{z}}{c} \\ &= (\bar{B}_\phi + \eta I_0)^{-1} A_\phi \bar{z} - \bar{z} \end{aligned} \tag{A.52}$$

Let us define $g_\infty := ((\bar{B}_\phi + \eta I_0)^{-1} A_\phi - I)\bar{z}$. The ODE $\dot{\bar{z}}(t) = g_\infty(\bar{z}(t))$ has origin as its unique globally asymptotically stable equilibrium if $I - (\bar{B}_\phi + \eta I_0)^{-1} A_\phi$ is positive definite matrix.

$\|\cdot\|$ refers to L2-norm. $\lambda_i$ are the eigenvalues of the matrix $C_\phi$. Let us assume the following:

$$\begin{aligned} &\max(1, \max_i(\frac{1}{\lambda_i + \eta})) = \|(\bar{B}_\phi + \eta I_0)^{-1}\| \leq \frac{1}{\|A_\phi\|} \\ &\implies \|(\bar{B}_\phi + \eta I_0)^{-1}\| \|A_\phi\| < 1 \\ &\implies \|x\| \|(\bar{B}_\phi + \eta I_0)^{-1}\| \|A_\phi\| \|x\| < \|x\|^2 \\ &\implies \|x^\intercal (\bar{B}_\phi + \eta I_0)^{-1} A_\phi x\| < \|x\|^2 \\ &\implies x^\intercal (\bar{B}_\phi + \eta I_0)^{-1} A_\phi x < \|x\|^2 \\ &\implies x^\intercal (I - (\bar{B}_\phi + \eta I_0)^{-1} A_\phi) x > 0 \end{aligned} \tag{A.53}$$

Hence, if $\max(1, \max_i(\frac{1}{\lambda_i + \eta})) < \frac{1}{\|A_\phi\|}$, then $I - (\bar{B}_\phi + \eta I_0)^{-1} A_\phi$ is positive definite matrix. Therefore, the ODE $\dot{\bar{z}}(t) = g_\infty(\bar{z}(t))$ has origin as its unique globally asymptotically stable. Condition A5 is satisfied.

Let us the consider the ODE $\dot{z}(t) = h(z(t), \bar{z})$. Here, $h(z(t), \bar{z}) = \bar{R}^\pi_\phi + \bar{A}_\phi \bar{z} - (\bar{B}_\phi + \eta I_0) z(t)$ As earlier, for $\eta + \lambda_{min}(C_\phi) > 0$, $\bar{B}_\phi + \eta I_0$ is positive definite matrix. Therefore, the ODE $\dot{z}(t) := h(z(t), \bar{z})$ has a unique globally asymptotically stable equilibrium point $\lambda(\bar{z}) = (B_\phi + \eta I_0)^{-1}(R^\pi_\phi + A_\phi \bar{z})$.

Now, consider the ODE $\dot{\bar{z}}(t) = g(\lambda(\bar{z}(t)), \bar{z}(t))$. Here, $g(\lambda(\bar{z}(t)), \bar{z}(t)) = \lambda(\bar{z}(t)) - \bar{z}(t) = (B_\phi + \eta I_0)^{-1} R^\pi_\phi - (I - (B_\phi + \eta I_0)^{-1} A_\phi)\bar{z}(t)$. For $\max(1, \max_i(\frac{1}{\lambda(C_\phi)_i + \eta})) < \frac{1}{\|A_\phi\|}$, $I - (\bar{B}_\phi + \eta I_0)^{-1} A_\phi$ is positive definite matrix. Therefor the ODE $\dot{\bar{z}}(t) = g(\lambda(\bar{z}(t)), \bar{z}(t))$ has a unique globally asymptotically stable equilibrium point $(B_\phi + \eta I_0 - A_\phi)^{-1} R^\pi_\phi$.

