

Supplementary Material

A PRELIMINARY LEMMAS

A.1 GEOMETRIC MIXING

The operation $p \otimes q$ denotes the tensor product between two distributions $p(x)$ and $q(y)$, i.e. $(p \otimes q)(x, y) = p(x) \cdot q(y)$.

Lemma 1. *Suppose Assumption 4 holds for a Markov chain generated by the rule $a_t \sim \pi_\theta(\cdot|s_t)$, $s_{t+1} \sim \tilde{\mathcal{P}}(\cdot|s_t, a_t)$. For any $\theta \in \mathbb{R}^d$, we have*

$$\sup_{s_0 \in \mathcal{S}} d_{TV} \left(\mathbb{P}((s_t, a_t, s_{t+1}) \in \cdot | s_0, \pi_\theta), \mu_\theta \otimes \pi_\theta \otimes \tilde{\mathcal{P}} \right) \leq \kappa \rho^t. \quad (19)$$

where $\mu_\theta(\cdot)$ is the stationary distribution with policy π_θ and transition kernel $\tilde{\mathcal{P}}(\cdot|s, a)$.

Proof. We start with

$$\begin{aligned} & \sup_{s_0 \in \mathcal{S}} d_{TV} \left(\mathbb{P}((s_t, a_t, s_{t+1}) \in \cdot | s_0, \pi_\theta), \mu_\theta \otimes \pi_\theta \otimes \tilde{\mathcal{P}} \right) \\ &= \sup_{s_0 \in \mathcal{S}} d_{TV} \left(\mathbb{P}(s_t = \cdot | s_0, \pi_\theta) \otimes \pi_\theta \otimes \tilde{\mathcal{P}}, \mu_\theta \otimes \pi_\theta \otimes \tilde{\mathcal{P}} \right) \\ &= \sup_{s_0 \in \mathcal{S}} \frac{1}{2} \int_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \int_{s' \in \mathcal{S}} \left| \mathbb{P}(s_t = ds | s_0, \pi_\theta) \pi_\theta(a|s) \tilde{\mathcal{P}}(ds' | s, a) - \mu_\theta(ds) \pi_\theta(a|s) \tilde{\mathcal{P}}(ds' | s, a) \right| \\ &= \sup_{s_0 \in \mathcal{S}} \frac{1}{2} \int_{s \in \mathcal{S}} \left| \mathbb{P}(s_t = ds | s_0, \pi_\theta) - \mu_\theta(ds) \right| \sum_{a \in \mathcal{A}} \pi_\theta(a|s) \int_{s' \in \mathcal{S}} \tilde{\mathcal{P}}(ds' | s, a) \\ &= \sup_{s_0 \in \mathcal{S}} d_{TV} (\mathbb{P}(s_t \in \cdot | s_0, \pi_\theta), \mu_\theta) \\ &\leq \kappa \rho^t, \end{aligned}$$

which completes the proof. \square

For the use in the later proof, given $K > 0$, we first define m_K as:

$$m_K := \min \{ m \in \mathbb{N}^+ \mid \kappa \rho^{m-1} \leq \min\{\alpha_k, \beta_k\} \}, \quad (20)$$

where κ and ρ are constants defined in (4). m_K is the minimum number of samples needed for the Markov chain to approach the stationary distribution so that the bias incurred by the Markovian sampling is small enough.

A.2 AUXILIARY MARKOV CHAIN

The auxiliary Markov chain is a virtual Markov chain with no policy drifting — a technique developed in [35] to analyze stochastic approximation algorithms in non-stationary settings.

Lemma 2. *Under Assumption 1 and Assumption 3, consider the update (9) in Algorithm 1 with Markovian sampling. For a given number of samples m , consider the Markov chain of the worker that contributes to the k th update:*

$$s_{t-m} \xrightarrow{\theta_{k-d_m}} a_{t-m} \xrightarrow{\tilde{\mathcal{P}}} s_{t-m+1} \xrightarrow{\theta_{k-d_{m-1}}} a_{t-m+1} \cdots s_{t-1} \xrightarrow{\theta_{k-d_1}} a_{t-1} \xrightarrow{\tilde{\mathcal{P}}} s_t \xrightarrow{\theta_{k-d_0}} a_t \xrightarrow{\tilde{\mathcal{P}}} s_{t+1},$$

where $(s_t, a_t, s_{t+1}) = (s^{(k)}, a^{(k)}, s'^{(k)})$, and $\{d_j\}_{j=0}^m$ is some increasing sequence with $d_0 := \tau_k$.

Given $(s_{t-m}, a_{t-m}, s_{t-m+1})$ and θ_{k-d_m} , we construct its auxiliary Markov chain by repeatedly applying $\pi_{\theta_{k-d_m}}$:

$$s_{t-m} \xrightarrow{\theta_{k-d_m}} a_{t-m} \xrightarrow{\tilde{\mathcal{P}}} s_{t-m+1} \xrightarrow{\theta_{k-d_m}} \tilde{a}_{t-m+1} \cdots \tilde{s}_{t-1} \xrightarrow{\theta_{k-d_m}} \tilde{a}_{t-1} \xrightarrow{\tilde{\mathcal{P}}} \tilde{s}_t \xrightarrow{\theta_{k-d_m}} \tilde{a}_t \xrightarrow{\tilde{\mathcal{P}}} \tilde{s}_{t+1}.$$

Define $x_t := (s_t, a_t, s_{t+1})$, then we have:

$$\begin{aligned} & d_{TV} (\mathbb{P}(x_t \in \cdot | \theta_{k-d_m}, s_{t-m+1}), \mathbb{P}(\tilde{x}_t \in \cdot | \theta_{k-d_m}, s_{t-m+1})) \\ & \leq \frac{1}{2} |\mathcal{A}| L_\pi \sum_{i=\tau_k}^{d_m} \mathbb{E} [\|\theta_{k-i} - \theta_{k-d_m}\|_2 | \theta_{k-d_m}, s_{t-m+1}]. \end{aligned} \quad (21)$$

Proof. Throughout the lemma, all expectations and probabilities are conditioned on θ_{k-d_m} and s_{t-m+1} . We omit this condition for convenience.

First we have

$$\begin{aligned}
& d_{TV}(\mathbb{P}(s_{t+1} \in \cdot), \mathbb{P}(\tilde{s}_{t+1} \in \cdot)) \\
&= \frac{1}{2} \int_{s' \in \mathcal{S}} |\mathbb{P}(s_{t+1} = ds') - \mathbb{P}(\tilde{s}_{t+1} = ds')| \\
&= \frac{1}{2} \int_{s' \in \mathcal{S}} \left| \int_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \mathbb{P}(s_t = ds, a_t = a, s_{t+1} = ds') - \mathbb{P}(\tilde{s}_t = ds, \tilde{a}_t = a, \tilde{s}_{t+1} = ds') \right| \\
&\leq \frac{1}{2} \int_{s' \in \mathcal{S}} \int_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} |\mathbb{P}(s_t = ds, a_t = a, s_{t+1} = ds') - \mathbb{P}(\tilde{s}_t = ds, \tilde{a}_t = a, \tilde{s}_{t+1} = ds')| \\
&= \frac{1}{2} \int_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \int_{s' \in \mathcal{S}} |\mathbb{P}(s_t = ds, a_t = a, s_{t+1} = ds') - \mathbb{P}(\tilde{s}_t = ds, \tilde{a}_t = a, \tilde{s}_{t+1} = ds')| \\
&= d_{TV}(\mathbb{P}(x_t \in \cdot), \mathbb{P}(\tilde{x}_t \in \cdot)), \tag{22}
\end{aligned}$$

where the last second equality is due to Tonelli's theorem. Next we have

$$\begin{aligned}
& d_{TV}(\mathbb{P}(x_t \in \cdot), \mathbb{P}(\tilde{x}_t \in \cdot)) \\
&= \frac{1}{2} \int_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \int_{s' \in \mathcal{S}} |\mathbb{P}(s_t = ds, a_t = a, s_{t+1} = ds') - \mathbb{P}(\tilde{s}_t = ds, \tilde{a}_t = a, \tilde{s}_{t+1} = ds')| \\
&= \frac{1}{2} \int_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} |\mathbb{P}(s_t = ds, a_t = a) - \mathbb{P}(\tilde{s}_t = ds, \tilde{a}_t = a)| \int_{s' \in \mathcal{S}} \tilde{\mathcal{P}}(s_{t+1} = ds' | s_t = ds, a_t = a) \\
&= \frac{1}{2} \int_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} |\mathbb{P}(s_t = ds, a_t = a) - \mathbb{P}(\tilde{s}_t = ds, \tilde{a}_t = a)| \\
&= d_{TV}(\mathbb{P}((s_t, a_t) \in \cdot), \mathbb{P}((\tilde{s}_t, \tilde{a}_t) \in \cdot)). \tag{23}
\end{aligned}$$

Due to the fact that $\theta_{k-\tau_k}$ is dependent on s_t , we need to write $\mathbb{P}(s_t, a_t)$ as

$$\begin{aligned}
\mathbb{P}(s_t, a_t) &= \int_{\theta_{k-\tau_k} \in \mathbb{R}^d} \mathbb{P}(s_t, \theta_{k-\tau_k}, a_t) \\
&= \int_{\theta \in \mathbb{R}^d} \mathbb{P}(s_t) \mathbb{P}(\theta_{k-\tau_k} = d\theta | s_t) \pi_{\theta_{k-\tau_k}}(a_t | s_t) \\
&= \mathbb{P}(s_t) \int_{\theta \in \mathbb{R}^d} \mathbb{P}(\theta_{k-\tau_k} = d\theta | s_t) \pi_{\theta_{k-\tau_k}}(a_t | s_t) \\
&= \mathbb{P}(s_t) \mathbb{E}[\pi_{\theta_{k-\tau_k}}(a_t | s_t) | s_t].
\end{aligned}$$

