OFF-POLICY AVERAGE REWARD ACTOR-CRITIC WITH DETERMINSITIC POLICY SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

The average reward criterion is relatively less explored 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 paper, 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 show a finite time analysis of the resulting three-timescale stochastic approximation scheme and obtain an ϵ-optimal stationary policy with a sample complexity of $\Omega(\epsilon^{-2.5})$. We compare the average reward performance of our proposed 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 the average reward criterion over the discounted reward (Dewanto & Gallagher, 2021; Naik et al., 2019). In the case of recurrent Markov Decision Processes (MDPs), average reward happens to be the most selective optimization criterion. Further, optimization in average reward setting is not dependent on the initial state distribution. Moreover, an obvious discrepancy between the objective function and the evaluation metric, that exists for discounted reward setting, is resolved by opting for the average reward criterion.

There are few algorithms 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 as well. 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 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 optimizing algorithm and the target network increases here.

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 perform non-asymptotic convergence analysis and provide a finite time analysis of our three timescale stochastic approximation based actor-critic algorithm using a linear function approximator.
- We show the results of implementations using our algorithm with other state-of-the-art algorithms in the literature.

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 algorithm. Section 4 then presents the main theoretical results related to the finite time 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 finite time 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. The policy π is defined as $\pi : S \mapsto A$. In the above, $\mathcal{B}(S)$ represents the Borel sigma algebra on S.

Assumption 1. *The Markov process obtained under any policy π is ergodic.*

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:

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

Here, ρ_0 is the initial state distribution and V^{π} is the value function. $V_{\pi}(s)$ denotes the long term reward acquired when starting in the state s.

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

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:

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

where $R^{\pi}(s) \triangleq R(s, \pi(s))$. The limit in the first equality in equation 3 exists because of Assumption 1. The quantity $d^{\pi}(s)$ in the second equality in equation 3 corresponds to the steady state probability of the Markov process being in state $s \in S$ and it exists and is unique given π from Assumption 1 as well.

Lemma 1. *There exists a unique constant $k(=\rho(\pi))$ which satisfies the following equation:*

$$V^{\pi}_{diff}(s_t) = \mathbb{E}^{\pi}[R(s_t, a_t) - k + V^{\pi}_{diff}(s_{t+1})|s_t] \tag{4}$$

Proof. See appendix for the proof. □

In (4), V^{π}_{diff} is the differential value function corresponding to the policy π and is defined in (5). Further, the differential Q-value or action-value function Q^{π}_{diff} is defined in (6).

$$V^{\pi}_{diff}(s_t) = \mathbb{E}^{\pi}[\sum_{k=t}^{\infty} R(s_k, a_k) - \rho(\pi)|s_t]. \tag{5}$$

$$Q^{\pi}_{diff}(s_t, a_t) = \mathbb{E}^{\pi}[\sum_{k=t}^{\infty} R(s_k, a_k) - \rho(\pi)|s_t, a_t]. \tag{6}$$

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.

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

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

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

$$\rho^{\pi}(s) = \frac{1}{\gamma}\sum_{t=0}^{\infty} \gamma^t P^{\pi}_t(s). \tag{9}$$

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

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

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

Silver et al. (2014) came up with the deterministic policy gradient theorem, 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).

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

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 (4) to obtain the average reward deterministic policy gradient theorem.

Theorem 1. *The gradient of $\rho(\pi)$ with respect to policy parameter θ 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 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 π. 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 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^π_{diff}.

Lemma 2. *Assume that the differential Q-value function (6) satisfies the following:*

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

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

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

Proof. See the appendix for a proof. □

An important implication of lemma 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 θ (used in the parameterized policy) and w (used to approximate the differential Q-value function) are the same. Lemma 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 2. *For the Markov chain obtained from the policy π, let $K(\cdot|\cdot)$ be the transition kernel and S^π the steady state measure. Then there exists $a > 0$ and $\kappa \in (0, 1)$ such that*

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

Theorem 2. *The approximate gradient of the average reward $\rho(\pi)$ with respect to the policy parameter θ is given by the following expression:*

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

Further, the approximation error is $\mathcal{E}(\pi, \mu) = \|\nabla_\theta \rho(\pi) - \nabla_\theta \bar{\rho}(\pi)\|$, where μ represents the behaviour policy. $\mathcal{E}$ satisfies

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

Here, $Z = 2^{m+1}C(\lceil \log_\kappa a^{-1} \rceil + 1/\kappa)L_t$ with L_t being the Lipchitz constant for the transition probability density function.

Proof. See the appendix for a proof. □

Theorem 2 suggests that the approximation error in the gradient increases as the difference between the target policy π and the behaviour policy μ 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 as in (20), which is a naive off-policy version of (3).

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

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

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

   The issue with update rule (21) is that it only considers the reward function and not the transition dynamics of the MDP. If we look at (13), 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.

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 μ but obtaining this information about $P(s'|s,\pi(a))$ is impossible as in the off-policy setting data from π is assumed to be simply unavailable.

3.5 Actor-Critic Update rule

The critic and average reward parameters are estimated using the TD(0) update rule but use target estimators. In the following, (24)-(27) present the critic and average reward parameter updates.The actor is updated using (13) and the practical update is presented in (28)-(29). The target parameters are updated using Polyak averaging in (30)– (32). Let $\{s_i, a_i, s_i'\}_{i=0}^{n-1}$ denote the batch of sampled data from the replay buffer.

$$\xi_t^j = \frac{1}{2}\sum_{i=0}^{n-1}\Big(R(s_i,a_i) - \overline{\rho_t} - Q_{diff}^{w_i}(s_i,a_i) + \min(\overline{Q}_{diff}^{w_1}, \overline{Q}_{diff}^{w_1})(s_i', \pi(s_i', \overline{\theta_t}))\Big)^2 \quad j \in \{1,2\} \tag{24}$$

$$\xi_t^3 = \frac{1}{2}\sum_{i=0}^{n-1}\Big(R(s_i,a_i) - \rho_t - \min\Big(\overline{Q}_{diff}^{w_1}(s_i,a_i), \overline{Q}_{diff}^{w_2}(s_i,a_i)\Big) + \min(\overline{Q}_{diff}^{w_1}, \overline{Q}_{diff}^{w_1})(s_i', \pi(s_i', \overline{\theta_t}))\Big)^2 \tag{25}$$

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

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

$$\nu_i = \nabla_a min(Q_{diff}^{w_1}, Q_{diff}^{w_2})(s_i,a)|_{a=\pi(s_i)} \nabla_\theta \pi(s_i,\theta_t) \tag{28}$$

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

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

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

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

4 Finite time Analysis

In this section we present the 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 finite time analysis followed by the main results.

Assumption 3. *$\phi^\pi(s)$ denotes the feature vector of state s and satisfies $\|\phi^\pi(s)\| \leq 1$.*

Assumption 4. *The reward function is uniformly bounded, viz., $|R^\pi(s)| \leq C_r < \infty$.*

Algorithm 1 ARO-DDPG Practical Algorithm

Initialize actor parameter θ 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 ρ. Initialize target average reward parameter $\rho \rightarrow \overline{\rho}$. Initialize Replay buffer = { }

$t = 0$, s_0 = env.reset()
while $t \leq$ total steps **do**
  $a_t = \pi(s_t) + \epsilon$ {ϵ 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

Assumption 5. *$Q_{diff}(s, a)$ is Lipchitz continuous w.r.t to a. Thus, $\|Q_{diff}(s, a_1) - Q_{diff}(s, a_2)\| \leq L_a \|a_1 - a_2\|$.*

Assumption 6. *Parameterised policy $\pi(s, \theta)$ is Lipchitz continuous w.r.t θ. Thus, $\|\pi(s, \theta_1) - \pi(s, \theta_2)\| \leq L_\pi \|\theta_1 - \theta_2\|$.*

Assumption 7. *The state feature mapping ($\phi_\pi(s)$) defined for a policy π with parameter θ is Lipschitz continuous w.r.t θ. Thus, $\max_s \|\phi^{\pi_1}(s) - \phi^{\pi_2}(s)\| \leq L_\phi \|\theta_1 - \theta_2\|$.*

4.1 On-Policy Analysis

In this section we present the theorem for finite time analysis of the on-policy version of the algorithm with linear function approximator and target estimator for the critic and average reward.

Theorem 3. *The on-policy average reward actor critic algorithm (Algorithm 2) obtains an ϵ-accurate optimal point with sample complexity of $\Omega(\epsilon^{-2.5})$. We obtain*

$$\min_{0 \leq t \leq T-1} E||\nabla_\theta \rho(\theta_t)||^2 = \mathcal{O}\left(\frac{1}{T^{0.4}}\right) + \mathcal{O}(1),$$
$$\leq \epsilon + \mathcal{O}(1).$$

Proof. We present a proof outline here. See the appendix for complete details of the proof.

1. Obtain a bound on $\frac{1}{T}\sum_{t=0}^{T-1} E||\nabla_\theta \rho(\theta_t)||^2$ using $\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}||\Delta w_t||^2$ (Lemma 3).
2. Obtain a bound on $\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}||\Delta w_t||^2$ using $\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}||\Delta \rho_t||^2$ (Lemma 4).
3. Obtain a bound on $\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}||\Delta \rho_t||^2$ (Lemma 5).
4. The claim follows upon combining the results from the above steps.

□

4.2 Off-Policy Analysis

In this section we present the theorem for finite time analysis of off-policy version of the algorithm with linear function approximator and target estimator for the critic and average reward.

Theorem 4. *The off-policy average reward actor critic algorithm (Algorithm 3) with behavior policy μ obtains an ϵ-accurate optimal point with sample complexity of $\Omega(\epsilon^{-2.5})$. We obtain*

$$\min_{0\leq t\leq T-1} E\|\nabla_\theta \bar{\rho}(\theta_t)\|^2 = \mathcal{O}\left(\frac{1}{T^{0.4}}\right) + \mathcal{O}(1) + \mathcal{O}(N_\theta^2)$$
$$\leq \epsilon + \mathcal{O}(1) + \mathcal{O}(N_\theta^2)$$
$$\textit{where } N_\theta := \max_t \|\theta_\mu - \theta_t\|.$$

Proof. See the appendix for a proof. □

The error bound in the off-policy algorithm is the same as the on-policy algorithm except for a single term $\mathcal{O}(N_\theta^2)$. The extra term denotes the error induced because of not using the samples from the current policy for performing updates.

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 to be superior than the other algorithms. 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 used the original hyper-parameters suggested by the author for ATRPO.

