Deterministic Policy Gradient: Convergence Analysis (Supplementary material)

Huaqing Xiong[*1] **Tengyu Xu**[*1] **Lin Zhao**[2] **Yingbin Liang**[1] **Wei Zhang**[3]

[1]Department of Electrical and Computer Engineering, The Ohio State University, Columbus, Ohio, USA
[2]Department of Electrical and Computer Engineering, National University of Singapore, Singapore, Republic of Singapore
[3]Department of Mechanical and Energy Engineering, Southern University of Science and Technology (SUSTech),
Shenzhen, Guangdong, China

1 PROOF OF LEMMA 1

1.1 SUPPORTING LEMMAS

We first provide some useful lemmas. The first lemma provides the Lipschitz continuity property of the state visitation measure.

Lemma 1. *Suppose Assumptions 1 and 2 hold. We define the total variation norm between two state visitation distributions respectively corresponding to two policies* $\mu_{\theta_1}, \mu_{\theta_2}$ *as* $\|\nu_{\theta_1}(\cdot) - \nu_{\theta_2}(\cdot)\|_{TV} = \int_{\mathcal{S}} |\nu_{\theta_1}(ds) - \nu_{\theta_2}(ds)|$. *Then there exists some constant* $L_\nu > 0$, *such that*

$$\|\nu_{\theta_1}(\cdot) - \nu_{\theta_2}(\cdot)\|_{TV} \leq L_\nu \|\theta_1 - \theta_2\| .$$

Proof. Since we consider ergodic Markov chains, Theorem 3.1 of Mitrophanov [2005] shows that there exists some constant $C_\nu > 1$, such that

$$\|\nu_{\theta_1}(\cdot) - \nu_{\theta_2}(\cdot)\|_{TV} \leq C_\nu \|P_{\theta_1} - P_{\theta_2}\|_{\text{op}} , \tag{1}$$

where P_θ denotes the state transition kernel corresponding to a policy μ_θ, and the operator norm $\|\cdot\|_{\text{op}}$ is given by $\|P\|_{\text{op}} = \sup_{\|q\|_{TV}=1} \|qP\|_{TV}$. Then we have

$$\begin{aligned}
\|P_{\theta_1} - P_{\theta_2}\|_{\text{op}} &= \sup_{\|q\|_{TV}=1} \left\| \int_{\mathcal{S}} (P_{\theta_1} - P_{\theta_2})(s, \cdot) q(ds) \right\|_{TV} \\
&= \frac{1}{2} \sup_{\|q\|_{TV}=1} \int_{s'} \left| \int_s (P_{\theta_1}(s, ds') - P_{\theta_2}(s, ds'))\, q(ds) \right| \\
&\leq \frac{1}{2} \sup_{\|q\|_{TV}=1} \int_{s'} \int_s |P_{\theta_1}(s, ds') - P_{\theta_2}(s, ds')|\, q(ds) \\
&= \frac{1}{2} \sup_{\|q\|_{TV}=1} \int_{s'} \int_s |P(ds'|s, \mu_{\theta_1}(s)) - P(ds'|s, \mu_{\theta_1}(s))|\, q(ds) \\
&\overset{\text{(i)}}{\leq} \frac{1}{2} \sup_{\|q\|_{TV}=1} \int_s L_P \|\mu_{\theta_1}(s) - \mu_{\theta_2}(s)\|\, q(ds) \\
&\overset{\text{(ii)}}{\leq} \frac{1}{2} L_P L_\mu \|\theta_1 - \theta_2\| ,
\end{aligned}$$

where (i) follows form Assumption 2, and (ii) follows from Assumption 1. Then, combining the above bound together with (1) completes the proof. □

[*]equal contribution

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

Next, we show that the value function of a deterministic policy is Lipschitz continuous.

Lemma 2. *Suppose Assumptions 1 and 2 hold. The value function is Lipschitz continuous w.r.t. the policies. That is, for any $\theta_1, \theta_2 \in \mathbb{R}^d$, $s \in \mathcal{S}$, we have*

$$\left\|V^{\mu_{\theta_1}}(s) - V^{\mu_{\theta_2}}(s)\right\| \le L_V \left\|\theta_1 - \theta_2\right\|,$$

where $L_V = R_{\max} L_\nu + \frac{L_r L_\mu}{1-\gamma}$.

Proof. By definition, we have $V^{\mu_\theta}(s_0) = \int_{\mathcal{S}} r(s, \mu_\theta(s)) \nu_{\mu_\theta}^{s_0}(ds)$, where $\nu_{\mu_\theta}^{s_0}(\cdot)$ is the discounted state visitation measure given the initial state, i.e., $\nu_{\mu_\theta}^{s_0}(s) = \int_{\mathcal{S}} \sum_{t=0}^{\infty} \gamma^t p(s_0 \to s, t, \mu_\theta) ds$. We then derive

$$\begin{aligned}
&\left|V^{\mu_{\theta_1}}(s_0) - V^{\mu_{\theta_2}}(s_0)\right| \\
&= \left| \int_{\mathcal{S}} r(s, \mu_{\theta_1}(s)) \nu_{\mu_{\theta_1}}^{s_0}(ds) - \int_{\mathcal{S}} r(s, \mu_{\theta_2}(s)) \nu_{\mu_{\theta_2}}^{s_0}(ds) \right| \\
&\le \left| \int_{\mathcal{S}} r(s, \mu_{\theta_1}(s)) \nu_{\mu_{\theta_1}}^{s_0}(ds) - \int_{\mathcal{S}} r(s, \mu_{\theta_1}(s)) \nu_{\mu_{\theta_2}}^{s_0}(ds) \right| \\
&\quad + \left| \int_{\mathcal{S}} r(s, \mu_{\theta_1}(s)) \nu_{\mu_{\theta_2}}^{s_0}(ds) - \int_{\mathcal{S}} r(s, \mu_{\theta_2}(s)) \nu_{\mu_{\theta_2}}^{s_0}(ds) \right| \\
&\le \int_{\mathcal{S}} \left| r(s, \mu_{\theta_1}(s)) \right| \cdot \left| \nu_{\mu_{\theta_1}}^{s_0}(ds) - \nu_{\mu_{\theta_2}}^{s_0}(ds) \right| + \int_{\mathcal{S}} \left| r(s, \mu_{\theta_1}(s)) - r(s, \mu_{\theta_2}(s)) \right| \nu_{\mu_{\theta_2}}^{s_0}(ds) \\
&\overset{\text{(i)}}{\le} R_{\max} \left\| \nu_{\mu_{\theta_1}}^{s_0}(\cdot) - \nu_{\mu_{\theta_2}}^{s_0}(\cdot) \right\|_{TV} + L_r \int_{\mathcal{S}} \left\| \mu_{\theta_1}(s) - \mu_{\theta_2}(s) \right\| \nu_{\mu_{\theta_2}}^{s_0}(ds) \\
&\overset{\text{(ii)}}{\le} R_{\max} L_\nu \left\|\theta_1 - \theta_2\right\| + L_r L_\mu \left\|\theta_1 - \theta_2\right\| \int_{\mathcal{S}} \nu_{\mu_{\theta_2}}^{s_0}(ds) \\
&= \left(R_{\max} L_\nu + \frac{L_r L_\mu}{1-\gamma} \right) \left\|\theta_1 - \theta_2\right\|,
\end{aligned}$$

where (i) follows from Assumption 2, and (ii) follows from Lemma 1 and Assumption 1.

□

The next lemma establishes the boundedness and Lipschitz continuity property for the gradient of Q-function.

Lemma 3. *Suppose Assumptions 1-3 hold. The gradient of Q-function w.r.t. action is uniformly bounded. That is, for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, $\theta \in \mathbb{R}^d$,*

$$\left\|\nabla_a Q^{\mu_\theta}(s, a)\right\| \le C_Q,$$

where $C_Q = L_r + L_P \cdot \frac{\gamma R_{\max}}{1-\gamma}$. Furthermore, $\nabla_a Q^{\mu_\theta}(s, a_\theta)$ is Lipschitz continuous w.r.t. θ, that is, for any $\theta_1, \theta_2 \in \mathbb{R}^d$, we have

$$\left\|\nabla_a Q^{\mu_{\theta_1}}(s, a_{\theta_1}) - \nabla_a Q^{\mu_{\theta_2}}(s, a_{\theta_2})\right\| \le L'_Q \left\|\theta_1 - \theta_2\right\|,$$

where $L'_Q = L_Q L_\mu + \gamma L_P L_V$.

Proof. For the boundedness property, we have

$$\begin{aligned}
\left\|\nabla_a Q^{\mu_\theta}(s, a)\right\| &= \left\| \nabla_a \int_{\mathcal{S}} \left(r(s, a) + \gamma P(s'|s, a) V^{\mu_\theta}(s') \right) ds' \right\| \\
&\le \left\|\nabla_a r(s, a)\right\| + \gamma \int_{\mathcal{S}} \left\|\nabla_a P(s'|s, a)\right\| \cdot \left|V^{\mu_\theta}(s')\right| ds' \\
&\le L_r + L_P \cdot \frac{\gamma R_{\max}}{1-\gamma},
\end{aligned}$$

where the last inequality follows from Assumptions 1, 2 and the fact that $\left|V^{\mu_\theta}(s')\right| \le \frac{R_{\max}}{1-\gamma}$.

We next show the Lipschitz property as follows.

$$\begin{aligned}
&\left\|\nabla_a Q^{\mu_{\theta_1}}(s,a_{\theta_1}) - \nabla_a Q^{\mu_{\theta_2}}(s,a_{\theta_2})\right\| \\
&\quad \le \left\|\nabla_a Q^{\mu_{\theta_1}}(s,a_{\theta_1}) - \nabla_a Q^{\mu_{\theta_1}}(s,a_{\theta_2})\right\| + \left\|\nabla_a Q^{\mu_{\theta_1}}(s,a_{\theta_2}) - \nabla_a Q^{\mu_{\theta_2}}(s,a_{\theta_2})\right\| \\
&\quad \overset{\text{(i)}}{\le} L_Q \left\|a_{\theta_1} - a_{\theta_2}\right\| + \left\|\nabla_a Q^{\mu_{\theta_1}}(s,a_{\theta_2}) - \nabla_a Q^{\mu_{\theta_2}}(s,a_{\theta_2})\right\| \\
&\quad = L_Q \left\|\mu_{\theta_1}(s) - \mu_{\theta_2}(s)\right\| + \left\|\nabla_a Q^{\mu_{\theta_1}}(s,a_{\theta_2}) - \nabla_a Q^{\mu_{\theta_2}}(s,a_{\theta_2})\right\| \\
&\quad \overset{\text{(ii)}}{\le} L_Q L_\mu \left\|\theta_1 - \theta_2\right\| + \left\|\int_{\mathcal{S}} \gamma \nabla_a P(s'|s,a)\left(V^{\mu_{\theta_1}}(s') - V^{\mu_{\theta_2}}(s')\right) ds'\right\| \\
&\quad \le L_Q L_\mu \left\|\theta_1 - \theta_2\right\| + \gamma \int_{\mathcal{S}} \left\|\nabla_a P(s'|s,a)\right\| \cdot \left|V^{\mu_{\theta_1}}(s') - V^{\mu_{\theta_2}}(s')\right| ds' \\
&\quad \overset{\text{(iii)}}{\le} \left(L_Q L_\mu + \gamma L_P L_V\right) \left\|\theta_1 - \theta_2\right\|,
\end{aligned}$$

where (i) follows from Assumption 3, (ii) follows from Assumption 1 and (iii) follows from Assumption 2 and Lemma 2.

□

1.2 PROOF OF LEMMA 1

To simplify the notation, we define $\psi_\theta(s) := \nabla_\theta \mu_\theta(s)$, $a_\theta = \mu_\theta(s)$ and $\nabla_a Q^{\mu_\theta}(s,a_\theta) = \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)}$ in the following proof.