Since condition A1-A5 are satisfied, using the conclusion of (Lakshminarayanan & Bhatnagar, 2017), $z_t (= [w_t \ \rho_t]^\intercal)$ and $\bar{z}_t (= [\bar{w}_t \ \bar{\rho}_t]^\intercal)$ converges to $(B_\phi + \eta I_0 - A_\phi)^{-1} R^\pi_\phi (= [w(\theta)^* \ \rho(\theta)^*]^\intercal = [\bar{w}(\theta)^* \ \bar{\rho}(\theta)^*]^\intercal)$. □

**Lemma A.33.** *For policy $\pi$ parameterized by $\theta$ and optimal critic parameter $w^* (= w(\theta_t)^*)$ according to TheoremA.32, the following condition holds true:*

$$\mathbb{E}[\nabla_a Q^{w^*}_{diff}(s,a)|_{a=\pi(s)} \nabla_\theta \pi(s) | \theta] = \nabla_\theta \rho(\pi) + e^\pi$$

*Here, $e^\pi$ denotes the error in gradient due to function approximation.*

$$e^\pi = \int_S d^\pi(s) \big( (\nabla_a Q^{w^*}_{diff}(s,a) - \nabla_a Q^\pi_{diff}(s,a))|_{a=\pi(s)} \big) \nabla_\theta \pi(s) \, ds$$

*Proof.*

$$\begin{aligned}
\mathbb{E}[\nabla_a Q^w_{diff}(s,a)|_{a=\pi(s)} \nabla_\theta \pi(s) | \theta] &= \int_S d^\pi(s) \nabla_a Q^w_{diff}(s,a) \nabla_\theta \big) \pi(s) \, ds \\
&= \int_S d^\pi(s) \big( (\nabla_a Q^w_{diff}(s,a) - \nabla_a Q^\pi_{diff}(s,a))|_{a=\pi(s)} \big) \nabla_\theta \pi(s) \, ds \\
&\quad + \int_S d^\pi(s) \nabla_a Q^\pi_{diff}(s,a)|_{a=\pi(s)} \nabla_\theta \pi(s) \, ds \\
&= \nabla_\theta \rho(\pi) + e^\pi \quad \text{(Using Theorem 3.1)}
\end{aligned}$$

□

We will now prove the convergence of policy parameter $\theta_t (\in \mathbb{R}^d)$ using the following update rule (M = 1):

$$\theta_{t+1} = \Gamma_{C_\theta} \Big( \theta_t + \gamma_t \nabla_a Q^w_{diff}(s_t, a)|_{a=\pi(s_t)} \nabla_\theta \pi(s_t) \Big)$$

Here, $s_t$ is the state sampled from the buffer at time step $t$. $\Gamma_{C_\theta} : \mathbb{R}^d \to C_\theta$ is a projection operator, where $C_\theta$ is compact convex set.

**Theorem A.34.** *$\Gamma_{C_\theta} : \mathbb{R}^d \to C_\theta$ is a projection operator, where $C_\theta$ is compact convex set and $\hat{\Gamma}_{C_\theta}(\theta) \nabla_\theta \rho(\theta)$ refers to directional derivative of $\Gamma_{C_\theta}(\cdot)$ in the direction $\nabla_\theta \rho(\theta)$ at $\theta$. Let $K = \{\theta \in C_\theta | \hat{\Gamma}_{C_\theta}(\theta) \nabla_\theta \rho(\theta) = 0\}$ and $K^\epsilon = \{\theta' \in C_\theta | \exists \ \theta \in K \ \|\theta' - \theta\| < \epsilon\}$. $\forall \epsilon > 0 \ \exists \delta$ such that if $\sup_\pi \|e^\pi\| < \delta$ then $\theta_t$ converges to $K^\epsilon$ as $t \to \infty$ with probability one. $e^\pi$ is the function approximation error defined in Lemma A.33.*

*Proof.* Let the $\sigma$-field $\mathcal{F}^2_t$ for actor iterate be defined as $\sigma(s_m, m < t; \theta_n, n \leq t)$. We will use the Theorem 5.3.1 from Chapter 5 of Kushner & Clark (2012) to prove convergence.

$$\begin{aligned}
\theta_{t+1} &= \Gamma_{C_\theta} \Big( \theta_t + \gamma_t \nabla_a Q^w_{diff}(s_t, a)|_{a=\pi(s_t)} \nabla_\theta \pi(s_t) \Big) \\
\theta_{t+1} &= \Gamma_{C_\theta} \Big( \theta_t + \gamma_t (h_2(\theta_t) + \mathcal{N}_{t+1} + \mathcal{M}^3_{t+1} \Big)
\end{aligned} \tag{A.54}$$