Then we have

$$\begin{aligned}
& d_{TV}(\mathbb{P}((s_t, a_t) \in \cdot), \mathbb{P}((\tilde{s}_t, \tilde{a}_t) \in \cdot)) \\
&= \frac{1}{2} \int_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \left| \mathbb{P}(s_t = ds) \mathbb{E}[\pi_{\theta_{k-\tau_k}}(a_t = a | s_t = ds) | s_t = ds] - \mathbb{P}(\tilde{s}_t = ds) \pi_{\theta_{k-d_m}}(\tilde{a}_t = a | \tilde{s}_t = ds) \right| \\
&\leq \frac{1}{2} \int_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \left| \mathbb{P}(s_t = ds) \mathbb{E}[\pi_{\theta_{k-\tau_k}}(a_t = a | s_t = ds) | s_t = ds] - \mathbb{P}(s_t = ds) \pi_{\theta_{k-d_m}}(a_t = a | s_t = ds) \right| \\
&\quad + \frac{1}{2} \int_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \left| \mathbb{P}(s_t = ds) \pi_{\theta_{k-d_m}}(\tilde{a}_t = a | \tilde{s}_t = ds) - \mathbb{P}(\tilde{s}_t = ds) \pi_{\theta_{k-d_m}}(\tilde{a}_t = a | \tilde{s}_t = ds) \right| \\
&= \frac{1}{2} \int_{s \in \mathcal{S}} \mathbb{P}(s_t = ds) \sum_{a \in \mathcal{A}} \left| \mathbb{E}[\pi_{\theta_{k-\tau_k}}(a_t = a | s_t = ds) | s_t = ds] - \pi_{\theta_{k-d_m}}(a_t = a | s_t = ds) \right| \\
&\quad + \frac{1}{2} \int_{s \in \mathcal{S}} |\mathbb{P}(s_t = ds) - \mathbb{P}(\tilde{s}_t = ds)|.
\end{aligned}$$

Using Jensen's inequality, we have

$$\begin{aligned}
& d_{TV}(\mathbb{P}((s_t, a_t) \in \cdot), \mathbb{P}((\tilde{s}_t, \tilde{a}_t) \in \cdot)) \\
& \leq \frac{1}{2} \int_{s \in \mathcal{S}} \mathbb{P}(s_t = ds) \sum_{a \in \mathcal{A}} \mathbb{E} \left[\left| \pi_{\theta_{k-\tau_k}}(a_t = a | s_t = ds) - \pi_{\theta_{k-d_m}}(a_t = a | s_t = ds) \right| \middle| s_t = ds \right] \\
& \quad + \frac{1}{2} \int_{s \in \mathcal{S}} |\mathbb{P}(s_t = ds) - \mathbb{P}(\tilde{s}_t = ds)| \\
& \leq \frac{1}{2} \int_{s \in \mathcal{S}} \mathbb{P}(s_t = ds) \sum_{a \in \mathcal{A}} \mathbb{E} [\|\theta_{k-\tau_k} - \theta_{k-d_m}\|_2 | s_t = ds] + \frac{1}{2} \int_{s \in \mathcal{S}} |\mathbb{P}(s_t = ds) - \mathbb{P}(\tilde{s}_t = ds)| \\
& = \frac{1}{2} |\mathcal{A}| L_\pi \mathbb{E} \|\theta_{k-\tau_k} - \theta_{k-d_m}\|_2 + d_{TV}(\mathbb{P}(s_t \in \cdot), \mathbb{P}(\tilde{s}_t \in \cdot)) \tag{24}
\end{aligned}$$

where the last inequality follows Assumption 3.

Now we start to prove (21).

$$\begin{aligned}
d_{TV}(\mathbb{P}(x_t \in \cdot), \mathbb{P}(\tilde{x}_t \in \cdot)) & \stackrel{(23)}{=} d_{TV}(\mathbb{P}((s_t, a_t) \in \cdot), \mathbb{P}((\tilde{s}_t, \tilde{a}_t) \in \cdot)) \\
& \stackrel{(24)}{\leq} d_{TV}(\mathbb{P}(s_t \in \cdot), \mathbb{P}(\tilde{s}_t \in \cdot)) + \frac{1}{2} |\mathcal{A}| L_\pi \mathbb{E} \|\theta_{k-\tau_k} - \theta_{k-d_m}\|_2 \\
& \stackrel{(22)}{\leq} d_{TV}(\mathbb{P}(x_{t-1} \in \cdot), \mathbb{P}(\tilde{x}_{t-1} \in \cdot)) + \frac{1}{2} |\mathcal{A}| L_\pi \mathbb{E} \|\theta_{k-\tau_k} - \theta_{k-d_m}\|_2.
\end{aligned}$$

Now we have

$$d_{TV}(\mathbb{P}(x_t \in \cdot), \mathbb{P}(\tilde{x}_t \in \cdot)) \leq d_{TV}(\mathbb{P}(x_{t-1} \in \cdot), \mathbb{P}(\tilde{x}_{t-1} \in \cdot)) + \frac{1}{2} |\mathcal{A}| L_\pi \mathbb{E} \|\theta_{k-\tau_k} - \theta_{k-d_m}\|_2. \tag{25}$$

Since $d_{TV}(\mathbb{P}(x_{t-m} \in \cdot), \mathbb{P}(x_{t-m} \in \cdot)) = 0$, recursively applying (25) for $\{t-1, \dots, t-m\}$ gives

$$\begin{aligned}
d_{TV}(\mathbb{P}(x_t \in \cdot), \mathbb{P}(\tilde{x}_t \in \cdot)) & \leq \frac{1}{2} |\mathcal{A}| L_\pi \sum_{j=0}^m \mathbb{E} \|\theta_{k-d_j} - \theta_{k-d_m}\|_2 \\
& \leq \frac{1}{2} |\mathcal{A}| L_\pi \sum_{i=\tau_k}^{d_m} \mathbb{E} \|\theta_{k-i} - \theta_{k-d_m}\|_2,
\end{aligned}$$

which completes the proof. \square

A.3 LIPSCHITZ CONTINUITY OF VALUE FUNCTION

Lemma 3. *Suppose Assumption 3 holds. For any $\theta_1, \theta_2 \in \mathbb{R}^d$ and $s \in \mathcal{S}$, we have*

$$\|\nabla V_{\pi_{\theta_1}}(s)\|_2 \leq L_V, \tag{26a}$$

$$|V_{\pi_{\theta_1}}(s) - V_{\pi_{\theta_2}}(s)| \leq L_V \|\theta_1 - \theta_2\|_2, \tag{26b}$$

where the constant is $L_V := C_\psi r_{\max}/(1-\gamma)$ with C_ψ defined as in Assumption 3.

Proof. First we have

$$\begin{aligned}
Q_\pi(s, a) & = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}) | s_0 = s, a_0 = a \right] \\
& \leq \sum_{t=0}^{\infty} \gamma^t r_{\max} = \frac{r_{\max}}{1-\gamma}.
\end{aligned}$$

By the policy gradient theorem [7], we have

$$\begin{aligned}\|\nabla V_{\pi_{\theta_1}}(s)\|_2 &= \|\mathbb{E}[Q_{\pi_{\theta_1}}(s, a)\psi_{\theta_1}(s, a)]\|_2 \\ &\leq \mathbb{E}\|Q_{\pi_{\theta_1}}(s, a)\psi_{\theta_1}(s, a)\|_2 \\ &\leq \mathbb{E}[|Q_{\pi_{\theta_1}}(s, a)|\|\psi_{\theta_1}(s, a)\|_2] \\ &\leq \frac{r_{\max}}{1-\gamma}C_\psi,\end{aligned}$$

where the first inequality is due to Jensen's inequality, and the last inequality follows Assumption 3 and the fact that $Q_\pi(s, a) \leq \frac{r_{\max}}{1-\gamma}$. By the mean value theorem, we immediately have

$$\|V_{\pi_{\theta_1}}(s) - V_{\pi_{\theta_2}}(s)\| \leq \sup_{\theta_1 \in \mathbb{R}^d} \|\nabla V_{\pi_{\theta_1}}(s)\|_2 \|\theta_1 - \theta_2\|_2 = L_V \|\theta_1 - \theta_2\|_2,$$

which completes the proof. \square

A.4 LIPSCHITZ CONTINUITY OF POLICY GRADIENT

We give a proposition regarding the L_J -Lipschitz of the policy gradient under proper assumptions, which has been shown by [33].

Proposition 1. *Suppose Assumption 3 and 4 hold. For any $\theta, \theta' \in \mathbb{R}^d$, we have $\|\nabla J(\theta) - \nabla J(\theta')\|_2 \leq L_J \|\theta - \theta'\|_2$, where L_J is a positive constant.*

A.5 LIPSCHITZ CONTINUITY OF OPTIMAL CRITIC PARAMETER

We provide a justification for Lipschitz continuity of ω_θ^* in the next proposition.

Proposition 2. *Suppose Assumption 3 and 4 hold. For any $\theta_1, \theta_2 \in \mathbb{R}^d$, we have*

$$\|\omega_{\theta_1}^* - \omega_{\theta_2}^*\|_2 \leq L_\omega \|\theta_1 - \theta_2\|_2,$$

where $L_\omega := 2r_{\max}|\mathcal{A}|L_\pi(\lambda^{-1} + \lambda^{-2}(1 + \gamma))(1 + \log_\rho \kappa^{-1} + (1 - \rho)^{-1})$.