We start from the form of the off-policy deterministic policy gradient given in (2), and have

$$\begin{aligned}
&\left\|\nabla J(\theta_1) - \nabla J(\theta_2)\right\| \\
&= \left\|\int_{\mathcal{S}} \psi_{\theta_1}(s) \nabla_a Q^{\mu_{\theta_1}}(s,a_{\theta_1}) \nu_{\theta_1}(ds) - \int_{\mathcal{S}} \psi_{\theta_2}(s) \nabla_a Q^{\mu_{\theta_2}}(s,a_{\theta_2}) \nu_{\theta_2}(ds)\right\| \\
&= \left\|\int_{\mathcal{S}} \psi_{\theta_1}(s) \nabla_a Q^{\mu_{\theta_1}}(s,a_{\theta_1}) \nu_{\theta_1}(ds) - \int_{\mathcal{S}} \psi_{\theta_1}(s) \nabla_a Q^{\mu_{\theta_1}}(s,a_{\theta_1}) \nu_{\theta_2}(ds)\right. \\
&\quad + \int_{\mathcal{S}} \psi_{\theta_1}(s) \nabla_a Q^{\mu_{\theta_1}}(s,a_{\theta_1}) \nu_{\theta_2}(ds) - \int_{\mathcal{S}} \psi_{\theta_1}(s) \nabla_a Q^{\mu_{\theta_2}}(s,a_{\theta_2}) \nu_{\theta_2}(ds) \\
&\quad \left. + \int_{\mathcal{S}} \psi_{\theta_1}(s) \nabla_a Q^{\mu_{\theta_2}}(s,a_{\theta_2}) \nu_{\theta_2}(ds) - \int_{\mathcal{S}} \psi_{\theta_2}(s) \nabla_a Q^{\mu_{\theta_2}}(s,a_{\theta_2}) \nu_{\theta_2}(ds)\right\| \\
&\le \left\|\int_{\mathcal{S}} \psi_{\theta_1}(s) \nabla_a Q^{\mu_{\theta_1}}(s,a_{\theta_1}) \nu_{\theta_1}(ds) - \int_{\mathcal{S}} \psi_{\theta_1}(s) \nabla_a Q^{\mu_{\theta_1}}(s,a_{\theta_1}) \nu_{\theta_2}(ds)\right\| \\
&\quad + \left\|\int_{\mathcal{S}} \psi_{\theta_1}(s) \nabla_a Q^{\mu_{\theta_1}}(s,a_{\theta_1}) \nu_{\theta_2}(ds) - \int_{\mathcal{S}} \psi_{\theta_1}(s) \nabla_a Q^{\mu_{\theta_2}}(s,a_{\theta_2}) \nu_{\theta_2}(ds)\right\| \\
&\quad + \left\|\int_{\mathcal{S}} \psi_{\theta_1}(s) \nabla_a Q^{\mu_{\theta_2}}(s,a_{\theta_2}) \nu_{\theta_2}(ds) - \int_{\mathcal{S}} \psi_{\theta_2}(s) \nabla_a Q^{\mu_{\theta_2}}(s,a_{\theta_2}) \nu_{\theta_2}(ds)\right\| \\
&\le \int_{\mathcal{S}} \left\|\psi_{\theta_1}(s)\right\| \cdot \left\|\nabla_a Q^{\mu_{\theta_1}}(s,a_{\theta_1})\right\| \left|\nu_{\theta_1}(ds) - \nu_{\theta_2}(ds)\right| \\
&\quad + \int_{\mathcal{S}} \left\|\psi_{\theta_1}(s)\right\| \cdot \left\|\nabla_a Q^{\mu_{\theta_1}}(s,a_{\theta_1}) - \nabla_a Q^{\mu_{\theta_2}}(s,a_{\theta_2})\right\| \nu_{\theta_2}(ds) \\
&\quad + \int_{\mathcal{S}} \left\|\psi_{\theta_1}(s) - \psi_{\theta_2}(s)\right\| \cdot \left\|\nabla_a Q^{\mu_{\theta_2}}(s,a_{\theta_2})\right\| \nu_{\theta_2}(ds) \\
&\overset{\text{(i)}}{\le} L_\mu C_Q \left\|\nu_{\theta_1}(\cdot) - \nu_{\theta_2}(\cdot)\right\|_{TV} + L_\mu \int_{\mathcal{S}} \left\|\nabla_a Q^{\mu_{\theta_1}}(s,a_{\theta_1}) - \nabla_a Q^{\mu_{\theta_2}}(s,a_{\theta_2})\right\| \nu_{\theta_2}(ds) \\
&\quad + C_Q \int_{\mathcal{S}} \left\|\psi_{\theta_1}(s) - \psi_{\theta_2}(s)\right\| \nu_{\theta_2}(ds) \\
&\overset{\text{(ii)}}{\le} L_\mu C_Q \left\|\nu_{\theta_1}(\cdot) - \nu_{\theta_2}(\cdot)\right\|_{TV} + L_\mu L'_Q \left\|\theta_1 - \theta_2\right\| \int_{\mathcal{S}} \nu_{\theta_2}(ds) + C_Q L_\psi \left\|\theta_1 - \theta_2\right\| \int_{\mathcal{S}} \nu_{\theta_2}(ds)
\end{aligned}$$

$$
\begin{aligned}
&\overset{\text{(iii)}}{=} L_\mu C_Q \left\|\nu_{\theta_1}(\cdot) - \nu_{\theta_2}(\cdot)\right\|_{TV} + \frac{L_\mu L'_Q}{1-\gamma}\left\|\theta_1-\theta_2\right\| + \frac{C_Q L_\psi}{1-\gamma}\left\|\theta_1-\theta_2\right\| \\
&\overset{\text{(iv)}}{\leq} \left(L_\mu C_Q L_\nu + \frac{L_\mu L'_Q}{1-\gamma} + \frac{C_Q L_\psi}{1-\gamma}\right)\left\|\theta_1-\theta_2\right\| \\
&:= L_J \left\|\theta_1-\theta_2\right\|,
\end{aligned}
$$

where (i) follows because $\|\psi_\theta(s)\| \leq L_\mu$ as indicated by Assumption 1 and $\|\nabla_a Q^{\mu_\theta}(s,a)\| \leq C_Q$ by Lemma 3, (ii) follows from Assumption 1 and Lemma 3, (iii) follows because $\int_{\mathcal{S}} \nu_\theta(ds) = \frac{1}{1-\gamma}$, and (iv) follows from Lemma 1.

2 PROOF OF THEOREM 1 AND THEOREM 1

2.1 SUPPORTING LEMMAS

In the following, we provide a few supporting lemmas that are used in the main proof of Theorem 1. The first lemma characterizes the properties of mini-batch sampling.

Lemma 4. *The following two properties hold.*

1. *Let $\hat{Y}, \bar{Y} \in \mathbb{R}^{d_1 \times d_2}$ be matrices satisfying $\left\|\hat{Y}\right\|_F \leq C_Y$, $\left\|\bar{Y}\right\|_F \leq C_Y$. If $\hat{Y}$ is an unbiased estimator of $\bar{Y}$ and $\{\hat{Y}_j\}_j$ are i.i.d. estimators, then we have*

$$
\mathbb{E}\left\|\frac{1}{M}\sum_{j=0}^{M-1}\hat{Y}_j - \bar{Y}\right\|_F^2 \leq \frac{4C_Y^2}{M}.
$$

2. *Let $\hat{y}, \bar{y} \in \mathbb{R}^d$ be vectors satisfying $\|\hat{y}\| \leq C_y$, $\|\bar{y}\| \leq C_y$. If $\hat{y}$ is an unbiased estimator of $\bar{y}$ and $\{y_j\}_j$ are i.i.d. estimators, then we have*

$$
\mathbb{E}\left\|\frac{1}{M}\sum_{j=0}^{M-1}\hat{y}_j - \bar{y}\right\|^2 \leq \frac{4C_y^2}{M}.
$$

Proof. We first prove the first statement of the matrix case as follows.

$$
\begin{aligned}
\mathbb{E}\left\|\frac{1}{M}\sum_{j=0}^{M-1}\hat{Y}_j - \bar{Y}\right\|_F^2 &= \frac{1}{M^2}\sum_{c}^{d_2}\sum_{i=0}^{M-1}\sum_{j=0}^{M-1}\mathbb{E}\langle \hat{Y}_i^c - \bar{Y}^c, Y_j^c - \bar{Y}^c\rangle \\
&= \frac{1}{M^2}\sum_{j=0}^{M-1}\mathbb{E}\left\|\hat{Y}_j - \bar{Y}\right\|_F^2 + \frac{1}{M^2}\sum_{c}^{d_2}\sum_{i\neq j}\mathbb{E}\langle \hat{Y}_i^c - \bar{Y}^c, Y_j^c - \bar{Y}^c\rangle \\
&= \frac{1}{M^2}\sum_{j=0}^{M-1}\mathbb{E}\left\|\hat{Y}_j - \bar{Y}\right\|_F^2 \\
&\leq \frac{2}{M^2}\sum_{j=0}^{M-1}\left(\mathbb{E}\left\|\hat{Y}_j\right\|_F^2 + \mathbb{E}\left\|\bar{Y}\right\|_F^2\right) \\
&\leq \frac{4C_Y^2}{M},
\end{aligned}
$$

where $\hat{Y}_j^c$ is the c-th column of $\hat{Y}_j$.

We next prove the second statement of the vector case as follows.

$$\begin{aligned}
\mathbb{E}\left\|\frac{1}{M}\sum_{j=0}^{M-1}\hat{y}_j-\bar{y}\right\|^2 &= \frac{1}{M^2}\sum_{i=0}^{M-1}\sum_{j=0}^{M-1}\mathbb{E}\langle\hat{y}_i-\bar{y},y_j-\bar{y}\rangle\\
&= \frac{1}{M^2}\sum_{j=0}^{M-1}\mathbb{E}\left\|\hat{y}_j-\bar{y}\right\|^2+\frac{1}{M^2}\sum_{i\neq j}\mathbb{E}\langle\hat{y}_i-\bar{y},y_j-\bar{y}\rangle\\
&= \frac{1}{M^2}\sum_{j=0}^{M-1}\mathbb{E}\left\|\hat{y}_j-\bar{y}\right\|^2\\
&\leq \frac{2}{M^2}\sum_{j=0}^{M-1}\left(\mathbb{E}\left\|\hat{y}_j\right\|^2+\mathbb{E}\left\|\bar{y}\right\|^2\right)\\
&\leq \frac{4C_y^2}{M}.
\end{aligned}$$

□

Next, we provide some important properties of $w^*_{\xi_\theta}$.

Lemma 5. *Let $w^*_{\xi_\theta}$ be defined in Proposition 1. Suppose Assumptions 1-3 hold. Then we have*

$$\left\|w^*_{\xi_\theta}\right\|\leq C_{w_\xi},$$

where $C_{w_\xi}=\frac{L_\mu C_Q}{\lambda_\Psi(1-\gamma)}$. Furthermore, for any θ_1,θ_2, we have

$$\left\|w^*_{\xi_{\theta_1}}-w^*_{\xi_{\theta_2}}\right\|\leq L_w\left\|\theta_1-\theta_2\right\|,$$

where $L_w=\frac{L_J}{\lambda_\Psi}+\frac{L_\mu C_Q}{\lambda^2_\Psi(1-\gamma)}\left(L^2_\mu L_\nu+\frac{2L_\mu L_\psi}{1-\gamma}\right)$.

Proof. We first show the boundedness of $\|\nabla J(\theta)\|$.

$$\begin{aligned}
\|\nabla J(\theta)\| &= \left\|\int_{\mathcal{S}}\nabla_\theta\mu_\theta(s)\nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)}\nu_\theta(ds)\right\|\\
&\leq \int_{\mathcal{S}}\|\nabla_\theta\mu_\theta(s)\|\left\|\nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)}\right\|\nu_\theta(ds)\\
&\overset{\text{(i)}}{\leq} L_\mu C_Q\int_{\mathcal{S}}\nu_\theta(ds)=\frac{L_\mu C_Q}{(1-\gamma)},
\end{aligned}$$

where (i) follows from Assumption 1 and Lemma 3.

