Supplementary Materials

A Useful Lemmas for Proving Theorem 1

In this subsection, we prove some useful Lemmas for our finite-sample analysis.

Before we start, we first introduce some nations. In the following proof, $\|a\|$ denotes the ℓ_2 norm if a is a vector; and $\|A\|$ denotes the operator norm if A is a matrix. Let λ be the smallest eigenvalue of the matrix C. Then the operator norm of C^{-1} is $\frac{1}{\lambda}$. We note that the Greedy-GQ algorithm in Algorithm 1 was shown to converge asymptotically, and θ_t and ω_t were shown to be bounded a.s. (see Proposition 4 in [18]). We then define R as the upper bound on both θ_t and ω_t. Specifically, for any t, $\|\theta_t\| \leq R$ and $\|\omega_t\| \leq R$ a.s..

We first prove that if the policy π_θ is smooth in θ, then the object function $J(\theta)$ is also smooth.

Lemma 2. *The objective function $J(\theta)$ is K-smooth for $\theta \in \{\theta : \|\theta\| \leq R\}$, i.e., for any $\|\theta_1\|, \|\theta_2\| \leq R$,*

$$\|\nabla J(\theta_1) - \nabla J(\theta_2)\| \leq K||\theta_1 - \theta_2||, \tag{38}$$

where $K = 2\gamma\frac{1}{\lambda}\left((k_1|\mathcal{A}|R+1)(1+\gamma+\gamma Rk_1|\mathcal{A}|) + |\mathcal{A}|(r_{\max}+R+\gamma R)(2k_1+k_2R)\right)$.

Proof. Recall the expression of $J\left(\theta\right)$:

$$J\left(\theta\right) = \mathbb{E}_\mu\left[\delta_{S,A,S'}\left(\theta\right)\phi_{S,A}\right]^\top C^{-1}\mathbb{E}_\mu\left[\delta_{S,A,S'}\left(\theta\right)\phi_{S,A}\right], \tag{39}$$

where $\delta_{S,A,S'} = r_{S,A,S'} + \gamma\sum_{a\in\mathcal{A}}\pi_\theta\left(a|S'\right)\theta^\top\phi_{S',a} - \theta^\top\phi_{S,A}$. Then,

$$\nabla J\left(\theta\right) = 2\nabla\left(\mathbb{E}_\mu\left[\delta_{S,A,S'}\left(\theta\right)\phi_{S,A}\right]\right)C^{-1}\mathbb{E}_\mu\left[\delta_{S,A,S'}\left(\theta\right)\phi_{S,A}\right], \tag{40}$$

where

$$\begin{aligned}\nabla\left(\mathbb{E}_\mu\left[\delta_{S,A,S'}\left(\theta\right)\phi_{S,A}\right]\right) &= \mathbb{E}_\mu\left[\left(\nabla\gamma\sum_{a\in\mathcal{A}}\pi_\theta\left(a|S'\right)\theta^\top\phi_{S',a}\right)\phi_{S,A}^\top\right]\\ &= \gamma\mathbb{E}_\mu\left[\left(\sum_{a\in\mathcal{A}}\nabla\left(\pi_\theta\left(a|S'\right)\right)\theta^\top\phi_{S',a} + \pi_\theta\left(a|S'\right)\phi_{S',a}\right)\phi_{S,A}^\top\right].\end{aligned} \tag{41}$$

It then follows that

$$\begin{aligned}&\nabla J\left(\theta_1\right) - \nabla J\left(\theta_2\right)\\ &= 2\nabla\left(\mathbb{E}_\mu\left[\delta_{S,A,S'}\left(\theta_1\right)\phi_{S,A}\right]\right)C^{-1}\mathbb{E}_\mu\left[\delta_{S,A,S'}\left(\theta_1\right)\phi_{S,A}\right] - 2\nabla\left(\mathbb{E}_\mu\left[\delta_{S,A,S'}\left(\theta_2\right)\phi_{S,A}\right]\right)C^{-1}\mathbb{E}_\mu\left[\delta_{S,A,S'}\left(\theta_2\right)\phi_{S,A}\right]\\ &= 2\nabla\left(\mathbb{E}_\mu\left[\delta_{S,A,S'}\left(\theta_1\right)\phi_{S,A}\right]\right)C^{-1}\mathbb{E}_\mu\left[\delta_{S,A,S'}\left(\theta_1\right)\phi_{S,A}\right] - 2\nabla\left(\mathbb{E}_\mu\left[\delta_{S,A,S'}\left(\theta_1\right)\phi_{S,A}\right]\right)C^{-1}\mathbb{E}_\mu\left[\delta_{S,A,S'}\left(\theta_2\right)\phi_{S,A}\right]\\ &+ 2\nabla\left(\mathbb{E}_\mu\left[\delta_{S,A,S'}\left(\theta_1\right)\phi_{S,A}\right]\right)C^{-1}\mathbb{E}_\mu\left[\delta_{S,A,S'}\left(\theta_2\right)\phi_{S,A}\right] - 2\nabla\left(\mathbb{E}_\mu\left[\delta_{S,A,S'}\left(\theta_2\right)\phi_{S,A}\right]\right)C^{-1}\mathbb{E}_\mu\left[\delta_{S,A,S'}\left(\theta_2\right)\phi_{S,A}\right].\end{aligned} \tag{42}$$

Since C^{-1} is positive definite, thus to show $\nabla J(\theta)$ is Lipschitz, it suffices to show both $\nabla\left(\mathbb{E}_\mu\left[\delta_{S,A,S'}\left(\theta\right)\phi_{S,A}\right]\right)$ and $\mathbb{E}_\mu\left[\delta_{S,A,S'}\left(\theta\right)\phi_{S,A}\right]$ are Lipschitz in θ and bounded.

We first show that

$$||\mathbb{E}_\mu\left[\delta_{S,A,S'}\left(\theta\right)\phi_{S,A}\right]|| \leq r_{\max} + (1+\gamma)R, \tag{43}$$

and

$$||\nabla\mathbb{E}_\mu\left[\delta_{S,A,S'}\left(\theta\right)\phi_{S,A}\right]|| = ||\mathbb{E}_\mu\left[\nabla\delta_{S,A,S'}\left(\theta\right)\phi_{S,A}\right]|| \leq \gamma(k_1|\mathcal{A}|R+1). \tag{44}$$

Following from (41), we then have that

$$\begin{aligned}
&\nabla\left(\mathbb{E}_{\mu}\left[\delta_{S,A,S'}\left(\theta_{1}\right)\phi_{S,A}\right]\right)-\nabla\left(\mathbb{E}_{\mu}\left[\delta_{S,A,S'}\left(\theta_{2}\right)\phi_{S,A}\right]\right)\\
&=\gamma\mathbb{E}_{\mu}\left[\left(\sum_{a\in\mathcal{A}}\nabla\left(\pi_{\theta_{1}}\left(a|S'\right)\right)\theta_{1}^{\top}\phi_{S',a}-\nabla\left(\pi_{\theta_{2}}\left(a|S'\right)\right)\theta_{2}^{\top}\phi_{S',a}+\pi_{\theta_{1}}\left(a|S'\right)\phi_{S',a}-\pi_{\theta_{2}}\left(a|S'\right)\phi_{S',a}\right)\phi_{S,A}^{\top}\right]\\
&=\gamma\mathbb{E}_{\mu}\Bigg[\Bigg(\sum_{a\in\mathcal{A}}\nabla\left(\pi_{\theta_{1}}\left(a|S'\right)\right)\theta_{1}^{\top}\phi_{S',a}-\nabla\left(\pi_{\theta_{2}}\left(a|S'\right)\right)\theta_{1}^{\top}\phi_{S',a}+\nabla\left(\pi_{\theta_{2}}\left(a|S'\right)\right)\theta_{1}^{\top}\phi_{S',a}\\
&\quad-\nabla\left(\pi_{\theta_{2}}\left(a|S'\right)\right)\theta_{2}^{\top}\phi_{S',a}\Bigg)\phi_{S,A}^{\top}\Bigg]+\gamma\mathbb{E}_{\mu}\left[\left(\sum_{a\in\mathcal{A}}\left(\pi_{\theta_{1}}\left(a|S'\right)\phi_{S',a}-\pi_{\theta_{2}}\left(a|S'\right)\phi_{S',a}\right)\right)\phi_{S,A}^{\top}\right].
\end{aligned}\tag{45}$$

This implies that

$$\begin{aligned}
&\left\|\nabla\left(\mathbb{E}_{\mu}\left[\delta_{S,A,S'}\left(\theta_{1}\right)\phi_{S,A}\right]\right)-\nabla\left(\mathbb{E}_{\mu}\left[\delta_{S,A,S'}\left(\theta_{2}\right)\phi_{S,A}\right]\right)\right\|\\
&\leq\gamma|\mathcal{A}|\left(2k_{1}+k_{2}R\right)\left\|\theta_{1}-\theta_{2}\right\|,
\end{aligned}\tag{46}$$

and thus $\nabla\left(\mathbb{E}_{\mu}\left[\delta_{S,A,S'}\left(\theta\right)\phi_{S,A}\right]\right)$ is Lipschitz in θ.

Following similar steps, we can also show that $\mathbb{E}_{\mu}\left[\delta_{S,A,S'}\left(\theta\right)\phi_{S,A}\right]$ is Lipschitz:

$$\left\|\mathbb{E}_{\mu}\left[\delta_{S,A,S'}\left(\theta_{1}\right)\phi_{S,A}\right]-\mathbb{E}_{\mu}\left[\delta_{S,A,S'}\left(\theta_{2}\right)\phi_{S,A}\right]\right\|\leq\left(\gamma(|\mathcal{A}|k_{1}R+1)+1\right)\left\|\theta_{1}-\theta_{2}\right\|.\tag{47}$$

Now by combining both parts in (46) and (47), we can show that

$$\begin{aligned}
&\left\|\nabla J\left(\theta_{1}\right)-\nabla J\left(\theta_{2}\right)\right\|\\
&\leq\left\|2\nabla\left(\mathbb{E}_{\mu}\left[\delta_{S,A,S'}\left(\theta_{1}\right)\phi_{S,A}\right]\right)C^{-1}\left(\mathbb{E}_{\mu}\left[\delta_{S,A,S'}\left(\theta_{1}\right)\phi_{S,A}\right]-\mathbb{E}_{\mu}\left[\delta_{S,A,S'}\left(\theta_{2}\right)\phi_{S,A}\right]\right)\right\|\\
&\quad+\left\|2\left(\nabla\left(\mathbb{E}_{\mu}\left[\delta_{S,A,S'}\left(\theta_{1}\right)\phi_{S,A}\right]\right)-\nabla\left(\mathbb{E}_{\mu}\left[\delta_{S,A,S'}\left(\theta_{2}\right)\phi_{S,A}\right]\right)\right)C^{-1}\mathbb{E}_{\mu}\left[\delta_{S,A,S'}\left(\theta_{2}\right)\phi_{S,A}\right]\right.\\
&\leq 2\gamma(k_{1}|\mathcal{A}|R+1)\frac{1}{\lambda}(1+\gamma(1+Rk_{1}|\mathcal{A}|)\|\theta_{1}-\theta_{2}\|\\
&\quad+2\frac{1}{\lambda}(r_{\max}+(1+\gamma)R)\gamma|\mathcal{A}|(2k_{1}+k_{2}R)\|\theta_{1}-\theta_{2}\|\\
&=2\gamma\frac{1}{\lambda}\left((k_{1}|\mathcal{A}|R+1)(1+\gamma+\gamma Rk_{1}|\mathcal{A}|)+|\mathcal{A}|(r_{\max}+R+\gamma R)(2k_{1}+k_{2}R)\right)\|\theta_{1}-\theta_{2}\|,
\end{aligned}\tag{48}$$

which implies that $\nabla J\left(\theta\right)$ is Lipschitz. This completes the proof. □