Proof. We use A_1, A_2, b_1 and b_2 as shorthand notations of $A_{\pi_{\theta_1}}, A_{\pi_{\theta_2}}, b_{\pi_{\theta_1}}$ and $b_{\pi_{\theta_2}}$ respectively. By Assumption 2, $A_{\theta, \phi}$ is invertible for any $\theta \in \mathbb{R}^d$, so we can write $\omega_\theta^* = -A_{\theta, \phi}^{-1}b_{\theta, \phi}$. Then we have

$$\begin{aligned}\|\omega_1^* - \omega_2^*\|_2 &= \|-A_1^{-1}b_1 + A_2^{-1}b_2\|_2 \\ &= \|-A_1^{-1}b_1 - A_1^{-1}b_2 + A_1^{-1}b_2 + A_2^{-1}b_2\|_2 \\ &= \|-A_1^{-1}(b_1 - b_2) - (A_1^{-1} - A_2^{-1})b_2\|_2 \\ &\leq \|A_1^{-1}(b_1 - b_2)\|_2 + \|(A_1^{-1} - A_2^{-1})b_2\|_2 \\ &\leq \|A_1^{-1}\|_2 \|b_1 - b_2\|_2 + \|A_1^{-1} - A_2^{-1}\|_2 \|b_2\|_2 \\ &= \|A_1^{-1}\|_2 \|b_1 - b_2\|_2 + \|A_1^{-1}(A_2 - A_1)A_2^{-1}\|_2 \|b_2\|_2 \\ &\leq \|A_1^{-1}\|_2 \|b_1 - b_2\|_2 + \|A_1^{-1}\|_2 \|A_2^{-1}\|_2 \|b_2\|_2 \|(A_2 - A_1)\|_2 \\ &\leq \lambda^{-1} \|b_1 - b_2\|_2 + \lambda^{-2} r_{\max} \|A_1 - A_2\|_2,\end{aligned}\tag{27}$$

where the last inequality follows Assumption 2, and the fact that

$$\|b_2\|_2 = \|\mathbb{E}[r(s, a, s')\phi(s)]\|_2 \leq \mathbb{E}\|r(s, a, s')\phi(s)\|_2 \leq \mathbb{E}[|r(s, a, s')|\|\phi(s)\|_2] \leq r_{\max}.$$

Denote (s^1, a^1, s'^1) and (s^2, a^2, s'^2) as samples drawn with θ_1 and θ_2 respectively, i.e. $s^1 \sim \mu_{\theta_1}$, $a^1 \sim \pi_{\theta_1}$, $s'^1 \sim \tilde{\mathcal{P}}$ and $s^2 \sim \mu_{\theta_2}$, $a^2 \sim \pi_{\theta_2}$, $s'^2 \sim \tilde{\mathcal{P}}$. Then we have

$$\begin{aligned}\|b_1 - b_2\|_2 &= \|\mathbb{E}[r(s^1, a^1, s'^1)\phi(s^1)] - \mathbb{E}[r(s^2, a^2, s'^2)\phi(s^2)]\|_2 \\ &\leq \sup_{s, a, s'} \|r(s, a, s')\phi(s)\|_2 \|\mathbb{P}((s^1, a^1, s'^1) \in \cdot) - \mathbb{P}((s^2, a^2, s'^2) \in \cdot)\|_{TV} \\ &\leq r_{\max} \|\mathbb{P}((s^1, a^1, s'^1) \in \cdot) - \mathbb{P}((s^2, a^2, s'^2) \in \cdot)\|_{TV} \\ &= 2r_{\max} d_{TV}(\mu_{\theta_1} \otimes \pi_{\theta_1} \otimes \tilde{\mathcal{P}}, \mu_{\theta_2} \otimes \pi_{\theta_2} \otimes \tilde{\mathcal{P}}) \\ &\leq 2r_{\max} |\mathcal{A}| L_\pi (1 + \log_\rho \kappa^{-1} + (1 - \rho)^{-1}) \|\theta_1 - \theta_2\|_2,\end{aligned}\tag{28}$$

where the first inequality follows the definition of total variation (TV) norm, and the last inequality follows Lemma A.1. in [17]. Similarly we have:

$$\begin{aligned} \|A_1 - A_2\|_2 &\leq 2(1 + \gamma)d_{TV}(\mu_{\theta_1} \otimes \pi_{\theta_1}, \mu_{\theta_2} \otimes \pi_{\theta_2}) \\ &= (1 + \gamma)|\mathcal{A}|L_\pi(1 + \log_\rho \kappa^{-1} + (1 - \rho)^{-1})\|\theta_1 - \theta_2\|_2. \end{aligned} \quad (29)$$

Substituting (28) and (29) into (27) completes the proof. \square

B PROOF OF MAIN THEOREMS

B.1 PROOF OF THEOREM 1

For brevity, we first define the following notations:

$$\begin{aligned} x &:= (s, a, s'), \\ \hat{\delta}(x, \omega) &:= r(s, a, s') + \gamma\phi(s')^\top \omega - \phi(s)^\top \omega, \\ g(x, \omega) &:= \hat{\delta}(x, \omega)\phi(s), \\ \bar{g}(\theta, \omega) &:= \mathbb{E}_{s \sim \mu_\theta, a \sim \pi_\theta, s' \sim \tilde{\mathcal{P}}} [g(x, \omega)]. \end{aligned}$$

We also define constant $C_\delta := r_{\max} + (1 + \gamma) \max\{\frac{r_{\max}}{1-\gamma}, R_\omega\}$, and we immediately have

$$\|g(x, \omega)\|_2 \leq |r(x) + \gamma\phi(s')^\top \omega - \phi(s)^\top \omega| \leq r_{\max} + (1 + \gamma)R_\omega \leq C_\delta \quad (30)$$

and likewise, we have $\|\bar{g}(x, \omega)\|_2 \leq C_\delta$.

The critic update in Algorithm 1 can be written compactly as:

$$\omega_{k+1} = \Pi_{R_\omega}(\omega_k + \beta_k g(x_{(k)}, \omega_{k-\tau_k})), \quad (31)$$

where τ_k is the delay of the parameters used in evaluating the k th stochastic gradient, and $x_{(k)} := (s_{(k)}, a_{(k)}, s'_{(k)})$ is the sample used to evaluate the stochastic gradient at k th update.

Proof. Using ω_k^* as shorthand notation of $\omega_{\theta_k}^*$, we start with the optimality gap

$$\begin{aligned} &\|\omega_{k+1} - \omega_{k+1}^*\|_2^2 \\ &= \|\Pi_{R_\omega}(\omega_k + \beta_k g(x_{(k)}, \omega_{k-\tau_k})) - \omega_{k+1}^*\|_2^2 \\ &\leq \|\omega_k + \beta_k g(x_{(k)}, \omega_{k-\tau_k}) - \omega_{k+1}^*\|_2^2 \\ &= \|\omega_k - \omega_k^*\|_2^2 + 2\beta_k \langle \omega_k - \omega_k^*, g(x_{(k)}, \omega_{k-\tau_k}) \rangle + 2 \langle \omega_k - \omega_k^*, \omega_k^* - \omega_{k+1}^* \rangle + \|\omega_k^* - \omega_{k+1}^* + \beta_k g(x_{(k)}, \omega_{k-\tau_k})\|_2^2 \\ &= \|\omega_k - \omega_k^*\|_2^2 + 2\beta_k \langle \omega_k - \omega_k^*, g(x_{(k)}, \omega_{k-\tau_k}) - g(x_{(k)}, \omega_k) \rangle + 2\beta_k \langle \omega_k - \omega_k^*, g(x_{(k)}, \omega_k) - \bar{g}(\theta_k, \omega_k) \rangle \\ &\quad + 2\beta_k \langle \omega_k - \omega_k^*, \bar{g}(\theta_k, \omega_k) \rangle + 2 \langle \omega_k - \omega_k^*, \omega_k^* - \omega_{k+1}^* \rangle + \|\omega_k^* - \omega_{k+1}^* + \beta_k g(x_{(k)}, \omega_{k-\tau_k})\|_2^2 \\ &\leq \|\omega_k - \omega_k^*\|_2^2 + 2\beta_k \langle \omega_k - \omega_k^*, g(x_{(k)}, \omega_{k-\tau_k}) - g(x_{(k)}, \omega_k) \rangle + 2\beta_k \langle \omega_k - \omega_k^*, g(x_{(k)}, \omega_k) - \bar{g}(\theta_k, \omega_k) \rangle \\ &\quad + 2\beta_k \langle \omega_k - \omega_k^*, \bar{g}(\theta_k, \omega_k) \rangle + 2 \langle \omega_k - \omega_k^*, \omega_k^* - \omega_{k+1}^* \rangle + 2 \|\omega_k^* - \omega_{k+1}^*\|_2^2 + 2C_\delta^2 \beta_k^2. \end{aligned} \quad (32)$$