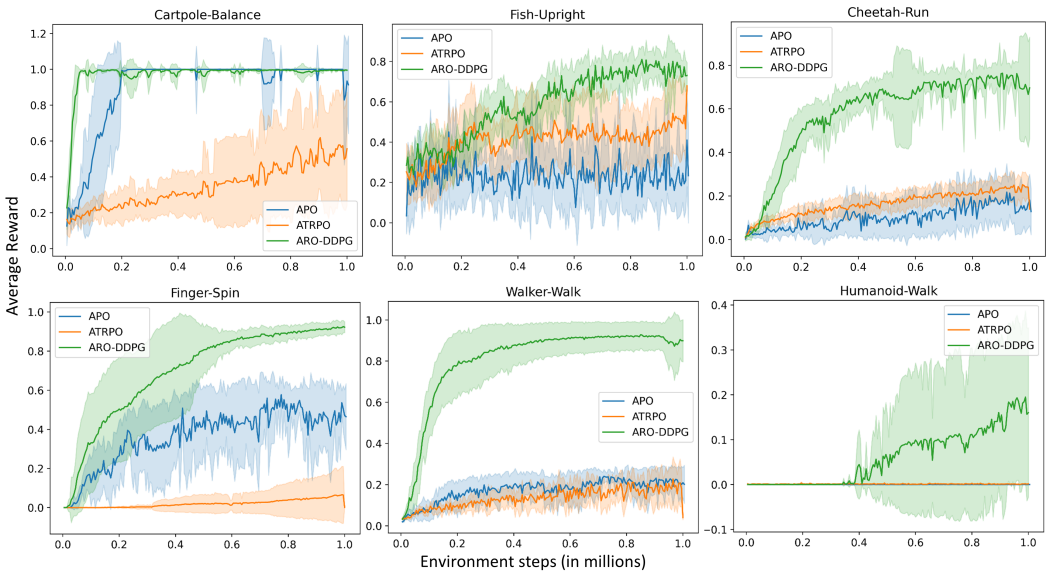

Figure 1: Comparison of performance of different average reward algorithms

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 reward performance. We do not reset the agent if it lands in a state before completing 10,000 steps from where it is unable to escape of its own, while continuing to give a penalty for the remaining length of the episode. That way the cost of failure is very 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 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. came up with 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). In 2007, Bhatnagar et al. came up with incremental natural policy gradient algorithms for the average reward setting and provided the asymptotic convergence proof of these.

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. came up with 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. came up with a convergent off-policy evaluation scheme inspired from the gradient temporal difference learning 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. 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 PPO). At the end we provided a finite time analysis for the on-policy and off-policy algorithms obtained from the proposed policy gradient theorem and obtained a 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).

REFERENCES

Shalabh Bhatnagar, Mohammad Ghavamzadeh, Mark Lee, and Richard S Sutton. Incremental natural actor-critic algorithms. *Advances in neural information processing systems*, 20, 2007.

Thomas Degris, Martha White, and Richard S. Sutton. 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.

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

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In Jennifer G. Dy and Andreas Krause (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.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer G. Dy and Andreas Krause (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.

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

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In Yoshua Bengio and Yann LeCun (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.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (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.

Xiaoteng Ma, Xiaohang Tang, Li Xia, Jun Yang, and Qianchuan Zhao. Average-reward reinforcement learning with trust region methods. In Zhi-Hua Zhou (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.

Sridhar Mahadevan. 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.

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

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (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.

Abhishek Naik, Roshan Shariff, Niko Yasui, and Richard S. Sutton. Discounted reinforcement learning is not an optimization problem. *CoRR*, abs/1910.02140, 2019. URL http://arxiv.org/abs/1910.02140.

John Schulman, Sergey Levine, Philipp Moritz, Michael Jordan, and Pieter Abbeel. 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.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin A. Riedmiller. 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.

Richard S. Sutton, David A. McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In Sara A. Solla, Todd K. Leen, and Klaus-Robert Müller (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.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy P. Lillicrap, and Martin A. Riedmiller. Deepmind control suite. *CoRR*, abs/1801.00690, 2018. URL http://arxiv.org/abs/1801.00690.

Yi Wan, Abhishek Naik, and Rich Sutton. Average-reward learning and planning with options. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (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.

Yi Wan, Abhishek Naik, and Richard S Sutton. Learning and planning in average-reward markov decision processes. In Marina Meila and Tong Zhang (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.

Yue Frank Wu, Weitong ZHANG, Pan Xu, and Quanquan Gu. A finite-time analysis of two time-scale actor-critic methods. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (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.

Huaqing Xiong, Tengyu Xu, Lin Zhao, Yingbin Liang, and Wei Zhang. Deterministic policy gradient: Convergence analysis. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022.

Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. 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.

Shangtong Zhang, Bo Liu, and Shimon Whiteson. 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.

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

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

Yiming Zhang and Keith W. Ross. On-policy deep reinforcement learning for the average-reward criterion. In Marina Meila and Tong Zhang (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.

A Appendix

We make the following additional assumptions.

Assumption 8. *The transition probability density function for a policy π with parameter θ is Lipschitz continuous w.r.t θ. Thus, $max_{s',s}|P^{\pi_1}(s'|s) - P^{\pi_2}(s'|s)| \leq L_t\|\theta_1 - \theta_2\|$.*

Assumption 9. *The reward function for a policy π with parameter θ is Lipschitz continuous w.r.t θ. Thus, $max_s|R^{\pi_1}(s) - R^{\pi_2}(s)| \leq L_r\|\theta_1 - \theta_2\|$.*

Assumption 10. *The initial value of target estimators is bounded. Thus, $\|\bar{w_0}\| \leq C_w$ and $\|\bar{\rho_0}\| \leq (Cr + 2C_w)$.*

Lemma 1. *There exists a unique constant $k = \rho(\pi)$ which satisfies the following equation:*

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

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^{π} of policy π :

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

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

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 1. *The gradient of $\rho(\pi)$ with respect to the policy parameter θ 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 1:

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

Differentiating w.r.t θ, we obtain

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

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

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

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

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

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

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

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

□

Lemma 2. *Assume that the differential Q-value function (6) satisfies the following:*

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

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,

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

□

Theorem 2. *The approximate gradient of the average reward $\rho(\pi)$ with respect to the policy parameter θ is given by the following expression:*

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

Further, the approximation error

$$\mathcal{E}(\pi,\mu) = \|\nabla_\theta \rho(\pi) - \nabla_\theta \bar{\rho}(\pi)\|$$

where μ represents the behaviour policy satisfies

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

Here, $Z = 2^{m+1}C(\lceil \log_\kappa a^{-1} \rceil + 1/\kappa)L_t$ with L_t being the lipchitz constant for the transition probability density function.

Proof.

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

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

$$\mathcal{E}(\pi,\mu) \leq CL_d\|\theta^\pi - \theta^\mu\| = Z\|\theta^\pi - \theta^\mu\| \text{ (using } Lemma 12).$$

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

□

Lemma 3. *Bound on the cumulative error of on-policy actor is proven using cumulative error of critic.*

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

$$R(s,a) \leq C_r, \gamma_t = \frac{C_\gamma}{(1+t)^v}, \|\nabla_\theta \pi(s)\| \leq C_\pi, \Delta w_t = w_t - w^*_t, \tau = \max_t \|w^*_t - w^*_{\epsilon,t}\|.$$

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

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

Now,

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

$$\begin{aligned} E\langle \nabla_\theta \rho(\theta_t), \theta_{t+1} - \theta_t \rangle &= \gamma_t E\langle \nabla_\theta \rho(\theta_t), h(B_t, w_t, \theta_t)\rangle \\ &= \gamma_t E\langle \nabla_\theta \rho(\theta_t), h(B_t, w_t, \theta_t) - \nabla_\theta \rho(\theta_t)\rangle + \gamma_t E\|\nabla_\theta \rho(\theta_t)\|^2. \end{aligned} \tag{A.3}$$

From (A.3), we have

$$\begin{aligned} E\langle \nabla_\theta \rho(\theta_t), h(B_t, w_t, \theta_t) - \nabla_\theta \rho(\theta_t)\rangle &\geq -\frac{1}{2}E\|\nabla_\theta \rho(\theta_t)\|^2 - \frac{1}{2}E\|h(B_t, w_t, \theta_t) - \nabla_\theta \rho(\theta_t)\|^2 \\ (\because x^\intercal y \geq -\|x\|^2/2 - \|y\|^2/2). \end{aligned} \tag{A.4}$$

From (A.4):

$$\begin{aligned} &E\|h(B_t, w_t, \theta_t) - \nabla_\theta \rho(\theta_t)\|^2 \\ &= E\|h(B_t, w_t, \theta_t) - h(B_t, w_t^*, \theta_t) + h(B_t, w_t^*, \theta_t) - h(B_t, w_{\epsilon,t}^*, \theta_t) + h(B_t, w_{\epsilon,t}^*, \theta_t) - \nabla_\theta \rho(\theta_t)\|^2 \\ &\leq 3(E\|h(B_t, w_t, \theta_t) - h(B_t, w_t^*, \theta_t)\|^2 \ \ ① \\ &\quad + E\|h(B_t, w_t^*, \theta_t) - h(B_t, w_{\epsilon,t}^*, \theta_t)\|^2 \ \ ② \\ &\quad + E\|h(B_t, w_{\epsilon,t}^*, \theta_t) - \nabla_\theta \rho(\theta_t)\|^2 \ \ ③ \end{aligned} \tag{A.5}$$

From (A.5):

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 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 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:

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

Combining ①,② and ③ and using in (A.5):

$$E||h(B_t, w_t, \theta_t) - \nabla_\theta \rho(\theta_t)||^2 \leq 3C_\pi^4(E||w_t - w_t^*||^2 + \tau^2 + \frac{4C_{w_\epsilon}^2}{M}). \tag{A.6}$$

Using (A.6) in (A.4):

$$\begin{aligned} E\langle \nabla_\theta \rho(\theta_t), h(B_t, w_t, \theta_t) - \nabla_\theta \rho(\theta_t)\rangle \geq & -\frac{1}{2}E||\nabla_\theta \rho(\theta_t)||^2 \\ & - \frac{3}{2}C_\pi^4(E||w_t - w_t^*||^2 + \tau^2 + \frac{4C_{w_\epsilon}^2}{M}). \end{aligned} \tag{A.7}$$

Using (A.7) in (A.3):

$$\begin{aligned} E\langle \nabla_\theta \rho(\theta_t), \theta_{t+1} - \theta_t\rangle \geq & \frac{\gamma_t}{2}E||\nabla_\theta \rho(\theta_t)||^2 \\ & - \frac{3\gamma_t}{2}C_\pi^4(E||w_t - w_t^*||^2 + \tau^2 + \frac{4C_{w_\epsilon}^2}{M}). \end{aligned} \tag{A.8}$$

Using (A.8) in (A.2):

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

$$\begin{aligned} \implies E||\nabla_\theta \rho(\theta_t)||^2 \geq & \frac{2}{\gamma_t}E[\rho(\theta_{t+1})] - E[\rho(\theta_t)] + 3C_\pi^4(E||w_t - w_t^*||^2) \\ & + 3C_\pi^4(\tau^2 + \frac{4C_{w_\epsilon}^2}{M}) + L_J\gamma_t G_\theta^2 \quad \text{(using lemma 7)} \end{aligned}$$

$$\begin{aligned} \implies \sum_{t=0}^{T-1} E||\nabla_\theta \rho(\theta_t)||^2 \geq & \sum_{t=0}^{T-1} \frac{2}{\gamma_t}E[\rho(\theta_{t+1})] - E[\rho(\theta_t)] \ \text{①} \\ & + \sum_{t=0}^{T-1} 3C_\pi^4(E||w_t - w_t^*||^2) \ \text{②} \\ & + \sum_{t=0}^{T-1} 3C_\pi^4(\tau^2 + \frac{4C_{w_\epsilon}^2}{M}) \ \text{③} \\ & + \sum_{t=0}^{T-1} L_J\gamma_t G_\theta^2 \ \text{④} \quad \text{(using lemma 7)} \end{aligned} \tag{A.9}$$

From equation A.9

①:

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

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:

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

Using ①-④ and dividing equation A.9 by T:

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

□

Lemma 4. *Bound on the cumulative error of critic is proven using cumulative error of average reward estimator.*