Recall that $G_{t+1}(\theta,\omega)=\delta_{t+1}(\theta)\phi_t-\gamma(\omega^T\phi_t)\hat{\phi}_{t+1}(\theta)$, where $\delta_{t+1}(\theta)=r_{t+1}+\gamma\bar{V}_{t+1}(\theta)-\theta^\top\phi_t$, $\bar{V}_{t+1}(\theta)=\bar{V}_\theta(S_{t+1})=\sum_{a\in\mathcal{A}}\pi_\theta(a|S_{t+1})\theta^\top\phi_{S_{t+1},a}$, and $\hat{\phi}_{t+1}(\theta)=\sum_{a\in\mathcal{A}}\theta^\top\phi_{S_{t+1},a}\nabla\pi_\theta(a|S_{t+1})+\pi_\theta(a|S_{t+1})\phi_{S_{t+1},a}$. The following Lemma shows that $G_{t+1}(\theta,\omega)$ is Lipschitz in ω, and $G_{t+1}(\theta,\omega^*(\theta))$ is Lipschitz in θ.

Lemma 3. *For any $\theta\in\{\theta:\|\theta\|\leq R\}$, $G_{t+1}(\theta,\omega)$ is Lipschitz in ω, and $G_{t+1}(\theta,\omega^*(\theta))$ is Lipschitz in θ. Specifically, for any w_1,w_2,*

$$\|G_{t+1}(\theta,\omega_1)-G_{t+1}(\theta,\omega_2)\|\leq\gamma(|\mathcal{A}|Rk_1+1)\|\omega_1-\omega_2\|,\tag{49}$$

and for any $\theta_1,\theta_2\in\{\theta:\|\theta\|\leq R\}$,

$$\|G_{t+1}(\theta_1,\omega^*(\theta_1))-G_{t+1}(\theta_2,\omega^*(\theta_2))\|\leq k_3\|\theta_1-\theta_2\|,\tag{50}$$

where $k_3=(1+\gamma+\gamma R|\mathcal{A}|k_1+\gamma\frac{1}{\lambda}|\mathcal{A}|(2k_1+k_2R)(r_{\max}+\gamma R+R)+\gamma\frac{1}{\lambda}(1+|\mathcal{A}|Rk_1)(1+\gamma+\gamma R|\mathcal{A}|k_1))$.

Proof. Following similar steps as those in (45) and (46), we can show that $\hat{\phi}_{t+1}(\theta)$ is Lipschitz in θ, i.e., for any θ_1,θ_2 $\in\{\theta:\|\theta\|\leq R\}$,

$$\|\hat{\phi}_{t+1}(\theta_1)-\hat{\phi}_{t+1}(\theta_2)\|\leq|\mathcal{A}|(2k_1+k_2R)\|\theta_1-\theta_2\|.\tag{51}$$

Under Assumption 4, it can be easily shown that

$$\|\hat{\phi}_{t+1}(\theta)\| \leq |\mathcal{A}|Rk_1 + 1. \tag{52}$$

It then follows that for any ω_1 and ω_2,

$$\begin{aligned}
&\|G_{t+1}(\theta, \omega_1)) - G_{t+1}(\theta, \omega_2))\| \\
&= \|\gamma(\omega_1 - \omega_2)^\top \phi_t)\hat{\phi}_{t+1}(\theta)\| \\
&\leq \gamma(|\mathcal{A}|Rk_1 + 1)\|\omega_1 - \omega_2\|.
\end{aligned} \tag{53}$$

To show that $G_{t+1}(\theta, \omega^*(\theta))$ is Lipschitz in θ, we have that

$$\begin{aligned}
&\|G_{t+1}(\theta_1, \omega^*(\theta_1)) - G_{t+1}(\theta_2, \omega^*(\theta_2))\| \\
&\leq |\delta_{t+1}(\theta_1) - \delta_{t+1}(\theta_2)| + \gamma\|(\omega^*(\theta_2))^\top \phi_t\hat{\phi}_{t+1}(\theta_2) - (\omega^*(\theta_1))^\top \phi_t\hat{\phi}_{t+1}(\theta_1)\| \\
&\overset{(a)}{\leq} \gamma\|(\omega^*(\theta_2))^\top \phi_t\hat{\phi}_{t+1}(\theta_2) - (\omega^*(\theta_1))^\top \phi_t\hat{\phi}_{t+1}(\theta_1) - (\omega^*(\theta_1))^\top \phi_t\hat{\phi}_{t+1}(\theta_2) + (\omega^*(\theta_1))^\top \phi_t\hat{\phi}_{t+1}(\theta_2)\| \\
&\quad + (1 + \gamma + \gamma R|\mathcal{A}|k_1)\|\theta_1 - \theta_2\| \\
&\leq \gamma(1 + |\mathcal{A}|Rk_1)\|\omega^*(\theta_2) - \omega^*(\theta_1)\| + \gamma\|\omega^*(\theta_1)\|\|\hat{\phi}_{t+1}(\theta_1) - \hat{\phi}_{t+1}(\theta_2)\ \| \\
&\quad + \gamma(1 + R|\mathcal{A}|k_1)\|\theta_1 - \theta_2\| + \|\theta_1 - \theta_2\| \\
&\overset{(b)}{\leq} \left(1 + \gamma + \gamma R|\mathcal{A}|k_1 + \gamma\frac{1}{\lambda}|\mathcal{A}|(2k_1 + k_2R)(r_{\max} + \gamma R + R) + \gamma\frac{1}{\lambda}(1 + |\mathcal{A}|Rk_1)(1 + \gamma + \gamma R|\mathcal{A}|k_1)\right) \\
&\quad \times \|\theta_1 - \theta_2\| \\
&\triangleq k_3\|\theta_1 - \theta_2\|,
\end{aligned} \tag{54}$$

where (a) can be shown following steps similar to those in (47), while (b) can be shown by combining

$$\|\omega^*(\theta)\| = \|C^{-1}\mathbb{E}[\delta_{t+1}(\theta)\phi_t]\| \leq \frac{1}{\lambda}(r_{\max} + \gamma R + R), \tag{55}$$

and

$$\|\omega^*(\theta_2) - \omega^*(\theta_1)\| \leq \frac{1}{\lambda}(1 + \gamma + \gamma R|\mathcal{A}|k_1)\|\theta_1 - \theta_2\|. \tag{56}$$

□

In the following lemma, we provide a decomposition of the stochastic bias, which is essential to our finite-sample analysis.

Lemma 4. *Consider the Greedy-GQ algorithm (see Algorithm 1), when the step-size α_t is constant, i.e., $\alpha_t = \alpha, \forall t \geq 0$, then*

$$\begin{aligned}
\sum_{t=0}^{T} \frac{\alpha_t}{2}\mathbb{E}[\|\nabla J(\theta_t)\|^2] \leq J(\theta_0) - J(\theta_{T+1}) + \gamma\alpha_t(1 + |\mathcal{A}|Rk_1)\sqrt{\sum_{t=0}^{T}\mathbb{E}[\|\nabla J(\theta_t)\|^2]}\sqrt{\sum_{t=0}^{T}\mathbb{E}[\|\omega^*(\theta_t) - \omega_t\|^2]} \\
+ \sum_{t=0}^{T}\alpha_t\mathbb{E}[\langle\nabla J(\theta_t), \frac{\nabla J(\theta_t)}{2} + G_{t+1}(\theta_t, \omega^*(\theta_t))\rangle] + \frac{K}{2}\sum_{t=0}^{T}\alpha_t^2\mathbb{E}[\|G_{t+1}(\theta_t, \omega_t)\|^2].
\end{aligned} \tag{57}$$

Proof. From Lemma 2, it follows that $J(\theta)$ is K-smooth. Then, by Taylor expansion, for any θ_1 and θ_2,

$$|J(\theta_1) - J(\theta_2) - \langle\nabla J(\theta_2), \theta_1 - \theta_2\rangle| \leq \frac{K}{2}||\theta_1 - \theta_2||^2. \tag{58}$$

Then, it can be shown that

$$\begin{aligned}
J(\theta_{t+1}) &\leq J(\theta_t) + \langle \nabla J(\theta_t), \theta_{t+1} - \theta_t \rangle + \frac{K}{2}\alpha_t^2 ||G_{t+1}(\theta_t, \omega_t)||^2 \\
&= J(\theta_t) + \alpha_t \langle \nabla J(\theta_t), G_{t+1}(\theta_t, \omega_t) \rangle + \frac{K}{2}\alpha_t^2 ||G_{t+1}(\theta_t, \omega_t)||^2 \\
&= J(\theta_t) - \alpha_t \langle \nabla J(\theta_t), -G_{t+1}(\theta_t, \omega_t) - \frac{\nabla J(\theta_t)}{2} + G_{t+1}(\theta_t, \omega^*(\theta_t)) - G_{t+1}(\theta_t, \omega^*(\theta_t)) \rangle \\
&\quad - \frac{\alpha_t}{2} ||\nabla J(\theta_t)||^2 + \frac{K}{2}\alpha_t^2 ||G_{t+1}(\theta_t, \omega_t)||^2 \\
&= J(\theta_t) - \alpha_t \langle \nabla J(\theta_t), -G_{t+1}(\theta_t, \omega_t) + G_{t+1}(\theta_t, \omega^*(\theta_t)) \rangle \\
&\quad + \alpha_t \langle \nabla J(\theta_t), \frac{\nabla J(\theta_t)}{2} + G_{t+1}(\theta_t, \omega^*(\theta_t)) \rangle - \frac{\alpha_t}{2} ||\nabla J(\theta_t)||^2 + \frac{K}{2}\alpha_t^2 ||G_{t+1}(\theta_t, \omega_t)||^2 \\
&\overset{(a)}{\leq} J(\theta_t) + \alpha_t \gamma \|\nabla J(\theta_t)\| (1 + |\mathcal{A}| R t_1) \|\omega^*(\theta_t) - \omega_t\| + \alpha_t \langle \nabla J(\theta_t), \frac{\nabla J(\theta_t)}{2} + G_{t+1}(\theta_t, \omega^*(\theta_t)) \rangle \\
&\quad - \frac{\alpha_t}{2} ||\nabla J(\theta_t)||^2 + \frac{K}{2}\alpha_t^2 ||G_{t+1}(\theta_t, \omega_t)||^2, \qquad (59)
\end{aligned}$$

where (a) follows from the fact that $G_{t+1}(\theta, \omega)$ is Lipschitz in ω (see Lemma 3).