We first bound $\langle \omega_k - \omega_k^*, \bar{g}(\theta_k, \omega_k) \rangle$ in (32) as

$$\begin{aligned} \langle \omega_k - \omega_k^*, \bar{g}(\theta_k, \omega_k) \rangle &= \langle \omega_k - \omega_k^*, \bar{g}(\theta_k, \omega_k) - \bar{g}(\theta_k, \omega_k^*) \rangle \\ &= \left\langle \omega_k - \omega_k^*, \mathbb{E} \left[(\gamma\phi(s') - \phi(s))^\top (\omega_k - \omega_k^*) \phi(s) \right] \right\rangle \\ &= \left\langle \omega_k - \omega_k^*, \mathbb{E} \left[\phi(s) (\gamma\phi(s') - \phi(s))^\top \right] (\omega_k - \omega_k^*) \right\rangle \\ &= \left\langle \omega_k - \omega_k^*, A_{\pi_{\theta_k}} (\omega_k - \omega_k^*) \right\rangle \\ &\leq -\lambda \|\omega_k - \omega_k^*\|_2^2, \end{aligned} \quad (33)$$

where the first equality is due to $\bar{g}(\theta, \omega_\theta^*) = A_{\theta, \phi} \omega_\theta^* + b = 0$, and the last inequality follows Assumption 2. Substituting (33) into (32), then taking expectation on both sides of (32) yield

$$\begin{aligned} \mathbb{E} \|\omega_{k+1} - \omega_{k+1}^*\|_2^2 &\leq (1 - 2\lambda\beta_k) \mathbb{E} \|\omega_k - \omega_k^*\|_2^2 + 2\beta_k \mathbb{E} \langle \omega_k - \omega_k^*, g(x_{(k)}, \omega_{k-\tau_k}) - g(x_{(k)}, \omega_k) \rangle \\ &\quad + 2\beta_k \mathbb{E} \langle \omega_k - \omega_k^*, g(x_{(k)}, \omega_k) - \bar{g}(\theta_k, \omega_k) \rangle + 2 \mathbb{E} \langle \omega_k - \omega_k^*, \omega_k^* - \omega_{k+1}^* \rangle \\ &\quad + 2 \mathbb{E} \|\omega_k^* - \omega_{k+1}^*\|_2^2 + 2C_\delta^2 \beta_k^2. \end{aligned} \quad (34)$$

We then bound the term $\mathbb{E} \langle \omega_k - \omega_k^*, g(x_{(k)}, \omega_{k-\tau_k}) - g(x_{(k)}, \omega_k) \rangle$ in (34) as

$$\begin{aligned}
\mathbb{E} \langle \omega_k - \omega_k^*, g(x_{(k)}, \omega_{k-\tau_k}) - g(x_{(k)}, \omega_k) \rangle &= \mathbb{E} \left\langle \omega_k - \omega_k^*, \left(\gamma \phi(s'_{(k)}) - \phi(s_{(k)}) \right)^\top (\omega_{k-\tau_k} - \omega_k) \phi(s_{(k)}) \right\rangle \\
&\leq (1 + \gamma) \mathbb{E} \left[\|\omega_k - \omega_k^*\|_2 \|\omega_{k-\tau_k} - \omega_k\|_2 \right] \\
&\leq (1 + \gamma) \mathbb{E} \left[\|\omega_k - \omega_k^*\|_2 \left\| \sum_{i=k-\tau_k}^{k-1} (\omega_{i+1} - \omega_i) \right\|_2 \right] \\
&\leq (1 + \gamma) \mathbb{E} \left[\|\omega_k - \omega_k^*\|_2 \sum_{i=k-\tau_k}^{k-1} \beta_i \|g(x_i, \omega_{i-\tau_i})\|_2 \right] \\
&\leq (1 + \gamma) \mathbb{E} \left[\|\omega_k - \omega_k^*\|_2 \sum_{i=k-\tau_k}^{k-1} \beta_{k-K_0} \|g(x_i, \omega_{i-\tau_i})\|_2 \right] \\
&\leq C_\delta (1 + \gamma) K_0 \beta_{k-K_0} \mathbb{E} \|\omega_k - \omega_k^*\|_2, \tag{35}
\end{aligned}$$

where the second last inequality is due to the monotonicity of step size, and the last inequality follows the definition of C_δ in (30).

Next we jointly bound the fourth and fifth term in (34) as

$$\begin{aligned}
&2 \mathbb{E} \langle \omega_k - \omega_k^*, \omega_k^* - \omega_{k+1}^* \rangle + 2 \mathbb{E} \|\omega_k^* - \omega_{k+1}^*\|_2^2 \\
&\leq 2 \mathbb{E} \left[\|\omega_k - \omega_k^*\|_2 \|\omega_k^* - \omega_{k+1}^*\|_2 \right] + 2 \mathbb{E} \|\omega_k^* - \omega_{k+1}^*\|_2^2 \\
&\leq 2L_\omega \mathbb{E} \left[\|\omega_k - \omega_k^*\|_2 \|\theta_k - \theta_{k+1}\|_2 \right] + 2L_\omega^2 \mathbb{E} \|\theta_k - \theta_{k+1}\|_2^2 \\
&= 2L_\omega \alpha_k \mathbb{E} \left[\|\omega_k - \omega_k^*\|_2 \left\| \hat{\delta}(x_{(k)}, \omega_{k-\tau_k}) \psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) \right\|_2 \right] + 2L_\omega^2 \alpha_k^2 \mathbb{E} \left\| \hat{\delta}(x_{(k)}, \omega_{k-\tau_k}) \psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) \right\|_2^2 \\
&\leq 2L_\omega C_p \alpha_k \mathbb{E} \|\omega_k - \omega_k^*\|_2 + 2L_\omega^2 C_p^2 \alpha_k^2, \tag{36}
\end{aligned}$$

where constant $C_p := C_\delta C_\psi$. The second inequality is due to the L_ω -Lipschitz of ω_θ^* shown in Proposition 2, and the last inequality follows the fact that

$$\left\| \hat{\delta}(x_{(k)}, \omega_{k-\tau_k}) \psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) \right\|_2 \leq C_\delta C_\psi = C_p. \tag{37}$$

Substituting (35) and (36) into (34) yields

$$\begin{aligned}
\mathbb{E} \|\omega_{k+1} - \omega_{k+1}^*\|_2^2 &\leq (1 - 2\lambda\beta_k) \mathbb{E} \|\omega_k - \omega_k^*\|_2^2 + 2\beta_k \left(C_1 \frac{\alpha_k}{\beta_k} + C_2 K_0 \beta_{k-K_0} \right) \mathbb{E} \|\omega_k - \omega_k^*\|_2 \\
&\quad + 2\beta_k \mathbb{E} \langle \omega_k - \omega_k^*, g(x_{(k)}, \omega_k) - \bar{g}(\theta_k, \omega_k) \rangle + C_q \beta_k^2, \tag{38}
\end{aligned}$$

where $C_1 := L_\omega C_p$, $C_2 := C_\delta (1 + \gamma)$ and $C_q := 2C_\delta^2 + 2L_\omega^2 C_p^2 \max_{(k)} \frac{\alpha_k}{\beta_k^2} = 2C_\delta^2 + 2L_\omega^2 C_p^2 \frac{c_1^2}{c_2^2}$.

For brevity, we use $x \sim \theta$ to denote $s \sim \mu_\theta$, $a \sim \pi_\theta$ and $s' \sim \tilde{\mathcal{P}}$ in this proof. Consider the third term in (38) conditioned on $\theta_k, \omega_k, \theta_{k-\tau_k}$. We bound it as

$$\begin{aligned}
&\mathbb{E} \left[\langle \omega_k - \omega_k^*, g(x_{(k)}, \omega_k) - \bar{g}(\theta_k, \omega_k) \rangle \mid \theta_k, \omega_k, \theta_{k-\tau_k} \right] \\
&= \left\langle \omega_k - \omega_k^*, \mathbb{E}_{x_{(k)} \sim \theta_{k-\tau_k}} [g(x_{(k)}, \omega_k) \mid \omega_k] - \bar{g}(\theta_k, \omega_k) \right\rangle \\
&= \left\langle \omega_k - \omega_k^*, \bar{g}(\theta_{k-\tau_k}, \omega_k) - \bar{g}(\theta_k, \omega_k) \right\rangle \\
&\leq \|\omega_k - \omega_k^*\|_2 \|\bar{g}(\theta_{k-\tau_k}, \omega_k) - \bar{g}(\theta_k, \omega_k)\|_2 \\
&\leq 2R_\omega \left\| \mathbb{E}_{x \sim \theta_{k-\tau_k}} [g(x, \omega_k)] - \mathbb{E}_{x \sim \theta_k} [g(x, \omega_k)] \right\|_2 \\
&\leq 2R_\omega \sup_x \|g(x, \omega_k)\|_2 \left\| \mu_{\theta_{k-\tau_k}} \otimes \pi_{\theta_{k-\tau_k}} \otimes \tilde{\mathcal{P}} - \mu_{\theta_k} \otimes \pi_{\theta_k} \otimes \tilde{\mathcal{P}} \right\|_{TV} \\
&\leq 4R_\omega C_\delta d_{TV} (\mu_{\theta_{k-\tau_k}} \otimes \pi_{\theta_{k-\tau_k}} \otimes \tilde{\mathcal{P}}, \mu_{\theta_k} \otimes \pi_{\theta_k} \otimes \tilde{\mathcal{P}}), \tag{39}
\end{aligned}$$

where second last inequality follows the definition of TV norm and the last inequality uses the definition of C_δ in (30).

Define constant $C_3 := 2R_\omega C_\delta |A| L_\pi (1 + \log_\rho \kappa^{-1} + (1 - \rho)^{-1})$. Then by following the third item in Lemma A.1. shown by [17], we can write (39) as

$$\begin{aligned}
& \mathbb{E} [\langle \omega_k - \omega_k^*, g(x_{(k)}, \omega_k) - \bar{g}(\theta_k, \omega_k) \rangle | \theta_k, \omega_k, \theta_{k-\tau_k}] \\
& \leq 4R_\omega C_\delta d_{TV}(\mu_{\theta_{k-\tau_k}} \otimes \pi_{\theta_{k-\tau_k}} \otimes \tilde{\mathcal{P}}, \mu_{\theta_k} \otimes \pi_{\theta_k} \otimes \tilde{\mathcal{P}}) \\
& \leq C_3 \|\theta_{k-\tau_k} - \theta_k\|_2 \\
& \leq C_3 \sum_{i=k-\tau_k}^{k-1} \alpha_i \|g(x_i, \omega_{i-\tau_i})\|_2 \\
& \leq C_3 C_\delta K_0 \alpha_{k-K_0},
\end{aligned} \tag{40}$$

where we used the monotonicity of α_k and Assumption 1.