Recall we define $\Psi_\theta=\mathbb{E}_{\nu_{\mu_\theta}}\left[\nabla_\theta\mu_\theta(s)\nabla_\theta\mu_\theta(s)^T\right]$. Assumption 1 implies that Ψ_θ is non-singular. Then by definition, we have

$$\left\|w^*_{\xi_\theta}\right\|=\left\|\Psi^{-1}_\theta\nabla J(\theta)\right\|\leq\frac{1}{\lambda_\Psi}\left\|\nabla J(\theta)\right\|\leq\frac{L_\mu C_Q}{\lambda_\Psi(1-\gamma)}.$$

Next, we show the Lipschitz continuity property.

$$
\begin{aligned}
&\left\| w^*_{\xi_{\theta_1}} - w^*_{\xi_{\theta_2}} \right\| \\
&\quad = \left\| \Psi_{\theta_1}^{-1} \nabla J(\theta_1) - \Psi_{\theta_2}^{-1} \nabla J(\theta_2) \right\| \\
&\quad = \left\| \Psi_{\theta_1}^{-1} \nabla J(\theta_1) - \Psi_{\theta_1}^{-1} \nabla J(\theta_2) + \Psi_{\theta_1}^{-1} \nabla J(\theta_2) - \Psi_{\theta_2}^{-1} \nabla J(\theta_2) \right\| \\
&\quad \leq \left\| \Psi_{\theta_1}^{-1} (\nabla J(\theta_1) - \nabla J(\theta_2)) \right\| + \left\| \left(\Psi_{\theta_1}^{-1} - \Psi_{\theta_2}^{-1} \right) \nabla J(\theta_2) \right\| \\
&\quad \overset{\text{(i)}}{\leq} \frac{L_J}{\lambda_\Psi} \|\theta_1 - \theta_2\| + \left\| \left(\Psi_{\theta_1}^{-1} - \Psi_{\theta_2}^{-1} \right) \nabla J(\theta_2) \right\| \\
&\quad = \frac{L_J}{\lambda_\Psi} \|\theta_1 - \theta_2\| + \left\| \left(\Psi_{\theta_1}^{-1} \Psi_{\theta_2} \Psi_{\theta_2}^{-1} - \Psi_{\theta_1}^{-1} \Psi_{\theta_1} \Psi_{\theta_2}^{-1} \right) \nabla J(\theta_2) \right\| \\
&\quad = \frac{L_J}{\lambda_\Psi} \|\theta_1 - \theta_2\| + \left\| \Psi_{\theta_1}^{-1} \left(\Psi_{\theta_2} - \Psi_{\theta_1} \right) \Psi_{\theta_2}^{-1} \nabla J(\theta_2) \right\| \\
&\quad \leq \frac{L_J}{\lambda_\Psi} \|\theta_1 - \theta_2\| + \frac{1}{\lambda_\Psi^2} \|\Psi_{\theta_2} - \Psi_{\theta_1}\| \|\nabla J(\theta_2)\| \\
&\quad \leq \frac{L_J}{\lambda_\Psi} \|\theta_1 - \theta_2\| + \frac{L_\mu C_Q}{\lambda_\Psi^2 (1-\gamma)} \|\Psi_{\theta_2} - \Psi_{\theta_1}\|,
\end{aligned}
$$

where (i) follows from Lemma 1 and Assumption 1.

Observe that

$$
\begin{aligned}
&\|\Psi_{\theta_2} - \Psi_{\theta_1}\| \\
&\quad = \left\| \int_{\mathcal{S}} \nabla_\theta \mu_{\theta_2}(s) \nabla_\theta \mu_{\theta_2}(s)^T \nu_{\theta_2}(ds) - \int_{\mathcal{S}} \nabla_\theta \mu_{\theta_1}(s) \nabla_\theta \mu_{\theta_1}(s)^T \nu_{\theta_1}(ds) \right\| \\
&\quad \leq \left\| \int_{\mathcal{S}} \nabla_\theta \mu_{\theta_2}(s) \nabla_\theta \mu_{\theta_2}(s)^T \nu_{\theta_2}(ds) - \int_{\mathcal{S}} \nabla_\theta \mu_{\theta_2}(s) \nabla_\theta \mu_{\theta_2}(s)^T \nu_{\theta_1}(ds) \right\| \\
&\qquad + \left\| \int_{\mathcal{S}} \nabla_\theta \mu_{\theta_2}(s) \nabla_\theta \mu_{\theta_2}(s)^T \nu_{\theta_1}(ds) - \int_{\mathcal{S}} \nabla_\theta \mu_{\theta_2}(s) \nabla_\theta \mu_{\theta_1}(s)^T \nu_{\theta_1}(ds) \right\| \\
&\qquad + \left\| \int_{\mathcal{S}} \nabla_\theta \mu_{\theta_2}(s) \nabla_\theta \mu_{\theta_1}(s)^T \nu_{\theta_1}(ds) - \int_{\mathcal{S}} \nabla_\theta \mu_{\theta_1}(s) \nabla_\theta \mu_{\theta_1}(s)^T \nu_{\theta_1}(ds) \right\| \\
&\quad \overset{\text{(i)}}{\leq} L_\mu^2 \|\nu_{\theta_1}(\cdot) - \nu_{\theta_2}(\cdot)\|_{TV} + 2L_\mu \int_{\mathcal{S}} \|\nabla_\theta \mu_{\theta_2}(s) - \nabla_\theta \mu_{\theta_1}(s)\| \nu_{\theta_1}(ds) \\
&\quad \overset{\text{(ii)}}{\leq} L_\mu^2 \|\nu_{\theta_1}(\cdot) - \nu_{\theta_2}(\cdot)\|_{TV} + \frac{2L_\mu L_\psi}{1-\gamma} \|\theta_1 - \theta_2\| \\
&\quad \overset{\text{(iii)}}{\leq} \left(L_\mu^2 L_\nu + \frac{2L_\mu L_\psi}{1-\gamma} \right) \|\theta_1 - \theta_2\|,
\end{aligned}
$$

where both (i) and (ii) follow from Assumption 1, and (iii) follows from Lemma 1.

Thus, we have

$$
\begin{aligned}
&\left\| w^*_{\xi_{\theta_1}} - w^*_{\xi_{\theta_2}} \right\| \\
&\quad \leq \frac{L_J}{\lambda_\Psi} \|\theta_1 - \theta_2\| + \frac{L_\mu C_Q}{\lambda_\Psi^2 (1-\gamma)} \|\Psi_{\theta_2} - \Psi_{\theta_1}\| \\
&\quad \leq \left[\frac{L_J}{\lambda_\Psi} + \frac{L_\mu C_Q}{\lambda_\Psi^2 (1-\gamma)} \left(L_\mu^2 L_\nu + \frac{2L_\mu L_\psi}{1-\gamma} \right) \right] \|\theta_1 - \theta_2\|.
\end{aligned}
$$

□

The next lemma provides an important bound for the difference between the gradient estimators and the true gradient.

Lemma 6. *Suppose Assumptions 1-3. Then we have*

$$\mathbb{E}\left\|h_{\theta_t}(w_t,\mathcal{B}_t)-\nabla J(\theta_t)\right\|^2 \le 3L_h^2\mathbb{E}\left\|w_t-w^*_{\theta_t}\right\|^2+3L_h^2\kappa^2+\frac{6L_\mu^4C_{w_\xi}^2}{M},$$

where $L_h=L_\mu^2$ and κ is defined in (7).

Proof. By definition, we have

$$\begin{aligned}
&\mathbb{E}\left\|h_{\theta_t}(w_t,\mathcal{B}_t)-\nabla J(\theta_t)\right\|^2\\
&\quad=\mathbb{E}\left\|h_{\theta_t}(w_t,\mathcal{B}_t)-h_{\theta_t}(w^*_{\theta_t},\mathcal{B}_t)+h_{\theta_t}(w^*_{\theta_t},\mathcal{B}_t)-h_{\theta_t}(w^*_{\xi_{\theta_t}},\mathcal{B}_t)+h_{\theta_t}(w^*_{\xi_{\theta_t}},\mathcal{B}_t)-\nabla J(\theta_t)\right\|^2\\
&\quad\le 3\mathbb{E}\left\|h_{\theta_t}(w_t,\mathcal{B}_t)-h_{\theta_t}(w^*_{\theta_t},\mathcal{B}_t)\right\|^2+3\mathbb{E}\left\|h_{\theta_t}(w^*_{\theta_t},\mathcal{B}_t)-h_{\theta_t}(w^*_{\xi_{\theta_t}},\mathcal{B}_t)\right\|^2\\
&\qquad+3\mathbb{E}\left\|h_{\theta_t}(w^*_{\xi_{\theta_t}},\mathcal{B}_t)-\nabla J(\theta_t)\right\|^2\\
&\quad\overset{\text{(i)}}{\le} 3L_h^2\mathbb{E}\left\|w_t-w^*_{\theta_t}\right\|^2+3L_h^2\mathbb{E}\left\|w^*_{\theta_t}-w^*_{\xi_{\theta_t}}\right\|^2+3\mathbb{E}\left\|h_{\theta_t}(w^*_{\xi_{\theta_t}},\mathcal{B}_t)-\nabla J(\theta_t)\right\|^2\\
&\quad\overset{\text{(ii)}}{\le} 3L_h^2\mathbb{E}\left\|w_t-w^*_{\theta_t}\right\|^2+3L_h^2\kappa^2+3\mathbb{E}\left\|h_{\theta_t}(w^*_{\xi_{\theta_t}},\mathcal{B}_t)-\nabla J(\theta_t)\right\|^2\\
&\quad\overset{\text{(iii)}}{\le} 3L_h^2\mathbb{E}\left\|w_t-w^*_{\theta_t}\right\|^2+3L_h^2\kappa^2+\frac{6L_\mu^4C_{w_\xi}^2}{M},
\end{aligned}$$

where (i) follows because for any $w_1,w_2,\theta\in\mathbb{R}^d$, we have

$$\begin{aligned}
\left\|h_\theta(w_1,\mathcal{B}_t)-h_\theta(w_2,\mathcal{B}_t)\right\|&=\left\|\frac{1}{M}\sum_{j=0}^{M-1}\nabla_\theta\mu_{\theta_t}(s'_{t,j})\nabla_\theta\mu_{\theta_t}(s'_{t,j})^T(w_1-w_2)\right\|\\
&\le L_\mu^2\left\|w_1-w_2\right\|:=L_h\left\|w_1-w_2\right\|,
\end{aligned}$$

(ii) follows from (7), and (iii) holds due to the fact that

$$\begin{aligned}
&\mathbb{E}\left\|h_{\theta_t}(w^*_{\xi_{\theta_t}},\mathcal{B}_t)-\nabla J(\theta_t)\right\|^2\\
&\quad=\mathbb{E}\left\|\frac{1}{M}\sum_{j=0}^{M-1}\nabla_\theta\mu_{\theta_t}(s'_{t,j})\nabla_\theta\mu_{\theta_t}(s'_{t,j})^Tw^*_{\xi_{\theta_t}}-\nabla J(\theta_t)\right\|^2\\
&\quad=\frac{1}{M^2}\sum_{i=0}^{M-1}\sum_{j=0}^{M-1}\mathbb{E}\langle\nabla_\theta\mu_{\theta_t}(s'_{t,i})\nabla_\theta\mu_{\theta_t}(s'_{t,i})^Tw^*_{\xi_{\theta_t}}-\nabla J(\theta_t),\\
&\qquad\nabla_\theta\mu_{\theta_t}(s'_{t,j})\nabla_\theta\mu_{\theta_t}(s'_{t,j})^Tw^*_{\xi_{\theta_t}}-\nabla J(\theta_t)\rangle\\
&\quad=\frac{1}{M^2}\sum_{j=0}^{M-1}\mathbb{E}\left\|\nabla_\theta\mu_{\theta_t}(s'_{t,j})\nabla_\theta\mu_{\theta_t}(s'_{t,j})^Tw^*_{\xi_{\theta_t}}-\nabla J(\theta_t)\right\|^2\\
&\quad\overset{\text{(i)}}{\le}\frac{1}{M^2}\sum_{j=0}^{M-1}2L_\mu^4C_{w_\xi}^2=\frac{2L_\mu^4C_{w_\xi}^2}{M},
\end{aligned}$$

where (i) follows from Assumption 1, Lemma 4 and Lemma 5. □

2.2 PROOF OF THEOREM 1

We use the following notations for the clarity of the presentation:

$$
\begin{aligned}
g_{\theta_t}(w_t, \mathcal{B}_t) &= \frac{1}{M}\sum_{j=0}^{M-1} \delta_{t,j}\phi(x_{t,j}) = \frac{1}{M}\sum_{j=0}^{M-1}(A_{t,j}w_t + b_{t,j}) := \hat{A}_t w_t + \hat{b}_t;\\
\bar{g}_{\theta_t}(w_t) &= \mathbb{E}_{d_{\theta_t}}\left[\delta_t\phi(x_t)\right] = \bar{A}_t w_t + \bar{b}_t;\\
\bar{g}_{\theta_t}(w^*_{\theta_t}) &= \bar{A}_t w^*_{\theta_t} + \bar{b}_t = 0;\\
h_{\theta_t}(w_t, \mathcal{B}_t) &= \frac{1}{M}\sum_{j=0}^{M-1} \nabla_\theta \mu_{\theta_t}(s'_{t,j})\nabla_\theta \mu_{\theta_t}(s'_{t,j})^T w_t.
\end{aligned}
$$

In this proof, we develop a new approach to analyzing the coupled actor and critic's stochastic approximation processes, due to their simultaneous updates both with constant stepsizes. The central idea is to cancel the critic's cumulative tracking error by the actor's overall positive progress to the stationary policy, which is different from the existing analysis of (stochastic) PG-type algorithms that mainly decouples or asymptotically decouples the critic's error from actor's error. Further, we develop a new analysis to bound the estimation error of the Fisher information of deterministic policy arising via the compatibility theorem, and then further capture how such a metric affects the convergence via its minimum eigenvalue.