By taking expectation of both sides, summing up the inequality from 0 to T, and rearranging the terms, we have that

$$\begin{aligned}
&\sum_{t=0}^{T} \frac{\alpha_t}{2} \mathbb{E}[\|\nabla J(\theta_t)\|^2] \\
&\leq J(\theta_0) - J(\theta_{T+1}) + \sum_{t=0}^{T} \gamma \alpha_t (1 + |\mathcal{A}| R k_1) \mathbb{E}[\|\nabla J(\theta_t)\| \|\omega^*(\theta_t) - \omega_t\|] \\
&\quad + \sum_{t=0}^{T} \alpha_t \mathbb{E}[\langle \nabla J(\theta_t), \frac{\nabla J(\theta_t)}{2} + G_{t+1}(\theta_t, \omega^*(\theta_t)) \rangle] + \frac{K}{2} \sum_{t=0}^{T} \alpha_t^2 \mathbb{E}[\|G_{t+1}(\theta_t, \omega_t)\|^2]. \qquad (60)
\end{aligned}$$

We then apply Cauchy-Schwarz's inequality, and we have that

$$\begin{aligned}
&\sum_{t=0}^{T} \mathbb{E}[\|\nabla J(\theta_t)\| \|\omega^*(\theta_t) - \theta_t\|] \\
&\leq \sum_{t=0}^{T} \sqrt{\mathbb{E}[\|\nabla J(\theta_t)\|^2] \mathbb{E}[\|\omega^*(\theta_t) - \theta_t\|^2]}. \qquad (61)
\end{aligned}$$

We further define two vectors a_E and a_z, where

$$a_E \triangleq \left(\sqrt{\mathbb{E}[\|\nabla J(\theta_0)\|^2]}, \sqrt{\mathbb{E}[\|\nabla J(\theta_1)\|^2]}, ..., \sqrt{\mathbb{E}[\|\nabla J(\theta_T)\|^2]} \right)^\top, \qquad (62)$$

$$a_z \triangleq \left(\sqrt{\mathbb{E}[\|\omega^*(\theta_0) - \theta_0\|^2]}, \sqrt{\mathbb{E}[\|\omega^*(\theta_1) - \theta_1\|^2]}, ..., \sqrt{\mathbb{E}[\|\omega^*(\theta_T) - \theta_T\|^2]} \right)^\top. \qquad (63)$$

Then, it follows that

$$\begin{aligned}
&\sum_{t=0}^{T} \sqrt{\mathbb{E}[\|\nabla J(\theta_t)\|^2] \mathbb{E}[\|\omega^*(\theta_t) - \theta_t\|^2]} \\
&= \langle a_E, a_z \rangle \\
&\leq \|a_E\| \|a_z\| \\
&= \sqrt{\sum_{t=0}^{T} \mathbb{E}[\|\nabla J(\theta_t)\|^2]} \sqrt{\sum_{t=0}^{T} \mathbb{E}[\|\omega^*(\theta_t) - \omega_t\|^2]}. \qquad (64)
\end{aligned}$$

Thus plugging (64) in (60), and since $\alpha_t = \alpha, \forall t \geq 0$ is constant, we have that

$$\begin{aligned}
&\sum_{t=0}^{T} \frac{\alpha_t}{2} \mathbb{E}[\|\nabla J(\theta_t)\|^2] \\
&\leq J(\theta_0) - J(\theta_{T+1}) + \gamma\alpha_t(1 + |\mathcal{A}|Rk_1) \sum_{t=0}^{T} \mathbb{E}[\|\nabla J(\theta_t)\| \|\omega^*(\theta_t) - \omega_t\|] \\
&\quad + \sum_{t=0}^{T} \alpha_t \mathbb{E}[\langle \nabla J(\theta_t), \frac{\nabla J(\theta_t)}{2} + G_{t+1}(\theta_t, \omega^*(\theta_t))\rangle] + \frac{K}{2} \sum_{t=0}^{T} \alpha_t^2 \mathbb{E}[\|G_{t+1}(\theta_t, \omega_t)\|^2] \\
&\leq J(\theta_0) - J(\theta_{T+1}) + \gamma\alpha_t(1 + |\mathcal{A}|Rk_1) \sqrt{\sum_{t=0}^{T} \mathbb{E}[\|\nabla J(\theta_t)\|^2]} \sqrt{\sum_{t=0}^{T} \mathbb{E}[\|\omega^*(\theta_t) - \omega_t\|^2]} \\
&\quad + \sum_{t=0}^{T} \alpha_t \mathbb{E}[\langle \nabla J(\theta_t), \frac{\nabla J(\theta_t)}{2} + G_{t+1}(\theta_t, \omega^*(\theta_t))\rangle] + \frac{K}{2} \sum_{t=0}^{T} \alpha_t^2 \mathbb{E}[\|G_{t+1}(\theta_t, \omega_t)\|^2]. \qquad (65)
\end{aligned}$$

□

We next derive the bounds on $\mathbb{E}[\langle \nabla J(\theta_t), \frac{\nabla J(\theta_t)}{2} + G_{t+1}(\theta_t, \omega^*(\theta_t))\rangle]$ and $\mathbb{E}[\|\omega^*(\theta_t) - \omega_t\|]$, where we refer to the second term as the "tracking error".

We first define $z_t = \omega_t - \omega^*(\theta_t)$, then the algorithm can be written as:

$$\theta_{t+1} = \theta_t + \alpha_t(f_1(\theta_t, O_t) + g_1(\theta_t, z_t, O_t)), \qquad (66)$$

$$z_{t+1} = z_t + \beta_t(f_2(\theta_t, O_t) + g_2(\theta_t, O_t)) + \omega^*(\theta_t) - \omega^*(\theta_{t+1}), \qquad (67)$$

where

$$\begin{cases} f_1(\theta_t, O_t) \triangleq \delta_{t+1}(\theta_t)\phi_t - \gamma\phi_t^\top \omega^*(\theta_t)\hat{\phi}_{t+1}(\theta_t), \\ g_1(\theta_t, z_t, O_t) \triangleq -\gamma\phi_t^\top z_t \hat{\phi}_{t+1}(\theta_t), \\ f_2(\theta_t, O_t) \triangleq (\delta_{t+1}(\theta_t) - \phi_t^\top \omega^*(\theta_t))\phi_t, \\ g_2(z_t, O_t) \triangleq -\phi_t^\top z_t \phi_t, \\ O_t \triangleq (s_t, a_t, r_t, s_{t+1}). \end{cases} \qquad (68)$$

We then develop some upper bounds of functions f_1, g_1, f_2, g_2 in the algorithm in the following lemma.

Lemma 5. *For $\|\theta\| \leq R$, $\|z\| \leq 2R$, there exist constants c_{f_1}, c_{g_1}, c_{g_2} and c_{f_2} such that $\|f_1(\theta, O_t)\| \leq c_{f_1}$, $\|g_1(\theta, z, O_t)\| \leq c_{g_1}$, $|f_2(\theta, O_t)| \leq c_{f_2}$ and $|g_2(\theta, O_t)| \leq c_{g_2}$, where $c_{f_1} = r_{\max} + (1+\gamma)R + \gamma\frac{1}{\lambda}(r_{\max} + (1+\gamma)R)(1 + R|\mathcal{A}|k_1)$, $c_{g_1} = 2\gamma R(1 + R|\mathcal{A}|k_1)$, $c_{f_2} = r_{\max} + (1+\gamma)R + \frac{1}{\lambda}(r_{\max} + (1+\gamma)R)$, and $c_{g_2} = 2R$.*

Proof. This Lemma can be shown easily using (43), (52) and (56). □

We further define $\zeta(\theta, O_t) \triangleq \langle \nabla J(\theta), \frac{\nabla J(\theta)}{2} + G_{t+1}(\theta, \omega^*(\theta))\rangle$, then we have that $\mathbb{E}_\mu[\zeta(\theta, O_t)] = 0$ for any fixed θ, where (S_t, A_t) in O_t follow the stationary distribution μ. In the following lemma, we provide upper bound on $\mathbb{E}[\zeta(\theta, O_t)]$.

Lemma 6. *Let $\tau_{\alpha_T} \triangleq \min\{k : m\rho^k \leq \alpha_T\}$. If $t \leq \tau_{\alpha_T}$, then*

$$\mathbb{E}[\zeta(\theta_t, O_t)] \leq c_\zeta(c_{f_1} + c_{g_1})\alpha_0\tau_{\alpha_T}, \qquad (69)$$

and if $t > \tau_{\alpha_T}$, then

$$\mathbb{E}[\zeta(\theta_t, O_t)] \leq k_\zeta\alpha_T + c_\zeta(c_{f_1} + c_{g_1})\tau_{\alpha_T}\alpha_{t-\tau_{\alpha_T}}. \qquad (70)$$

Where $c_\zeta = 2\gamma(1+k_1|\mathcal{A}|R)\frac{1}{\lambda}(r_{\max}+R+\gamma R)(\frac{K}{2}+k_3)+K(r_{\max}+R+\gamma R)(\gamma\frac{1}{\lambda}(1+k_1|\mathcal{A}|R)+1+\gamma\frac{1}{\lambda}(1+Rk_1|\mathcal{A}|))$ and $k_\zeta = 4\gamma(1 + k_1R|\mathcal{A}|)\frac{1}{\lambda}(r_{\max} + R + \gamma R)^2(2\gamma(1 + k_1|\mathcal{A}|R)\frac{1}{\lambda} + 1)$.