Taking total expectation on both sides of (40) and substituting it into (38) yield

$$\begin{aligned}
\mathbb{E} \|\omega_{k+1} - \omega_{k+1}^*\|_2^2 & \leq (1 - 2\lambda\beta_k) \mathbb{E} \|\omega_k - \omega_k^*\|_2^2 + 2\beta_k \left(C_1 \frac{\alpha_k}{\beta_k} + C_2 K_0 \beta_{k-K_0} \right) \mathbb{E} \|\omega_k - \omega_k^*\|_2 \\
& \quad + 2C_3 C_\delta K_0 \beta_k \alpha_{k-K_0} + C_q \beta_k^2.
\end{aligned} \tag{41}$$

Taking summation on both sides of (41) and rearranging yield

$$\begin{aligned}
2\lambda \sum_{k=K_0}^K \mathbb{E} \|\omega_k - \omega_k^*\|_2^2 & \leq \underbrace{\sum_{k=K_0}^K \frac{1}{\beta_k} \left(\mathbb{E} \|\omega_k - \omega_k^*\|_2^2 - \mathbb{E} \|\omega_{k+1} - \omega_{k+1}^*\|_2^2 \right)}_{I_1} + C_q \underbrace{\sum_{k=K_0}^K \beta_k}_{I_2} \\
& \quad + 2 \underbrace{\sum_{k=K_0}^K 2C_3 C_\delta K_0 \alpha_{k-K_0}}_{I_3} + 2 \underbrace{\sum_{k=K_0}^K \left(C_1 \frac{\alpha_k}{\beta_k} + C_2 K_0 \beta_{k-K_0} \right) \mathbb{E} \|\omega_k - \omega_k^*\|_2}_{I_4}.
\end{aligned} \tag{42}$$

We bound I_1 as

$$\begin{aligned}
I_1 & = \sum_{k=M_K}^K \frac{1}{\beta_k} \left(\mathbb{E} \|\omega_k - \omega_k^*\|_2^2 - \mathbb{E} \|\omega_{k+1} - \omega_{k+1}^*\|_2^2 \right) \\
& = \sum_{k=M_K}^K \left(\frac{1}{\beta_k} - \frac{1}{\beta_{k-1}} \right) \mathbb{E} \|\omega_k - \omega_k^*\|_2^2 + \frac{1}{\beta_{M_K-1}} \mathbb{E} \|\omega_{M_K} - \omega_{M_K}^*\|_2^2 - \frac{1}{\beta_k} \mathbb{E} \|\omega_{K+1} - \omega_{K+1}^*\|_2^2 \\
& \leq \sum_{k=M_K}^K \left(\frac{1}{\beta_k} - \frac{1}{\beta_{k-1}} \right) \mathbb{E} \|\omega_k - \omega_k^*\|_2^2 + \frac{1}{\beta_{M_K-1}} \mathbb{E} \|\omega_{M_K} - \omega_{M_K}^*\|_2^2 \\
& \leq 4R_\omega^2 \left(\sum_{k=M_K}^K \left(\frac{1}{\beta_k} - \frac{1}{\beta_{k-1}} \right) + \frac{1}{\beta_{M_K-1}} \right) = \frac{4R_\omega^2}{\beta_k} = \mathcal{O}(K^{\sigma_2}),
\end{aligned} \tag{43}$$

where the last inequality is due to the fact that

$$\|\omega_k - \omega_\theta^*\|_2 \leq \|\omega_k\|_2 + \|\omega_\theta^*\|_2 \leq 2R_\omega.$$

We bound I_2 as

$$\sum_{k=M_K}^K \beta_k = \sum_{k=M_K}^K \frac{c_2}{(1+k)^{\sigma_2}} = \mathcal{O}(K^{1-\sigma_2}) \tag{44}$$

where the inequality follows from the integration rule $\sum_{k=a}^b k^{-\sigma} \leq \frac{b^{1-\sigma}}{1-\sigma}$.

We bound I_3 as

$$I_3 = \sum_{k=K_0}^K 2C_3C_\delta K_0 \alpha_{k-K_0} = 2C_3C_\delta c_1 K_0 \sum_{k=0}^{K-K_0} (1+k)^{-\sigma_1} = \mathcal{O}(K_0 K^{1-\sigma_1}). \quad (45)$$

For the last term I_4 , we have

$$\begin{aligned} I_4 &= \sum_{k=K_0}^K \left(C_1 \frac{\alpha_k}{\beta_k} + C_2 K_0 \beta_{k-K_0} \right) \mathbb{E} \|\omega_k - \omega_k^*\|_2 \\ &\leq \sqrt{\sum_{k=K_0}^K \left(C_1 \frac{\alpha_k}{\beta_k} + C_2 K_0 \beta_{k-K_0} \right)^2} \sqrt{\sum_{k=K_0}^K (\mathbb{E} \|\omega_k - \omega_k^*\|_2)^2} \\ &\leq \sqrt{\sum_{k=K_0}^K \left(C_1 \frac{\alpha_k}{\beta_k} + C_2 K_0 \beta_{k-K_0} \right)^2} \sqrt{\sum_{k=K_0}^K \mathbb{E} \|\omega_k - \omega_k^*\|_2^2}, \end{aligned} \quad (46)$$

where the first inequality follows Cauchy–Schwartz inequality, and the second inequality follows Jensen’s inequality. In (46), we have

$$\begin{aligned} \sum_{k=K_0}^K \left(C_1 \frac{\alpha_k}{\beta_k} + C_2 K_0 \beta_{k-K_0} \right)^2 &\leq \sum_{k=0}^{K-K_0} \left(C_1 \frac{\alpha_k}{\beta_k} + C_2 K_0 \beta_k \right)^2 \\ &= C_1^2 \sum_{k=0}^{K-K_0} \frac{\alpha_k^2}{\beta_k^2} + 2C_1 C_2 K_0 \sum_{k=0}^{K-K_0} \alpha_k + C_2^2 K_0^2 \sum_{k=0}^{K-K_0} \beta_k^2 \\ &= \mathcal{O}\left(K^{2(\sigma_2-\sigma_1)+1}\right) + \mathcal{O}\left(K_0 K^{-\sigma_1+1}\right) + \mathcal{O}\left(K_0^2 K^{1-2\sigma_2}\right) \end{aligned} \quad (47)$$

where the first inequality is due to the fact that $\frac{\alpha_k}{\beta_k}$ and β_{k-K_0} are monotonically decreasing.

Substituting (47) into (46) gives

$$I_4 \leq \sqrt{\mathcal{O}\left(K^{2(\sigma_2-\sigma_1)+1}\right) + \mathcal{O}\left(K_0 K^{-\sigma_1+1}\right) + \mathcal{O}\left(K_0^2 K^{1-2\sigma_2}\right)} \sqrt{\sum_{k=K_0}^K \mathbb{E} \|\omega_k - \omega_k^*\|_2^2}. \quad (48)$$

Substituting (43), (44), (45) and (48) into (42), and dividing both sides of (42) by $K - K_0 + 1$ give

$$\begin{aligned} &2\lambda \frac{1}{K - K_0 + 1} \sum_{k=K_0}^K \mathbb{E} \|\omega_k - \omega_k^*\|_2^2 \\ &\leq \frac{\sqrt{\mathcal{O}\left(K^{2(\sigma_2-\sigma_1)+1}\right) + \mathcal{O}\left(K_0 K^{-\sigma_1+1}\right) + \mathcal{O}\left(K_0^2 K^{1-2\sigma_2}\right)}}{K - K_0 + 1} \sqrt{\sum_{k=K_0}^K \mathbb{E} \|\omega_k - \omega_k^*\|_2^2} \\ &\quad + \mathcal{O}\left(\frac{1}{K^{1-\sigma_2}}\right) + \mathcal{O}\left(\frac{1}{K^{\sigma_2}}\right) + \mathcal{O}\left(\frac{K_0}{K^{\sigma_1}}\right). \end{aligned} \quad (49)$$

We define the following functions:

$$\begin{aligned} T_1(K) &:= \frac{1}{K - K_0 + 1} \sum_{k=K_0}^K \mathbb{E} \|\omega_k - \omega_k^*\|_2^2, \\ T_2(K) &:= \mathcal{O}\left(\frac{1}{K^{1-\sigma_2}}\right) + \mathcal{O}\left(\frac{1}{K^{\sigma_2}}\right) + \mathcal{O}\left(\frac{K_0}{K^{\sigma_1}}\right), \\ T_3(K) &:= \frac{\mathcal{O}\left(K^{2(\sigma_2-\sigma_1)+1}\right) + \mathcal{O}\left(K_0 K^{-\sigma_1+1}\right) + \mathcal{O}\left(K_0^2 K^{1-2\sigma_2}\right)}{K - K_0 + 1}. \end{aligned}$$

Then (49) can be written as:

$$T_1(K) - \frac{1}{2\lambda} \sqrt{T_1(K)} \sqrt{T_3(K)} \leq \frac{1}{2\lambda} T_2(K).$$

Solving this quadratic inequality in terms of $T_1(K)$, we obtain

$$T_1(K) \leq \frac{1}{\lambda} T_2(K) + \frac{1}{2\lambda^2} T_3(K), \quad (50)$$

which implies

$$\begin{aligned} & \frac{1}{K - K_0 + 1} \sum_{k=K_0}^K \mathbb{E} \|\omega_k - \omega_k^*\|_2^2 \\ &= \mathcal{O}\left(\frac{1}{K^{1-\sigma_2}}\right) + \mathcal{O}\left(\frac{1}{K^{2(\sigma_1-\sigma_2)}}\right) + \mathcal{O}\left(\frac{K_0^2}{K^{2\sigma_2}}\right) + \mathcal{O}\left(\frac{K_0}{K^{\sigma_1}}\right) + \mathcal{O}\left(\frac{1}{K^{\sigma_2}}\right). \end{aligned}$$