Here,

$$\begin{aligned}
h_2(\theta_t) &= \mathbb{E}[\nabla_a Q^{w^*}_{diff}(s_t, a)|_{a=\pi(s_t)} \nabla_\theta \pi(s_t) | \mathcal{F}^2_t] \\
\mathcal{N}_{t+1} &= \mathbb{E}[\nabla_a Q^w_{diff}(s_t, a)|_{a=\pi(s_t)} \nabla_\theta \pi(s_t) | \mathcal{F}^2_t] - \mathbb{E}[\nabla_a Q^{w^*}_{diff}(s_t, a)|_{a=\pi(s_t)} \nabla_\theta \pi(s_t) | \mathcal{F}^2_t] \\
\mathcal{M}^3_{t+1} &= \nabla_a Q^w_{diff}(s_t, a)|_{a=\pi(s_t)} \nabla_\theta \pi(s_t) - \mathbb{E}[\nabla_a Q^w_{diff}(s_t, a)|_{a=\pi(s_t)} \nabla_\theta \pi(s_t) | \mathcal{F}^2_t]
\end{aligned}$$

**Condition B1**:

We have, $\sum_t \alpha_t = \sum_t \frac{C_\alpha}{(1+t)^\sigma} = \infty$, $\sum_t \beta_t = \sum_t \frac{C_\beta}{(1+t)^u} = \infty$, $\sum_t \gamma_t = \sum_t \frac{C_\gamma}{(1+t)^v} = \infty$ and $\sum_t(\alpha_t^2 + \beta_t^2 + \gamma_t^2) = \sum_t \Big( \big(\frac{C_\alpha}{(1+t)^\sigma}\big)^2 + \big(\frac{C_\beta}{(1+t)^u}\big)^2 + \big(\frac{C_\gamma}{(1+t)^v}\big)^2 \Big) < \infty$. We can carefully set the value of $\sigma$, $u$, and $v$ to satisfy the conditions on step sizes. Further if $\sigma < u < v$ then $\beta_t = o(\alpha_t)$ and $\gamma_t = o(\beta_t)$.

**Condition B2**: We will now prove that $h_2(\theta)$ is Lipchitz continuous in $\theta$.

$$
\begin{aligned}
\nabla_\theta h_2(\theta) =& \nabla_\theta \int_S d^\pi(s)(w(\theta)^*)^\intercal \nabla_a \phi(s,a)|_{a=\pi(s)} \nabla_\theta \pi(s)\, ds \\
=& \int_S \nabla_\theta d^\pi(s)(w(\theta)^*)^\intercal \nabla_a \phi(s,a)|_{a=\pi(s)} \nabla_\theta \pi(s)\, ds \;\; \textcircled{1} \\
&+ \int_S d^\pi(s)(\nabla_\theta w(\theta)^*)^\intercal \nabla_a \phi(s,a)|_{a=\pi(s)} \nabla_\theta \pi(s)\, ds \;\; \textcircled{2} \\
&+ \int_S d^\pi(s)(w(\theta)^*)^\intercal (\nabla_\theta \nabla_a \phi(s,a)|_{a=\pi(s)}) \nabla_\theta \pi(s)\, ds \;\; \textcircled{3} \\
&+ \int_S d^\pi(s)(w(\theta)^*)^\intercal \nabla_a \phi(s,a)|_{a=\pi(s)} \nabla^2_\theta \pi(s)\, ds \;\; \textcircled{4}
\end{aligned} \tag{A.55}
$$

Using Assumption4.4, $\pi(s,\theta)$ is Lipchitz continuous in $\theta$ and hence $\nabla_\theta \pi(s)$ is bounded. By Assumption 4.5, $\phi(s,a)$ is Lipchitz continuous in $a$ and therefore $\nabla_a \phi(s,a)$ is bounded. $\nabla_\theta d^\pi(s)$ is bounded by application of Theorem 2.1 of Mao & Song (2020). Further, $\nabla_\theta w(\theta)^*$ is bounded because $w(\theta)^*$ is Lipchitz continuous in $\theta$ (Lemma A.21). By Assumption A.8, $\nabla^2_\theta \pi(s)$ exists and is bounded because $\theta \in C_\theta$. All the terms in A.55 are bounded. Consequently, $\nabla_\theta h_2(\theta)$ is bounded and Lipchtiz continuous in $\theta$.