Proof. We note that when θ is fixed, $\mathbb{E}[G_{t+1}(\theta, \omega^*(\theta))] = -\frac{1}{2}\nabla J(\theta)$. We will use this fact and the Markov mixing property to show this Lemma. Note that for any θ_1 and θ_2, it follows that

$$\begin{aligned}
&|\zeta(\theta_1, O_t) - \zeta(\theta_2, O_t)| \\
&= |\langle \nabla J(\theta_1), \frac{\nabla J(\theta_1)}{2} + G_{t+1}(\theta_1, \omega^*(\theta_1))\rangle - \langle \nabla J(\theta_1), \frac{\nabla J(\theta_2)}{2} + G_{t+1}(\theta_2, \omega^*(\theta_2))\rangle \\
&\quad + \langle \nabla J(\theta_1), \frac{\nabla J(\theta_2)}{2} + G_{t+1}(\theta_2, \omega^*(\theta_2))\rangle - \langle \nabla J(\theta_2), \frac{\nabla J(\theta_2)}{2} + G_{t+1}(\theta_2, \omega^*(\theta_2))\rangle|. \quad (71)
\end{aligned}$$

Since $J(\theta)$ and $\|\nabla J(\theta)\|$ are Lipschitz in θ by Lemma 2, thus $\zeta(\theta, O_t)$ is also Lipschitz in θ. We then denote its Lipschitz constant by c_ζ, i.e.,

$$|\zeta(\theta_1, O_t) - \zeta(\theta_2, O_t)| \leq c_\zeta \|\theta_1 - \theta_2\|, \quad (72)$$

where

$$\begin{aligned}
c_\zeta &= 2\gamma(1 + k_1|\mathcal{A}|R)\frac{1}{\lambda}(r_{\max} + R + \gamma R)(\frac{K}{2} + k_3) \\
&\quad + K(r_{\max} + R + \gamma R)(\gamma\frac{1}{\lambda}(1 + k_1|\mathcal{A}|R) + 1 + \gamma\frac{1}{\lambda}(1 + Rk_1|\mathcal{A}|)). \quad (73)
\end{aligned}$$

Thus from (71), it follows that for any $\tau \geq 0$,

$$|\zeta(\theta_t, O_t) - \zeta(\theta_{t-\tau}, O_t)| \leq c_\zeta\|\theta_t - \theta_{t-\tau}\| \leq c_\zeta(c_{f_1} + c_{g_1})\sum_{k=t-\tau}^{t-1}\alpha_k. \quad (74)$$

We define an independent random variable $\hat{O} = (\hat{S}, \hat{A}, \hat{R}, \hat{S}')$, where $(\hat{S}, \hat{A}) \sim \mu$, $\hat{S}'$ is the subsequent state and $\hat{R}$ is the reward. Then $\mathbb{E}[\zeta(\theta_{t-\tau}, \hat{O})] = 0$ by the fact that $\mathbb{E}_\mu[G_{t+1}(\theta, \omega^*(\theta))] = -\frac{1}{2}\nabla J(\theta)$. Thus,

$$\mathbb{E}[\zeta(\theta_{t-\tau}, O_t)] \leq |\mathbb{E}[\zeta(\theta_{t-\tau}, O_t)] - \mathbb{E}[\zeta(\theta_{t-\tau}, O')]| \leq k_\zeta m\rho^\tau, \quad (75)$$

which follows from the Markov Mixing property in Assumption 3, where $k_\zeta = 4\gamma(1 + k_1R|\mathcal{A}|)\frac{1}{\lambda}(r_{\max} + R + \gamma R)^2(2\gamma(1 + k_1|\mathcal{A}|R)\frac{1}{\lambda} + 1)$.

If $t \leq \tau_{\alpha_T}$, then we choose $\tau = t$ in (74). Then we have that

$$\mathbb{E}[\zeta(\theta_t, O_t)] \leq \mathbb{E}[\zeta(\theta_0, O_t)] + c_\zeta(c_{f_1} + c_{g_1})\sum_{k=0}^{t-1}\alpha_k \leq c_\zeta(c_{f_1} + c_{g_1})t\alpha_0 \overset{(a)}{\leq} c_\zeta(c_{f_1} + c_{g_1})\alpha_0\tau_{\alpha_T}, \quad (76)$$

where (a) is due to the fact that α_t is non-increasing. If $t > \tau_{\alpha_T}$, we choose $\tau = \tau_{\alpha_T}$, and then

$$\begin{aligned}
&\mathbb{E}[\zeta(\theta_t, O_t)] \leq \mathbb{E}[\zeta(\theta_{t-\tau_{\alpha_T}}, O_t)] + c_\zeta(c_{f_1} + c_{g_1})\sum_{k=t-\tau_{\alpha_T}}^{t-1}\alpha_k \\
&\leq k_\zeta m\rho^{\tau_{\alpha_T}} + c_\zeta(c_{f_1} + c_{g_1})\tau_{\alpha_T}\alpha_{t-\tau_{\alpha_T}} \leq k_\zeta\alpha_T + c_\zeta(c_{f_1} + c_{g_1})\tau_{\alpha_T}\alpha_{t-\tau_{\alpha_T}}. \quad (77)
\end{aligned}$$

□

We next bound the tracking error $\mathbb{E}[\|z_t\|]$. Define $\zeta_{f_2}(\theta, z, O_t) \triangleq \langle z, f_2(\theta, O_t)\rangle$, and $\zeta_{g_2}(z, O_t) \triangleq \langle z, g_2(z, O_t) - \bar{g_2}(z)\rangle$, where $\bar{g_2}(z) \triangleq \mathbb{E}[g_2(z, O_t)] = \mathbb{E}[-\phi_t^\top z\phi_t]$.

Lemma 7. *Consider any $\theta, \theta_1, \theta_2 \in \{\theta : \|\theta\| \leq R\}$ and any $z, z_1, z_2 \in \{z : \|z\| \leq 2R\}$. Then 1) $|\zeta_{f_2}(\theta, z, O_t)| \leq 2Rc_{f_2}$; 2) $|\zeta_{f_2}(\theta_1, z_1, O_t) - \zeta_{f_2}(\theta_2, z_2, O_t)| \leq k_{f_2}\|\theta_1 - \theta_2\| + k'_{f_2}\|z_1 - z_2\|$, where $k_{f_2} = 2R(1 + \gamma + \gamma Rk_1|\mathcal{A}|)(1 + \frac{1}{\lambda})$ and $k'_{f_2} = c_{f_2}$; 3) $|\zeta_{g_2}(z, O_t)| \leq 8R^2$; and 4) $|\zeta_{g_2}(z_1, O_t) - \zeta_{g_2}(z_2, O_t)| \leq 8R\|z_1 - z_2\|$.*

Proof. To prove 1), it can be shown that $|\zeta_{f_2}(\theta, z, O_t)| = |\langle z, f_2(\theta, O_t)\rangle| \leq 2Rc_{f_2}$.

For 2), it can be shown that

$$\begin{aligned}
&|\zeta_{f_2}(\theta_1, z_1, O_t) - \zeta_{f_2}(\theta_2, z_2, O_t)| \\
&= |\langle z_1, f_2(\theta_1, O_t)\rangle - \langle z_2, f_2(\theta_2, O_t)\rangle| \\
&\leq |\langle z_1, f_2(\theta_1, O_t)\rangle - \langle z_1, f_2(\theta_2, O_t)| + |\langle z_1, f_2(\theta_2, O_t) - \langle z_2, f_2(\theta_2, O_t)\rangle| \\
&\leq 2R\|f_2(\theta_1, O_t) - f_2(\theta_2, O_t)\| + \|f_2(\theta_2, O_t)\|\|z_1 - z_2\| \\
&\leq 2R(|\delta_{t+1}(\theta_1) - \delta_{t+1}(\theta_2)| + \|\omega^*(\theta_1) - \omega^*(\theta_2)\|) + c_{f_2}\|z_1 - z_2\| \\
&\overset{(a)}{\leq} k_{f_2}\|\theta_1 - \theta_2\| + k'_{f_2}\|z_1 - z_2\|,
\end{aligned} \tag{78}$$

where (a) is from both $\delta(\theta)$ and $\omega^*(\theta_t)(\theta)$ are Lipschitz, $k_{f_2} = 2R(1 + \gamma + \gamma Rk_1|\mathcal{A}|)(1 + \frac{1}{\lambda})$, and $k'_{f_2} = c_{f_2}$.

For 3), we have that $\zeta_{g_2}(z, O_t) = \langle z, -\phi_t^\top z\phi_t + \mathbb{E}[\phi_t^\top z\phi_t]\rangle \leq 8R^2$.

To prove 4), we have that

$$\begin{aligned}
&|\zeta_{g_2}(z_1, O_t) - \zeta_{g_2}(z_2, O_t)| \\
&= |\langle z_1, -\phi_t^\top z_1\phi_t + \mathbb{E}[\phi_t^\top z_1\phi_t]\rangle - \langle z_1, -\phi_t^\top z_2\phi_t + \mathbb{E}[\phi_t^\top z_2\phi_t]\rangle + \langle z_1, -\phi_t^\top z_2\phi_t \\
&\quad + \mathbb{E}[\phi_t^\top z_2\phi_t]\rangle - \langle z_2, -\phi_t^\top z_2\phi_t + \mathbb{E}[\phi_t^\top z_2\phi_t]\rangle| \\
&\leq 8R\|z_1 - z_2\|.
\end{aligned} \tag{79}$$

□

In the following lemma, we derive bounds on $\mathbb{E}[\zeta_{f_2}(\theta_1, z_t, O_t)]$ and $\mathbb{E}[\zeta_{g_2}(z_t, O_t)]$.

Lemma 8. *Define* $\tau_{\beta_T} = \min\{k : m\rho^k \leq \beta_T\}$. *If* $t \leq \tau_{\beta_T}$, *then*

$$\mathbb{E}[\zeta_{f_2}(\theta_t, z_t, O_t)] \leq 4Rc_{f_2}\beta_T + a_{f_2}\tau_{\beta_T}, \tag{80}$$

where $a_{f_2} = (k'_{f_2}(c_{f_2} + c_{g_2})\beta_0 + (k_{f_2}(c_{f_1} + c_{g_1}) + k'_{f_2}\frac{1}{\lambda}(1 + \gamma + \gamma R|\mathcal{A}|k_1)(c_{f_1} + c_{g_1}))\alpha_0)$*; and if* $t > \tau_{\beta_T}$, *then*

$$\mathbb{E}[\zeta_{f_2}(\theta_t, z_t, O_t)] \leq 4Rc_{f_2}\beta_T + b_{f_2}\tau_{\beta_T}\beta_{t-\tau_{\beta_T}}, \tag{81}$$

where $b_{f_2} = (k'_{f_2}(c_{f_2} + c_{g_2}) + (k_{f_2}(c_{f_1} + c_{g_1}) + k'_{f_2}\frac{1}{\lambda}(1 + \gamma + \gamma R|\mathcal{A}|k_1)(c_{f_1} + c_{g_1})))$.

Proof. We first note that

$$\begin{aligned}
&\|z_{t+1} - z_t\| \\
&= \|\beta_t(f_2(\theta_t, O_t) + g_2(z_t, O_t)) + \omega^*(\theta_t) - \omega^*(\theta_{t+1})\| \\
&\leq (c_{f_2} + c_{g_2})\beta_t + \frac{1}{\lambda}(1 + \gamma + \gamma R|\mathcal{A}|k_1)(c_{f_1} + c_{g_1})\alpha_t,
\end{aligned} \tag{82}$$

where the last step is due to (56). Furthermore, due to part 2) in Lemma 7, ζ_{f_2} is Lipschitz in both θ and z, then we have that for any $\tau \geq 0$

$$\begin{aligned}
&|\zeta_{f_2}(\theta_t, z_t, O_t) - \zeta_{f_2}(\theta_{t-\tau}, z_{t-\tau}, O_t)| \\
&\overset{(a)}{\leq} k_{f_2}(c_{f_1} + c_{g_1})\sum_{i=t-\tau}^{t-1}\alpha_i + k'_{f_2}(c_{f_2} + c_{g_2})\sum_{i=t-\tau}^{t-1}\beta_i + \sum_{i=t-\tau}^{t-1} k'_{f_2}\frac{1}{\lambda}(1 + \gamma + \gamma R|\mathcal{A}|k_1)(c_{f_1} + c_{g_1})\alpha_i \\
&= k'_{f_2}(c_{f_2} + c_{g_2})\sum_{i=t-\tau}^{t-1}\beta_i + (k_{f_2}(c_{f_1} + c_{g_1}) + k'_{f_2}\frac{1}{\lambda}(1 + \gamma + \gamma R|\mathcal{A}|k_1)(c_{f_1} + c_{g_1}))\sum_{i=t-\tau}^{t-1}\alpha_i,
\end{aligned} \tag{83}$$

where in (a), we apply (56) and Lemma 5 to obtain the third term.