We further have

$$\begin{aligned} \frac{1}{K} \sum_{k=1}^K \mathbb{E} \|\omega_k - \omega_k^*\|_2^2 &\leq \frac{1}{K} \left(\sum_{k=1}^{K_0-1} 4R_\omega^2 + \sum_{k=K_0}^K \mathbb{E} \|\omega_k - \omega_k^*\|_2^2 \right) \\ &= \frac{K_0-1}{K} 4R_\omega^2 + \frac{K-K_0+1}{K} \frac{1}{K-K_0+1} \sum_{k=K_0}^K \mathbb{E} \|\omega_k - \omega_k^*\|_2^2 \\ &= \mathcal{O}\left(\frac{K_0}{K}\right) + \mathcal{O}\left(\frac{1}{K-K_0+1} \sum_{k=K_0}^K \mathbb{E} \|\omega_k - \omega_k^*\|_2^2\right) \\ &= \mathcal{O}\left(\frac{1}{K-K_0+1} \sum_{k=K_0}^K \mathbb{E} \|\omega_k - \omega_k^*\|_2^2\right) \end{aligned} \quad (51)$$

which completes the proof. \square

B.2 PROOF OF THEOREM 2

We first clarify the notations:

$$\begin{aligned} x &:= (s, a, s'), \\ \hat{\delta}(x, \omega) &:= r(s, a, s') + \gamma \phi(s')^\top \omega - \phi(s)^\top \omega, \\ \delta(x, \theta) &:= r(s, a, s') + \gamma V_{\pi_\theta}(s') - V_{\pi_\theta}(s), \end{aligned}$$

The update in Algorithm 1 can be written compactly as:

$$\theta_{k+1} = \theta_k + \alpha_k \hat{\delta}(x_{(k)}, \omega_{k-\tau_k}) \psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}). \quad (52)$$

For brevity, we use ω_k^* as shorthand notation of $\omega_{\theta_k}^*$ in this proof. Then we are ready to give the convergence proof.

Proof. From L_J -Lipschitz of policy gradient shown in Proposition 1, we have:

$$\begin{aligned} J(\theta_{k+1}) &\geq J(\theta_k) + \langle \nabla J(\theta_k), \theta_{k+1} - \theta_k \rangle - \frac{L_J}{2} \|\theta_{k+1} - \theta_k\|_2^2 \\ &= J(\theta_k) + \alpha_k \left\langle \nabla J(\theta_k), \left(\hat{\delta}(x_{(k)}, \omega_{k-\tau_k}) - \hat{\delta}(x_{(k)}, \omega_k^*) \right) \psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) \right\rangle \\ &\quad + \alpha_k \left\langle \nabla J(\theta_k), \hat{\delta}(x_{(k)}, \omega_k^*) \psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) \right\rangle - \frac{L_J}{2} \alpha_k^2 \|\hat{\delta}(x_{(k)}, \omega_{k-\tau_k}) \psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)})\|_2^2 \\ &\geq J(\theta_k) + \alpha_k \left\langle \nabla J(\theta_k), \left(\hat{\delta}(x_{(k)}, \omega_{k-\tau_k}) - \hat{\delta}(x_{(k)}, \omega_k^*) \right) \psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) \right\rangle \\ &\quad + \alpha_k \left\langle \nabla J(\theta_k), \hat{\delta}(x_{(k)}, \omega_k^*) \psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) \right\rangle - \frac{L_J}{2} C_p^2 \alpha_k^2, \end{aligned}$$

where the last inequality follows the definition of C_p in (37).

Taking expectation on both sides of the last inequality yields

$$\begin{aligned} \mathbb{E}[J(\theta_{k+1})] &\geq \mathbb{E}[J(\theta_k)] + \alpha_k \underbrace{\mathbb{E} \left\langle \nabla J(\theta_k), \left(\hat{\delta}(x_{(k)}, \omega_{k-\tau_k}) - \hat{\delta}(x_{(k)}, \omega_k^*) \right) \psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) \right\rangle}_{I_1} \\ &\quad + \alpha_k \underbrace{\mathbb{E} \left\langle \nabla J(\theta_k), \hat{\delta}(x_{(k)}, \omega_k^*) \psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) \right\rangle}_{I_2} - \frac{L_J}{2} C_p^2 \alpha_k^2. \end{aligned} \quad (53)$$

We first decompose I_1 as

$$\begin{aligned} I_1 &= \mathbb{E} \left\langle \nabla J(\theta_k), \left(\hat{\delta}(x_{(k)}, \omega_{k-\tau_k}) - \hat{\delta}(x_{(k)}, \omega_k^*) \right) \psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) \right\rangle \\ &= \mathbb{E} \left\langle \nabla J(\theta_k), \underbrace{\left(\hat{\delta}(x_{(k)}, \omega_{k-\tau_k}) - \hat{\delta}(x_{(k)}, \omega_k) \right) \psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)})}_{I_1^{(1)}} \right\rangle \\ &\quad + \mathbb{E} \left\langle \nabla J(\theta_k), \underbrace{\left(\hat{\delta}(x_{(k)}, \omega_k) - \hat{\delta}(x_{(k)}, \omega_k^*) \right) \psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)})}_{I_1^{(2)}} \right\rangle. \end{aligned}$$

We bound $I_1^{(1)}$ as

$$\begin{aligned} I_1^{(1)} &= \mathbb{E} \left\langle \nabla J(\theta_k), \left(\gamma \phi(s'_{(k)}) - \phi(s_{(k)}) \right)^\top (\omega_{k-\tau_k} - \omega_k) \psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) \right\rangle \\ &\geq -\mathbb{E} \left[\|\nabla J(\theta_k)\|_2 \|\gamma \phi(s'_{(k)}) - \phi(s_{(k)})\|_2 \|\omega_k - \omega_{k-\tau_k}\|_2 \|\psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)})\|_2 \right] \\ &\geq -(1 + \gamma) C_\psi \mathbb{E} [\|\nabla J(\theta_k)\|_2 \|\omega_k - \omega_{k-\tau_k}\|_2] \\ &\geq -(1 + \gamma) C_\psi C_\delta K_0 \beta_{k-1} \mathbb{E} \|\nabla J(\theta_k)\|_2, \end{aligned}$$

where the last inequality follows

$$\begin{aligned} \|\omega_k - \omega_{k-\tau_k}\|_2 &= \left\| \sum_{i=k-\tau_k}^{k-1} (\omega_{i+1} - \omega_i) \right\|_2 \\ &\leq \sum_{i=k-\tau_k}^{k-1} \|\beta_i g(x_i, \omega_{i-\tau_i})\|_2 \\ &\leq \beta_{k-1} \sum_{i=k-\tau_k}^{k-1} \|g(x_i, \omega_{i-\tau_i})\|_2 \\ &\leq \beta_{k-1} K_0 C_\delta, \end{aligned}$$

where the second inequality is due to the monotonicity of step size, and the third one follows (30).

Then we bound $I_1^{(2)}$ as

$$\begin{aligned} I_1^{(2)} &= \mathbb{E} \left\langle \nabla J(\theta_k), \left(\hat{\delta}(x_{(k)}, \omega_k) - \hat{\delta}(x_{(k)}, \omega_k^*) \right) \psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) \right\rangle \\ &= -\mathbb{E} \left\langle \nabla J(\theta_k), \left(\gamma \phi(s'_{(k)}) - \phi(s_{(k)}) \right)^\top (\omega_k^* - \omega_k) \psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) \right\rangle \\ &\geq -\mathbb{E} \left[\|\nabla J(\theta_k)\|_2 \|\gamma \phi(s'_{(k)}) - \phi(s_{(k)})\|_2 \|\omega_k - \omega_k^*\|_2 \|\psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)})\|_2 \right] \\ &\geq -(1 + \gamma) C_\psi \mathbb{E} [\|\nabla J(\theta_k)\|_2 \|\omega_k - \omega_k^*\|_2]. \end{aligned}$$

Collecting lower bounds of $I_1^{(1)}$ and $I_1^{(2)}$ gives

$$\begin{aligned}
I_1 &\geq -(1+\gamma)C_\psi \mathbb{E} [\|\nabla J(\theta_k)\|_2 (C_\delta K_0 \beta_{k-1} + \|\omega_k - \omega_k^*\|_2)] \\
&= -(1+\gamma)C_\psi \sqrt{(\mathbb{E} [\|\nabla J(\theta_k)\|_2 (C_\delta K_0 \beta_{k-1} + \|\omega_k - \omega_k^*\|_2)]^2)} \\
&\geq -(1+\gamma)C_\psi \sqrt{\mathbb{E} \|\nabla J(\theta_k)\|_2^2} \mathbb{E} [(C_\delta K_0 \beta_{k-1} + \|\omega_k - \omega_k^*\|_2)^2] \\
&\geq -(1+\gamma)C_\psi \sqrt{\mathbb{E} \|\nabla J(\theta_k)\|_2^2} \sqrt{\mathbb{E} [2C_\delta^2 K_0^2 \beta_{k-1}^2 + 2\|\omega_k - \omega_k^*\|_2^2]} \\
&= -\sqrt{2}(1+\gamma)C_\psi \sqrt{\mathbb{E} \|\nabla J(\theta_k)\|_2^2} \sqrt{C_\delta^2 K_0^2 \beta_{k-1}^2 + \mathbb{E} \|\omega_k - \omega_k^*\|_2^2}, \tag{54}
\end{aligned}$$

where the second inequality follows Cauchy-Schwartz inequality, and the third inequality follows Young's inequality.