The main proof consists of three steps.

Step I: Characterizing dynamics of critic's error via coupling with actor.

In the first step, we characterize the propagation of the dynamics of critic's dynamic tracking error based on its coupling with actor's updates. That is, we develop the relationship between $\left\|w_{t+1} - w^*_{\theta_{t+1}}\right\|^2$ and $\left\|w_t - w^*_{\theta_t}\right\|^2$ by their coupling with actor's updates.

We first use the dynamics of the critic to obtain

$$
\begin{aligned}
&\left\|w_{t+1} - w^*_{\theta_t}\right\|^2\\
&= \left\|w_t + \alpha_w g_{\theta_t}(w_t, \mathcal{B}_t) - w^*_{\theta_t}\right\|^2\\
&= \left\|w_t - w^*_{\theta_t}\right\|^2 + 2\alpha_w\langle w_t - w^*_{\theta_t}, g_{\theta_t}(w_t, \mathcal{B}_t)\rangle + \alpha_w^2\left\|g_{\theta_t}(w_t, \mathcal{B}_t)\right\|^2\\
&= \left\|w_t - w^*_{\theta_t}\right\|^2 + 2\alpha_w\langle w_t - w^*_{\theta_t}, \bar{g}_{\theta_t}(w_t)\rangle + 2\alpha_w\langle w_t - w^*_{\theta_t}, g_{\theta_t}(w_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(w_t)\rangle\\
&\quad + \alpha_w^2\left\|g_{\theta_t}(w_t, \mathcal{B}_t)\right\|^2\\
&= \left\|w_t - w^*_{\theta_t}\right\|^2 + 2\alpha_w(w_t - w^*_{\theta_t})^T\bar{A}_t(w_t - w^*_{\theta_t}) + 2\alpha_w\langle w_t - w^*_{\theta_t}, g_{\theta_t}(w_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(w_t)\rangle\\
&\quad + \alpha_w^2\left\|g_{\theta_t}(w_t, \mathcal{B}_t)\right\|^2\\
&\overset{\text{(i)}}{\le} (1 - 2\alpha_w\lambda)\left\|w_t - w^*_{\theta_t}\right\|^2 + 2\alpha_w\langle w_t - w^*_{\theta_t}, g_{\theta_t}(w_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(w_t)\rangle + \alpha_w^2\left\|g_{\theta_t}(w_t, \mathcal{B}_t)\right\|^2\\
&\le (1 - 2\alpha_w\lambda)\left\|w_t - w^*_{\theta_t}\right\|^2 + 2\alpha_w\langle w_t - w^*_{\theta_t}, g_{\theta_t}(w_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(w_t)\rangle\\
&\quad + 2\alpha_w^2\left\|g_{\theta_t}(w_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(w_t)\right\|^2 + 2\alpha_w^2\left\|\bar{g}_{\theta_t}(w_t)\right\|^2\\
&= (1 - 2\alpha_w\lambda)\left\|w_t - w^*_{\theta_t}\right\|^2 + 2\alpha_w\langle w_t - w^*_{\theta_t}, g_{\theta_t}(w_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(w_t)\rangle\\
&\quad + 2\alpha_w^2\left\|g_{\theta_t}(w_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(w_t)\right\|^2 + 2\alpha_w^2\left\|\bar{g}_{\theta_t}(w_t) - \bar{g}_{\theta_t}(w^*_{\theta_t})\right\|^2\\
&= (1 - 2\alpha_w\lambda)\left\|w_t - w^*_{\theta_t}\right\|^2 + 2\alpha_w\langle w_t - w^*_{\theta_t}, g_{\theta_t}(w_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(w_t)\rangle\\
&\quad + 2\alpha_w^2\left\|g_{\theta_t}(w_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(w_t)\right\|^2 + 2\alpha_w^2\left\|\bar{A}_t(w_t - w^*_{\theta_t})\right\|^2\\
&\le (1 - 2\alpha_w\lambda)\left\|w_t - w^*_{\theta_t}\right\|^2 + 2\alpha_w\langle w_t - w^*_{\theta_t}, g_{\theta_t}(w_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(w_t)\rangle\\
&\quad + 2\alpha_w^2\left\|g_{\theta_t}(w_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(w_t)\right\|^2 + 2\alpha_w^2\left\|\bar{A}_t\right\|^2\left\|(w_t - w^*_{\theta_t})\right\|^2\\
&\overset{\text{(ii)}}{\le} (1 - 2\alpha_w\lambda + 2\alpha_w^2 C_A^2)\left\|w_t - w^*_{\theta_t}\right\|^2 + 2\alpha_w\langle w_t - w^*_{\theta_t}, g_{\theta_t}(w_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(w_t)\rangle\\
&\quad + 2\alpha_w^2\left\|g_{\theta_t}(w_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(w_t)\right\|^2,
\end{aligned}
$$

where (i) follows from the property $(w_t - w^*_{\theta_t})^T \bar{A}_t (w_t - w^*_{\theta_t}) \le -\lambda \left\| w_t - w^*_{\theta_t} \right\|^2$ with some constant $\lambda > 0$ for any policy, which has been proved in Tsitsiklis and Van Roy [1997], Bhandari et al. [2018], Tu and Recht [2019], Xiong et al. [2020], and (ii) follows because $\|A\|^2 \le 2(1+\gamma^2)C_\phi^4 \le 4C_\phi^4 := C_A^2$.

Taking the expectation on both sides yields

$$
\begin{aligned}
&\mathbb{E}\left\| w_{t+1} - w^*_{\theta_t} \right\|^2 \\
&\le (1 - 2\alpha_w\lambda + 2\alpha_w^2 C_A^2)\mathbb{E}\left\| w_t - w^*_{\theta_t} \right\|^2 + 2\alpha_w \mathbb{E}\langle w_t - w^*_{\theta_t}, g_{\theta_t}(w_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(w_t)\rangle \\
&\quad + 2\alpha_w^2 \mathbb{E}\left\| g_{\theta_t}(w_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(w_t) \right\|^2 \\
&= (1 - 2\alpha_w\lambda + 2\alpha_w^2 C_A^2)\mathbb{E}\left\| w_t - w^*_{\theta_t} \right\|^2 + 2\alpha_w^2 \mathbb{E}\left\| g_{\theta_t}(w_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(w_t) \right\|^2. \qquad (2)
\end{aligned}
$$

Observe that

$$
\begin{aligned}
&\mathbb{E}\left\| g_{\theta_t}(w_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(w_t) \right\|^2 \\
&= \mathbb{E}\left\| \hat{A}_t w_t + \hat{b}_t - \bar{A}_t w_t - \bar{b}_t \right\|^2 \\
&\overset{\text{(i)}}{\le} 3\mathbb{E}\left\| (\hat{A}_t - \bar{A}_t)(w_t - w^*_{\theta_t}) \right\|^2 + 3\mathbb{E}\left\| (\hat{A}_t - \bar{A}_t) w^*_{\theta_t} \right\|^2 + 3\mathbb{E}\left\| \hat{b}_t - \bar{b}_t \right\|^2 \\
&\le 3\mathbb{E}\left\| \hat{A}_t - \bar{A}_t \right\|_F^2 \left\| w_t - w^*_{\theta_t} \right\|^2 + 3\mathbb{E}\left\| \hat{A}_t - \bar{A}_t \right\|_F^2 \left\| w^*_{\theta_t} \right\|^2 + 3\mathbb{E}\left\| \hat{b}_t - \bar{b}_t \right\|^2 \\
&\overset{\text{(ii)}}{\le} \frac{12C_A^2}{M}\mathbb{E}\left\| w_t - w^*_{\theta_t} \right\|^2 + \frac{12(C_A^2 \mathbb{E}\left\| w^*_{\theta_t} \right\|^2 + C_b^2)}{M} \\
&\overset{\text{(iii)}}{\le} \frac{12C_A^2}{M}\mathbb{E}\left\| w_t - w^*_{\theta_t} \right\|^2 + \frac{12(C_A^2 C_w^2 + C_b^2)}{M},
\end{aligned}
$$

where (i) follows because $(x+y+z)^2 \le 3x^2 + 3y^2 + 3z^2$, (ii) follows from Lemma 4 and $C_b := R_{\max} C_\phi \ge \|b\|$, and (iii) follows because $\left\| w^*_{\theta_t} \right\|^2 = \left\| \bar{A}_t^{-1} \bar{b}_t \right\|^2 \le C_b / \lambda_A = R_{\max} C_\phi / \lambda_A := C_w$ by Assumption 4.

Substituting the above bound into (2), we have

$$
\begin{aligned}
&\mathbb{E}\left\| w_{t+1} - w^*_{\theta_t} \right\|^2 \\
&\le (1 - 2\alpha_w\lambda + 2\alpha_w^2 C_A^2)\mathbb{E}\left\| w_t - w^*_{\theta_t} \right\|^2 + 2\alpha_w^2 \mathbb{E}\left\| g_{\theta_t}(w_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(w_t) \right\|^2 \\
&\le \left(1 - 2\alpha_w\lambda + 2\alpha_w^2 C_A^2 + \frac{24\alpha_w^2 C_A^2}{M}\right)\mathbb{E}\left\| w_t - w^*_{\theta_t} \right\|^2 + \frac{24\alpha_w^2(C_A^2 C_w^2 + C_b^2)}{M}.
\end{aligned}
$$

Since $\alpha_w \le \frac{\lambda}{2C_A^2}; M \ge \frac{48\alpha_w C_A^2}{\lambda}$, we further obtain

$$
\begin{aligned}
&\mathbb{E}\left\| w_{t+1} - w^*_{\theta_t} \right\|^2 \\
&\le \left(1 - 2\alpha_w\lambda + 2\alpha_w^2 C_A^2 + \frac{24\alpha_w^2 C_A^2}{M}\right)\mathbb{E}\left\| w_t - w^*_{\theta_t} \right\|^2 + \frac{24\alpha_w^2(C_A^2 C_w^2 + C_b^2)}{M} \\
&\le \left(1 - \frac{\alpha_w\lambda}{2}\right)\mathbb{E}\left\| w_t - w^*_{\theta_t} \right\|^2 + \frac{24\alpha_w^2(C_A^2 C_w^2 + C_b^2)}{M}. \qquad (3)
\end{aligned}
$$

Next, we use Young's inequality, and obtain

$$
\begin{aligned}
&\mathbb{E}\left\| w_{t+1} - w^*_{\theta_{t+1}} \right\|^2 \\
&\le \left(1 + \frac{1}{2(2/\lambda\alpha_w - 1)}\right)\mathbb{E}\left\| w_{t+1} - w^*_{\theta_t} \right\|^2 + (1 + 2(2/\lambda\alpha_w - 1))\,\mathbb{E}\left\| w^*_{\theta_t} - w^*_{\theta_{t+1}} \right\|^2 \\
&\overset{\text{(i)}}{\le} \left(1 - \frac{\lambda\alpha_w}{4}\right)\mathbb{E}\left\| w_t - w^*_{\theta_t} \right\|^2 + \frac{4 - \lambda\alpha_w}{4 - 2\lambda\alpha_w}\cdot\frac{24\alpha_w^2(C_A^2 C_w^2 + C_b^2)}{M} + \frac{4}{\lambda\alpha_w}\mathbb{E}\left\| w^*_{\theta_t} - w^*_{\theta_{t+1}} \right\|^2 \\
&\overset{\text{(ii)}}{\le} \left(1 - \frac{\lambda\alpha_w}{4}\right)\mathbb{E}\left\| w_t - w^*_{\theta_t} \right\|^2 + \frac{4 - \lambda\alpha_w}{4 - 2\lambda\alpha_w}\cdot\frac{24\alpha_w^2(C_A^2 C_w^2 + C_b^2)}{M} + \frac{4L_w^2}{\lambda\alpha_w}\mathbb{E}\left\| \theta_{t+1} - \theta_t \right\|^2, \qquad (4)
\end{aligned}
$$

where (i) follows from the bound derived in (3), and (ii) follows from Lemma 5.

Step II: Bounding cumulative tracking error via compatibility theorem for DPG.

In this step, we bound the cumulative tracking error based on the dynamics of the tracking error from the last step. To this end, we need to first bound the difference between two consecutive actor parameters.