Define an independent random variable $\hat{O} = (\hat{S}, \hat{A}, \hat{R}, \hat{S}')$, where $(\hat{S}, \hat{A}) \sim \mu$, $\hat{S}' \sim \mathsf{P}(\cdot|\hat{S}, \hat{A})$ is the subsequent state, and $\hat{R}$ is the reward. Then it can be shown that

$$\begin{aligned}&\mathbb{E}[\zeta_{f_2}(\theta_{t-\tau}, z_{t-\tau}, O_t)]\\&\overset{(a)}{\le} |\mathbb{E}[\zeta_{f_2}(\theta_{t-\tau}, z_{t-\tau}, O_t)] - \mathbb{E}[\zeta_{f_2}(\theta_{t-\tau}, z_{t-\tau}, \hat{O})]|\\&\le 4Rc_{f_2} m\rho^{\tau},\end{aligned} \tag{84}$$

where (a) is due to the fact that $\mathbb{E}[\zeta_{f_2}(\theta_{t-\tau}, z_{t-\tau}, \hat{O})] = 0$, and the last inequality follows from Assumption 3.

If $t \le \tau_{\beta_T}$, we choose $\tau = t$ in (83). Then it can be shown that

$$\begin{aligned}&\mathbb{E}[\zeta_{f_2}(\theta_t, z_t, O_t)]\\&\le \mathbb{E}[\zeta_{f_2}(\theta_0, z_0, O_t)] + k'_{f_2}(c_{f_2} + c_{g_2})\sum_{i=0}^{t-1}\beta_i + (k_{f_2}(c_{f_1} + c_{g_1})\\&\quad + k'_{f_2}\frac{1}{\lambda}(1+\gamma+\gamma R|\mathcal{A}|k_1)(c_{f_1} + c_{g_1}))\sum_{i=0}^{t-1}\alpha_i\\&\le 4Rc_{f_2} m\rho^t + k'_{f_2}(c_{f_2} + c_{g_2})t\beta_0 + (k_{f_2}(c_{f_1} + c_{g_1}) + k'_{f_2}\frac{1}{\lambda}(1+\gamma+\gamma R|\mathcal{A}|k_1)(c_{f_1} + c_{g_1}))t\alpha_0\\&\le 4Rc_{f_2}\beta_T + (k'_{f_2}(c_{f_2} + c_{g_2})\beta_0 + (k_{f_2}(c_{f_1} + c_{g_1}) + k'_{f_2}\frac{1}{\lambda}(1+\gamma+\gamma R|\mathcal{A}|k_1)(c_{f_1} + c_{g_1}))\alpha_0)\tau_{\beta_T}.\end{aligned} \tag{85}$$

If $t > \tau_{\beta_T}$, we choose $\tau = \tau_{\beta_T}$ in (83). Then, it can be shown that

$$\begin{aligned}&\mathbb{E}[\zeta_{f_2}(\theta_t, z_t, O_t)]\\&\le \mathbb{E}[\zeta_{f_2}(\theta_{t-\tau_{\beta_T}}, z_{t-\tau_{\beta_T}}, O_t)]\\&\quad + k'_{f_2}(c_{f_2} + c_{g_2})\sum_{i=t-\tau_{\beta_T}}^{t-1}\beta_i + (k_{f_2}(c_{f_1} + c_{g_1}) + k'_{f_2}\frac{1}{\lambda}(1+\gamma+\gamma R|\mathcal{A}|k_1)(c_{f_1} + c_{g_1}))\sum_{i=t-\tau_{\beta_T}}^{t-1}\alpha_i\\&\le 4Rc_{f_2} m\rho^{\tau_{\beta_T}} + k'_{f_2}(c_{f_2} + c_{g_2})\tau_{\beta_T}\beta_{t-\tau_{\beta_T}} + (k_{f_2}(c_{f_1} + c_{g_1}) + k'_{f_2}\frac{1}{\lambda}(1+\gamma+\gamma R|\mathcal{A}|k_1)(c_{f_1} + c_{g_1}))\tau_{\beta_T}\alpha_{t-\tau_{\beta_T}}\\&\le 4Rc_{f_2}\beta_T + (k'_{f_2}(c_{f_2} + c_{g_2}) + (k_{f_2}(c_{f_1} + c_{g_1}) + k'_{f_2}\frac{1}{\lambda}(1+\gamma+\gamma R|\mathcal{A}|k_1)(c_{f_1} + c_{g_1})))\tau_{\beta_T}\beta_{t-\tau_{\beta_T}},\end{aligned} \tag{86}$$

where in the last step we upper bound α_t using β_t. Note that this will not change the order of the bound. □

Similarly, in the following lemma, we derive a bound on $\mathbb{E}[\zeta_{g_2}(z_t, O_t)]$.

Lemma 9. *If $t \le \tau_{\beta_T}$, then*

$$\mathbb{E}[\zeta_{g_2}(z_t, O_t)] \le a_{g_2}\tau_{\beta_T}; \tag{87}$$

and if $t > \tau_{\beta_T}$, then

$$\mathbb{E}[\zeta_{g_2}(z_t, O_t)] \le b_{g_2}\beta_T + b'_{g_2}\tau_{\beta_T}\beta_{t-\tau_{\beta_T}}, \tag{88}$$

where $a_{g_2} = 8R(c_{f_2} + c_{g_2})\beta_0 + \frac{1}{\lambda}(1+\gamma+\gamma R|\mathcal{A}|k_1)(c_{f_1} + c_{g_1})\alpha_0)$, $b_{g_2} = 16R^2$, and $b'_{g_2} = 8R(c_{f_2} + c_{g_2})\beta_0 + \frac{1}{\lambda}(1+\gamma+\gamma R|\mathcal{A}|k_1)(c_{f_1} + c_{g_1})\alpha_0$.

Proof. The proof is similar to the one for Lemma 8. □

We then bound the tracking error as follows:

$$
\begin{aligned}
&||z_{t+1}||^2\\
&= ||z_t + \beta_t(f_2(\theta_t, O_t) + g_2(z_t, O_t)) + \omega^*(\theta_t) - \omega^*(\theta_{t+1})||^2\\
&= ||z_t||^2 + 2\beta_t\langle z_t, f_2(\theta_t, O_t)\rangle + 2\beta_t\langle z_t, g_2(z_t, O_t)\rangle + 2\langle z_t, \omega^*(\theta_t) - \omega^*(\theta_{t+1})\rangle\\
&\quad + ||\beta_t f_2(\theta_t, O_t) + \beta_t g_2(z_t, O_t) + \omega^*(\theta_t) - \omega^*(\theta_{t+1})||^2\\
&\leq ||z_t||^2 + 2\beta_t\langle z_t, f_2(\theta_t, O_t)\rangle + 2\beta_t\langle z_t, g_2(z_t, O_t)\rangle + 2\langle z_t, \omega^*(\theta_t) - \omega^*(\theta_{t+1})\rangle\\
&\quad + 3\beta_t^2||f_2(\theta_t, O_t)||^2 + 3\beta_t^2||g_2(z_t, O_t)||^2 + 3||\omega^*(\theta_t) - \omega^*(\theta_{t+1})||^2\\
&\overset{(a)}{\leq} ||z_t||^2 + 2\beta_t\langle z_t, f_2(\theta_t, O_t)\rangle + 2\beta_t\langle z_t, \bar{g_2}(z_t)\rangle + 2\langle z_t, \omega^*(\theta_t) - \omega^*(\theta_{t+1})\rangle + 2\beta_t\langle z_t, g_2(z_t, O_t) - \bar{g_2}(z_t)\rangle\\
&\quad + 3\beta_t^2 c_{f_2}^2 + 3\beta_t^2 c_{g_2}^2 + 6\frac{1}{\lambda^2}(1 + \gamma + \gamma R|\mathcal{A}|k_1)^2\alpha_t^2(c_{f_1}^2 + c_{g_1}^2),
\end{aligned}\tag{89}
$$

where (a) follows from Lemma 5 and (56).

Note that $\langle z_t, \bar{g_2}(z_t)\rangle = -z_t^\top C z_t$, and C is a positive definite matrix. Recall the minimal eigenvalue of C is denoted by λ, then (89) can be further bounded as follows:

$$
\begin{aligned}
||z_{t+1}||^2 \leq{}& (1 - 2\beta_t\lambda)\|z_t\|^2 + 2\beta_t\zeta_{f_2} + 2\beta_t\zeta_{g_2} + 2\langle z_t, \omega^*(\theta_t) - \omega^*(\theta_{t+1})\rangle + 3\beta_t^2 c_{f_2}^2\\
&+ 3\beta_t^2 c_{g_2}^2 + 6\frac{1}{\lambda^2}(1 + \gamma + \gamma R|\mathcal{A}|k_1)^2\alpha_t^2(c_{f_1}^2 + c_{g_1}^2).
\end{aligned}\tag{90}
$$

Taking expectation on both sides of the (90), and applying it recursively, we obtain that

$$
\begin{aligned}
\mathbb{E}[||z_{t+1}||^2] \leq{}& \prod_{i=0}^{t}(1 - 2\beta_i\lambda)||z_0||^2\\
&+ 2\sum_{i=0}^{t}\prod_{k=i+1}^{t}(1 - 2\beta_k\lambda)\beta_i\mathbb{E}[\zeta_{f_2}(z_i, \theta_i, O_i)]\\
&+ 2\sum_{i=0}^{t}\prod_{k=i+1}^{t}(1 - 2\beta_k\lambda)\beta_i\mathbb{E}[\zeta_{g_2}(z_i, O_i)]\\
&+ 2\sum_{i=0}^{t}\prod_{k=i+1}^{t}(1 - 2\beta_k\lambda)\mathbb{E}\langle z_i, \omega^*(\theta_i) - \omega^*(\theta_{i+1})\rangle + 3(c_{f_2}^2 + c_{g_2}^2)\sum_{i=0}^{t}\prod_{k=i+1}^{t}(1 - 2\beta_k\lambda)\beta_i^2\\
&+ 6\frac{1}{\lambda^2}(1 + \gamma + \gamma R|\mathcal{A}|k_1)^2(c_{f_1}^2 + c_{g_1}^2)\sum_{i=0}^{t}\prod_{k=i+1}^{t}(1 - 2\beta_k\lambda)\alpha_i^2.
\end{aligned}\tag{91}
$$