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2 \le 2\Bigg(\sqrt{\frac{2C_w^2}{\lambda C_\alpha}T^{\sigma-1}+\frac{C_gC_\alpha}{1-\sigma}T^{-\sigma}}+\\ \frac{LwG_\theta}{\lambda}\left(\frac{1}{T}\sum\Big(\frac{\gamma_t}{\alpha_t}\Big)^2\right)^{1/2}+\frac{2(C_r+3C_w)}{\lambda}\Bigg)^2+\frac{2}{\lambda^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta\rho_t||^2\\ \alpha_t=\frac{C_\alpha}{(1+t)^\sigma},0<\sigma<v<1\\ \gamma_t=\frac{C_\gamma}{(1+t)^v}\end{aligned}$$

Proof.

$$
\begin{aligned}
w_{t+1} &= w_t + \alpha_t \frac{1}{M} \sum_{i=0}^{M-1} \Big(R^\pi(s_{t,i}) - \bar{\rho}_t + \phi^\pi(s'_{t,i})^\intercal \bar{w}_t - \phi^\pi(s_{t,i})^\intercal w_t \Big) \phi^\pi(s_{t,i}) - \alpha_t \eta w_t \\
\implies w_{t+1} - w_{t+1}^* &= w_t - w_t^* + w_t^* - w_{t+1}^* - \alpha_t \eta w_t \quad ① \\
&\quad + \alpha_t \frac{1}{M} \sum_{i=0}^{M-1} \Big(R^\pi(s_{t,i}) - \rho_t^* + \phi^\pi(s'_{t,i})^\intercal \bar{w}_t - \phi^\pi(s_{t,i})^\intercal w_t \Big) \phi^\pi(s_{t,i}) \quad ② \\
&\quad + \alpha_t \frac{1}{M} \sum_{i=0}^{M-1} \Big(\rho_t^* - \rho_t \Big) \phi^\pi(s_{t,i}) \quad ③ \\
&\quad + \alpha_t \frac{1}{M} \sum_{i=0}^{M-1} \Big(\rho_t - \bar{\rho}_t \Big) \phi^\pi(s_{t,i}) \quad ④
\end{aligned}
\tag{A.10}
$$

From equation A.10:
②:

$$
\begin{aligned}
&\frac{1}{M} \sum_{i=0}^{M-1} \Big(R^\pi(s_{t,i}) - \rho_t^* + \phi^\pi(s'_{t,i})^\intercal \bar{w}_t - \phi^\pi(s_{t,i})^\intercal w_t \Big) \phi^\pi(s_{t,i}) \\
&\qquad = \frac{1}{M} \sum_{i=0}^{M-1} \Big(R^\pi(s_{t,i}) - \rho_t^* + \phi^\pi(s'_{t,i})^\intercal w_t - \phi^\pi(s_{t,i})^\intercal w_t \Big) \phi^\pi(s_{t,i}) - \eta w_t + \eta w_t \\
&\qquad\quad + \frac{1}{M} \sum_{i=0}^{M-1} \phi^\pi(s_{t,i}) \phi^\pi(s'_{t,i})^\intercal (\bar{w}_t - w_t) \\
&\qquad = \frac{1}{M} \sum_{i=0}^{M-1} \phi^\pi(s_{t,i}) \phi^\pi(s'_{t,i})^\intercal (\bar{w}_t - w_t) + g(B_t, w_t, \theta_t) - \bar{g}(w_t, \theta_t) \\
&\qquad\quad + \bar{g}(w_t, \theta_t) - \bar{g}(w_t^*, \theta_t)
\end{aligned}
\tag{A.11}
$$

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

Using equation A.11 in equation A.10:

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

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

$$
\begin{aligned}
||w_{t+1} - w_{t+1}^*||^2 &= ||(w_t - w_t^*) + (w_t^* - w_{t+1}^*) + \alpha_t f(B_t, w_t, \theta_t)||^2 \\
&= ||w_t - w_t^*||^2 + ||w_t^* - w_{t+1}^*||^2 \\
&\quad + \alpha_t^2||f(B_t, w_t, \theta_t)||^2 \\
&\quad + 2\langle \Delta w_t, w_t^* - w_{t+1}^*\rangle + 2\alpha_t\langle \Delta w_t, f(B_t, w_t, \theta_t)\rangle \\
&\quad + 2\alpha_t\langle w_t^* - w_{t+1}^*, f(B_t, w_t, \theta_t)\rangle
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}||w_{t+1} - w_{t+1}^*||^2 \leq& \mathbb{E}||\Delta w_t||^2 + 2\mathbb{E}||w_t^* - w_{t+1}^*||^2 \\
&+ 2\alpha_t^2 E||f(B_t, w_t, \theta_t)||^2 \\
&+ 2\mathbb{E}\langle \Delta w_t, w_t^* - w_{t+1}^*\rangle \\
&+ 2\alpha_t\mathbb{E}\langle \Delta w_t, f(B_t, w_t, \theta_t)\rangle \\
=& \mathbb{E}||\Delta w_t||^2 + 2\mathbb{E}||w_t^* - w_{t+1}^*||^2 \quad ① \\
&+ 2\alpha_t^2\mathbb{E}||f(B_t, w_t, \theta_t)||^2 \quad ② \\
&+ 2\mathbb{E}\langle \Delta w_t, w_t^* - w_{t+1}^*\rangle \quad ③ \\
&+ 2\alpha_t\mathbb{E}\langle \Delta w_t, \frac{1}{M}\sum_{i=0}^{M-1}(\rho_t^* - \rho_t)\phi^\pi(s_{t,i})\rangle \quad ④ \\
&+ 2\alpha_t\mathbb{E}\langle \Delta w_t, \frac{1}{M}\sum_{i=0}^{M-1}(\rho_t - \bar{\rho}_t)\phi^\pi(s_{t,i})\rangle \quad ⑤ \\
&+ 2\alpha_t\mathbb{E}\langle \Delta w_t, \frac{1}{M}\sum_{i=0}^{M-1}\phi^\pi(s_{t,i})\phi^\pi(s'_{t,i})^\intercal(\bar{w}_t - w_t)\rangle \quad ⑥ \\
&+ 2\alpha_t\mathbb{E}\langle \Delta w_t, g(B_t, w_t, \theta_t) - \bar{g}(w_t, \theta_t)\rangle \quad ⑦ \\
&+ 2\alpha_t\mathbb{E}\langle \Delta w_t, \bar{g}(w_t, \theta_t) - \bar{g}(w_t^*, \theta_t)\rangle \quad ⑧
\end{aligned} \tag{A.12}
$$

From equation A.12:
①:

$$
\begin{aligned}
\mathbb{E}||w_t^* - w_{t+1}^*||^2 &\leq L_w^2\mathbb{E}||\theta_{t+1} - \theta_t||^2 \quad \text{(using lemma 6)} \\
&\leq L_w^2\gamma_t^2 G_\theta^2 \quad \text{(using lemma 7)}
\end{aligned}
$$

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,

$$
\begin{aligned}
||\phi^\pi(s)|| &< 1 \quad \text{(Assumption 3)} \\
|R^\pi(s)| &\leq C_r \quad \text{(Assumption 4)} \\
||w_t|| &\leq C_w \quad \text{(Algorithm 2, step 8)} \\
|\rho_t| &\leq C_r + 2C_w \quad \text{(lemma 8)} \\
||\bar{w}_t|| &\leq C_w \quad \text{(lemma 9)} \\
||\bar{\rho}_t|| &\leq C_r + 2C_w \quad \text{(lemma 10)}
\end{aligned}
$$

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

③:

$$
\begin{aligned}
\mathbb{E}\langle \Delta w_t, w_t^* - w_{t+1}^* \rangle &\leq \mathbb{E}||\Delta w_t||\,||w_t^* - w_{t+1}^*|| \\
&\leq L_w \mathbb{E}||\Delta w_t||\,||\theta_{t+1} - \theta_t|| \quad \text{(using Lemma 6)}
\end{aligned}
$$

④:

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

⑤:

$$
\begin{aligned}
\mathbb{E}[\langle \Delta w_t, \frac{1}{M}\sum_{i=0}^{M-1}(\rho_t - \bar{\rho}_t)\phi^\pi(s_{t,i})\rangle] &= \mathbb{E}[\frac{1}{M}\sum_{i=0}^{M-1}\langle \Delta w_t, \phi^\pi(s_{t,i})\rangle(\rho_t - \bar{\rho_t})] \\
&\leq \mathbb{E}[\frac{1}{M}\sum_{i=0}^{M-1}||\Delta w_t||\,||\phi^\pi(s_{t,i})|||\rho_t - \bar{\rho}_t|] \\
&\leq \mathbb{E}[||\Delta w_t|||\rho_t - \bar{\rho}_t|] \\
&\leq \mathbb{E}[||\Delta w_t||(|\rho_t| + |\bar{\rho}_t|)] \\
&\leq \mathbb{E}[||\Delta w_t||]2(C_r + 2C_w) \quad \text{(using Lemma 8, 10)}
\end{aligned}
$$

⑥:

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

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:

$$
\begin{aligned}
\mathbb{E}[\langle \Delta w_t, \bar{g}(w_t,\theta_t) - \bar{g}(w_t^*,\theta_t)\rangle] & \\
A(\theta_t) &= \int_S d^\pi(s,\theta_t)(\phi^\pi(s)(\mathbb{E}[\phi^\pi(s')] - \phi^\pi(s)^\intercal - \eta I)ds \\
b(\theta_t) &= \int_S d^\pi(s,\theta_t) r^\pi(s)\phi^\pi(s)ds \\
\bar{g}(w_t,\theta_t) - \bar{g}(w_t^*,\theta_t) &= b(\theta_t) + A(\theta_t)w_t - b(\theta_t) - A(\theta_t)w_t^* \\
&= A(\theta_t)(w_t - w_t^*) \\
\text{Now,}\quad \mathbb{E}[\langle \Delta w_t, \bar{g}(w_t,\theta_t) - \bar{g}(w_t^*,\theta_t)\rangle] &= \mathbb{E}[\langle \Delta w_t, A(\theta_t)\Delta w_t\rangle] \\
&= \mathbb{E}[\Delta w_t^\intercal A(\theta_t)\Delta w_t] \\
&\leq -\lambda \mathbb{E}||\Delta w_t||^2 \qquad \text{(Lemma 11)}
\end{aligned}
$$