**Condition B3**: Now, we will prove the noise terms $\mathcal{N}_{t+1}$ and $\mathcal{M}^3_{t+1}$ converges asymptotically. $\mathcal{N}_{t+1}$ is $o(1)$ term because $w_t$ converges $w(\theta_t)^*$ according to Lemma A.32. Further,

$$
\begin{aligned}
\xi_T &= \sum_{t=0}^{T-1} \gamma_t \mathcal{M}^3_{t+1} \\
&= \sum_{t=0}^{T-1} \gamma_t \Big( \nabla_a Q^w_{diff}(s_t,a)|_{a=\pi(s_t)} \nabla_\theta \pi(s_t) - \mathbb{E}[\nabla_a Q^w_{diff}(s_t,a)|_{a=\pi(s_t)} \nabla_\theta \pi(s_t) | \mathcal{F}^2_t] \Big)
\end{aligned}
$$

We will now prove that $\xi_t$ is a martingale process.

$$
\begin{aligned}
&\mathbb{E}\big[\mathcal{M}^3_{t+1} | \mathcal{F}^2_t\big] \\
&= \mathbb{E}\big[\big(\nabla_a Q^w_{diff}(s_t,a)|_{a=\pi(s_t)} \nabla_\theta \pi(s_t) - \mathbb{E}[\nabla_a Q^w_{diff}(s_t,a)|_{a=\pi(s_t)} \nabla_\theta \pi(s_t) | \mathcal{F}^2_t]\big) | \mathcal{F}^2_t\big] \\
&= \mathbb{E}\big[\big(\nabla_a Q^w_{diff}(s_t,a)|_{a=\pi(s_t)} \nabla_\theta \pi(s_t)\big) | \mathcal{F}^2_t\big] - \mathbb{E}[\nabla_a Q^w_{diff}(s_t,a)|_{a=\pi(s_t)} \nabla_\theta \pi(s_t) | \mathcal{F}^2_t] \\
&= 0
\end{aligned} \tag{A.56}
$$

$$
\begin{aligned}
\mathbb{E}\big[\xi_T | \mathcal{F}^2_{T-1}\big] =& \mathbb{E}\big[\sum_{t=0}^{T-1} \gamma_t \mathcal{M}^3_{t+1} | \mathcal{F}^2_{T-1}\big] \\
=& \sum_{t=0}^{T-2} \gamma_t \mathcal{M}^3_{t+1} + \mathbb{E}\big[\mathcal{M}^3_T | \mathcal{F}^2_{T-1}\big] \\
=& \xi_{T-1} \quad \text{(Using A.56)}
\end{aligned} \tag{A.57}
$$

$$\begin{aligned}\mathbb{E}\big[\|\xi_T\|^2\big] =&\mathbb{E}\Big[\Big(\sum_{n=0}^{T-1}\gamma_n\mathcal{M}_{n+1}^3\Big)^\intercal\Big(\sum_{m=0}^{T-1}\gamma_m\mathcal{M}_{m+1}^3\Big)\Big]\\ =&\mathbb{E}\Big[\sum_{n=0}^{T-1}\Big\|\gamma_n\mathcal{M}_{n+1}^3\Big\|^2\Big]\\ \Big(\because \text{For } n>m\ \mathbb{E}\Big[(\mathcal{M}_n^3)^\intercal\mathcal{M}_m^3\Big] = \mathbb{E}\Big[(\mathcal{M}_n^3)^\intercal E[\mathcal{M}_m^3|\mathcal{F}_{m-1}^2]\Big] = 0\Big)&\\ \leq&\Big(\sum_{n=0}^{T-1}\gamma_n^2\Big)\sup_n\mathbb{E}\Big[\Big\|\mathcal{M}_{n+1}^3\Big\|^2\Big] < \infty\end{aligned} \tag{A.58}$$

We have $\sum_n \gamma_n^2 < \infty$ from condition B1. From Assumption 4.4 and 4.5 it can be proved that $\|\mathcal{M}_t^3\|$ is bounded. Therefore $\mathbb{E}\big[\|\xi_T\|^2\big] < \infty$. Using A.57 and A.58 we have $\xi_t$ is martingale process.