Also note that $1 - 2\beta_i\lambda \leq e^{-2\beta_i\lambda}$, which further implies that

$$
\begin{aligned}
\mathbb{E}[||z_{t+1}||^2 \leq{}& A_t||z_0||^2 + 2\sum_{i=0}^{t}B_{it} + 2\sum_{i=0}^{t}C_{it} + 2\sum_{i=0}^{t}D_{it}\\
&+ 3(c_{f_2}^2 + c_{g_2}^2 + 2\frac{1}{\lambda^2}(1 + \gamma + \gamma R|\mathcal{A}|k_1)^2(c_{f_1}^2 + c_{g_1}^2))\sum_{i=0}^{t}E_{it},
\end{aligned}\tag{92}
$$

where

$$
\begin{aligned}
A_t &= e^{-2\lambda\sum_{i=0}^{t}\beta_i},\\
B_{it} &= e^{-2\lambda\sum_{k=i+1}^{t}\beta_k}\beta_i\mathbb{E}[\zeta_{f_2}(z_i, \theta_i, O_i)],\\
C_{it} &= e^{-2\lambda\sum_{k=i+1}^{t}\beta_k}\beta_i\mathbb{E}[\zeta_{g_2}(z_i, O_i)],\\
D_{it} &= e^{-2\lambda\sum_{k=i+1}^{t}\beta_k}\mathbb{E}[\langle z_t, \omega^*(\theta_i) - \omega^*(\theta_{i+1})\rangle],\\
E_{it} &= e^{-2\lambda\sum_{k=i+1}^{t}\beta_k}\beta_i^2.
\end{aligned}\tag{93}
$$

Consider the second term in (92). Using Lemma 8, it can be further bounded as follows:

$$\begin{aligned}\sum_{i=0}^{t} B_{it} &= \sum_{i=0}^{t} e^{-2\lambda \sum_{k=i+1}^{t} \beta_k} \beta_i \mathbb{E}[\zeta_{f_2}(z_i, \theta_i, O_i)] \\ &\leq \sum_{i=0}^{\tau_{\beta_T}} (a_{f_2} \tau_{\beta_T} + 4Rc_{f_2}\beta_T) e^{-2\lambda \sum_{k=i+1}^{t} \beta_k} \beta_i + 4Rc_{f_2}\beta_T \sum_{i=\tau_{\beta_T}+1}^{t} e^{-2\lambda \sum_{k=i+1}^{t} \beta_k} \beta_i \\ &\quad + b_{f_2}\tau_{\beta_T} \sum_{i=\tau_{\beta_T}+1}^{t} e^{-2\lambda \sum_{k=i+1}^{t} \beta_k} \beta_{i-\tau_{\beta_T}} \beta_i. \end{aligned} \tag{94}$$

Further analysis of the bound will be made when we specify the step-sizes α_t, β_t, which will be provided later.

Similarly, using Lemma 9, we can bound the third term in (92) as follows:

$$\begin{aligned}\sum_{i=0}^{t} C_{it} &= \sum_{i=0}^{t} e^{-2\lambda \sum_{k=i+1}^{t} \beta_k} \beta_i \mathbb{E}[\zeta_{g_2}(z_i, O_i)] \\ &\leq \tau_{\beta_T} a_{g_2} \sum_{i=0}^{\tau_{\beta_T}} e^{-2\lambda \sum_{k=i+1}^{t} \beta_k} \beta_i + b_{g_2}\beta_T \sum_{i=\tau_{\beta_T}+1}^{t} e^{-2\lambda \sum_{k=i+1}^{t} \beta_k} \beta_i \\ &\quad + b'_{g_2}\tau_{\beta_T} \sum_{i=\tau_{\beta_T}+1}^{t} e^{-2\lambda \sum_{k=i+1}^{t} \beta_k} \beta_{i-\tau_{\beta_T}} \beta_i. \end{aligned} \tag{95}$$

The last step in bounding the tracking error is to bound $\mathbb{E}[\langle z_i, \omega^*(\theta_i) - \omega^*(\theta_{i+1})\rangle]$, which is shown in the following lemma.

Lemma 10.

$$\begin{aligned}&\sum_{i=0}^{t} e^{-2\lambda \sum_{k=i+1}^{t} \beta_k} \mathbb{E}[\langle z_i, \omega^*(\theta_i) - \omega^*(\theta_{i+1})\rangle] \\ &\leq 2\frac{1}{\lambda}(1+\gamma+\gamma R|\mathcal{A}|k_1)R(c_{f_1}+c_{g_1}) \sum_{i=0}^{t} e^{-2\lambda \sum_{k=i+1}^{t} \beta_k} \alpha_i. \end{aligned} \tag{96}$$

Proof. From (56), we first have that

$$||\omega^*(\theta_i) - \omega^*(\theta_{i+1})|| \leq \frac{1}{\lambda}(1+\gamma+\gamma R|\mathcal{A}|k_1)||\theta_i - \theta_{i+1}||. \tag{97}$$

Then it follows that

$$\begin{aligned}&\sum_{i=0}^{t} e^{-2\lambda \sum_{k=i+1}^{t} \beta_k} \mathbb{E}[\langle z_i, \omega^*(\theta_i) - \omega^*(\theta_{i+1})\rangle] \\ &\leq \sum_{i=0}^{t} e^{-2\lambda \sum_{k=i+1}^{t} \beta_k} \mathbb{E}[\frac{1}{\lambda}(1+\gamma+\gamma R|\mathcal{A}|k_1)\|z_i\|\|\theta_i - \theta_{i+1}\|] \\ &\leq 2\frac{1}{\lambda}(1+\gamma+\gamma R|\mathcal{A}|k_1)R(c_{f_1}+c_{g_1}) \sum_{i=0}^{t} e^{-2\lambda \sum_{k=i+1}^{t} \beta_k} \alpha_i. \end{aligned} \tag{98}$$

□

B Proof of Theorem 1

In this section, we will use the lemmas in Appendix A to prove Theorem 1.

In Appendix A, we have developed bounds on both the tracking error and $\mathbb{E}[\zeta(\theta_t, O_t)]$. We then plug them both into (60),

$$\begin{aligned}
&\frac{\sum_{t=0}^{T} \alpha_t \mathbb{E}[\|\nabla J(\theta_t)\|^2]}{2\sum_{t=0}^{T} \alpha_t} \\
&\leq \frac{1}{\sum_{t=0}^{T} \alpha_t} \Bigg(J(\theta_0) - J^* + \gamma \alpha_t (1 + |\mathcal{A}| R k_1) \sqrt{\sum_{t=0}^{T} \mathbb{E}[\|\nabla J(\theta_t)\|^2]} \sqrt{\sum_{t=0}^{T} \mathbb{E}[\|z_t\|^2]} \\
&\quad + \sum_{t=0}^{T} \alpha_t \mathbb{E}[\zeta(\theta_t, O_t)] + \sum_{t=0}^{T} \alpha_t^2 (c_{f_1} + c_{g_1}) \Bigg), \qquad (99)
\end{aligned}$$

where J^* denotes $\min_\theta J(\theta)$, and is positive and finite.

By Lemma 6, for large T, we have that

$$\begin{aligned}
&\sum_{t=0}^{T} \alpha_t \mathbb{E}[\zeta(\theta_t, O_t)] \\
&\leq \sum_{t=0}^{\tau_{\alpha_T}} c_\zeta (c_{f_1} + c_{g_1}) \alpha_0 \alpha_t \tau_{\alpha_T} + \sum_{t=\tau_{\alpha_T}+1}^{T} k_\zeta \alpha_T \alpha_t + c_\zeta (c_{f_1} + c_{g_1}) \tau_{\alpha_T} \alpha_{t-\tau_{\alpha_T}} \alpha_t. \qquad (100)
\end{aligned}$$

Here, $\tau_{\alpha_T} = \mathcal{O}(|\log \alpha_T|)$ by its definition. Therefore, for non-increasing sequence $\{\alpha_t\}_{t=0}^{\infty}$, (100) can be further upper bounded as follows:

$$\sum_{t=0}^{T} \alpha_t \mathbb{E}[\zeta(\theta_t, O_t)] = \mathcal{O}\left(|\log \alpha_T|^2 \alpha_0^2 + \sum_{t=0}^{T} \left(\alpha_t \alpha_T + |\log \alpha_T| \alpha_t^2 \right) \right). \qquad (101)$$

We note that we can also specify the constants for (101), which, however, will be cumbersome. How those constants affect the finite-sample bound can be easily inferred from (100), and thus is not explicitly analyzed in the following steps. Also, at the beginning we bound $\sqrt{\frac{\sum_{t=0}^{T} \mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}}$ by some constant that does not scale with T: $\gamma \|C^{-1}\| (k_1 + |\mathcal{A}| R + 1)(r_{\max} + R + \gamma R)$.

Hence, we have that

$$\begin{aligned}
&\frac{\sum_{t=0}^{T} \alpha_t \mathbb{E}[\|\nabla J(\theta_t)\|^2]}{\sum_{t=0}^{T} \alpha_t} \\
&= \mathcal{O}\Bigg(\frac{1}{\sum_{t=0}^{T} \alpha_t} \Bigg(J(\theta_0) - J^* + \sum_{t=0}^{T} \alpha_t^2 + \alpha_t \sqrt{T} \sqrt{\sum_{t=0}^{T} \mathbb{E}[\|z_t\|^2]} + \alpha_0^2 |\log(\alpha_T)|^2 + \sum_{t=0}^{T} \alpha_t \alpha_T \\
&\quad + \sum_{t=0}^{T} |\log(\alpha_T)| \alpha_t^2 \Bigg) \Bigg). \qquad (102)
\end{aligned}$$

In the following, we focus on the case with constant step-sizes. For other possible choices of step-sizes, the convergence rate can also be derived using (102). Let $\alpha_t = \frac{1}{T^a} = \alpha$ and $\beta_t = \frac{1}{T^b} = \beta$. In this case, (102) can be written as follows:

$$\begin{aligned}
\frac{\sum_{t=0}^{T} \alpha \mathbb{E}[\|\nabla J(\theta_t)\|^2]}{\sum_{t=0}^{T} \alpha} &= \mathcal{O}\left(\frac{1}{T} \left(\sqrt{T} \sqrt{\sum_{t=0}^{T} \mathbb{E}[\|z_t\|^2]} + \alpha \log(\alpha)^2 + T\alpha + T\alpha |\log(\alpha)| \right) + \frac{J(\theta_0) - J^*}{T\alpha} \right) \\
&= \mathcal{O}\left(\sqrt{\frac{\sum_{t=0}^{T} \mathbb{E}[\|z_t\|^2]}{T}} \right) + \mathcal{O}\left(\frac{\log T^2}{T^{1+a}} + \frac{1}{T^a} + \frac{\log T}{T^a} + \frac{1}{T^{1-a}} \right). \qquad (103)
\end{aligned}$$