Combining ① - ⑧ into equation A.12:

$$
\begin{aligned}
\mathbb{E}||w_{t+1} - w_{t+1}^*||^2 \leq (1 - 2\lambda\alpha_t)\mathbb{E}||\Delta w_t||^2 + 2L_w^2\gamma_t^2 G_\theta^2 + 2\alpha_t^2 C_\delta^2 \\
+ 2L_w \mathbb{E}||\Delta w_t||||\theta_{t+1} - \theta_t|| + 2\alpha_t \mathbb{E}||\Delta w_t|||\Delta\rho_t| \\
+ 4\alpha_t \mathbb{E}||\Delta w_t||(2C_w + C_r) + 4\alpha_t C_w \mathbb{E}||\Delta w_t||
\end{aligned}
$$

$$
\begin{aligned}
\implies 2\lambda\alpha_t \mathbb{E}||\Delta w_t||^2 \leq \mathbb{E}[||\Delta w_t||^2] - \mathbb{E}||\Delta w_{t+1}||^2 + 2L_w^2\gamma_t^2 G_\theta^2 + 2\alpha_t^2 C_\delta^2 \\
+ 2L_w\gamma_t G_\theta \mathbb{E}||\Delta w_t|| + 2\alpha_t \mathbb{E}||\Delta w_t|||\Delta\rho_t| \\
+ 4\alpha_t(C_r + 3C_w)\mathbb{E}||\Delta w_t||
\end{aligned}
$$

$$
\begin{aligned}
\implies \mathbb{E}||\Delta w_t||^2 \leq \frac{1}{2\lambda\alpha_t}(\mathbb{E}||\Delta w_t||^2 - \mathbb{E}||w_{t+1}||^2) \\
+ \Big(\frac{L_w^2\gamma_t^2}{\lambda\alpha_t}G_\theta^2 + \frac{\alpha_t}{\lambda}C_\delta^2\Big) \\
+ \frac{L_w}{\lambda}\frac{\gamma_t}{\alpha_t}G_\theta \mathbb{E}||\Delta w_t|| \\
+ \frac{\mathbb{E}||\Delta w_t|||\Delta\rho_t|}{\lambda} \\
+ \frac{2}{\lambda}(C_r + 3C_w)\mathbb{E}||\Delta w_t||
\end{aligned}
$$

$$
\begin{aligned}
\implies \sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2 \leq \sum_{t=0}^{T-1}\frac{1}{2\lambda\alpha_t}(\mathbb{E}||\Delta w_t||^2 - \mathbb{E}||\Delta w_{t+1}||^2) \ ① \\
+ \sum_{t=0}^{T-1}\Big(\frac{L_w}{\lambda}\frac{\gamma_t^2}{\alpha_t}G_\theta^2 + \frac{\alpha_t}{\lambda}C_\delta^2\Big) \ ② \\
+ \sum_{t=0}^{T-1}\frac{L_w}{\lambda}\frac{\gamma_t}{\alpha_t}G_\theta \mathbb{E}||\Delta w_t|| \ ③ \\
+ \sum_{t=0}^{T-1}\frac{\mathbb{E}||\Delta w_t|||\Delta\rho_t|}{\lambda} \ ④ \\
+ \sum_{t=0}^{T-1}\frac{2}{\lambda}(C_r + 3C_w)\mathbb{E}||\Delta w_t|| \ ⑤
\end{aligned} \tag{A.13}
$$

From equation A.13:
①:

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

②:

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

③:

$$\sum_{t=0}^{T-1}\frac{L_w}{\lambda}\frac{\gamma_t}{\alpha_t}G_\theta\mathbb{E}||\Delta w_t||=\frac{L_w}{\lambda}G_\theta\sum_{t=0}^{T-1}\frac{\gamma_t}{\alpha_t}\mathbb{E}||\Delta w_t||$$
$$\leq\frac{L_w}{\lambda}G_\theta\left(\sum_{t=0}^{T-1}\Big(\frac{\gamma_t}{\alpha_t}\Big)^2\right)^{\frac{1}{2}}\Big(\sum_{t=0}^{T-1}(\mathbb{E}||\Delta w_t||)^2\Big)^{\frac{1}{2}}$$
(Using Cauchy Schwartz inequality)
$$\leq\frac{L_w}{\lambda}G_\theta\left(\sum_{t=0}^{T-1}\Big(\frac{\gamma_t}{\alpha_t}^2\Big)^{\frac{1}{2}}\Big(\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2\Big)^{\frac{1}{2}}\right.$$
(Using Jensen's inequality)

④:

$$\frac{1}{\lambda}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t|||\Delta\rho_t|\leq\frac{1}{\lambda}\Big(\sum_{t=0}^{T-1}(\mathbb{E}||\Delta w_t||)^2\Big)^{\frac{1}{2}}\Big(\sum_{t=0}^{T-1}(\mathbb{E}|\Delta\rho_t|)^2\Big)^{\frac{1}{2}}$$
$$\leq\frac{1}{\lambda}\Big(\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2\Big)^{\frac{1}{2}}\Big(\sum_{t=0}^{T-1}\mathbb{E}|\Delta\rho_t|^2\Big)^{\frac{1}{2}}$$

⑤:

$$\sum_{t=0}^{T-1}\frac{2(C_r+3C_w)}{\lambda}\mathbb{E}||\Delta w_t||\leq\frac{2(C_r+3C_w)}{\lambda}\Big(\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2\Big)^{\frac{1}{2}}\Big(\sum_{t=0}^{T-1}1\Big)^{\frac{1}{2}}$$
$$\leq\frac{2(C_r+3C_w)}{\lambda}T^{\frac{1}{2}}\Big(\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2\Big)^{\frac{1}{2}}$$

Combining ① - ⑤ into equation A.13:

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2 \leq \frac{2C_w^2}{\lambda C_\alpha}T^{\sigma-1} + \frac{C_g C^\alpha}{1-\sigma}T^{-\sigma} \\ + \frac{L_w G_\theta}{\lambda}\left(\frac{1}{T}\sum_{t=0}^{T-1}\left(\frac{\gamma_t}{\alpha_t}\right)^2\right)^{\frac{1}{2}}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2\right)^{\frac{1}{2}} \\ + \frac{1}{\lambda}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2\right)^{\frac{1}{2}}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta \rho_t||^2\right)^{\frac{1}{2}} \\ + \frac{2(C_r+3C_w)}{\lambda}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2\right)^{\frac{1}{2}}\end{aligned}$$

Let,

$$\begin{aligned}M(T) &= \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2 \\ N(T) &= \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}|\Delta \rho_t|^2\end{aligned}$$

$$\begin{aligned}M(T) &\leq K_1 + K_2\sqrt{M(T)} + K_3\sqrt{M(T)}\sqrt{N(T)} \\ K_1 &:= \frac{2C_w^2}{\lambda C_\alpha}T^{\sigma-1} + \frac{C_g C_\alpha}{1-\sigma}T^{-\sigma} \\ K_2 &:= \frac{L_w G_\theta}{\lambda}\left(\frac{1}{T}\sum_{t=0}^{T-1}\left(\frac{\gamma_t}{\alpha_t}\right)^2\right)^{\frac{1}{2}} + \frac{2(C_r+3C_w)}{\lambda} \\ K_3 &:= \frac{1}{\lambda}\end{aligned}$$

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

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2 \leq 2\left(\sqrt{\frac{2C_w^2}{\lambda C_\alpha}T^{\sigma-1} + \frac{C_g C_\alpha}{1-\sigma}T^{-\sigma}} + \frac{L_w G_\theta}{\lambda}\left(\frac{1}{T}\sum_{t=0}^{T-1}\left(\frac{\gamma_t}{\alpha_t}\right)^2\right)^{\frac{1}{2}} + \frac{2(C_r+3C_w)}{\lambda}\right)^2 \\ + \frac{2}{\lambda^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta \rho_t||^2\end{aligned}$$

□

Lemma 5. *Bound on the cumulative error of average reward estimator is proven using cumulative error of critic.*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}|\Delta\rho_t|^2 \leq 2\Bigg(\sqrt{\frac{2(C_r+2C_w)^2T^{\sigma-1}}{C_\alpha}+\frac{C_sC_\alpha T^{-\sigma}}{1-\sigma}} + L_pG_\theta\Big(\frac{1}{T}\sum_{t=0}^{T-1}\Big(\frac{\gamma_t}{\alpha_t}\Big)^2\Big)^{\frac{1}{2}}+4C_w\Bigg)^2 + \frac{8}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2$$

$\Delta\rho_t = \rho_t - \rho_t^*$

Proof.

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

Here,

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

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

$$
\begin{aligned}
\mathbb{E}||\Delta\rho_{t+1}||^2 \leq\ & \mathbb{E}||\Delta\rho_t||^2 + 2\mathbb{E}||\rho_t^* - \rho_{t+1}^*||^2 \ \ ① \\
&+ 2\alpha_t^2\mathbb{E}||l(B_t, \bar{w}_t, \rho_t, \theta_t)||^2 \ \ ② \\
&+ 2\mathbb{E}\langle\Delta\rho_t, \rho_t^* - \rho_{t+1}^*\rangle \ \ ③ \\
&+ 2\alpha_t\mathbb{E}\langle\Delta\rho_{t1} - \Delta\rho_t\rangle \ \ ④ \\
&+ 2\alpha_t\mathbb{E}\langle\Delta\rho_t, \frac{1}{M}\sum_{i=0}^{M-1}(\phi^\pi(s'_{t,i}) - \phi^\pi(s_{t,i}))^\intercal(\bar{w}_t - w_t)\rangle \ \ ⑤ \\
&+ 2\alpha_t\mathbb{E}\langle\Delta\rho_t, l(B_t, w_t, \theta_t) - \bar{l}(w_t, \theta_t)\rangle \ \ ⑥ \\
&+ 2\alpha_t\mathbb{E}\langle\Delta\rho_t, \bar{l}(w_t, \theta_t) - \bar{l}(w_t^*, \theta_t)\rangle \ \ ⑦
\end{aligned}
\tag{A.14}
$$

From equation A.14:

①:

$$\mathbb{E}||\rho_t^* - \rho_{t+1}^*||^2 \leq L_p^2\mathbb{E}||\theta_{t+1} - \theta_t||^2 \text{(Lemma 14)}$$

②:

$$
\begin{aligned}
\mathbb{E}||l(B_t, \rho_t, \bar{w}_t, \theta_t)||^2 &= \mathbb{E}||\frac{1}{M}\sum_{i=0}^{M-1}\big(R^\pi(s_{t,i}) - \rho_t + \big(\phi^\pi(s'_{t,i}) - \phi^\pi(s_{t,i})\big)^\intercal\bar{w}_t\big)||^2 \\
&\leq \mathbb{E}\Big(\frac{1}{M}\sum_{i=0}^{M-1}(C_r + C_r + 2C_w + 2C_w)\Big)^2 \\
&= 4(C_r + 2C_2)^2
\end{aligned}
$$

③:

$$
\begin{aligned}
\mathbb{E}\langle\Delta\rho_t, \rho_t^* - \rho_{t+1}^*\rangle &\leq \mathbb{E}||\Delta\rho_t||\,|\rho_t^* - \rho_{t+1}^*| \\
&\leq L_p\mathbb{E}|\Delta\rho_t|\,||\theta_{t+1} - \theta_t||
\end{aligned}
$$

④:

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

⑤:

$$\begin{aligned}
&\mathbb{E}\langle \Delta\rho_t, \frac{1}{M}\sum_{i=0}^{M-1}\big(\phi^\pi(s'_{t,i})^\intercal - \phi^\pi(s_{t,i})^\intercal\big)(\bar{w}_t - w_t)\rangle \\
&\leq \mathbb{E}\Big[\frac{1}{M}\sum_{i=0}^{M-1}||\phi^\pi(s'_{t,i}) - \phi^\pi(s_{t,i})||\,||\bar{w}_t - w_t||\,|\Delta\rho_t|\Big] \\
&\leq 4C_w\mathbb{E}|\Delta\rho_t|
\end{aligned}$$

⑥:

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

⑦:

$$\begin{aligned}
\mathbb{E}[\langle \Delta\rho_t, \bar{l}(w_t, \theta_t) - \bar{l}(w_t^*, \theta_t)\rangle] &= \mathbb{E}[\langle \Delta\rho_t, \big(\mathbb{E}[\phi^\pi(s')] - \phi^\pi(s)\big)^\intercal(w_t - w_t^*)\rangle] \\
&\leq \mathbb{E}[\langle \phi^\pi(s') - \phi^\pi(s), \Delta w_t\rangle|\Delta\rho_t|] \\
&\leq \mathbb{E}[||\phi^\pi(s') - \phi^\pi(s)||\,||\Delta w_t||\,|\Delta\rho_t|] \\
&\leq 2\mathbb{E}(||\Delta w_t||\,|\Delta\rho_t|)
\end{aligned}$$