Now,

$$\begin{aligned}\sum_t\mathbb{E}\Big[\|\xi_{t+1}-\xi_t\|\Big|\mathcal{F}_t^2\Big] &= \sum_t\mathbb{E}\Big[\gamma_t^2\|\mathcal{M}_{t+1}^3\|^2\Big|\mathcal{F}_t^2\Big]\\ &\leq\Big(\sum_{t=0}\gamma_t^2\Big)\sup_n\mathbb{E}\Big[\Big\|\mathcal{M}_{n+1}^3\Big\|^2\Big|\mathcal{F}_n^2\Big] < \infty\end{aligned} \tag{A.59}$$

By martingale convergence theorem of Chapter 11 of Borkar (2009) and using A.58 and A.59 it can proved that martingale $\xi_t$ converges and $\sum_{n=t}^{\infty}\gamma_n\mathcal{M}_{n+1}^3 \to 0$ as $t\to\infty$.

Hence the noise terms $\mathcal{N}_{t+1}$ and $\mathcal{M}_{t+1}^3$ converge asymptotically.

**Condition B4**: $\|\theta_t\|$ is bounded because of projection operator $\Gamma_{C_\theta}$.

Using Theorem 5.3.1 of Kushner & Clark (2012) with the satisfaction of condition B1-B4 ensures that A.54 tracks the ODE given in A.60 and $\theta_t$ converges to $K^\epsilon$ as $t\to\infty$ where $\hat{\Gamma}_{C_\theta}(y)(x) = \lim_{\delta\to\infty}\dfrac{\Gamma_{C_\theta}(x+\delta y)-\Gamma_{C_\theta}(x)}{\delta}$.

$$\dot{\theta}(t) = \hat{\Gamma}_{C_\theta}(\theta(t))h_2(\theta(t)) = \hat{\Gamma}_{C_\theta}(\theta(t))(\nabla_\theta\rho(\theta(t)) + e^{\pi(t)}) \quad \text{(Using Lemma A.33)} \tag{A.60}$$

Further, as $\sup_\pi\|e^\pi\|\to 0$, A.54 tracks the ODE given in A.61 and $\theta_t$ converges to $K$ as $t\to\infty$.

$$\dot{\theta}(t) = \hat{\Gamma}_{C_\theta}(\theta(t))(\nabla_\theta\rho(\theta(t))) \quad \text{(Using Lemma A.33)} \tag{A.61}$$

□

# B. Algorithm and Hyperparameters

## B.1. (Off-Policy) ARO-DDPG Practical Algorithm

**Algorithm 1** (Off-Policy) ARO-DDPG Practical Algorithm

Initialize actor parameter $\theta$ and critic parameters $w_1, w_2$.Initialize actor target parameter $\theta \rightarrow \overline{\theta}$
Initialize critic target parameters $w_1 \rightarrow \overline{w_1}, w_2 \rightarrow \overline{w_2}$. Initialize average reward parameter $\rho$.
Initialize target average reward parameter $\rho \rightarrow \overline{\rho}$. Initialize Replay buffer = $\{\}$

$t = 0$, $s_0$ = env.reset()
**while** $t \leq$ total steps **do**
  $a_t = \pi(s_t) + \epsilon$ {$\epsilon$ denotes the noise}
  $s_{t+1} \sim P(\cdot|s_t, a_t)$ and $r_t = R(s_t, a_t)$
  Store $\{s_t, a_t, s_{t+1}\}$ in the Replay Buffer
  **if** $t \,\%\, eval\_freq == 0$ **then**
    Evaluate(agent)
  **end if**
  **if** $t \,\%\, critic\_update\_freq == 0$ **then**
    Update critic according to (24) - (27)
  **end if**
  **if** $t \,\%\, actor\_update\_freq == 0$ **then**
    Update actor according to (28) - (29)
    Update target estimators according to (30) - (32)
  **end if**
  **if** $s_{t+1}$ is terminal **then**
    $s_t =$ env.reset()
  **else**
    $s_t = s_{t+1}$
  **end if**
**end while**