We then consider the tracking error $\mathbb{E}[\|z_t\|^2]$. Applying (92), (94), (95) and (98), we obtain that for $t > \tau_{\beta_T}$,

$$\begin{aligned}
&\mathbb{E}[\|z_t\|^2] \\
&\leq \|z_0\|^2 e^{-2\lambda t\beta} \\
&\quad + 2(4Rc_{f_2}\beta + (a_{f_2}+a_{g_2})\tau_{\beta_T})\beta \sum_{i=0}^{\tau_{\beta_T}} e^{-2\lambda(t-i)\beta} + (8Rc_{f_2}+2b_{g_2})\beta^2 \sum_{i=\tau_{\beta_T}+1}^{t} e^{-2\lambda(t-i)\beta} \\
&\quad + (2b_{f_2}+2b'_{g_2})\tau_{\beta_T}\beta^2 \sum_{i=\tau_{\beta_T}+1}^{t} e^{-2\lambda(t-i)\beta} + \frac{4}{\lambda}(1+\gamma+\gamma R|\mathcal{A}|k_1)R(c_{f_1}+c_{g_1})\alpha \sum_{i=0}^{t} e^{-2\lambda(t-i)\beta} \\
&\quad + 3(c_{f_2}^2 + c_{g_2}^2 + 2\frac{1}{\lambda^2}(1+\gamma+\gamma R|\mathcal{A}|k_1)^2(c_{f_1}^2+c_{g_1}^2)) \sum_{i=0}^{t} e^{-2\lambda(t-i)\beta}\beta^2 \\
&= \mathcal{O}\left(e^{-2\lambda t\beta} + \tau\beta \sum_{i=0}^{\tau} e^{-2\lambda(t-i)\beta} + \tau\beta^2 \sum_{i=1+\tau}^{t} e^{-2\lambda(t-i)\beta} + (\alpha+\beta^2)\sum_{i=0}^{t} e^{-2\lambda(t-i)\beta}\right) \\
&= \mathcal{O}\left(e^{-2\lambda t\beta} + \tau\beta e^{-2\lambda t\beta}\frac{1-e^{2\lambda\beta(\tau+1)}}{1-e^{2\lambda\beta}} + \tau\beta^2(e^{-2\lambda t\beta} - e^{-2\lambda\beta\tau})\frac{e^{2\lambda\beta(\tau+1)}}{1-e^{2\lambda\beta}} + (\alpha+\beta^2)\frac{e^{-2\lambda t\beta}-e^{2\lambda\beta}}{1-e^{2\lambda\beta}}\right).
\end{aligned} \tag{104}$$

Similarly, for $t \leq \tau_{\beta_T}$, we obtain that

$$\begin{aligned}
\mathbb{E}[\|z_t\|^2] &\leq \|z_0\|^2 e^{-2\lambda t\beta} + 2(4Rc_{f_2}\beta + (a_{f_2}+a_{g_2})\tau_{\beta_T})\beta \sum_{i=0}^{t} e^{-2\lambda(t-i)\beta} \\
&\quad + \frac{4}{\lambda}(1+\gamma+\gamma R|\mathcal{A}|k_1)R(c_{f_1}+c_{g_1})\alpha \sum_{i=0}^{t} e^{-2\lambda(t-i)\beta} \\
&\quad + 3(c_{f_2}^2 + c_{g_2}^2 + 2\frac{1}{\lambda^2}(1+\gamma+\gamma R|\mathcal{A}|k_1)^2(c_{f_1}^2+c_{g_1}^2)) \sum_{i=0}^{t} e^{-2\lambda(t-i)\beta}\beta^2 \\
&= \mathcal{O}\left(e^{-2\lambda\beta t} + \tau\beta\sum_{i=0}^{t} e^{-2\lambda(t-i)\beta}\right) = \mathcal{O}\left(e^{-2\lambda\beta t} + \tau\beta\frac{e^{-2\lambda\beta t}-e^{2\lambda\beta}}{1-e^{2\lambda\beta}}\right).
\end{aligned} \tag{105}$$

We then bound $\sum_{t=0}^{T}\mathbb{E}[\|z_t\|^2]$. The sum is divided into two parts: $\sum_{t=0}^{\tau}\mathbb{E}[\|z_t\|^2]$ and $\sum_{t=\tau+1}^{T}\mathbb{E}[\|z_t\|^2]$, thus

$$
\begin{aligned}
&\sum_{t=0}^{T}\mathbb{E}[\|z_t\|^2] \\
&=\sum_{t=0}^{\tau}\mathbb{E}[\|z_t\|^2]+\sum_{t=\tau+1}^{T}\mathbb{E}[\|z_t\|^2] \\
&=\sum_{t=0}^{\tau}(e^{-2\lambda\beta t}+\tau\beta\frac{e^{-2\lambda\beta t}-e^{2\lambda\beta}}{1-e^{2\lambda\beta}})+\sum_{t=\tau+1}^{T}\Bigg(e^{-2\lambda t\beta}+\tau\beta e^{-2\lambda t\beta}\frac{1-e^{2\lambda\beta(\tau+1)}}{1-e^{2\lambda\beta}} \\
&\quad+\tau\beta^2(e^{-2\lambda t\beta}-e^{-2\lambda\beta\tau})\frac{e^{2\lambda\beta(\tau+1)}}{1-e^{2\lambda\beta}}+(\alpha+\beta^2)\frac{e^{-2\lambda t\beta}-e^{2\lambda\beta}}{1-e^{2\lambda\beta}}\Bigg) \\
&=\frac{1-e^{-2\lambda\beta(T+1)}}{1-e^{-2\lambda\beta}}+\tau\beta\left((\tau+1)\frac{-e^{2\lambda\beta}}{1-e^{2\lambda\beta}}+\frac{1-e^{-2\lambda\beta(\tau+1)}}{(1-e^{2\lambda\beta})(1-e^{-2\lambda\beta})}\right) \\
&\quad+\tau\beta\frac{1-e^{2\lambda\beta(\tau+1)}}{1-e^{2\lambda\beta}}e^{-2\lambda\beta(\tau+1)}\frac{1-e^{-2\lambda\beta(T-\tau)}}{1-e^{-2\lambda\beta}}+\tau\beta^2\frac{e^{2\lambda\beta(\tau+1)}}{1-e^{2\lambda\beta}}\Bigg(e^{-2\lambda\beta(\tau+1)}\frac{1-e^{-2\lambda\beta(T-\tau)}}{1-e^{-2\lambda\beta}} \\
&\quad-(T-\tau)e^{-2\lambda\beta\tau}\Bigg)+(\alpha+\beta^2)\frac{1}{1-e^{2\lambda\beta}}\Bigg(e^{-2\lambda\beta(\tau+1)}\frac{1-e^{-2\lambda\beta(T-\tau)}}{1-e^{-2\lambda\beta}}-(T-\tau)e^{2\lambda\beta}\Bigg) \\
&=\mathcal{O}\left(\frac{1}{\beta}+\tau^2+\tau+\tau\beta T+\frac{\alpha+\beta^2}{\beta}T\right). \qquad (106)
\end{aligned}
$$

Thus, we have that

$$
\frac{\sum_{t=0}^{T}\mathbb{E}[\|z_t\|^2]}{T}=\mathcal{O}\left(\frac{1}{T^{1-b}}+\frac{(\log T)^2}{T}+\frac{\log T}{T^b}+\frac{1}{T^{a-b}}+\frac{1}{T^b}\right)=\mathcal{O}\left(\frac{\log T}{T^{\min\{a-b,b\}}}\right). \qquad (107)
$$

We then plug the tracking error (107) in (103), and we have that

$$
\frac{\sum_{t=0}^{T}\alpha\mathbb{E}[\|\nabla J(\theta_t)\|^2]}{\sum_{t=0}^{T}\alpha}=\mathcal{O}\left(\frac{1}{T^{1-a}}\right)+\mathcal{O}\left(\frac{\log T}{T^{\min\{a-b,b\}}}\right). \qquad (108)
$$

In the following we will recursively refine our bounds on the tracking error using the bound in (103).

Recall (65), and denote $D=J(\theta_0-J^*)$, then

$$
\begin{aligned}
\frac{\sum_{t=0}^{T}\mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}&=\frac{D}{T\alpha}+\mathcal{O}\left(\frac{\sum_{t=0}^{T}\sqrt{\mathbb{E}[\|\nabla J(\theta_t)\|^2]\mathbb{E}[\|z_t\|^2]}}{T}\right) \\
&=\mathcal{O}\left(\frac{1}{T\alpha}+\sqrt{\frac{\sum_{t=0}^{T}\mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}}\sqrt{\frac{\sum_{t=0}^{T}\mathbb{E}[\|z_t\|^2]}{T}}\right). \qquad (109)
\end{aligned}
$$

In the first round, we upper bound $\frac{\sum_{t=0}^{T}\mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}$ by a constant. It then follows that

$$
\frac{\sum_{t=0}^{T}\mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}=\mathcal{O}\left(\frac{1}{T^{1-a}}\right)+\sqrt{\mathcal{O}\left(\frac{\log T}{T^b}+\frac{1}{T^{a-b}}\right)}=\mathcal{O}\left(\frac{1}{T^{1-a}}\right)+\mathcal{O}\left(\frac{\sqrt{\log T}}{T^{\min\{b/2,a/2-b/2\}}}\right), \qquad (110)
$$

where we denote $\min\{b/2,a/2-b/2\}$ by $c/2$. We then plug (110) into (109), and we obtain that

$$
\frac{\sum_{t=0}^{T}\mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}=\mathcal{O}\left(\frac{1}{T^{1-a}}\right)+\mathcal{O}\left(\frac{\sqrt{\log T}}{T^{c/2}}\sqrt{\frac{\sum_{t=0}^{T}\mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}}\right). \qquad (111)
$$

Case 1. If $1-a<c/2$, then bound in (110) is $\mathcal{O}\left(\frac{1}{T^{1-a}}\right)$: $\frac{\sum_{t=0}^{T}\mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}=\mathcal{O}\left(\frac{1}{T^{1-a}}\right)$. Then

$$\frac{\sum_{t=0}^{T}\mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}=\mathcal{O}\left(\frac{1}{T^{1-a}}+\frac{\sqrt{\log T}}{T^{c/2}}\frac{1}{T^{1/2-a/2}}\right). \tag{112}$$

Note that $c/2>1-a$, then $c/2+1/2-a/2>1-a$, thus the order would be

$$\frac{\sum_{t=0}^{T}\mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}=\mathcal{O}\left(\frac{1}{T^{1-a}}\right). \tag{113}$$

Therefore, such a recursive refinement will not improve the convergence rate if $1-a<\frac{c}{2}$.