Combining ①-⑦ into equation A.14:

$$\begin{aligned}
\mathbb{E}||\Delta\rho_{t+1}||^2 \leq (1 - 2\alpha_t)\mathbb{E}||\Delta\rho_t||^2 &+ 2L_p^2\mathbb{E}||\theta_{t+1} - \theta_t||^2 \\
&+ 8\alpha_t^2(C_r + 2C_w)^2 + 2L_p\mathbb{E}|\Delta\rho_t|\,||\theta_{t+1} - \theta_t|| \\
&+ 8\alpha_t C_w\mathbb{E}|\Delta\rho_t| + 4\alpha_t\mathbb{E}||\Delta w_t||\,|\Delta\rho_t|
\end{aligned}$$

$$\begin{aligned}
\implies \sum_{t=0}^{T-1}\mathbb{E}||\Delta\rho_t||^2 \leq &\sum_{t=0}^{T-1}\frac{1}{2\alpha_t}\Big(\mathbb{E}||\Delta\rho_t||^2 - \mathbb{E}||\Delta\rho_{t+1}||^2\Big) \;\; ① \\
&+ \sum_{t=0}^{T-1}\Big(\frac{L_p^2\gamma_t^2}{\alpha_t}G_\theta^2 + 4\alpha_t(C_r + 2C_w)^2\Big) \;\; ② \\
&+ \sum_{t=0}^{T-1}\Big(L_p\frac{\gamma_t}{\alpha_t}G_\theta + 4C_w\Big)\mathbb{E}|\Delta\rho_t| \;\; ③ \\
&+ \sum_{t=0}^{T-1}2\mathbb{E}||\Delta w_t||\,|\Delta\rho_t| \;\; ④
\end{aligned} \tag{A.15}$$

From equation A.15:
①:

$$\begin{aligned}
\frac{1}{2}\sum_{t=0}^{T-1}\frac{1}{\alpha_t}(\mathbb{E}||\Delta\rho_t||^2 - \mathbb{E}||\Delta\rho_{t+1}||^2) &= \frac{1}{2}\left(\sum_{t=0}^{T-1}\Big(\frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}}\Big)\mathbb{E}|\Delta\rho_t|^2 + \frac{1}{\alpha_0}\mathbb{E}|\Delta\rho_0|^2 - \frac{1}{\alpha_{T-1}}\mathbb{E}|\Delta\rho_t|^2\right) \\
&\leq \frac{1}{2}\left(\sum_{t=0}^{T-1}\Big(\frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}}\Big) + \frac{1}{\alpha_0}\right)4(C_r + 2C_w)^2 \\
&\leq \frac{2(C_r + 2C_2)^2}{C_\alpha}T^\sigma
\end{aligned}$$

②:

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

③:

$$\sum_{t=0}^{T-1}\Big(L_pG_\theta\frac{\gamma_t}{\alpha_t}+4C_w\Big)\mathbb{E}||\Delta\rho_t||=\sum_{t=0}^{T-1}L_pG_\theta\frac{\gamma_t}{\alpha_t}\mathbb{E}||\Delta\rho_t||+4C_w\sum_{t=0}^{T-1}\mathbb{E}||\Delta\rho_t||$$
$$\leq L_pG_\theta\left(\sum_{t=0}^{T-1}\Big(\frac{\gamma_t}{\alpha_t}\Big)^2\right)^{\frac{1}{2}}\Big(\sum_{t=0}^{T-1}\mathbb{E}|\Delta\rho_t|^2\Big)^{\frac{1}{2}}+4C_w\Big(\sum_{t=0}^{T-1}\mathbb{E}|\Delta\rho_t|^2\Big)^{\frac{1}{2}}T^{\frac{1}{2}}$$
(using cauchy schwarz inequality)

④:

$$2\sum_{t-0}^{T-1}\mathbb{E}||\Delta w_t||\,|\Delta\rho_t|\leq 2(\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2)^{\frac{1}{2}}(\sum_{t=0}^{T-1}\mathbb{E}|\Delta\rho_t|^2)^{\frac{1}{2}}$$
(using cauchy schwarz inequality)

Combining ①-④ into equation A.15

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta\rho_t||^2\leq\frac{2(C_r+2C_w)^2T^{\sigma-1}}{C_\alpha}+\frac{C_sC_\alpha T^{-\sigma}}{1-\sigma}$$
$$+L_pG_\theta\left(\frac{1}{T}\sum_{t=0}^{T-1}\Big(\frac{\gamma_t}{\alpha_t}\Big)^2\right)^{\frac{1}{2}}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}|\Delta\rho_t|^2\right)^{\frac{1}{2}}$$
$$+4C_w\Big(\frac{1}{T}\sum_{t=0}T-1\mathbb{E}|\Delta\rho_t|^2\Big)^{\frac{1}{2}}$$
$$+2\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2\right)^{\frac{1}{2}}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}|\Delta\rho_t|^2\right)^{\frac{1}{2}}$$

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

Here,

$$K_1=\frac{2(C_r+2C_2)^2T^{\sigma-1}}{C_\alpha}+\frac{C_sC_\alpha T^{-\sigma}}{1-\sigma}$$
$$K_2=L_pG_\theta\left(\frac{1}{T}\sum_{t=0}^{T-1}\Big(\frac{\gamma_t}{\alpha_t}\Big)^2\right)^{\frac{1}{2}}+4C_w$$
$$K_3=2$$

From Lemma 4, we know that

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

Hence,

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}|\Delta\rho_t|^2 \leq 2\Big(&\sqrt{\frac{2(C_r+2C_w)^2}{C_\alpha}T^{\sigma-1} + \frac{C_sC_\alpha}{1-\sigma}T^{-\sigma}} \\ &+ L_pG_\theta\left(\frac{1}{T}\sum_{t=0}^{T-1}\left(\frac{\gamma_t}{\alpha_t}\right)^2\right)^{\frac{1}{2}} + 4C_w\Big)^2 \\ &+ 8\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\Delta w_t||^2\end{aligned}$$

□

Theorem 3. *The on-policy average reward actor critic algorithm obtain ϵ-accurate optimal point with sample complexity of $\Omega(\epsilon^{-2.5})$.*

$$\min_{0\leq t\leq T-1} E||\nabla_\theta\rho(\theta_t)||^2 = \mathcal{O}\left(\frac{1}{T^{0.4}}\right) + \mathcal{O}(1)$$
$$\min_{0\leq t\leq T-1} E||\nabla_\theta\rho(\theta_t)||^2 \leq \epsilon + \mathcal{O}(1)$$

Proof. Using lemma 4 and lemma 5 we obtain,

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}E||\Delta w_t||^2 \leq 2\Bigg(&\sqrt{\frac{2C_w^2}{\lambda C_\alpha}T^{\sigma-1} + \frac{C_gC_\alpha}{1-\sigma}T^{-\sigma}} + \frac{L_wG_\theta}{\lambda}\Big(\frac{1}{T}\sum_{t=0}^{T-1}(\frac{\gamma_t}{\alpha_t})^2\Big)^{1/2} + \frac{2(C_r+3C_w}{\lambda}\Bigg)^2 \\ &+ \frac{4}{\lambda^2}\Bigg(\sqrt{\frac{2(C_r+2C_w)^2}{C_\alpha}T^{\sigma-1} + \frac{C_sC_\alpha}{1-\sigma}T^{-\sigma}} + L_pG_\theta\Big(\frac{1}{T}\sum_{t=0}^{T-1}(\frac{\gamma_t}{\alpha_t})^2\Big)^{1/2} + 4C_w\Bigg)^2 \\ &+ \frac{16}{\lambda^2 T}\sum_{t=0}^{T-1}E||\Delta w_t||^2\end{aligned}$$

$$\begin{aligned}\implies \frac{1}{T}\sum_{t=0}^{T-1}E||\Delta w_t||^2 \leq \frac{2\lambda^2}{\lambda^2-16}\Bigg(&\sqrt{\frac{2C_w^2}{\lambda C_\alpha}T^{\sigma-1} + \frac{C_gC_\alpha}{1-\sigma}T^{-\sigma}} \\ &+ \frac{L_wG_\theta}{\lambda}\Big(\frac{1}{T}\sum_{t=0}^{T-1}(\frac{\gamma_t}{\alpha_t})^2\Big)^{1/2} + \frac{2(C_r+3C_w}{\lambda}\Bigg)^2 \quad ① \\ &+ \frac{4}{\lambda^2-16}\Bigg(\sqrt{\frac{2(C_r+2C_w)^2}{C_\alpha}T^{\sigma-1} + \frac{C_sC_\alpha}{1-\sigma}T^{-\sigma}} \\ &+ L_pG_\theta\Big(\frac{1}{T}\sum_{t=0}^{T-1}(\frac{\gamma_t}{\alpha_t})^2\Big)^{1/2} + 4C_w\Bigg)^2 \quad ②\end{aligned} \tag{A.16}$$

From equation A.16

①:

$$\frac{1}{T}\sum_{t=0}^{T-1}(\frac{\gamma_t}{\alpha_t})^2 \leq \frac{1}{T}\sum_{t=0}^{T-1}(\frac{C_\gamma}{C_\alpha})^2\frac{1}{(1+t)^{2(v-\sigma)}} \leq \frac{T^{-2(v-\sigma)}}{1-2(v-\sigma)}$$
$$\Big(\because \sum_{t=0}^{T-1}\frac{1}{1+t^v} \leq \int_0^T \frac{1}{t^v}\,dt = \frac{T^{1-v}}{1-v}\Big)$$

$$
\begin{aligned}
&\Big(\sqrt{\frac{2C_w^2}{\lambda C_\alpha}T^{\sigma-1}+\frac{C_gC_\alpha}{1-\sigma}T^{-\sigma}}+\frac{L_wG_\theta}{\lambda}\Big(\frac{1}{T}\sum_{t=0}^{T-1}(\frac{\gamma_t}{\alpha_t})^2\Big)^{1/2}+\frac{2(C_r+3C_w}{\lambda}\Big)^2\\
&\leq\Big(\sqrt{\frac{2C_w^2}{\lambda C_\alpha}T^{\sigma-1}+\frac{C_gC_\alpha}{1-\sigma}T^{-\sigma}}+\frac{L_wG_\theta}{\lambda}\Big(\frac{T^{-2(v-\sigma)}}{1-2(v-\sigma)}\Big)^{1/2}+\frac{2(C_r+3C_w}{\lambda}\Big)^2\\
&\leq 3\Big(\frac{2C_w^2}{\lambda C_\alpha}T^{\sigma-1}+\frac{C_gC_\alpha}{1-\sigma}T^{-\sigma}+\Big(\frac{L_wG_\theta}{\lambda}\Big)^2\Big(\frac{T^{-2(v-\sigma)}}{1-2(v-\sigma)}\Big)+\Big(\frac{2(C_r+3C_w}{\lambda}\Big)^2\Big)\\
&(\because (a+b+c)^2\leq 3(a^2+b^2+c^2))\\
&=\mathcal{O}\Big(\frac{1}{T^{1-\sigma}}\Big)+\mathcal{O}\Big(\frac{1}{T^{\sigma}}\Big)+\mathcal{O}\Big(\frac{1}{T^{2(v-\sigma)}}\Big)+\mathcal{O}(1)
\end{aligned}
$$