## B.2. Finite time analysis algorithm

Here we present the algorithm with linear function approximator for which finite time analysis was done. $\mathcal{B}_t$ denotes the batch of tuple of the form $\{s_i, a_i, s_i'\}$ sampled from the buffer at timestep $t$. $\Gamma_{C_w}$ is a projection operator defined as $\Gamma_{C_w} : \mathbb{R}^k \rightarrow B$, where $B(\subset \mathbb{R}^k)$ is a compact convex set. Here, the critic parameter $w \in \mathbb{R}^k$.

**Algorithm 2** On-policy AR-DPG with Linear FA

Initialize actor parameter $\theta$ and critic parameters $w$. Initialize actor target parameter $\theta \rightarrow \overline{\theta}$.
Initialize critic target parameters $w \rightarrow \overline{w}$.Initialize average reward parameter $\rho$
Initialize target average reward parameter $\rho \rightarrow \overline{\rho}$
Initialize buffer = $\{\}$

$t = 0$, $s_0$ = env.reset()
**while** $t \leq$ total steps **do**
  $a_t = \pi(s_t) + \epsilon$ $\{\epsilon$ is the noise$\}$
  $s_{t+1} \sim P(\cdot|s_t, a_t)$ and $r_t = R(s_t, a_t)$
  Store $\{s_t, a_t, s_{t+1}\}$ in the Buffer
  **if** $t \,\%\, critic\_update\_freq == 0$ **then**
    Sample $\mathcal{B}_t = \{s_i, a_i, s'_i\}_{i=0}^{M-1}$ from the Replay Buffer
    $w_{t+1} = \Gamma_{C_w}\Big(w_t + \frac{\alpha_t}{M}\sum_{i=0}^{M-1}\Big(R^\pi(s_i) - \bar{\rho}_t + \phi^\pi(s'_i)^\intercal \bar{w}_t - \phi^\pi(s_i)^\intercal w_t\Big)\phi^\pi(s_i) - \alpha_t \eta w_t\Big)$
    $\rho_{t+1} = \rho_t + \frac{\alpha_t}{M}\sum_{i=0}^{M-1}\Big(R^\pi(s_i) - \rho_t + \phi^\pi(s'_i)^\intercal \bar{w}_t - \phi^\pi(s_i)^\intercal \bar{w}_t\Big)$
    $\overline{w}_{t+1} = \overline{w}_t + \beta_t(w_{t+1} - \overline{w}_{t+1})$
    $\overline{\rho}_{t+1} = \overline{\rho}_t + \beta_t(\rho_{t+1} - \overline{\rho}_{t+1})$
    $\theta_{t+1} = \theta_t + \frac{\gamma_t}{M}\sum_{i=0}^{M-1}\nabla_a Q^w_{diff}(s_i, a)|_{a=\pi(s_i)}\nabla_\theta \pi(s_i)$
    buffer = $\{\}$
  **end if**
  **if** $s_{t+1}$ is terminal **then**
    $s_t =$ env.reset()
  **else**
    $s_t = s_{t+1}$
  **end if**
**end while**

**Algorithm 3** Off-policy AR-DPG with Linear FA

Initialize actor parameter $\theta$ and critic parameters $w$. Initialize actor target parameter $\theta \rightarrow \overline{\theta}$
Initialize critic target parameters $w \rightarrow \overline{w}$. Initialize average reward parameter $\rho$
Initialize target average reward parameter $\rho \rightarrow \overline{\rho}$. $\mu$ is the behavior policy
Initialize Replay buffer = $\{\}$