Observe that $\theta_{t+1}-\theta_t = \frac{1}{M}\sum_{j=0}^{M-1}\nabla_\theta\mu_{\theta_t}(s'_{t,j})\nabla_\theta\mu_{\theta_t}(s'_{t,j})^T w_t := h_{\theta_t}(w_t,\mathcal{B}_t)$ and $\mathbb{E}\left\|h_{\theta_t}(w_t,\mathcal{B}_t)\right\|^2 \le 2\mathbb{E}\left\|\nabla J(\theta_t)\right\|^2 + 2\mathbb{E}\left\|h_{\theta_t}(w_t,\mathcal{B}_t)-\nabla J(\theta_t)\right\|^2$. We proceed to bound (4) as follows

$$\begin{aligned}
&\mathbb{E}\left\|w_{t+1}-w^*_{\theta_{t+1}}\right\|^2\\
&\le \left(1-\frac{\lambda\alpha_w}{4}\right)\mathbb{E}\left\|w_t-w^*_{\theta_t}\right\|^2 + \frac{4-\lambda\alpha_w}{4-2\lambda\alpha_w}\cdot\frac{24\alpha_w^2(C_A^2C_w^2+C_b^2)}{M} + \frac{4L_w^2}{\lambda\alpha_w}\mathbb{E}\left\|\theta_{t+1}-\theta_t\right\|^2\\
&\le \left(1-\frac{\lambda\alpha_w}{4}\right)\mathbb{E}\left\|w_t-w^*_{\theta_t}\right\|^2 + \frac{48\alpha_w^2(C_A^2C_w^2+C_b^2)}{M} + \frac{8L_w^2\alpha_\theta^2}{\lambda\alpha_w}\mathbb{E}\left\|\nabla J(\theta_t)\right\|^2\\
&\quad + \frac{8L_w^2\alpha_\theta^2}{\lambda\alpha_w}\mathbb{E}\left\|h_{\theta_t}(w_t,\mathcal{B}_t)-\nabla J(\theta_t)\right\|^2\\
&\overset{\text{(i)}}{\le} \left(1-\frac{\lambda\alpha_w}{4}+\frac{24L_h^2L_w^2\alpha_\theta^2}{\lambda\alpha_w}\right)\mathbb{E}\left\|w_t-w^*_{\theta_t}\right\|^2 + \frac{48\alpha_w^2(C_A^2C_w^2+C_b^2)}{M} + \frac{8L_w^2\alpha_\theta^2}{\lambda\alpha_w}\mathbb{E}\left\|\nabla J(\theta_t)\right\|^2\\
&\quad + \frac{8L_w^2\alpha_\theta^2}{\lambda\alpha_w}\left(3L_h^2\kappa^2+\frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\\
&\overset{\text{(ii)}}{\le} \left(1-\frac{\lambda\alpha_w}{8}\right)\mathbb{E}\left\|w_t-w^*_{\theta_t}\right\|^2 + \frac{8L_w^2\alpha_\theta^2}{\lambda\alpha_w}\mathbb{E}\left\|\nabla J(\theta_t)\right\|^2 + \frac{48\alpha_w^2(C_A^2C_w^2+C_b^2)}{M}\\
&\quad + \frac{8L_w^2\alpha_\theta^2}{\lambda\alpha_w}\left(3L_h^2\kappa^2+\frac{6L_\mu^4C_{w_\xi}^2}{M}\right),
\end{aligned}\tag{5}$$

where (i) follows from Lemma 6, and (ii) follows because $\alpha_\theta \le \frac{\lambda\alpha_w}{\sqrt{96}L_hL_w}$.

We further take the summation over all iterations on both sides of (5) and have

$$\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}\left\|w_t-w^*_{\theta_t}\right\|^2\\
&\le \sum_{t=0}^{T-1}\left(1-\frac{\lambda\alpha_w}{8}\right)^t\left\|w_0-w^*_{\theta_0}\right\|^2 + \frac{8L_w^2\alpha_\theta^2}{\lambda\alpha_w}\sum_{t=0}^{T-1}\sum_{i=0}^{t-1}\left(1-\frac{\lambda\alpha_w}{8}\right)^{t-1-i}\mathbb{E}\left\|\nabla J(\theta_t)\right\|^2\\
&\quad + \left[\frac{48\alpha_w^2(C_A^2C_w^2+C_b^2)}{M} + \frac{8L_w^2\alpha_\theta^2}{\lambda\alpha_w}\left(3L_h^2\kappa^2+\frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\right]\sum_{t=0}^{T-1}\sum_{i=0}^{t-1}\left(1-\frac{\lambda\alpha_w}{8}\right)^{t-1-i}\\
&\le \frac{8\left\|w_0-w^*_{\theta_0}\right\|^2}{\lambda\alpha_w} + \left[\frac{48\alpha_w^2(C_A^2C_w^2+C_b^2)}{M} + \frac{8L_w^2\alpha_\theta^2}{\lambda\alpha_w}\left(3L_h^2\kappa^2+\frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\right]\cdot\frac{8T}{\lambda\alpha_w}\\
&\quad + \frac{64L_w^2\alpha_\theta^2}{\lambda^2\alpha_w^2}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla J(\theta_t)\right\|^2.
\end{aligned}\tag{6}$$

Step III: Overall convergence by canceling tracking error via actor's positive progress.

In this step, we establish the overall convergence to a stationary policy by novel cancellation of the above cumulative tracking error via actor's update progress.

Based on Lemma 1, we have

$$\begin{aligned}
&\mathbb{E}[J(\theta_{t+1})]-\mathbb{E}[J(\theta_t)]\\
&\ge \mathbb{E}\langle\nabla J(\theta_t),\theta_{t+1}-\theta_t\rangle - \frac{L_J}{2}\mathbb{E}\left\|\theta_{t+1}-\theta_t\right\|^2
\end{aligned}$$

$$
\begin{aligned}
&= \alpha_\theta \mathbb{E}\langle \nabla J(\theta_t), h_{\theta_t}(w_t, \mathcal{B}_t)\rangle - \frac{L_J \alpha_\theta^2}{2} \mathbb{E}\left\| h_{\theta_t}(w_t, \mathcal{B}_t)\right\|^2 \\
&= \alpha_\theta \mathbb{E}\left\| \nabla J(\theta_t)\right\|^2 + \alpha_\theta \mathbb{E}\langle \nabla J(\theta_t), h_{\theta_t}(w_t, \mathcal{B}_t) - \nabla J(\theta_t)\rangle - \frac{L_J \alpha_\theta^2}{2} \mathbb{E}\left\| h_{\theta_t}(w_t, \mathcal{B}_t)\right\|^2 \\
&\overset{\text{(i)}}{\geq} \frac{\alpha_\theta}{2} \mathbb{E}\left\| \nabla J(\theta_t)\right\|^2 - \frac{\alpha_\theta}{2} \mathbb{E}\left\| h_{\theta_t}(w_t, \mathcal{B}_t) - \nabla J(\theta_t)\right\|^2 \\
&\quad - \frac{L_J \alpha_\theta^2}{2} \mathbb{E}\left\| h_{\theta_t}(w_t, \mathcal{B}_t) - \nabla J(\theta_t) + \nabla J(\theta_t)\right\|^2 \\
&\geq \left(\frac{\alpha_\theta}{2} - L_J \alpha_\theta^2\right) \mathbb{E}\left\| \nabla J(\theta_t)\right\|^2 - \left(\frac{\alpha_\theta}{2} + L_J \alpha_\theta^2\right) \mathbb{E}\left\| h_{\theta_t}(w_t, \mathcal{B}_t) - \nabla J(\theta_t)\right\|^2 \\
&\overset{\text{(ii)}}{\geq} \left(\frac{\alpha_\theta}{2} - L_J \alpha_\theta^2\right) \mathbb{E}\left\| \nabla J(\theta_t)\right\|^2 - \left(\frac{\alpha_\theta}{2} + L_J \alpha_\theta^2\right) \left(3L_h^2 \mathbb{E}\left\| w_t - w^*_{\theta_t}\right\|^2 + 3L_h^2 \kappa^2 + \frac{6L_\mu^4 C_{w_\xi}^2}{M}\right) \\
&\overset{\text{(iii)}}{\geq} \frac{\alpha_\theta}{4} \mathbb{E}\left\| \nabla J(\theta_t)\right\|^2 - \frac{3\alpha_\theta}{4} \left(3L_h^2 \mathbb{E}\left\| w_t - w^*_{\theta_t}\right\|^2 + 3L_h^2 \kappa^2 + \frac{6L_\mu^4 C_{w_\xi}^2}{M}\right),
\end{aligned}
\tag{7}
$$

where (i) follows because $x^T y \geq -\frac{1}{2}x^2 - \frac{1}{2}y^2$, (ii) follows from Lemma 6, and (iii) follows from the condition $\alpha_\theta \leq \frac{1}{4L_J}$. We next take the summation over all iterations on both sides of the above bound and obtain

$$
\begin{aligned}
&\frac{\alpha_\theta}{4} \sum_{t=0}^{T-1} \mathbb{E}\left\| \nabla J(\theta_t)\right\|^2 \\
&\quad \leq \mathbb{E}[J(\theta_{T+1})] - \mathbb{E}[J(\theta_0)] + \frac{3\alpha_\theta}{4} \left(3L_h^2 \kappa^2 + \frac{6L_\mu^4 C_{w_\xi}^2}{M}\right) \cdot T + \frac{9\alpha_\theta L_h^2}{4} \sum_{t=0}^{T-1} \mathbb{E}\left\| w_t - w^*_{\theta_t}\right\|^2 \\
&\quad \leq \frac{R_{\max}}{1-\gamma} + \frac{3\alpha_\theta}{4} \left(3L_h^2 \kappa^2 + \frac{6L_\mu^4 C_{w_\xi}^2}{M}\right) \cdot T + \frac{9\alpha_\theta L_h^2}{4} \sum_{t=0}^{T-1} \mathbb{E}\left\| w_t - w^*_{\theta_t}\right\|^2 .
\end{aligned}
\tag{8}
$$

Substituting the cumulative tracking error bound derived in (6) into (8) yields

$$
\begin{aligned}
&\frac{\alpha_\theta}{8} \sum_{t=0}^{T-1} \mathbb{E}\left\| \nabla J(\theta_t)\right\|^2 \\
&\quad \overset{\text{(i)}}{\leq} \left(\frac{\alpha_\theta}{4} - \frac{144 L_h^2 L_w^2 \alpha_\theta^3}{\lambda^2 \alpha_w^2}\right) \sum_{t=0}^{T-1} \mathbb{E}\left\| \nabla J(\theta_t)\right\|^2 \\
&\quad \leq \frac{R_{\max}}{1-\gamma} + \frac{3\alpha_\theta}{4} \left(3L_h^2 \kappa^2 + \frac{6L_\mu^4 C_{w_\xi}^2}{M}\right) \cdot T + \frac{18\alpha_\theta L_h^2}{\lambda \alpha_w} \left\| w_0 - w^*_{\theta_0}\right\|^2 \\
&\qquad + \left[\frac{48\alpha_w^2 (C_A^2 C_w^2 + C_b^2)}{M} + \frac{8L_w^2 \alpha_\theta^2}{\lambda \alpha_w} \left(3L_h^2 \kappa^2 + \frac{6L_\mu^4 C_{w_\xi}^2}{M}\right)\right] \cdot \frac{18\alpha_\theta L_h^2 T}{\lambda \alpha_w},
\end{aligned}
$$

where (i) follows from the condition $\alpha_\theta \leq \frac{\lambda \alpha_w}{24 L_h L_w}$.

Finally, we have

$$
\begin{aligned}
\min_{t\in[T]} \mathbb{E}\left\| \nabla J(\theta_t)\right\|^2 &\leq \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left\| \nabla J(\theta_t)\right\|^2 \\
&\leq \left(\frac{8R_{\max}}{\alpha_\theta(1-\gamma)} + \frac{144L_h^2}{\lambda \alpha_w} \left\| w_0 - w^*_{\theta_0}\right\|^2\right) \cdot \frac{1}{T} + 6\left(3L_h^2 \kappa^2 + \frac{6L_\mu^4 C_{w_\xi}^2}{M}\right) \\
&\quad + \left[\frac{48\alpha_w^2 (C_A^2 C_w^2 + C_b^2)}{M} + \frac{8L_w^2 \alpha_\theta^2}{\lambda \alpha_w} \left(3L_h^2 \kappa^2 + \frac{6L_\mu^4 C_{w_\xi}^2}{M}\right)\right] \cdot \frac{144L_h^2}{\lambda \alpha_w} \\
&= \frac{c_1}{T} + \frac{c_2}{M} + c_3 \kappa^2,
\end{aligned}
$$

where

$$c_1 = \frac{8R_{\max}}{\alpha_\theta(1-\gamma)} + \frac{144L_h^2}{\lambda\alpha_w}\left\|w_0 - w^*_{\theta_0}\right\|^2, \tag{9}$$

$$c_2 = 36L_\mu^4 C_{w_\xi}^2 + \left[48\alpha_w^2(C_A^2 C_w^2 + C_b^2) + \frac{48L_w^2 L_\mu^4 C_{w_\xi}^2 \alpha_\theta^2}{\lambda\alpha_w}\right]\cdot\frac{144L_h^2}{\lambda\alpha_w}, \tag{10}$$

$$c_3 = 18L_h^2 + \frac{24L_w^2 L_h^2 \alpha_\theta^2}{\lambda\alpha_w}. \tag{11}$$

2.3 PROOF OF COROLLARY 1

Following from the upper bound in Theorem 1, we let $\frac{c_1}{T} \le \frac{\epsilon}{2}$ and $\frac{c_2}{M} \le \frac{\epsilon}{2}$ to achieve the ϵ-accuracy. Then we obtain $T \ge \frac{2c_1}{\epsilon}$ and $M \ge \frac{2c_2}{\epsilon}$. Further, since we generate M samples in the update steps of both critic and actor in Algorithm 1, the total number of samples we use is thus $2MT = \frac{8c_1c_2}{\epsilon^2}$.