Case 2. If $c>1-a\geq c/2$, then

$$\frac{\sum_{t=0}^{T}\mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}=\mathcal{O}\left(\frac{\sqrt{\log T}}{T^{c/2}}\right). \tag{114}$$

Also plug this order in (109), and we obtain that

$$\frac{\sum_{t=0}^{T}\mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}=\mathcal{O}\left(\frac{1}{T^{1-a}}\right)+\mathcal{O}\left(\frac{\sqrt{\log T}}{T^{c/2}}\frac{(\log T)^{1/4}}{T^{c/4}}\right)=\mathcal{O}\left(\frac{1}{T^{1-a}}+\frac{(\log T)^{\frac{3}{4}}}{T^{3c/4}}\right). \tag{115}$$

Here, we start the second iteration. If $1-a\geq\frac{3c}{4}$, we know that the order is improved as follows

$$\frac{\sum_{t=0}^{T}\mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}=\mathcal{O}\left(\frac{(\log T)^{\frac{3}{4}}}{T^{3c/4}}\right). \tag{116}$$

And if $1-a<\frac{3c}{4}$, then order of (110) will still be $\mathcal{O}\left(\frac{1}{T^{1-a}}\right)$. Thus we will stop the recursion, and we have that

$$\frac{\sum_{t=0}^{T}\mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T}=\mathcal{O}\left(\frac{1}{T^{1-a}}\right). \tag{117}$$

This implies that if the recursion stops after some step until there is no further rate improvement, then the convergence rate will be $\mathcal{O}\left(\frac{1}{T^{1-a}}\right)$. Note in this case, since $1-a<c$, then there exists some integral n, such that $1-a<\frac{2^n-1}{2^n}c$, and after round n, the recursion will stop. Thus the final rate is $\mathcal{O}\left(\frac{1}{T^{1-a}}\right)$.

Case 3. If $1-a\geq c$, then after a number of recursions, the order of the bound will be sufficiently close to $\mathcal{O}\left(\frac{\log T}{T^c}\right)$.

To conclude the three cases, when $1-a<c$, the recursion will stop after finite number of iterations, and the rate would be $\mathcal{O}\left(\frac{1}{T^{1-a}}\right)$; While when $1-a\geq c$, the recursion will always continue, and the fastest rate we can obtain is $\mathcal{O}\left(\frac{\log T}{T^c}\right)$. Thus the overall rate we can obtain can be written as

$$\mathcal{O}\left(\frac{1}{T^{1-a}}+\frac{\log T}{T^c}\right). \tag{118}$$

B.1 Proof of Corollary 1

We next look for suitable a and b, such that the rate obtained is the fastest. It can be seen that the best rate is achieved when $1-a=c$, and at the same time $0.5<a\le 1$ and $0<b<a$. Thus, the best choices are $a=\frac{2}{3}$ and $b=\frac{1}{3}$, and the best rate we can obtain is

$$\mathbb{E}[\|\nabla J(\theta_M)\|^2] = \frac{\sum_{t=0}^{T}\mathbb{E}[\|\nabla J(\theta_t)\|^2]}{T} = \mathcal{O}\left(\frac{\log T}{T^{1-a}}\right) = \mathcal{O}\left(\frac{\log T}{T^{\frac{1}{3}}}\right). \tag{119}$$

C Softmax Is Lipschitz and Smooth

We first restate Lemma 1 as follows, and then derive its proof.

Lemma 11. *The softmax policy is 2σ-Lipschitz and $8\sigma^2$-smooth, i.e., for any $(s,a)\in\mathcal{S}\times\mathcal{A}$, and for any $\theta_1,\theta_2\in\mathbb{R}^N$, $|\pi_{\theta_1}(a|s)-\pi_{\theta_2}(a|s)|\le 2\sigma\|\theta_1-\theta_2\|$ and $\|\nabla\pi_{\theta_1}(a|s)-\nabla\pi_{\theta_2}(a|s)\|\le 8\sigma^2\|\theta_1-\theta_2\|$.*

Proof. By the definition of the softmax policy, for any $a\in\mathcal{A}$, $s\in\mathcal{S}$ and $\theta\in\mathbb{R}^N$,

$$\pi_\theta(a|s) = \frac{e^{\sigma\theta^\top\phi_{s,a}}}{\sum_{a'\in\mathcal{A}} e^{\sigma\theta^\top\phi_{s,a'}}}, \tag{120}$$

where $\sigma>0$ is a constant. Then, it can be shown that

$$\begin{aligned}\nabla\pi_\theta(a|s) &= \frac{1}{\left(\sum_{a'\in\mathcal{A}} e^{\sigma\theta^\top\phi_{s,a'}}\right)^2}\left(\sigma e^{\sigma\theta^\top\phi_{s,a}}\phi_{s,a}\left(\sum_{a'\in\mathcal{A}} e^{\sigma\theta^\top\phi_{s,a'}}\right) - \left(\sum_{a'\in\mathcal{A}}\sigma e^{\sigma\theta^\top\phi_{s,a'}}\phi_{s,a'}\right)e^{\sigma\theta^\top\phi_{s,a}}\right)\\ &= \frac{\sigma}{(\sum_{a'\in\mathcal{A}} e^{\sigma\theta^\top\phi_{s,a'}})^2}\left(\sum_{a'\in\mathcal{A}}\phi_{s,a}e^{\sigma\theta^\top(\phi_{s,a}+\phi_{s,a'})} - \phi_{s,a'}e^{\sigma\theta^\top(\phi_{s,a}+\phi_{s,a'})}\right)\\ &= \frac{\sigma\sum_{a'\in\mathcal{A}}(\phi_{s,a}-\phi_{s,a'})e^{\sigma\theta^\top(\phi_{s,a}+\phi_{s,a'})}}{(\sum_{a'\in\mathcal{A}} e^{\sigma\theta^\top\phi_{s,a'}})^2}. \end{aligned}\tag{121}$$

Thus,

$$||\nabla\pi_\theta(a|s)|| \le 2\sigma\frac{\sum_{a'\in\mathcal{A}} e^{\sigma\theta^\top(\phi_{s,a}+\phi_{s,a'})}}{\left(\sum_{a'\in\mathcal{A}} e^{\sigma\theta^\top\phi_{s,a'}}\right)^2} = 2\sigma\frac{e^{\sigma\theta^\top\phi_{s,a}}}{\sum_{a'\in\mathcal{A}} e^{\sigma\theta^\top\phi_{s,a'}}} \le 2\sigma, \tag{122}$$

where the last step is due to the fact that $\frac{e^{\sigma\theta^\top\phi_{s,a}}}{\sum_{a'\in\mathcal{A}} e^{\sigma\theta^\top\phi_{s,a'}}}\le 1$.

Note that for any θ_1 and θ_2, there exists some $\alpha\in(0,1)$ and $\bar\theta=\alpha\theta_1+(1-\alpha)\theta_2$, such that

$$\|\nabla\pi_{\theta_1}(a|s)-\nabla\pi_{\theta_2}(a|s)\| \le \|\nabla^2\pi_{\bar\theta}(a|s)\|\times\|\theta_1-\theta_2\|, \tag{123}$$

which follows from the mean-value theorem. Here, $\nabla^2\pi_\theta(a|s)$ denotes the Hessian matrix of $\pi_\theta(a|s)$ at θ. Thus it suffices to find an universal bound of $\|\nabla^2\pi_\theta(a|s)\|$ for any θ and $(a,s)\in\mathcal{A}\times\mathcal{S}$.

Note that $\nabla\pi_\theta(a|s) = \frac{\sigma\sum_{a'\in\mathcal{A}}(\phi_{s,a}-\phi_{s,a'})e^{\sigma\theta^\top(\phi_{s,a}+\phi_{s,a'})}}{\left(\sum_{a'\in\mathcal{A}} e^{\sigma\theta^\top\phi_{s,a'}}\right)^2}$ is a sum of vectors $(\phi_{s,a}-\phi_{s,a'})$ with each entry multiplied by $\frac{\sigma e^{\sigma\theta^\top(\phi_{s,a}+\phi_{s,a'})}}{\left(\sum_{a'\in\mathcal{A}} e^{\sigma\theta^\top\phi_{s,a'}}\right)^2}$. Then it follows that

$$\nabla^2\pi_\theta(a|s) = \sigma\sum_{a'\in\mathcal{A}}(\phi_{s,a}-\phi_{s,a'})\left(\nabla\frac{e^{\sigma\theta^\top(\phi_{s,a}+\phi_{s,a'})}}{\left(\sum_{a'\in\mathcal{A}} e^{\sigma\theta^\top\phi_{s,a'}}\right)^2}\right)^\top. \tag{124}$$

Thus, to bound $\|\nabla^2\pi_\theta(a|s)\|$, we compute the following:

$$\begin{aligned}
&\nabla \frac{e^{\sigma\theta^\top(\phi_{s,a}+\phi_{s,a'})}}{(\sum_{a'\in\mathcal{A}} e^{\sigma\theta^\top\phi_{s,a'}})^2}\\
&=\sigma\frac{e^{\sigma\theta^\top(\phi_{s,a}+\phi_{s,a'})}\left((\sum_{a'\in\mathcal{A}} e^{\sigma\theta^\top\phi_{s,a'}})(\phi_{s,a}+\phi_{s,a'})-2(\sum_{a'\in\mathcal{A}} e^{\sigma\theta^\top\phi_{s,a'}}\phi_{s,a'})\right)}{(\sum_{a'\in\mathcal{A}} e^{\sigma\theta^\top\phi_{s,a'}})^3}. \qquad (125)
\end{aligned}$$

Then the norm of (125) can be bounded as follows:

$$\begin{aligned}
&\left\|\nabla\left(\frac{e^{\sigma\theta^\top(\phi_{s,a}+\phi_{s,a'})}}{(\sum_{a'\in\mathcal{A}} e^{\sigma\theta^\top\phi_{s,a'}})^2}\right)\right\|\\
&\leq\sigma\frac{2e^{\sigma\theta^\top(\phi_{s,a}+\phi_{s,a'})}\left(\sum_{a'\in\mathcal{A}} e^{\sigma\theta^\top\phi_{s,a'}}+(\sum_{a'\in\mathcal{A}} e^{\sigma\theta^\top\phi_{s,a'}})\right)}{(\sum_{a'\in\mathcal{A}} e^{\theta^\top\phi_{s,a'}})^3}\\
&=4\sigma\frac{e^{\sigma\theta^\top(\phi_{s,a}+\phi_{s,a'})}}{\left(\sum_{a'\in\mathcal{A}} e^{\sigma\theta^\top\phi_{s,a'}}\right)^2}\\
&\leq 4\sigma. \qquad (126)
\end{aligned}$$

Plug this in the expression of $\nabla^2\pi_\theta(a|s)$, we obtain that

$$||\nabla^2\pi_\theta(a|s)||\leq 8\sigma^2. \qquad (127)$$

Thus the softmax policy is 2σ-Lipschitz and $8\sigma^2$-smooth. This completes the proof. □