$t = 0$, $s_0$ = env.reset()
**while** $t \leq$ total steps **do**
  $a_t = \mu(s_t) + \epsilon$ $\{\epsilon$ is the noise$\}$
  $s_{t+1} \sim P(\cdot|s_t, a_t)$ and $r_t = R(s_t, a_t)$
  Store $\{s_t, a_t, s_{t+1}\}$ in the Replay Buffer
  Sample $\mathbb{B}_t = \{s_i, a_i, s'_i\}_{i=0}^{M-1}$ from the Replay Buffer
  $w_{t+1} = \Gamma_{C_w}\Big(w_t + \frac{\alpha_t}{M}\sum_{i=0}^{M-1}\Big(R^\mu(s_i) - \bar{\rho}_t + \phi^\pi(s'_i)^\intercal \bar{w}_t - \phi^\pi(s_i)^\intercal w_t\Big)\phi^\pi(s_i) - \alpha_t \eta w_t\Big)$
  $\rho_{t+1} = \rho_t + \frac{\alpha_t}{M}\sum_{i=0}^{M-1}\Big(R^\mu(s_i) - \rho_t + \phi^\pi(s'_i)^\intercal \bar{w}_t - \phi^\pi(s_i)^\intercal \bar{w}_t\Big)$
  $\overline{w}_{t+1} = \overline{w}_t + \beta_t(w_{t+1} - \overline{w}_{t+1})$
  $\overline{\rho}_{t+1} = \overline{\rho}_t + \beta_t(\rho_{t+1} - \overline{\rho}_{t+1})$
  $\theta_{t+1} = \theta_t + \frac{\gamma_t}{M}\sum_{i=0}^{M-1}\nabla_a Q^w_{diff}(s_i, a)|_{a=\pi(s_i)}\nabla_\theta \pi(s_i)$
  **if** $s_{t+1}$ is terminal **then**
    $s_t =$ env.reset()
  **else**
    $s_t = s_{t+1}$
  **end if**
**end while**

### B.3. Asymptotic analysis algorithm

Here we present the algorithm with linear function approximator for which asymptotic analysis was done. $\mathcal{B}_t$ denotes the batch of tuple of the form $\{s_i, a_i, s_i'\}$ sampled from the buffer at timestep $t$. $\Gamma_{C_\theta}$ is a projection operator defined as $\Gamma_{C_\theta} : \mathbb{R}^d \to C_\theta$, where $C_\theta (\subset \mathbb{R}^d)$ is a compact convex set. Here, the actor parameter $\theta \in \mathbb{R}^d$.

**Algorithm 4** On-policy AR-DPG with Linear FA

Initialize actor parameter $\theta$ and critic parameters $w$. Initialize actor target parameter $\theta \to \overline{\theta}$.
Initialize critic target parameters $w \to \overline{w}$.Initialize average reward parameter $\rho$
Initialize target average reward parameter $\rho \to \overline{\rho}$
Initialize buffer = $\{\}$

$t = 0$, $s_0$ = env.reset()
**while** $t \leq$ total steps **do**
  $a_t = \pi(s_t) + \epsilon$ $\{\epsilon$ is the noise$\}$
  $s_{t+1} \sim P(\cdot|s_t, a_t)$ and $r_t = R(s_t, a_t)$
  Store $\{s_t, a_t, s_{t+1}\}$ in the Buffer
  **if** $t \,\%\, critic\_update\_freq == 0$ **then**
    Sample $\mathcal{B}_t = \{s_i, a_i, s_i'\}_{i=0}^{M-1}$ from the Replay Buffer
    $w_{t+1} = w_t + \frac{\alpha_t}{M} \sum_{i=0}^{M-1} \Big( R^\pi(s_i) - \bar{\rho}_t + \phi^\pi(s_i')^\intercal \bar{w}_t - \phi^\pi(s_i)^\intercal w_t \Big) \phi^\pi(s_i) - \alpha_t \eta w_t$
    $\rho_{t+1} = \rho_t + \frac{\alpha_t}{M} \sum_{i=0}^{M-1} \Big( R^\pi(s_i) - \rho_t + \phi^\pi(s_i')^\intercal \bar{w}_t - \phi^\pi(s_i)^\intercal \bar{w}_t \Big)$
    $\overline{w}_{t+1} = \overline{w}_t + \beta_t (w_{t+1} - \overline{w}_{t+1})$
    $\overline{\rho}_{t+1} = \overline{\rho}_t + \beta_t (\rho_{t+1} - \overline{\rho}_{t+1})$
    $\theta_{t+1} = \Gamma_{C_\theta} \Big( \theta_t + \frac{\gamma_t}{M} \sum_{i=0}^{M-1} \nabla_a Q_{diff}^w(s_i, a)|_{a=\pi(s_i)} \nabla_\theta \pi(s_i) \Big)$
    buffer = $\{\}$
  **end if**
  **if** $s_{t+1}$ is terminal **then**
    $s_t =$ env.reset()
  **else**
    $s_t = s_{t+1}$
  **end if**
**end while**