3 PROOF OF LEMMA 2

We use the notations $\psi_\theta(s) := \nabla_\theta\mu_\theta(s)$, $a_\theta = \mu_\theta(s)$ and $\nabla_a Q^{\mu_\theta}(s, a_\theta) = \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)}$ in the following proof.

We start from the form of the deterministic policy gradient given in (4), and have

$$\begin{aligned}
&\|\nabla J_\beta(\theta_1) - \nabla J_\beta(\theta_2)\| \\
&= \left\|\int_{\mathcal{S}} \psi_{\theta_1}(s)\nabla_a Q^{\mu_{\theta_1}}(s,a_{\theta_1})\nu_\beta(ds) - \int_{\mathcal{S}} \psi_{\theta_2}(s)\nabla_a Q^{\mu_{\theta_2}}(s,a_{\theta_2})\nu_\beta(ds)\right\| \\
&= \left\|\int_{\mathcal{S}} \psi_{\theta_1}(s)\nabla_a Q^{\mu_{\theta_1}}(s,a_{\theta_1})\nu_\beta(ds) - \int_{\mathcal{S}} \psi_{\theta_1}(s)\nabla_a Q^{\mu_{\theta_2}}(s,a_{\theta_2})\nu_\beta(ds)\right. \\
&\quad \left. + \int_{\mathcal{S}} \psi_{\theta_1}(s)\nabla_a Q^{\mu_{\theta_2}}(s,a_{\theta_2})\nu_\beta(ds) - \int_{\mathcal{S}} \psi_{\theta_2}(s)\nabla_a Q^{\mu_{\theta_2}}(s,a_{\theta_2})\nu_\beta(ds)\right\| \\
&\le \left\|\int_{\mathcal{S}} \psi_{\theta_1}(s)\nabla_a Q^{\mu_{\theta_1}}(s,a_{\theta_1})\nu_\beta(ds) - \int_{\mathcal{S}} \psi_{\theta_1}(s)\nabla_a Q^{\mu_{\theta_2}}(s,a_{\theta_2})\nu_\beta(ds)\right\| \\
&\quad + \left\|\int_{\mathcal{S}} \psi_{\theta_1}(s)\nabla_a Q^{\mu_{\theta_2}}(s,a_{\theta_2})\nu_\beta(ds) - \int_{\mathcal{S}} \psi_{\theta_2}(s)\nabla_a Q^{\mu_{\theta_2}}(s,a_{\theta_2})\nu_\beta(ds)\right\| \\
&\le \int_{\mathcal{S}} \|\psi_{\theta_1}(s)\|\cdot\|\nabla_a Q^{\mu_{\theta_1}}(s,a_{\theta_1}) - \nabla_a Q^{\mu_{\theta_2}}(s,a_{\theta_2})\|\,\nu_\beta(ds) \\
&\quad + \int_{\mathcal{S}} \|\psi_{\theta_1}(s) - \psi_{\theta_2}(s)\|\cdot\|\nabla_a Q^{\mu_{\theta_2}}(s,a_{\theta_2})\|\,\nu_\beta(ds) \\
&\overset{\text{(i)}}{\le} L_\mu \int_{\mathcal{S}} \|\nabla_a Q^{\mu_{\theta_1}}(s,a_{\theta_1}) - \nabla_a Q^{\mu_{\theta_2}}(s,a_{\theta_2})\|\,\nu_\beta(ds) + C_Q\int_{\mathcal{S}} \|\psi_{\theta_1}(s) - \psi_{\theta_2}(s)\|\,\nu_\beta(ds) \\
&\overset{\text{(ii)}}{\le} L_\mu L'_Q\|\theta_1-\theta_2\|\int_{\mathcal{S}}\nu_\beta(ds) + C_Q L_\psi\|\theta_1-\theta_2\|\int_{\mathcal{S}}\nu_\beta(ds) \\
&\overset{\text{(iii)}}{=} \left(\frac{L_\mu L'_Q}{1-\gamma} + \frac{C_Q L_\psi}{1-\gamma}\right)\|\theta_1-\theta_2\| \\
&:= L_{J_\beta}\|\theta_1-\theta_2\|,
\end{aligned}$$

where (i) follows because $\|\psi_\theta(s)\| \le L_\mu$ as indicated by Assumption 1 and $\|\nabla_a Q^{\mu_\theta}(s,a)\| \le C_Q$ by Lemma 3, (ii) follows from Assumption 1 and Lemma 3, and (iii) follows because $\int_{\mathcal{S}}\nu_\beta(ds) = \frac{1}{1-\gamma}$.

4 PROOF OF THEOREM 2 AND COROLLARY 3

4.1 SUPPORTING LEMMAS

The following lemma provides the important properties of w^*_{β,ξ_θ}.

Lemma 7. *Let w^*_{β,ξ_θ} be defined in* (5). *Suppose Assumptions 1-3 hold. Then we have*

$$\left\| w^*_{\beta,\xi_\theta} \right\| \leq C_{w_\xi},$$

where $C_{w_\xi} = \frac{L_\mu C_Q}{\lambda_\Psi(1-\gamma)}$. Furthermore, for any θ_1, θ_2, we have

$$\left\| w^*_{\beta,\xi_{\theta_1}} - w^*_{\beta,\xi_{\theta_2}} \right\| \leq L_{w'} \left\| \theta_1 - \theta_2 \right\|,$$

where $L_{w'} = \frac{L_{J_\beta}}{\lambda_\Psi} + \frac{2L_\mu^2 L_\psi C_Q}{\lambda_\Psi^2(1-\gamma)^2}$.

Proof. We first show the boundedness of $\|\nabla J_\beta(\theta)\|$.

$$\begin{aligned}
\|\nabla J_\beta(\theta)\| &= \left\| \int_{\mathcal{S}} \nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} \nu_\beta(ds) \right\| \\
&\leq \int_{\mathcal{S}} \|\nabla_\theta \mu_\theta(s)\| \left\| \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} \right\| \nu_\beta(ds) \\
&\overset{\text{(i)}}{\leq} L_\mu C_Q \int_{\mathcal{S}} \nu_\beta(ds) = \frac{L_\mu C_Q}{(1-\gamma)},
\end{aligned}$$

where (i) follows from Assumption 1 and Lemma 3.

We define $\Psi_{\beta,\theta} = \mathbb{E}_{\nu_{\mu_\beta}}\left[\nabla_\theta \mu_\theta(s) \nabla_\theta \mu_\theta(s)^T\right]$. Assumption 1 implies that $\Psi_{\beta,\theta}$ is non-singular. Then by definition, we have

$$\left\| w^*_{\beta,\xi_\theta} \right\| = \left\| \Psi^{-1}_{\beta,\theta} \nabla J_\beta(\theta) \right\| \leq \frac{1}{\lambda_\Psi} \|\nabla J_\beta(\theta)\| \leq \frac{L_\mu C_Q}{\lambda_\Psi(1-\gamma)}.$$

Next, we show the Lipschitz continuity property.

$$\begin{aligned}
&\left\| w^*_{\xi_{\theta_1}} - w^*_{\xi_{\theta_2}} \right\| \\
&= \left\| \Psi^{-1}_{\beta,\theta_1} \nabla J_\beta(\theta_1) - \Psi^{-1}_{\beta,\theta_2} \nabla J_\beta(\theta_2) \right\| \\
&= \left\| \Psi^{-1}_{\beta,\theta_1} \nabla J_\beta(\theta_1) - \Psi^{-1}_{\beta,\theta_1} \nabla J_\beta(\theta_2) + \Psi^{-1}_{\beta,\theta_1} \nabla J_\beta(\theta_2) - \Psi^{-1}_{\beta,\theta_2} \nabla J_\beta(\theta_2) \right\| \\
&\leq \left\| \Psi^{-1}_{\beta,\theta_1} (\nabla J_\beta(\theta_1) - \nabla J_\beta(\theta_2)) \right\| + \left\| \left(\Psi^{-1}_{\beta,\theta_1} - \Psi^{-1}_{\beta,\theta_2} \right) \nabla J_\beta(\theta_2) \right\| \\
&\overset{\text{(i)}}{\leq} \frac{L_J}{\lambda_\Psi} \|\theta_1 - \theta_2\| + \left\| \left(\Psi^{-1}_{\beta,\theta_1} - \Psi^{-1}_{\beta,\theta_2} \right) \nabla J_\beta(\theta_2) \right\| \\
&= \frac{L_J}{\lambda_\Psi} \|\theta_1 - \theta_2\| + \left\| \left(\Psi^{-1}_{\beta,\theta_1} \Psi_{\beta,\theta_2} \Psi^{-1}_{\beta,\theta_2} - \Psi^{-1}_{\beta,\theta_1} \Psi_{\beta,\theta_1} \Psi^{-1}_{\beta,\theta_2} \right) \nabla J_\beta(\theta_2) \right\| \\
&= \frac{L_J}{\lambda_\Psi} \|\theta_1 - \theta_2\| + \left\| \Psi^{-1}_{\beta,\theta_1} \left(\Psi_{\beta,\theta_2} - \Psi_{\beta,\theta_1} \right) \Psi^{-1}_{\beta,\theta_2} \nabla J_\beta(\theta_2) \right\| \\
&\leq \frac{L_J}{\lambda_\Psi} \|\theta_1 - \theta_2\| + \frac{1}{\lambda_\Psi^2} \|\Psi_{\beta,\theta_2} - \Psi_{\beta,\theta_1}\| \|\nabla J_\beta(\theta_2)\| \\
&\leq \frac{L_J}{\lambda_\Psi} \|\theta_1 - \theta_2\| + \frac{L_\mu C_Q}{\lambda_\Psi^2(1-\gamma)} \|\Psi_{\beta,\theta_2} - \Psi_{\beta,\theta_1}\|,
\end{aligned}$$

where (i) follows from Lemma 1 and Assumption 1.

We further derive the following bound.

$$\begin{aligned}
&\|\Psi_{\beta,\theta_2} - \Psi_{\beta,\theta_1}\| \\
&= \left\| \int_{\mathcal{S}} \nabla_\theta \mu_{\theta_2}(s) \nabla_\theta \mu_{\theta_2}(s)^T \nu_\beta(ds) - \int_{\mathcal{S}} \nabla_\theta \mu_{\theta_1}(s) \nabla_\theta \mu_{\theta_1}(s)^T \nu_\beta(ds) \right\| \\
&\leq \left\| \int_{\mathcal{S}} \nabla_\theta \mu_{\theta_2}(s) \nabla_\theta \mu_{\theta_2}(s)^T \nu_\beta(ds) - \int_{\mathcal{S}} \nabla_\theta \mu_{\theta_2}(s) \nabla_\theta \mu_{\theta_1}(s)^T \nu_\beta(ds) \right\| \\
&\quad + \left\| \int_{\mathcal{S}} \nabla_\theta \mu_{\theta_2}(s) \nabla_\theta \mu_{\theta_1}(s)^T \nu_\beta(ds) - \int_{\mathcal{S}} \nabla_\theta \mu_{\theta_1}(s) \nabla_\theta \mu_{\theta_1}(s)^T \nu_\beta(ds) \right\| \\
&\overset{\text{(i)}}{\leq} 2L_\mu \int_{\mathcal{S}} \|\nabla_\theta \mu_{\theta_2}(s) - \nabla_\theta \mu_{\theta_1}(s)\| \nu_\beta(ds) \\
&\overset{\text{(ii)}}{\leq} \frac{2L_\mu L_\psi}{1-\gamma} \|\theta_1 - \theta_2\|,
\end{aligned}$$

where both (i) and (ii) follow from Assumption 1.