Now we consider I_2 . We first decompose I_2 as

$$\begin{aligned}
I_2 &= \mathbb{E} \left\langle \nabla J(\theta_k), \hat{\delta}(x(k), \omega_k^*) \psi_{\theta_{k-\tau_k}}(s(k), a(k)) \right\rangle \\
&= \underbrace{\mathbb{E} \left\langle \nabla J(\theta_k), \left(\hat{\delta}(x(k), \omega_k^*) - \hat{\delta}(x(k), \omega_{k-\tau_k}^*) \right) \psi_{\theta_{k-\tau_k}}(s(k), a(k)) \right\rangle}_{I_2^{(1)}} \\
&\quad + \underbrace{\mathbb{E} \left\langle \nabla J(\theta_k), \left(\hat{\delta}(x(k), \omega_{k-\tau_k}^*) - \delta(x(k), \theta_{k-\tau_k}) \right) \psi_{\theta_{k-\tau_k}}(s(k), a(k)) \right\rangle}_{I_2^{(2)}} \\
&\quad + \underbrace{\mathbb{E} \left\langle \nabla J(\theta_k), \delta(x(k), \theta_{k-\tau_k}) \psi_{\theta_{k-\tau_k}}(s(k), a(k)) - \nabla J(\theta_k) \right\rangle + \|\nabla J(\theta_k)\|_2^2}_{I_2^{(3)}}.
\end{aligned}$$

We bound $I_2^{(1)}$ as

$$\begin{aligned}
I_2^{(1)} &= \mathbb{E} \left\langle \nabla J(\theta_k), \left(\hat{\delta}(x(k), \omega_k^*) - \hat{\delta}(x(k), \omega_{k-\tau_k}^*) \right) \psi_{\theta_{k-\tau_k}}(s(k), a(k)) \right\rangle \\
&= \mathbb{E} \left\langle \nabla J(\theta_k), \left(\gamma \phi(s'(k)) - \phi(s(k)) \right)^\top (\omega_k^* - \omega_{k-\tau_k}^*) \psi_{\theta_{k-\tau_k}}(s(k), a(k)) \right\rangle \\
&\geq -\mathbb{E} \left[\|\nabla J(\theta_k)\|_2 \left\| \left(\gamma \phi(s'(k)) - \phi(s(k)) \right)^\top \right\|_2 \|\omega_k^* - \omega_{k-\tau_k}^*\|_2 \|\psi_{\theta_{k-\tau_k}}(s(k), a(k))\|_2 \right] \\
&\geq -L_V C_\psi (1+\gamma) \mathbb{E} \|\omega_k^* - \omega_{k-\tau_k}^*\|_2 \\
&\geq -L_V L_\omega C_\psi (1+\gamma) \mathbb{E} \|\theta_k - \theta_{k-\tau_k}\|_2 \\
&\geq -L_V L_\omega C_\psi C_p (1+\gamma) K_0 \alpha_{k-K_0},
\end{aligned}$$

where the second last inequality follows from Proposition 2 and the last inequality uses (37) as

$$\begin{aligned}
\|\theta_k - \theta_{k-\tau_k}\|_2 &\leq \sum_{i=k-\tau_k}^{k-1} \|\theta_{i+1} - \theta_i\|_2 \\
&= \sum_{i=k-\tau_k}^{k-1} \alpha_i \|\hat{\delta}(x_i, \omega_{i-\tau_i}) \psi_{\theta_{i-\tau_i}}(s_i, a_i)\|_2 \\
&\leq \sum_{i=k-\tau_k}^{k-1} \alpha_{k-\tau_k} C_p \leq C_p K_0 \alpha_{k-K_0}. \tag{55}
\end{aligned}$$

We bound $I_2^{(2)}$ as

$$\begin{aligned}
I_2^{(2)} &= \mathbb{E} \left\langle \nabla J(\theta_k), \left(\hat{\delta}(x(k), \omega_{k-\tau_k}^*) - \delta(x(k), \theta_{k-\tau_k}) \right) \psi_{\theta_{k-\tau_k}}(s(k), a(k)) \right\rangle \\
&\geq -\mathbb{E} \left[\|\nabla J(\theta_k)\|_2 \left| \hat{\delta}(x(k), \omega_{k-\tau_k}^*) - \delta(x(k), \theta_{k-\tau_k}) \right| \|\psi_{\theta_{k-\tau_k}}(s(k), a(k))\|_2 \right] \\
&\geq -L_V C_\psi \mathbb{E} \left| \hat{\delta}(x(k), \omega_{k-\tau_k}^*) - \delta(x(k), \theta_{k-\tau_k}) \right| \\
&= -L_V C_\psi \mathbb{E} \left| \left(\phi(s'(k))^\top \omega_{k-\tau_k}^* - V_{\pi_{\theta_{k-\tau_k}}}(s'(k)) \right) + V_{\pi_{\theta_{k-\tau_k}}}(s(k)) - \phi(s(k))^\top \omega_{k-\tau_k}^* \right| \\
&\geq -L_V C_\psi \left(\gamma \mathbb{E} \left| \phi(s'(k))^\top \omega_{k-\tau_k}^* - V_{\pi_{\theta_{k-\tau_k}}}(s'(k)) \right| + \mathbb{E} \left| V_{\pi_{\theta_{k-\tau_k}}}(s(k)) - \phi(s(k))^\top \omega_{k-\tau_k}^* \right| \right) \\
&\geq -L_V C_\psi \left(\gamma \sqrt{\mathbb{E} \left| \phi(s'(k))^\top \omega_{k-\tau_k}^* - V_{\pi_{\theta_{k-\tau_k}}}(s'(k)) \right|^2} + \sqrt{\mathbb{E} \left| V_{\pi_{\theta_{k-\tau_k}}}(s(k)) - \phi(s(k))^\top \omega_{k-\tau_k}^* \right|^2} \right) \\
&\geq -L_V C_\psi (1 + \gamma) \epsilon_{app}.
\end{aligned}$$

We bound $I_2^{(3)}$ as

$$\begin{aligned}
I_2^{(3)} &= \mathbb{E} \left\langle \nabla J(\theta_k), \delta(x(k), \theta_{k-\tau_k}) \psi_{\theta_{k-\tau_k}}(s(k), a(k)) - \nabla J(\theta_k) \right\rangle \\
&= \mathbb{E} \left[\mathbb{E} \left[\left\langle \nabla J(\theta_k), \delta(x(k), \theta_{k-\tau_k}) \psi_{\theta_{k-\tau_k}}(s(k), a(k)) - \nabla J(\theta_k) \right\rangle \middle| \theta_{k-\tau_k}, \theta_k \right] \right] \\
&= \mathbb{E} \left\langle \nabla J(\theta_k), \mathbb{E} \left[\delta(x(k), \theta_{k-\tau_k}) \psi_{\theta_{k-\tau_k}}(s(k), a(k)) \middle| \theta_{k-\tau_k}, \theta_k \right] - \nabla J(\theta_k) \right\rangle \\
&= \mathbb{E} \left\langle \nabla J(\theta_k), \mathbb{E}_{\substack{s(k) \sim \mu_{\theta_{k-\tau_k}} \\ a(k) \sim \pi_{\theta_{k-\tau_k}}}} \left[A_{\pi_{\theta_{k-\tau_k}}}(s(k), a(k)) \psi_{\theta_{k-\tau_k}}(s(k), a(k)) \right] - \nabla J(\theta_k) \right\rangle
\end{aligned}$$

where we used the fact that

$$\begin{aligned}
&\mathbb{E} \left[\delta(x(k), \theta_{k-\tau_k}) \psi_{\theta_{k-\tau_k}}(s(k), a(k)) \middle| \theta_{k-\tau_k}, \theta_k \right] \\
&= \mathbb{E}_{\substack{s(k) \sim \mu_{\theta_{k-\tau_k}} \\ a(k) \sim \pi_{\theta_{k-\tau_k}} \\ s'(k) \sim \tilde{\mathcal{P}}}} \left[\left(r(s(k), a(k), s'(k)) + \gamma V_{\pi_{\theta_{k-\tau_k}}}(s'(k)) - V_{\pi_{\theta_{k-\tau_k}}}(s(k)) \right) \psi_{\theta_{k-\tau_k}}(s(k), a(k)) \middle| \theta_{k-\tau_k}, \theta_k \right] \\
&= \mathbb{E}_{\substack{s(k) \sim \mu_{\theta_{k-\tau_k}} \\ a(k) \sim \pi_{\theta_{k-\tau_k}}}} \left[\left(\mathbb{E}_{s'(k) \sim \tilde{\mathcal{P}}} \left[r(s(k), a(k), s'(k)) + \gamma V_{\pi_{\theta_{k-\tau_k}}}(s'(k)) \right] - V_{\pi_{\theta_{k-\tau_k}}}(s(k)) \right) \psi_{\theta_{k-\tau_k}}(s(k), a(k)) \middle| \theta_{k-\tau_k}, \theta_k \right] \\
&= \mathbb{E}_{\substack{s(k) \sim \mu_{\theta_{k-\tau_k}} \\ a(k) \sim \pi_{\theta_{k-\tau_k}}}} \left[\left(Q_{\pi_{\theta_{k-\tau_k}}}(s(k), a(k)) - V_{\pi_{\theta_{k-\tau_k}}}(s(k)) \right) \psi_{\theta_{k-\tau_k}}(s(k), a(k)) \middle| \theta_{k-\tau_k}, \theta_k \right] \\
&= \mathbb{E}_{\substack{s(k) \sim \mu_{\theta_{k-\tau_k}} \\ a(k) \sim \pi_{\theta_{k-\tau_k}}}} \left[A_{\pi_{\theta_{k-\tau_k}}}(s(k), a(k)) \psi_{\theta_{k-\tau_k}}(s(k), a(k)) \middle| \theta_{k-\tau_k}, \theta_k \right].
\end{aligned}$$