**Algorithm 5** Off-policy AR-DPG with Linear FA

Initialize actor parameter $\theta$ and critic parameters $w$
Initialize actor target parameter $\theta \rightarrow \overline{\theta}$ and
Initialize critic target parameters $w \rightarrow \overline{w}$
Initialize average reward parameter $\rho$ and
Initialize target average reward parameter $\rho \rightarrow \overline{\rho}$
$\mu$ is the behavior policy
Initialize Replay buffer = $\{\}$

$t = 0$, $s_0$ = env.reset()
**while** $t \leq$ total steps **do**
  $a_t = \mu(s_t) + \epsilon$ $\{\epsilon$ is the noise$\}$
  $s_{t+1} \sim P(\cdot|s_t, a_t)$ and $r_t = R(s_t, a_t)$
  Store $\{s_t, a_t, s_{t+1}\}$ in the Replay Buffer
  Sample $\mathbb{B}_t = \{s_i, a_i, s'_i\}_{i=0}^{M-1}$ from the Replay Buffer
  $w_{t+1} = w_t + \frac{\alpha_t}{M} \sum_{i=0}^{M-1} \Big( R^{\mu}(s_i) - \bar{\rho}_t + \phi^{\pi}(s'_i)^{\intercal} \bar{w}_t - \phi^{\pi}(s_i)^{\intercal} w_t \Big) \phi^{\pi}(s_i) - \alpha_t \eta w_t$
  $\rho_{t+1} = \rho_t + \frac{\alpha_t}{M} \sum_{i=0}^{M-1} \Big( R^{\mu}(s_i) - \rho_t + \phi^{\pi}(s'_i)^{\intercal} \bar{w}_t - \phi^{\pi}(s_i)^{\intercal} \bar{w}_t \Big)$
  $\overline{w}_{t+1} = \overline{w}_t + \beta_t (w_{t+1} - \overline{w}_{t+1})$
  $\overline{\rho}_{t+1} = \overline{\rho}_t + \beta_t (\rho_{t+1} - \overline{\rho}_{t+1})$
  $\theta_{t+1} = \Gamma_{C_\theta} \Big( \theta_t + \frac{\gamma_t}{M} \sum_{i=0}^{M-1} \nabla_a Q^w_{diff}(s_i, a)|_{a=\pi(s_i)} \nabla_\theta \pi(s_i) \Big)$
  **if** $s_{t+1}$ is terminal **then**
    $s_t =$ env.reset()
  **else**
    $s_t = s_{t+1}$
  **end if**
**end while**

### B.4. Hyperparameters

The hyper-parameters mentioned in this section produces good performance for all the environment save for "fish-upright" where we used GeLU activation function.

| Hyperparameter | Value |
|---|---|
| Buffer Size | 1e6 |
| Total Environment Steps | 1e6 |
| Batch size | 256 |
| Evaluation Frequency | 5000 |
| Training Episode Length | 1000 |
| Evaluation Episode Length | 10000 |
| Activation Function | ReLU |
| Learning rate Actor | 3e-4 |
| Learning rate Critic | 3e-4 |
| Learning rate Average reward parameter | 3e-4 |
| No. of Hidden Layers | 2 |
| No. of Nodes in Hidden Layer | 128 |
| Update frequency | 10 steps |
| No. of Critic updates | 10 |
| No. of Actor updates | 5 |
| Polyak averaging constant | 0.995 |