Thus, we have

$$\begin{aligned}
&\left\| w^*_{\beta,\xi_{\theta_1}} - w^*_{\beta,\xi_{\theta_2}} \right\| \\
&\leq \frac{L_{J_\beta}}{\lambda_\Psi} \|\theta_1 - \theta_2\| + \frac{L_\mu C_Q}{\lambda_\Psi^2(1-\gamma)} \|\Psi_{\beta,\theta_2} - \Psi_{\beta,\theta_1}\| \\
&\leq \left(\frac{L_{J_\beta}}{\lambda_\Psi} + \frac{2L_\mu^2 L_\psi C_Q}{\lambda_\Psi^2 (1-\gamma)^2} \right) \|\theta_1 - \theta_2\|.
\end{aligned}$$

□

4.2 PROOF OF THEOREM 2

The main difference here from the proof of Theorem 1 lies in the fact that we apply TDC to update critic in Algorithm 2 due to the off-policy sampling, which introduces an extra correction parameter u_t. Thus, we introduce a grouped vector $z_t = [w_t^T u_t^T]^T \in \mathbb{R}^{2d}$ and rewrite the dynamics of critic as a lifted linear system:

$$\begin{aligned}
z_{t+1} &= z_t + \alpha_w \begin{bmatrix} \hat{A}_t & \hat{C}_t \\ \eta \hat{A}_t & \eta \hat{D}_t \end{bmatrix} z_t + \alpha_w \begin{bmatrix} \hat{b}_t \\ \eta \hat{b}_t \end{bmatrix} \\
&:= z_t + \alpha_w \left[\hat{G}_t z_t + \hat{\ell}_t \right] \\
&:= z_t + \alpha_w g_{\theta_t}(z_t, \mathcal{B}_t).
\end{aligned}$$

TDC algorithm is designed to find the fixed point $w^*_{\beta,\theta}$ satisfying $\bar{A} w^*_{\beta,\theta} + \bar{b} = 0$, where $\bar{A} = \mathbb{E}_{d_\beta}\left[\hat{A}\right], \bar{b} = \mathbb{E}_{d_\beta}\left[\hat{b}\right]$. Correspondingly, if we let $z^*_\theta = [{w^*_{\beta,\theta}}^T \mathbf{0}^T]^T$, then we have

$$\bar{g}_\theta(z^*_\theta) = \bar{G} z^*_\theta + \bar{\ell} = 0,$$

where $\bar{G} = \mathbb{E}_{d_\beta}\left[\hat{G}\right], \bar{\ell} = \mathbb{E}_{d_\beta}\left[\hat{\ell}\right]$. Based on the above lifted linear system, we proceed our proof as follows.

The main proof consists of three steps.

Step I: Characterizing dynamics of critic's error via coupling with actor.

In the following, we first characterize the relationship between $\left\| z_{t+1} - z^*_{\theta_{t+1}} \right\|^2$ and $\left\| z_t - z^*_{\theta_t} \right\|^2$.

We first use the dynamics of the above linear system to obtain

$$
\begin{aligned}
&\left\|z_{t+1} - z^*_{\theta_t}\right\|^2 \\
&= \left\|z_t + \alpha_w g_{\theta_t}(z_t, \mathcal{B}_t) - z^*_{\theta_t}\right\|^2 \\
&= \left\|z_t - z^*_{\theta_t}\right\|^2 + 2\alpha_w \langle z_t - z^*_{\theta_t}, g_{\theta_t}(z_t, \mathcal{B}_t)\rangle + \alpha_w^2 \left\|g_{\theta_t}(z_t, \mathcal{B}_t)\right\|^2 \\
&= \left\|z_t - z^*_{\theta_t}\right\|^2 + 2\alpha_w \langle z_t - z^*_{\theta_t}, \bar{g}_{\theta_t}(z_t)\rangle + 2\alpha_w \langle z_t - z^*_{\theta_t}, g_{\theta_t}(z_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(z_t)\rangle \\
&\quad + \alpha_w^2 \left\|g_{\theta_t}(z_t, \mathcal{B}_t)\right\|^2 \\
&\overset{\text{(i)}}{\le} (1 - 2\alpha_w \lambda') \left\|z_t - z^*_{\theta_t}\right\|^2 + 2\alpha_w \langle z_t - z^*_{\theta_t}, g_{\theta_t}(z_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(z_t)\rangle + \alpha_w^2 \left\|g_{\theta_t}(z_t, \mathcal{B}_t)\right\|^2 \\
&\le (1 - 2\alpha_w \lambda') \left\|z_t - z^*_{\theta_t}\right\|^2 + 2\alpha_w \langle z_t - z^*_{\theta_t}, g_{\theta_t}(z_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(z_t)\rangle \\
&\quad + 2\alpha_w^2 \left\|g_{\theta_t}(z_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(z_t)\right\|^2 + 2\alpha_w^2 \left\|\bar{g}_{\theta_t}(z_t)\right\|^2 \\
&= (1 - 2\alpha_w \lambda') \left\|z_t - z^*_{\theta_t}\right\|^2 + 2\alpha_w \langle z_t - z^*_{\theta_t}, g_{\theta_t}(z_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(z_t)\rangle \\
&\quad + 2\alpha_w^2 \left\|g_{\theta_t}(z_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(z_t)\right\|^2 + 2\alpha_w^2 \left\|\bar{g}_{\theta_t}(z_t) - \bar{g}_{\theta_t}(z^*_{\theta_t})\right\|^2 \\
&= (1 - 2\alpha_w \lambda') \left\|z_t - z^*_{\theta_t}\right\|^2 + 2\alpha_w \langle z_t - z^*_{\theta_t}, g_{\theta_t}(z_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(z_t)\rangle \\
&\quad + 2\alpha_w^2 \left\|g_{\theta_t}(z_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(z_t)\right\|^2 + 2\alpha_w^2 \left\|\bar{G}_t(z_t - z^*_{\theta_t})\right\|^2 \\
&\le (1 - 2\alpha_w \lambda') \left\|z_t - z^*_{\theta_t}\right\|^2 + 2\alpha_w \langle z_t - z^*_{\theta_t}, g_{\theta_t}(z_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(z_t)\rangle \\
&\quad + 2\alpha_w^2 \left\|g_{\theta_t}(z_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(z_t)\right\|^2 + 2\alpha_w^2 \left\|\bar{G}_t\right\|_F^2 \left\|(z_t - z^*_{\theta_t})\right\|^2 \\
&\overset{\text{(ii)}}{\le} (1 - 2\alpha_w \lambda' + 2\alpha_w^2 C_G^2) \left\|z_t - z^*_{\theta_t}\right\|^2 + 2\alpha_w \langle z_t - z^*_{\theta_t}, g_{\theta_t}(z_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(z_t)\rangle \\
&\quad + 2\alpha_w^2 \left\|g_{\theta_t}(z_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(z_t)\right\|^2,
\end{aligned}
$$

where (i) follows from the property $\langle z_t - z^*_{\theta_t}, \bar{g}_{\theta_t}(z_t)\rangle \le -\lambda' \left\|z_t - z^*_{\theta_t}\right\|^2$ with some constant $\lambda' > 0$ for any policy which has been proved in Theorem 3 of **?** as long as $\eta > \max\left\{0, \sigma_{\min}\left(D^{-1} \cdot \frac{A+A^T}{2}\right)\right\}$, and (ii) follows from $\left\|\bar{G}\right\|_F^2 = (1+\eta^2)\left\|\bar{A}\right\|_F^2 + \left\|\bar{C}\right\|_F^2 + \eta^2 \left\|\bar{D}\right\|_F^2 \le (1+\eta^2)C_A^2 + C_C^2 + \eta^2 C_D^2 \le 5(1+\eta^2)C_\phi^4 := C_G^2$.
Taking the expectation on both sides of the above bound yields

$$
\begin{aligned}
&\mathbb{E}\left\|z_{t+1} - z^*_{\theta_t}\right\|^2 \\
&\le (1 - 2\alpha_w \lambda' + 2\alpha_w^2 C_G^2)\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2 + 2\alpha_w \mathbb{E}\langle z_t - z^*_{\theta_t}, g_{\theta_t}(z_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(z_t)\rangle \\
&\quad + 2\alpha_w^2 \mathbb{E}\left\|g_{\theta_t}(z_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(z_t)\right\|^2 \\
&= (1 - 2\alpha_w \lambda' + 2\alpha_w^2 C_G^2)\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2 + 2\alpha_w^2 \mathbb{E}\left\|g_{\theta_t}(z_t, \mathcal{B}_t) - \bar{g}_{\theta_t}(z_t)\right\|^2 \\
&= (1 - 2\alpha_w \lambda' + 2\alpha_w^2 C_G^2)\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2 + 2\alpha_w^2 \mathbb{E}\left\|\hat{G}_t z_t + \hat{\ell}_t - \bar{G}_t z_t - \bar{\ell}_t\right\|^2 \\
&\le (1 - 2\alpha_w \lambda' + 2\alpha_w^2 C_G^2)\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2 \\
&\quad + 6\alpha_w^2 \left(\mathbb{E}\left\|(\hat{G}_t - \bar{G}_t)(z_t - z^*_{\theta_t})\right\|^2 + \mathbb{E}\left\|(\hat{G}_t - \bar{G}_t) z^*_{\theta_t}\right\|^2 + \mathbb{E}\left\|\hat{\ell}_t - \bar{\ell}_t\right\|^2\right) \\
&\le (1 - 2\alpha_w \lambda' + 2\alpha_w^2 C_G^2)\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2 \\
&\quad + 6\alpha_w^2 \left(\mathbb{E}\left\|\hat{G}_t - \bar{G}_t\right\|_F^2 \left\|z_t - z^*_{\theta_t}\right\|^2 + \mathbb{E}\left\|\hat{G}_t - \bar{G}_t\right\|_F^2 \left\|z^*_{\theta_t}\right\|^2 + \mathbb{E}\left\|\hat{\ell}_t - \bar{\ell}_t\right\|^2\right) \\
&\overset{\text{(i)}}{\le} (1 - 2\alpha_w \lambda' + 2\alpha_w^2 C_G^2)\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2 + 6\alpha_w^2 \left(\frac{4C_G^2}{M}\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2 + \frac{4(C_G^2 \mathbb{E}\left\|z^*_{\theta_t}\right\|^2 + C_\ell^2)}{M}\right) \\
&\overset{\text{(ii)}}{=} \left(1 - 2\alpha_w \lambda' + 2\alpha_w^2 C_G^2 + \frac{24\alpha_w^2 C_G^2}{M}\right)\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2 + \frac{24\alpha_w^2 \left(C_G^2 \mathbb{E}\left\|w^*_{\beta,\theta_t}\right\|^2 + C_\ell^2\right)}{M}
\end{aligned}
$$

$$
\begin{aligned}
&\overset{\text{(iii)}}{\leq} \left(1 - 2\alpha_w\lambda' + 2\alpha_w^2 C_G^2 + \frac{24\alpha_w^2 C_G^2}{M}\right)\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2 + \frac{24\alpha_w^2(C_G^2 C_w^2 + C_\ell^2)}{M} \\
&\overset{\text{(iv)}}{\leq} \left(1 - \frac{\alpha_w\lambda'}{2}\right)\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2 + \frac{24\alpha_w^2(C_G^2 C_w^2 + C_\ell^2)}{M},
\end{aligned} \tag{12}
$$

where (i) follows from Lemma 4, (ii) follows from $\left\|z^*_{\theta_t}\right\|^2 = \left\|w^*_{\beta,\theta_t}\right\|^2$, (iii) follows because $\left\|w^*_{\beta,\theta_t}\right\|^2 = \left\|\bar{A}_t^{-1}\bar{b}_t\right\|^2 \leq C_b/\lambda_A = R_{\max}C_\phi/\lambda_A := C_w$ by Assumption 4, and (iv) follows from the conditions $\alpha_w \leq \frac{\lambda'}{2C_G^2}$ and $M \geq \frac{48\alpha_w C_G^2}{\lambda'}$. We further derive that

$$
\begin{aligned}
&\mathbb{E}\left\|z_{t+1} - z^*_{\theta_{t+1}}\right\|^2 \\
&\quad\overset{\text{(i)}}{\leq} \left(1 + \frac{1}{2(2/\lambda'\alpha_w - 1)}\right)\mathbb{E}\left\|z_{t+1} - z^*_{\theta_t}\right\|^2 + (1 + 2(2/\lambda'\alpha_w - 1))\,\mathbb{E}\left\|z^*_{\theta_t} - z^*_{\theta_{t+1}}\right\|^2 \\
&\quad\overset{\text{(ii)}}{\leq} \left(1 - \frac{\lambda'\alpha_w}{4}\right)\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2 + \frac{48\alpha_w^2(C_G^2 C_w^2 + C_\ell^2)}{M} + \frac{4}{\lambda'\alpha_w}\mathbb{E}\left\|z^*_{\theta_t} - z^*_{\theta_{t+1}}\right\|^2 \\
&\quad\overset{\text{(iii)}}{=} \left(1 - \frac{\lambda'\alpha_w}{4}\right)\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2 + \frac{48\alpha_w^2(C_G^2 C_w^2 + C_\ell^2)}{M} + \frac{4}{\lambda'\alpha_w}\mathbb{E}\left\|w^*_{\beta,\theta_t} - w^*_{\beta,\theta_{t+1}}\right\|^2 \\
&\quad\overset{\text{(iv)}}{\leq} \left(1 - \frac{\lambda'\alpha_w}{4}\right)\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2 + \frac{48\alpha_w^2(C_G^2 C_w^2 + C_\ell^2)}{M} + \frac{4L_{w'}^2}{\lambda'\alpha_w}\mathbb{E}\left\|\theta_{t+1} - \theta_t\right\|^2,
\end{aligned} \tag{13}
$$

where (i) follows from Young's inequality, (ii) follows from the bound derived in (12), (iii) follows because $\left\|z^*_{\theta_t} - z^*_{\theta_{t+1}}\right\|^2 = \left\|w^*_{\beta,\theta_t} - w^*_{\beta,\theta_{t+1}}\right\|^2$, and (iv) follows from Lemma 7.