② (similar to ①):

$$
\begin{aligned}
&\Big(\sqrt{\frac{2(C_r+2C_w)^2}{C_\alpha}T^{\sigma-1}+\frac{C_sC_\alpha}{1-\sigma}T^{-\sigma}}+L_pG_\theta\Big(\frac{1}{T}\sum_{t=0}^{T-1}(\frac{\gamma_t}{\alpha_t})^2\Big)^{1/2}+4C_w\Big)^2\\
&=\mathcal{O}\Big(\frac{1}{T^{1-\sigma}}\Big)+\mathcal{O}\Big(\frac{1}{T^{\sigma}}\Big)+\mathcal{O}\Big(\frac{1}{T^{2(v-\sigma)}}\Big)+\mathcal{O}(1)
\end{aligned}
$$

Combining ① and ②:

$$
\frac{1}{T}\sum_{t=0}^{T-1}E||\Delta w_t)||^2=\mathcal{O}\Big(\frac{1}{T^{1-\sigma}}\Big)+\mathcal{O}\Big(\frac{1}{T^{\sigma}}\Big)+\mathcal{O}\Big(\frac{1}{T^{2(v-\sigma)}}\Big)+\mathcal{O}(1) \tag{A.17}
$$

Using lemma 3 and equation A.17

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}E||\nabla_\theta\rho(\theta_t)||^2&=\mathcal{O}\Big(\frac{1}{T^{1-v}}\Big)+\mathcal{O}\Big(\frac{1}{T^{v}}\Big)+\mathcal{O}\Big(\frac{1}{T^{1-\sigma}}\Big)+\mathcal{O}\Big(\frac{1}{T^{\sigma}}\Big)\\
&+\mathcal{O}\Big(\frac{1}{T^{2(v-\sigma)}}\Big)+\mathcal{O}(1)
\end{aligned}
$$

$$
\begin{aligned}
\implies \min_{0\leq t\leq T-1}E||\nabla_\theta\rho(\theta_t)||^2&=\mathcal{O}\Big(\frac{1}{T^{1-v}}\Big)+\mathcal{O}\Big(\frac{1}{T^{v}}\Big)+\mathcal{O}\Big(\frac{1}{T^{1-\sigma}}\Big)+\mathcal{O}\Big(\frac{1}{T^{\sigma}}\Big)\\
&+\mathcal{O}\Big(\frac{1}{T^{2(v-\sigma)}}\Big)+\mathcal{O}(1)\\
&\Big(\because \min_t E||\nabla_\theta\rho(\theta_t)||^2\leq\frac{1}{T}\sum_{t=0}^{T-1}E||\nabla_\theta\rho(\theta_t)||^2\Big)
\end{aligned}
$$

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

$$
\min_{0\leq t\leq T-1}E||\nabla_\theta\rho(\theta_t)||^2=\mathcal{O}\Big(\frac{1}{T^{0.4}}\Big)+\mathcal{O}(1)
$$

$$
\mathcal{O}\Big(\frac{1}{T^{0.4}}\Big)\leq\epsilon
$$

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

Lemma 6. *The optimal critic parameter $w(\theta_t)^*$ as a function of actor parameter θ_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. For on-policy TD(0) with l2-regularization and target estimators, the following condition holds true for optimal critic parameter w_t^*:

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

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

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

$$\begin{aligned} ||w_t^* - w_{t+1}^*|| &= ||A(\theta_t)^{-1}b(\theta_t) - A(\theta_{t+1})^{-1}b(\theta_{t+1})|| \\ &\leq ||A(\theta_t)^{-1}b(\theta_t) - A(\theta_{t+1})^{-1}b(\theta_t) + A(\theta_{t+1})^{-1}b(\theta_t) - A(\theta_{t+1})^{-1}b(\theta_{t+1})|| \\ &\leq ||A(\theta_t)^{-1} - A(\theta_{t+1})^{-1}||\ ||b(\theta_t)||\ \textcircled{1} \\ &\quad + ||A(\theta_{t+1})^{-1}||\ ||b(\theta_t) - b(\theta_{t+1})||\textcircled{2} \end{aligned} \tag{A.18}$$

From equation A.18:

①:

$$\begin{aligned} ||A(\theta_t)^{-1} - A(\theta_{t+1})^{-1}|| &= ||A(\theta_t)^{-1}A(\theta_{t+1})A(\theta_{t+1})^{-1} - A(\theta_t)^{-1}A(\theta_t)^A(\theta_{t+1})^{-1}|| \\ &\leq ||A(\theta_t)^{-1}||\ ||A(\theta_t) - A(\theta_{t+1})||\ ||A(\theta_{t+1})^{-1}|| \end{aligned} \tag{A.19}$$

From equation A.19:

Here, π' and π represents the policy with parameter θ_{t+1} and θ_t respectively.

$$\begin{aligned} ||A(\theta_t) - A(\theta_{t+1})|| &\leq ||\int d^{\pi'}(s)(\phi^{\pi'}(s)(\int P^{\pi'}(s'|s)\phi^{\pi'}(s')\,ds' - \phi^{\pi'}(s))^\intercal - \eta I)\,ds \\ &\quad - \int d^\pi(s)(\phi^\pi(s)(\int P^\pi(s'|s)\phi^\pi(s')\,ds' - \phi^\pi(s))^\intercal - \eta I)\,ds|| \\ &\leq ||\int d^{\pi'}(s)(\phi^{\pi'}(s)(\int P^{\pi'}(s'|s)\phi^{\pi'}(s')\,ds')^\intercal)\,ds \\ &\quad - \int d^\pi(s)(\phi^\pi(s)(\int P^\pi(s'|s)\phi^\pi(s')\,ds')^\intercal)\,ds||\ \textcircled{1} \\ &\leq ||\int d^\pi(s)(\phi^\pi(s)(\phi^\pi(s))^\intercal)\,ds - \int d^{\pi'}(s)(\phi^{\pi'}(s)(\phi^{\pi'}(s))^\intercal)\,ds||\ \textcircled{2} \end{aligned} \tag{A.20}$$

From equation A.20:

①:

$$
\begin{aligned}
&|| \int d^{\pi'}(s)(\phi^{\pi'}(s)(\int P^{\pi'}(s'|s)\phi^{\pi'}(s')\,ds')^\intercal)\,ds - \int d^{\pi}(s)(\phi^{\pi}(s)(\int P^{\pi}(s'|s)\phi^{\pi}(s')\,ds')^\intercal)\,ds|| \\
&\leq || \int (d^{\pi'}(s) - d^{\pi}(s))\phi^{\pi'}(s)(\int P^{\pi'}(s'|s)\phi^{\pi'}(s')\,ds')^\intercal\,ds|| \\
&\qquad + || \int d^{\pi}(s)(\phi^{\pi'}(s) - \phi^{\pi}(s))(\int P^{\pi'}(s'|s)\phi^{\pi'}(s')\,ds')^\intercal\,ds|| \\
&\qquad + || \int d^{\pi}(s)\phi^{\pi}(s)(\int (P^{\pi'}(s'|s) - P^{\pi}(s'|s))\phi^{\pi'}(s')\,ds')^\intercal\,ds|| \\
&\qquad + || \int d^{\pi}(s)\phi^{\pi}(s)(\int P^{\pi}(s'|s)(\phi^{\pi'}(s') - \phi^{\pi}(s'))\,ds'^\intercal)\,ds \\
&\leq L_d||\theta_{t+1} - \theta_t|| \qquad \text{(lemma 12)} \\
&+ L_\phi||\theta_{t+1} - \theta_t|| \qquad \text{(assumption 7)} \\
&+ L_t||\theta_{t+1} - \theta_t|| \qquad \text{(assumption 8)} \\
&+ L_\phi||\theta_{t+1} - \theta_t|| \qquad \text{(assumption 7)}
\end{aligned}
$$

$$
\begin{aligned}
&|| \int d^{\pi'}(s)(\phi^{\pi'}(s)(\int P^{\pi'}(s'|s)\phi^{\pi'}(s')\,ds')^\intercal)\,ds - \int d^{\pi}(s)(\phi^{\pi}(s)(\int P^{\pi}(s'|s)\phi^{\pi}(s')\,ds')^\intercal)\,ds|| \\
&\leq (L_d + L_t + 2L_\phi)||\theta_{t+1} - \theta_t||
\end{aligned} \tag{A.21}
$$

From equation A.20:

②:

$$
\begin{aligned}
&|| \int d^{\pi}(s)(\phi^{\pi}(s)(\phi^{\pi}(s))^\intercal)\,ds - \int d^{\pi'}(s)(\phi^{\pi'}(s)(\phi^{\pi'}(s))^\intercal)\,ds|| \\
&\qquad \leq || \int (d^{\pi}(s) - d^{\pi'}(s))\phi^{\pi}(s)(\phi^{\pi}(s))^\intercal\,ds|| \\
&\qquad\qquad + || \int d^{\pi'}(s)(\phi^{\pi}(s) - \phi^{\pi'}(s))(\phi^{\pi}(s))^\intercal\,ds|| \\
&\qquad\qquad + || \int d^{\pi'}(s)\phi^{\pi'}(s)(\phi^{\pi}(s) - \phi^{\pi'}(s))^\intercal\,ds|| \\
&\qquad \leq (L_d + 2L_\phi)||\theta_{t+1} - \theta_t||
\end{aligned} \tag{A.22}
$$

Using equation A.21 and equation A.22 in equation A.20

$$
||A(\theta_t) - A(\theta_{t+1})|| \leq (2L_d + 4L_\phi + L_t)||\theta_{t+1} - \theta_t|| \tag{A.23}
$$

From equation A.18:

②:

$$
\begin{aligned}
||b(\theta_t) - b(\theta_{t+1})|| &= || \int d^{\pi'}(s)((R^{\pi'}(s) - \rho^*_{t+1})\phi^{\pi'}(s)\, ds - \int d^{\pi}(s)(R^{\pi}(s) - \rho^*_{t})\phi^{\pi}(s)\, ds|| \\
&\leq || \int d^{\pi'}(s)(R^{\pi'}(s)\phi^{\pi'}(s)\, ds - \int d^{\pi}(s)R^{\pi}(s)\phi^{\pi}(s)\, ds|| \\
&+ || \int d^{\pi'}(s)\rho^*_{t+1}\phi^{\pi'}(s)\, ds - \int d^{\pi}(s)\rho^*_{t}\phi^{\pi}(s)\, ds|| \\
&\leq || \int (d^{\pi'}(s) - d^{\pi}(s))R^{\pi'}(s)\phi^{\pi'}(s)\, ds|| \\
&+ || \int d^{\pi}(s)(R^{\pi'}(s) - R^{\pi}(s))\phi^{\pi'}(s)\, ds|| \\
&+ || \int d^{\pi}(s)R^{\pi}(s)(\phi^{\pi'}(s) - \phi^{\pi}(s)\, ds|| \\
&+ || \int (d^{\pi'}(s) - d^{\pi}(s))\rho^*_{t+1}\phi^{\pi'}(s)\, ds|| \\
&+ || \int d^{\pi}(s)(\rho^*_{t+1} - \rho^*_{t})\phi^{\pi'}(s)\, ds|| \\
&+ || \int d^{\pi}(s)\rho^*_{t}(\phi^{\pi'}(s) - \phi^{\pi}(s)\, ds|| \\
&\leq C_r L_d ||\theta_{t+1} - \theta_t|| \qquad \text{(Assumption 4, Lemma 12)} \\
&+ L_r ||\theta_{t+1} - \theta_t|| \qquad \text{(Assumtion 9)} \\
&+ C_r L_\phi ||\theta_{t+1} - \theta_t|| \qquad \text{(Assumption 4,Assumption 7)} \\
&+ C_r L_d ||\theta_{t+1} - \theta_t|| \qquad \text{(Assumption 4, Lemma 12)} \\
&+ L_p ||\theta_{t+1} - \theta_t|| \qquad \text{(Lemma 14)} \\
&+ C_r L_\phi ||\theta_{t+1} - \theta_t|| \qquad \text{(Assumption 4,Assumption 7)}
\end{aligned}
$$

$$
\implies ||b(\theta_t) - b(\theta_{t+1})|| \leq (2L_d C_r + 2C_r L_\phi + L_r + L_p)||\theta_{t+1} - \theta_t|| \tag{A.24}
$$

Using equation A.19, equation A.23 and equation A.24 in equation A.18:

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

Note:

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

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

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 7.

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

Proof.

$$\begin{aligned} &||Q^\pi(s,a_1) - Q^\pi(s,a_2)|| \leq ||a_1 - a_2|| \;\; \text{(Assumption 5)} \\ &\implies ||\nabla_a Q(s,a)|| \leq L_a \end{aligned} \tag{A.25}$$

$$\begin{aligned} &||\pi(s,\theta_1) - \pi(s,\theta_2)|| \leq L_\pi ||\theta_1 - \theta_2|| \;\; \text{(Assumption 6)} \\ &\implies ||\nabla_\theta \pi(s)|| \leq L_\pi \end{aligned} \tag{A.26}$$

Using equation A.25 and equation A.26:

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

□

Lemma 8. *The average reward estimate is bounded.*

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

Proof.

$$|\rho_0| \leq C_r + 2C_w \quad \text{(Assumption 10)}$$

For t = 1:

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

$$\begin{aligned} |\rho_1| &\leq (1-\alpha_0)|\rho_0| + \alpha_0 || \Big(\frac{1}{M} \sum_{i=0}^{M-1} R^\pi(s_{0,i}) + \phi^\pi(s'_{0,i})^\intercal \bar{w}_0 - \phi^\pi(s_{0,i})^\intercal \bar{w}_0 \Big) || \\ &\leq (1-\alpha_0)|\rho_0| + \alpha_0 \Big(\frac{1}{M} \sum_{i=0}^{M-1} |R^\pi(s_{0,i})| + ||\phi^\pi(s'_{0,i})|| \; ||\bar{w}_0|| + ||\phi^\pi(s_{0,i})|| \; ||\bar{w}_0|| \Big) \\ &\leq (1-\alpha_0)(C_r + 2C_w) + (\alpha_0)(C_r + 2C_w) = (C_r + 2C_w) \quad \text{(Assumption 10)} \end{aligned}$$

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

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

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

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 9. *The norm of target critic estimator is bounded*

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

Proof. For t=1:

$$\bar{w}_1 = (1-\beta_0)\bar{w}_0 + \beta_0 w_1$$
$$||\bar{w}_1|| \leq (1-\beta_0)||\bar{w}_0|| + \beta_0||w_1||$$
$$||\bar{w}_1|| \leq (1-\beta_0)C_w + \beta_0 C_w \quad \text{(Assumption 10)}$$
$$||\bar{w}_1|| \leq C_w$$

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

$$\bar{w}_{k+1} = (1-\beta_k)\bar{w}_k + \beta_k w_{k+1}$$
$$||\bar{w}_{k+1}|| \leq (1-\beta_k)||\bar{w}_k|| + \beta_k||w_{k+1}||$$
$$||\bar{w}_{k+1}|| \leq (1-\beta_k)C_w + \beta_k C_w \quad \text{(Assumption 10)}$$
$$||\bar{w}_{k+1}|| \leq C_w$$

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 10. *The norm of target average reward estimator is bounded*

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

Proof. For t=1:

$$\bar{\rho}_1 = (1-\beta_0)\bar{\rho}_0 + \beta_0 \rho_1$$
$$||\bar{\rho}_1|| \leq (1-\beta_0)||\bar{\rho}_0|| + \beta_0||\rho_1||$$
$$||\bar{\rho}_1|| \leq (1-\beta_0)(C_r + 2C_w) + \beta_0(C_r + 2C_w) \quad \text{(Assumption 10)}$$
$$||\bar{\rho}_1|| \leq C_r + 2C_w$$

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

$$\begin{aligned}
&\bar{\rho}_{k+1} = (1-\beta_k)\bar{\rho}_k + \beta_k \rho_{k+1} \\
&||\bar{\rho}_{k+1}|| \leq (1-\beta_k)||\bar{\rho}_k|| + \beta_k||\rho_{k+1}|| \\
&||\bar{\rho}_{k+1}|| \leq (1-\beta_k)(C_r + 2C_w) + \beta_k(C_r + 2C_w) \quad \text{(Assumption 10)} \\
&||\bar{\rho}_{k+1}|| \leq C_r + 2C_w
\end{aligned}$$

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

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

□

Lemma 11. *The $A(\theta)$ matrix defined below is negative definite for all values of θ (θ is the policy parameter).*

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

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

Proof. Let:

$$A'(\theta) = \int d^\pi(s)(\phi^\pi(s)(\int P^\pi(s'|s)\phi^\pi(s')\,ds' - \phi^\pi(s))^\intercal)\,ds = A(\theta) + \eta I \tag{A.27}$$

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

Here, $\Big(\frac{A(\theta) + A'(\theta)}{2}\Big)x$ is a symmetric matrix and λ_{max} is the maximum eigen value of the $\Big(\frac{A(\theta) + A'(\theta)}{2}\Big)x$.

$$\begin{aligned}
&x^\intercal A'(\theta)x \leq \lambda_{max}||x||^2 \\
&x^\intercal (A'(\theta) - \eta I)x \leq (\lambda_{max} - \eta)||x||^2 \\
&x^\intercal A(\theta)x \leq (\lambda_{max} - \eta)||x||^2 \;\; \text{(using A.27)}
\end{aligned}$$

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

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

□

Lemma 12. *Let θ_1 and θ_2 be the policy parameter for π' and π respectively. $d^{\pi'}(\cdot)$ and $d^\pi(\cdot)$ be the stationary state distribution for π' and π respectively. We have:*

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

Here, $L_d = 2^{m+1}C(\lceil \log_\kappa a^{-1}\rceil + 1/\kappa)L_t$. Constants a and κ are defined in Assumption 2

Proof.

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

Let μ_1 and μ_2 be the stationary state probability measure for π' and π respectively. Then we have :

$$d\mu_1 = d^{pi'}(s)\, ds$$
$$d\mu_2 = d^{pi}(s)\, ds$$

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

$$2D_{TV}(\mu_1, \mu_2) \leq 2\Big(\lceil \log_\kappa a^{-1} \rceil + \frac{1}{\kappa}\Big)||K_1 - K_2|| \tag{A.28}$$

where K_1 and K_2 are probability transition kernel for markov chain induced by policy π' and π. From equation A.28:

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

$$\begin{aligned}
|| \int g(ds)(K_1(\cdot|s) - K_2(\cdot|s))||_{TV} &\leq \sup_{|f|\leq 1} | \iint f(s')(K_1 - k_2)(ds'|s)g(ds)| \\
&\leq \sup_{|f|\leq 1} | \iint f(s')(P^{\pi'}(s'|s) - P^{\pi}(s'|s))(s'|s)g(ds)ds'| \\
&\leq \sup_{|f|\leq 1} \iint |f(s')|\, |(P^{\pi'}(s'|s) - P^{\pi}(s'|s)|g(ds)ds' \\
&\leq L_t||\theta_1 - \theta_2|| \int g(ds) \int ds' \\
&\leq 2^m L_t||\theta_1 - \theta_2||
\end{aligned}$$

$$\implies ||K_1 - K_2|| \leq 2^m L_t||\theta_1 - \theta_2|| \tag{A.29}$$

From equation A.28 and equation A.29:

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

□

Lemma 13. *The optimal critic parameter according to compatible function approximation lemma (2) is bounded.*

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

Proof. From Lemma2:

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

Here,

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

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

By assuming the minimum eigenvalue of H_θ is λ_{min}^ϵ and using Assumption 4 :

$$||w_\epsilon^*|| \leq \frac{C_r}{\lambda_{min}^\epsilon} = C_{w_\epsilon^*}$$

□

Lemma 14. *The performance metric $\rho(\pi)(\rho(\theta))$ is lipchitz continuous wrt to the policy parameter θ.*

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

Proof. Let θ_1 and θ_2 be the policy parameters of policy π' and π.

$$\begin{aligned}
||\rho(\theta_1) - \rho(\theta_2)|| &= ||\rho(\pi') - \rho(\pi)|| \\
&= ||\int_S d^{\pi'}(s) R^{\pi'}(s)\, ds - \int_S d^\pi(s) R^\pi(s)\, ds|| \\
&\leq ||\int_S (d^{\pi'}(s) - d^\pi(s)) R^{\pi'}(s)\, ds|| + ||\int_S d^\pi(s)(R^{\pi'}(s) - R^\pi(s))\, ds|| \\
&\leq L_d ||\theta_1 - \theta_2|| \quad \text{(Lemma 12)} \\
&\quad + L_r ||\theta_1 - \theta_2|| \quad \text{(Assumption 9)} \\
&\leq (L_d + L_r)||\theta_1 - \theta_2|| = L_p ||\theta_1 - \theta_2|| \quad (L_d + L_r = L_p)
\end{aligned}$$

□

Lemma 15. *The optimal critic parameter $w(\theta_t)^*$ as a function of actor parameter θ_t is Lipchitz continuous with constant L_v for off-policy case. Note: $w_t^* = w(\theta_t)^*$.*

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

Proof. For off-policy TD(0) with l2-regularization and target estimators, the following condition holds true for optimal critic parameter w_t^*:

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

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

$$\begin{aligned}
||w_t^* - w_{t+1}^*|| &= ||A(\theta_t)^{-1} b(\theta_t) - A(\theta_{t+1})^{-1} b(\theta_{t+1})|| \\
&\leq ||A(\theta_t)^{-1} b(\theta_t) - A(\theta_{t+1})^{-1} b(\theta_t) + A(\theta_{t+1})^{-1} b(\theta_t) - A(\theta_{t+1})^{-1} b(\theta_{t+1})|| \\
&\leq ||A(\theta_t)^{-1} - A(\theta_{t+1})^{-1}||\ ||b(\theta_t)||\ \ ① \\
&\quad + ||A(\theta_{t+1})^{-1}||\ ||b(\theta_t) - b(\theta_{t+1})||\ \ ②
\end{aligned} \tag{A.30}$$

From equation A.30:

①:

$$\begin{aligned}
||A(\theta_t)^{-1} - A(\theta_{t+1})^{-1}|| &= ||A(\theta_t)^{-1} A(\theta_{t+1}) A(\theta_{t+1})^{-1} - A(\theta_t)^{-1} A(\theta_t)^A (\theta_{t+1})^{-1}|| \\
&\leq ||A(\theta_t)^{-1}||\ ||A(\theta_t) - A(\theta_{t+1})||\ ||A(\theta_{t+1})^{-1}||
\end{aligned} \tag{A.31}$$

From equation A.31:

Here, π' and π represents the policy with parameter θ_{t+1} and θ_t respectively and μ be the behaviour policy .