Step II: Bounding cumulative tracking error via compatibility theorem for DPG.

Recall that we define $h_{\theta_t}(w_t, \mathcal{B}_t) := \frac{1}{M}\sum_{j=0}^{M-1}\nabla_\theta\mu_{\theta_t}(s'_{t,j})\nabla_\theta\mu_{\theta_t}(s'_{t,j})^T w_t$. We continue with (13) and have

$$
\begin{aligned}
&\mathbb{E}\left\|z_{t+1} - z^*_{\theta_{t+1}}\right\|^2 \\
&\quad\leq \left(1 - \frac{\lambda'\alpha_w}{4}\right)\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2 + \frac{48\alpha_w^2(C_G^2 C_w^2 + C_\ell^2)}{M} + \frac{4L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\mathbb{E}\left\|h_{\theta_t}(w_t, \mathcal{B}_t)\right\|^2 \\
&\quad\leq \left(1 - \frac{\lambda'\alpha_w}{4}\right)\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2 + \frac{48\alpha_w^2(C_G^2 C_w^2 + C_\ell^2)}{M} + \frac{8L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\mathbb{E}\left\|\nabla J_\beta(\theta_t)\right\|^2 \\
&\qquad + \frac{8L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\mathbb{E}\left\|h_{\theta_t}(w_t, \mathcal{B}_t) - \nabla J_\beta(\theta_t)\right\|^2 \\
&\quad\overset{\text{(i)}}{\leq} \left(1 - \frac{\lambda'\alpha_w}{4}\right)\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2 + \frac{48\alpha_w^2(C_G^2 C_w^2 + C_\ell^2)}{M} + \frac{8L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\mathbb{E}\left\|\nabla J_\beta(\theta_t)\right\|^2 \\
&\qquad + \frac{8L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\left(3L_h^2\mathbb{E}\left\|w_t - w^*_{\theta_t}\right\|^2 + 3L_h^2\kappa^2 + \frac{6L_\mu^4 C_{w_\xi}^2}{M}\right) \\
&\quad\overset{\text{(ii)}}{\leq} \left(1 - \frac{\lambda'\alpha_w}{4}\right)\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2 + \frac{48\alpha_w^2(C_G^2 C_w^2 + C_\ell^2)}{M} + \frac{8L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\mathbb{E}\left\|\nabla J_\beta(\theta_t)\right\|^2 \\
&\qquad + \frac{8L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\left(3L_h^2\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2 + 3L_h^2\kappa^2 + \frac{6L_\mu^4 C_{w_\xi}^2}{M}\right) \\
&\quad= \left(1 - \frac{\lambda'\alpha_w}{4} + \frac{24L_h^2 L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\right)\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2 + \frac{48\alpha_w^2(C_G^2 C_w^2 + C_\ell^2)}{M} \\
&\qquad + \frac{8L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\mathbb{E}\left\|\nabla J_\beta(\theta_t)\right\|^2 + \frac{8L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\left(3L_h^2\kappa^2 + \frac{6L_\mu^4 C_{w_\xi}^2}{M}\right) \\
&\quad\overset{\text{(iii)}}{\leq} \left(1 - \frac{\lambda'\alpha_w}{8}\right)\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2 + \frac{48\alpha_w^2(C_G^2 C_w^2 + C_\ell^2)}{M} + \frac{8L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\mathbb{E}\left\|\nabla J_\beta(\theta_t)\right\|^2
\end{aligned}
$$

$$+ \frac{8L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\left(3L_h^2\kappa^2 + \frac{6L_\mu^4 C_{w_\xi}^2}{M}\right), \tag{14}$$

where (i) follows from Lemma 6, (ii) follows since $\left\|w_t - w^*_{\theta_t}\right\|^2 \le \left\|z_t - z^*_{\theta_t}\right\|^2$, and (iii) follows because $\alpha_\theta \le \frac{\lambda'\alpha_w}{\sqrt{96}L_hL_{w'}}$.

Then, taking the summation over all iterations on both sides of (14) yields

$$\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2\\
&\le \sum_{t=0}^{T-1}\left(1-\frac{\lambda'\alpha_w}{8}\right)^t\left\|z_0 - z^*_{\theta_0}\right\|^2 + \frac{8L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\sum_{t=0}^{T-1}\sum_{i=0}^{t-1}\left(1-\frac{\lambda'\alpha_w}{8}\right)^{t-1-i}\mathbb{E}\left\|\nabla J_\beta(\theta_t)\right\|^2\\
&\quad + \left[\frac{48\alpha_w^2(C_G^2C_w^2 + C_\ell^2)}{M} + \frac{8L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\left(3L_h^2\kappa^2 + \frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\right]\sum_{t=0}^{T-1}\sum_{i=0}^{t-1}\left(1-\frac{\lambda'\alpha_w}{8}\right)^{t-1-i}\\
&\le \frac{8\left\|z_0 - z^*_{\theta_0}\right\|^2}{\lambda'\alpha_w} + \left[\frac{48\alpha_w^2(C_G^2C_w^2 + C_\ell^2)}{M} + \frac{8L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\left(3L_h^2\kappa^2 + \frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\right]\cdot\frac{8T}{\lambda'\alpha_w}\\
&\quad + \frac{64L_{w'}^2\alpha_\theta^2}{\lambda'^2\alpha_w^2}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla J_\beta(\theta_t)\right\|^2.
\end{aligned}\tag{15}$$

Step III: Overall convergence by canceling tracking error via actor's positive progress.

Similarly to the on-policy case, we use the Lipschitz continuity property to obtain (see (7))

$$\begin{aligned}
&\mathbb{E}[J_\beta(\theta_{t+1})] - \mathbb{E}[J_\beta(\theta_t)]\\
&\quad\ge \frac{\alpha_\theta}{4}\mathbb{E}\left\|\nabla J_\beta(\theta_t)\right\|^2 - \frac{3\alpha_\theta}{4}\left(3L_h^2\mathbb{E}\left\|w_t - w^*_{\beta,\theta_t}\right\|^2 + 3L_h^2\kappa^2 + \frac{6L_\mu^4C_{w_\xi}^2}{M}\right),
\end{aligned}$$

where we use the condition $\alpha_\theta \le \frac{1}{4L_{J_\beta}}$.

Further, we take the summation over all iterations on both sides of the above bound and have

$$\begin{aligned}
&\frac{\alpha_\theta}{4}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla J_\beta(\theta_t)\right\|^2\\
&\le \mathbb{E}[J_\beta(\theta_{T+1})] - \mathbb{E}[J_\beta(\theta_0)] + \frac{3\alpha_\theta}{4}\left(3L_h^2\kappa^2 + \frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\cdot T + \frac{9\alpha_\theta L_h^2}{4}\sum_{t=0}^{T-1}\mathbb{E}\left\|w_t - w^*_{\beta,\theta_t}\right\|^2\\
&\le \frac{R_{\max}}{1-\gamma} + \frac{3\alpha_\theta}{4}\left(3L_h^2\kappa^2 + \frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\cdot T + \frac{9\alpha_\theta L_h^2}{4}\sum_{t=0}^{T-1}\mathbb{E}\left\|w_t - w^*_{\beta,\theta_t}\right\|^2\\
&\le \frac{R_{\max}}{1-\gamma} + \frac{3\alpha_\theta}{4}\left(3L_h^2\kappa^2 + \frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\cdot T + \frac{9\alpha_\theta L_h^2}{4}\sum_{t=0}^{T-1}\mathbb{E}\left\|z_t - z^*_{\theta_t}\right\|^2.
\end{aligned}$$

Then, we substitute the cumulative error from (15) into the above bound and have

$$\begin{aligned}
&\frac{\alpha_\theta}{8}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla J_\beta(\theta_t)\right\|^2\\
&\overset{\text{(i)}}{\le} \left(\frac{\alpha_\theta}{4} - \frac{144L_h^2L_{w'}^2\alpha_\theta^3}{\lambda'^2\alpha_w^2}\right)\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla J_\beta(\theta_t)\right\|^2\\
&\le \frac{R_{\max}}{1-\gamma} + \frac{3\alpha_\theta}{4}\left(3L_h^2\kappa^2 + \frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\cdot T + \frac{18\alpha_\theta L_h^2}{\lambda'\alpha_w}\left\|z_0 - z^*_{\theta_0}\right\|^2
\end{aligned}$$

$$+\left[\frac{48\alpha_w^2(C_G^2C_w^2+C_\ell^2)}{M}+\frac{8L_{w'}^2\alpha_\theta^2}{\lambda'\alpha_w}\left(3L_h^2\kappa^2+\frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\right]\cdot\frac{18\alpha_\theta L_h^2T}{\lambda'\alpha_w},$$

where (i) follows from the condition $\alpha_\theta \le \frac{\lambda'\alpha_w}{24L_hL_{w'}}$.

Finally, we obtain

$$\begin{aligned}
\min_{t\in[T]}\mathbb{E}\left\|\nabla J_\beta(\theta_t)\right\|^2 &\le \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla J_\beta(\theta_t)\right\|^2\\
&\le \left(\frac{8R_{\max}}{\alpha_\theta(1-\gamma)}+\frac{144L_h^2}{\lambda\alpha_w}\left\|z_0-z_{\theta_0}^*\right\|^2\right)\cdot\frac{1}{T}+6\left(3L_h^2\kappa^2+\frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\\
&\quad+\left[\frac{48\alpha_w^2(C_G^2C_w^2+C_\ell^2)}{M}+\frac{8L_{w'}^2\alpha_\theta^2}{\lambda\alpha_w}\left(3L_h^2\kappa^2+\frac{6L_\mu^4C_{w_\xi}^2}{M}\right)\right]\cdot\frac{144L_h^2}{\lambda\alpha_w}\\
&=\frac{c_4}{T}+\frac{c_5}{M}+c_6\kappa^2,
\end{aligned}$$

where

$$c_4=\frac{8R_{\max}}{\alpha_\theta(1-\gamma)}+\frac{144L_h^2}{\lambda'\alpha_w}\left\|z_0-z_{\theta_0}^*\right\|^2, \tag{16}$$

$$c_5=36L_\mu^4C_{w_\xi}^2+\left[48\alpha_w^2(C_G^2C_w^2+C_\ell^2)+\frac{48L_{w'}^2L_\mu^4C_{w_\xi}^2\alpha_\theta^2}{\lambda'\alpha_w}\right]\cdot\frac{144L_h^2}{\lambda'\alpha_w}, \tag{17}$$

$$c_6=18L_h^2+\frac{24L_{w'}^2L_h^2\alpha_\theta^2}{\lambda'\alpha_w}. \tag{18}$$

4.3 PROOF OF COROLLARY 3

Following from the upper bound in Theorem 2, we let $\frac{c_4}{T}\le\frac{\epsilon}{2}$ and $\frac{c_5}{M}\le\frac{\epsilon}{2}$ to achieve the target ϵ-accuracy. Then we obtain $T\ge\frac{2c_4}{\epsilon}$ and $M\ge\frac{2c_5}{\epsilon}$. Further, since we generate M samples in the update steps of both critic and actor in Algorithm 2, the total number of samples we use is thus $2MT=\frac{8c_4c_5}{\epsilon^2}$.