$$
\begin{aligned}
||A(\theta_t) - A(\theta_{t+1})|| &\leq || \int d^{\mu}(s)(\phi^{\pi'}(s)(\int P^{\mu'}(s'|s)\phi^{\pi'}(s')\, ds' - \phi^{\pi'}(s))^{\intercal} - \eta I)\, ds \\
&\quad - \int d^{\mu}(s)(\phi^{\pi}(s)(\int P^{\mu}(s'|s)\phi^{\pi}(s')\, ds' - \phi^{\pi}(s))^{\intercal} - \eta I)\, ds|| \\
&\leq || \int d^{\mu}(s)(\phi^{\pi'}(s)(\int P^{\mu}(s'|s)\phi^{\pi'}(s')\, ds')^{\intercal})\, ds \\
&\quad - \int d^{\mu}(s)(\phi^{\pi}(s)(\int P^{\mu}(s'|s)\phi^{\pi}(s')\, ds')^{\intercal})\, ds|| \quad ① \\
&\leq || \int d^{\mu}(s)(\phi^{\pi}(s)(\phi^{\pi}(s))^{\intercal})\, ds - \int d^{\mu}(s)(\phi^{\pi'}(s)(\phi^{\pi'}(s))^{\intercal})\, ds|| \quad ②
\end{aligned}
\tag{A.32}
$$

From equation A.32:

①:

$$
\begin{aligned}
&|| \int d^{\mu}(s)(\phi^{\pi'}(s)(\int P^{\mu}(s'|s)\phi^{\pi'}(s')\, ds')^{\intercal})\, ds - \int d^{\mu}(s)(\phi^{\pi}(s)(\int P^{\mu}(s'|s)\phi^{\pi}(s')\, ds')^{\intercal})\, ds|| \\
&\leq || \int (d^{\mu}(s) - d^{\mu}(s))\phi^{\pi'}(s)(\int P^{\mu}(s'|s)\phi^{\pi'}(s')\, ds')^{\intercal}\, ds|| \\
&\quad + || \int d^{\mu}(s)(\phi^{\pi'}(s) - \phi^{\pi}(s))(\int P^{\mu}(s'|s)\phi^{\pi'}(s')\, ds')^{\intercal}\, ds|| \\
&\quad + || \int d^{\mu}(s)\phi^{\pi}(s)(\int (P^{\mu}(s'|s) - P^{\mu}(s'|s))\phi^{\pi'}(s')\, ds')^{\intercal}\, ds|| \\
&\quad + || \int d^{\mu}(s)\phi^{\pi}(s)(\int P^{\mu}(s'|s)(\phi^{\pi'}(s') - \phi^{\pi}(s'))\, ds'^{\intercal})\, ds \\
&\leq L_{\phi}||\theta_{t+1} - \theta_t|| \qquad \text{(Assumption 7)} \\
&+ L_{\phi}||\theta_{t+1} - \theta_t|| \qquad \text{(Assumption 7)}
\end{aligned}
$$

$$
\begin{aligned}
&|| \int d^{\mu}(s)(\phi^{\pi'}(s)(\int P^{\mu}(s'|s)\phi^{\pi'}(s')\, ds')^{\intercal})\, ds - \int d^{\mu}(s)(\phi^{\pi}(s)(\int P^{\mu}(s'|s)\phi^{\pi}(s')\, ds')^{\intercal})\, ds|| \\
&\leq 2L_{\phi}||\theta_{t+1} - \theta_t||
\end{aligned}
\tag{A.33}
$$

From equation A.32:

②:

$$
\begin{aligned}
&|| \int d^{\mu}(s)(\phi^{\pi}(s)(\phi^{\pi}(s))^{\intercal})\, ds - \int d^{\mu}(s)(\phi^{\pi'}(s)(\phi^{\pi'}(s))^{\intercal})\, ds|| \\
&\leq || \int d^{\mu}(s)(\phi^{\pi}(s) - \phi^{\pi'}(s))(\phi^{\pi}(s))^{\intercal}\, ds|| \\
&\quad + || \int d^{\mu}(s)\phi^{\pi'}(s)(\phi^{\pi}(s) - \phi^{\pi'}(s))^{\intercal}\, ds|| \\
&\leq 2L_{\phi}||\theta_{t+1} - \theta_t||
\end{aligned}
\tag{A.34}
$$

Using equation A.33 and equation A.34 in equation A.32

$$
||A(\theta_t) - A(\theta_{t+1})|| \leq (4L_{\phi} + L_t)||\theta_{t+1} - \theta_t|| \tag{A.35}
$$

From equation A.30:

②:

$$
\begin{aligned}
||b(\theta_t) - b(\theta_{t+1})|| &= || \int d^\mu(s)((R^\mu(s) - \rho^*_{t+1})\phi^{\pi'}(s)\, ds - \int d^\mu(s)(R^\mu(s) - \rho^*_t)\phi^\pi(s)\, ds|| \\
&\leq || \int d^\mu(s)(R^\mu(s)\phi^{\pi'}(s)\, ds - \int d^\mu(s)R^\mu(s)\phi^\pi(s)\, ds|| \\
&+ || \int d^\mu(s)\rho^*_{t+1}\phi^{\pi'}(s)\, ds - \int d^\mu(s)\rho^*_t\phi^\pi(s)\, ds|| \\
&\leq || \int d^\mu(s)(R^\mu(s) - R^\mu(s))\phi^{\pi'}(s)\, ds|| \\
&+ || \int d^\mu(s)R^\mu(s)(\phi^{\pi'}(s) - \phi^\pi(s)\, ds|| \\
&+ || \int d^\mu(s)(\rho^*_{t+1} - \rho^*_t)\phi^{\pi'}(s)\, ds|| \\
&+ || \int d^\mu(s)\rho^*_t(\phi^{\pi'}(s) - \phi^\pi(s)\, ds|| \\
&\leq C_r L_\phi ||\theta_{t+1} - \theta_t|| \qquad \text{(Assumption 4)} \\
&+ L_p ||\theta_{t+1} - \theta_t|| \qquad \text{(Lemma 14)} \\
&+ C_r L_\phi ||\theta_{t+1} - \theta_t|| \qquad \text{(Assumption 4)}
\end{aligned}
$$

$$
\implies ||b(\theta_t) - b(\theta_{t+1})|| \leq (2C_r L_\phi + L_p)||\theta_{t+1} - \theta_t|| \tag{A.36}
$$

Using equation A.31, equation A.35 and equation A.36 in equation A.30:

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

Note:

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

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

where,

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

□

Lemma 16. *Bound on the cumulative error of off-policy actor with behaviour policy μ is proven using cumulative error of critic. θ^{μ} is the policy parameter for μ.*

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

$R(s,a) \leq C_r, \gamma_t = \dfrac{C_\gamma}{(1+t)^v}, ||\nabla_\theta \pi(s)|| \leq C_\pi, \Delta w_t = w_t - w_t^*, \tau = \max_t ||w_t^* - w_{\epsilon,t}^*||, Z = 2^{m+1}C(\lceil \log_\kappa a^{-1} \rceil + 1/\kappa)L_t$

Proof.

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

Using Theorem 2 and Lemma 3:

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

□

Theorem 4. *The off-policy average reward actor critic algorithm with behavior policy μ obtains ϵ-accurate optimal point with sample complexity of $\Omega(\epsilon^{-2.5})$.*

$$\min_{0\leq t\leq T-1} E||\nabla_\theta \bar{\rho}(\theta_t)||^2 = \mathcal{O}\left(\frac{1}{T^{0.4}}\right) + \mathcal{O}(1) + \mathcal{O}(N_\theta^2)$$
$$\min_{0\leq t\leq T-1} E||\nabla_\theta \bar{\rho}(\theta_t)||^2 \leq \epsilon + \mathcal{O}(1) + \mathcal{O}(N_\theta^2)$$
$$N_\theta := \max_t ||\theta_\mu - \theta_t||$$

Proof. Lemma 4 and Lemma 5 will hold in the case of off-policy update. Lemma 4 will require Lemma 15 instead of Lemma 6.

Using Lemma 4 and Lemma 5 and using the procedure followed in Theorem 3 to obtain asymptotic notations, we have:

$$\frac{1}{T}\sum_{t=0}^{T-1} E||\Delta w_t)||^2 = \mathcal{O}\left(\frac{1}{T^{1-\sigma}}\right) + \mathcal{O}\left(\frac{1}{T^{\sigma}}\right) + \mathcal{O}\left(\frac{1}{T^{2(v-\sigma)}}\right) + \mathcal{O}(1) \tag{A.37}$$

Using Lemma 16 and equation A.37:

$$\min_{0\leq t\leq T-1} E||\nabla_\theta \bar{\rho}(\theta_t)||^2 = \mathcal{O}\left(\frac{1}{T^{1-v}}\right) + \mathcal{O}\left(\frac{1}{T^{v}}\right) + \mathcal{O}\left(\frac{1}{T^{1-\sigma}}\right) + \mathcal{O}\left(\frac{1}{T^{\sigma}}\right) + \mathcal{O}\left(\frac{1}{T^{2(v-\sigma)}}\right) + \mathcal{O}(1) + \frac{Z}{T}\sum_{t=0}^{T-1} E||\theta^\mu - \theta^t||^2$$

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

$$\begin{aligned}\min_{0\leq t\leq T-1} E||\nabla_\theta \bar{\rho}(\theta_t)||^2 &= \mathcal{O}\left(\frac{1}{T^{0.4}}\right) + \mathcal{O}(1) + \frac{Z}{T}\sum_{t=0}^{T-1} E||\theta^\mu - \theta^t||^2 \\ &= \mathcal{O}\left(\frac{1}{T^{0.4}}\right) + \mathcal{O}(1) + ZN_\theta^2 \\ &= \mathcal{O}\left(\frac{1}{T^{0.4}}\right) + \mathcal{O}(1) + \mathcal{O}(N_\theta^2).\end{aligned}$$

Further,

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

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

B ALGORITHM AND HYPERPARAMETERS

B.1 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. Γ_{C_w} is a projection operator defined as $\Gamma_{C_w} : \mathbb{R}^k \to B$, 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 θ and critic parameters w. Initialize actor target parameter $\theta \rightarrow \overline{\theta}$.
Initialize critic target parameters $w \rightarrow \overline{w}$.Initialize average reward parameter ρ
Initialize target average reward parameter $\rho \rightarrow \overline{\rho}$
Initialize buffer = { }

$t = 0$, s_0 = env.reset()
while $t \leq$ total steps **do**
  $a_t = \pi(s_t) + \epsilon$ {ϵ 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 $\mathcal{B}_t = \{s_i, a_i, s'_i\}_{i=0}^{N-1}$ from the Replay Buffer
    $w_{t+1} = \Gamma_{C_w}\Big(w_t + \frac{\alpha_t}{M}\sum_{i=0}^{M-1}\Big(R^\pi(s_i) - \bar{\rho}_t + \phi^\pi(s'_i)^\intercal \bar{w}_t - \phi^\pi(s_i)^\intercal w_t\Big)\phi^\pi(s_i) - \alpha_t \eta w_t\Big)$
    $\rho_{t+1} = \rho_t + \frac{\alpha_t}{M}\sum_{i=0}^{M-1}\Big(R^\pi(s_i) - \rho_t + \phi^\pi(s'_i)^\intercal \bar{w}_t - \phi^\pi(s_i)^\intercal \bar{w}_t\Big)$
    $\overline{w}_{t+1} = \overline{w}_t + \beta_t(w_{t+1} - \overline{w}_{t+1})$
    $\overline{\rho}_{t+1} = \overline{\rho}_t + \beta_t(\rho_{t+1} - \overline{\rho}_{t+1})$
    $\theta_{t+1} = \theta_t + \frac{\gamma_t}{M}\sum_{i=0}^{M-1}\nabla_a Q^\pi_{diff}(s_i, a)|_{a=\pi(s_i)}\nabla_\theta \pi(s_i)$
    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 θ 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 ρ and
Initialize target average reward parameter $\rho \rightarrow \overline{\rho}$
μ is the behavior policy
Initialize Replay buffer = { }

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

B.2 HYPERPARAMETERS

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