According to [6], if μ_θ is the stationary distribution of an artificial MDP with transition kernel $\tilde{\mathcal{P}}(\cdot|s, a)$ and policy π_θ , then we have $\mu_\theta(\cdot) = d_\theta(\cdot)$. Therefore, it follows that

$$\begin{aligned}
I_2^{(3)} &= \mathbb{E} \left\langle \nabla J(\theta_k), \mathbb{E}_{\substack{s(k) \sim \mu_{\theta_{k-\tau_k}} \\ a(k) \sim \pi_{\theta_{k-\tau_k}}}} \left[A_{\pi_{\theta_{k-\tau_k}}}(s(k), a(k)) \psi_{\theta_{k-\tau_k}}(s(k), a(k)) \middle| \theta_{k-\tau_k}, \theta_k \right] - \nabla J(\theta_k) \right\rangle \\
&= \mathbb{E} \left\langle \nabla J(\theta_k), \mathbb{E}_{\substack{s(k) \sim d_{\theta_{k-\tau_k}} \\ a(k) \sim \pi_{\theta_{k-\tau_k}}}} \left[A_{\pi_{\theta_{k-\tau_k}}}(s(k), a(k)) \psi_{\theta_{k-\tau_k}}(s(k), a(k)) \middle| \theta_{k-\tau_k}, \theta_k \right] - \nabla J(\theta_k) \right\rangle \\
&\geq -\mathbb{E} [\|\nabla J(\theta_k)\|_2 \|\nabla J(\theta_{k-\tau_k}) - \nabla J(\theta_k)\|_2] \\
&\geq -L_V L_J \mathbb{E} \|\theta_{k-\tau_k} - \theta_k\|_2 \\
&\geq -L_V L_J C_p K_0 \alpha_{k-K_0},
\end{aligned}$$

where the second last inequality is due to L_J -Lipschitz of policy gradient shown in Proposition 1, and the last inequality follows (55).

Collecting lower bounds of $I_2^{(1)}$, $I_2^{(2)}$ and $I_2^{(3)}$ gives

$$I_2 \geq -D_1 K_0 \alpha_{k-K_0} - L_V C_\psi (1 + \gamma) \epsilon_{app}, \quad (56)$$

where constant $D_1 := L_V L_\omega C_\psi C_p (1 + \gamma) + L_V L_J C_p$.

Substituting (54) and (56) into (53) yields

$$\begin{aligned} \mathbb{E}[J(\theta_{k+1})] &\geq \mathbb{E}[J(\theta_k)] - \alpha_k \sqrt{2}(1 + \gamma) C_\psi \sqrt{\mathbb{E} \|\nabla J(\theta_k)\|_2^2} \sqrt{C_\delta^2 K_0^2 \beta_{k-1}^2 + \mathbb{E} \|\omega_k - \omega_k^*\|_2^2} \\ &\quad - \alpha_k D_1 K_0 \alpha_{k-K_0} - \alpha_k L_V C_\psi (1 + \gamma) \epsilon_{app} + \alpha_k \|\nabla J(\theta_k)\|_2^2 - \frac{L_J}{2} C_p^2 \alpha_k^2. \end{aligned} \quad (57)$$

Dividing both sides of (57) by α_k , then rearranging and taking summation on both sides give

$$\begin{aligned} \sum_{k=K_0}^K \mathbb{E} \|\nabla J(\theta_k)\|_2^2 &\leq \underbrace{\sum_{k=K_0}^K \frac{1}{\alpha_k} (\mathbb{E}[J(\theta_{k+1})] - \mathbb{E}[J(\theta_k)])}_{I_3} + \underbrace{\sum_{k=K_0}^K \left(D_1 K_0 \alpha_{k-K_0} + \frac{L_J}{2} C_p^2 \alpha_k \right)}_{I_4} \\ &\quad + \underbrace{\sqrt{2}(1 + \gamma) C_\psi \sum_{k=K_0}^K \sqrt{\mathbb{E} \|\nabla J(\theta_k)\|_2^2} \sqrt{C_\delta^2 K_0^2 \beta_{k-1}^2 + \mathbb{E} \|\omega_k - \omega_k^*\|_2^2}}_{I_5} \\ &\quad + (K - K_0 + 1)(1 + \gamma) L_V C_\psi \epsilon_{app}, \end{aligned} \quad (58)$$

We bound I_3 as

$$\begin{aligned} I_3 &= \sum_{k=M_K}^K \frac{1}{\alpha_k} (\mathbb{E}[J(\theta_{k+1})] - \mathbb{E}[J(\theta_k)]) \\ &= \sum_{k=M_K}^K \left(\frac{1}{\alpha_{k-1}} - \frac{1}{\alpha_k} \right) \mathbb{E}[J(\theta_k)] - \frac{1}{\alpha_{M_K-1}} \mathbb{E}[J(\theta_{M_K})] + \frac{1}{\alpha_K} \mathbb{E}[J(\theta_{K+1})] \\ &\leq \frac{1}{\alpha_K} \mathbb{E}[J(\theta_{K+1})] \\ &\leq \frac{r_{\max}}{1 - \gamma} \frac{1}{\alpha_K} = \mathcal{O}(K^{\sigma_1}), \end{aligned} \quad (59)$$

where the first inequality is due to the α_k is monotonic decreasing and positive, and last inequality is due to $V_{\pi_\theta}(s) \leq \frac{r_{\max}}{1 - \gamma}$ for any $s \in \mathcal{S}$ and π_θ .

We bound I_4 as

$$I_4 = \sum_{k=K_0}^K \left(D_1 K_0 \alpha_{k-K_0} + \frac{L_J}{2} C_p^2 \alpha_k \right) \leq \sum_{k=0}^{K-K_0} \left(D_1 K_0 \alpha_k + \frac{L_J}{2} C_p^2 \alpha_k \right) = \mathcal{O}(K_0 K^{1-\sigma_1}).$$

We bound I_5 as

$$\begin{aligned} I_5 &= \sum_{k=M_K}^K \sqrt{\mathbb{E} \|\nabla J(\theta_k)\|_2^2} \sqrt{C_\delta^2 K_0^2 \beta_{k-1}^2 + \mathbb{E} \|\omega_k - \omega_k^*\|_2^2} \\ &\leq \sqrt{\sum_{k=M_K}^K \mathbb{E} \|\nabla J(\theta_k)\|_2^2} \sqrt{\sum_{k=M_K}^K (C_\delta^2 K_0^2 \beta_{k-1}^2 + \mathbb{E} \|\omega_k - \omega_k^*\|_2^2)} \\ &= \sqrt{\sum_{k=M_K}^K \mathbb{E} \|\nabla J(\theta_k)\|_2^2} \sqrt{C_\delta^2 K_0^2 \sum_{k=M_K}^K \beta_{k-1}^2 + \sum_{k=M_K}^K \mathbb{E} \|\omega_k - \omega_k^*\|_2^2}, \end{aligned} \quad (60)$$

where the first inequality follows Cauchy-Schwartz inequality. In (60), we have

$$\sum_{k=M_K}^K \beta_{k-1}^2 \leq \sum_{k=0}^{K-M_K} \beta_k^2 = \sum_{k=0}^{K-M_K} c_2^2 (1+k)^{-2\sigma_2} = \mathcal{O}(K^{1-2\sigma_2}).$$

Substituting the last equality into (60) gives

$$I_5 \leq \sqrt{\sum_{k=M_K}^K \mathbb{E} \|\nabla J(\theta_k)\|_2^2} \sqrt{\mathcal{O}(K_0^2 K^{1-2\sigma_2}) + \sum_{k=M_K}^K \mathbb{E} \|\omega_k - \omega_k^*\|_2^2}. \quad (61)$$

Dividing both sides of (57) by $K - K_0 + 1$ and collecting upper bounds of I_3 , I_4 and I_5 give

$$\begin{aligned} & \frac{1}{K - K_0 + 1} \sum_{k=K_0}^K \mathbb{E} \|\nabla J(\theta_k)\|_2^2 \\ & \leq \frac{\sqrt{2}(1+\gamma)C_\psi}{K - K_0 + 1} \sqrt{\sum_{k=K_0}^K \mathbb{E} \|\nabla J(\theta_k)\|_2^2} \sqrt{\mathcal{O}(K_0^2 K^{1-2\sigma_2}) + \sum_{k=K_0}^K \mathbb{E} \|\omega_k - \omega_k^*\|_2^2} \\ & \quad + \mathcal{O}\left(\frac{1}{K^{1-\sigma_1}}\right) + \mathcal{O}\left(\frac{K_0}{K^{\sigma_1}}\right) + \mathcal{O}(\epsilon_{app}). \end{aligned} \quad (62)$$

Define the following functions

$$\begin{aligned} T_4(K) &:= \frac{1}{K - K_0 + 1} \sum_{k=K_0}^K \mathbb{E} \|\nabla J(\theta_k)\|_2^2, \\ T_5(K) &:= \frac{1}{K - K_0 + 1} \left(\mathcal{O}(K_0^2 K^{1-2\sigma_2}) + \sum_{k=K_0}^K \mathbb{E} \|\omega_k - \omega_k^*\|_2^2 \right), \\ T_6(K) &:= \mathcal{O}\left(\frac{1}{K^{1-\sigma_1}}\right) + \mathcal{O}\left(\frac{K_0}{K^{\sigma_1}}\right) + \mathcal{O}(\epsilon_{app}). \end{aligned}$$

Then (81) can be rewritten as

$$T_4(K) \leq T_6(K) + \sqrt{2}(1+\gamma)C_\psi \sqrt{T_4(K)} \sqrt{T_5(K)}.$$

Solving this quadratic inequality in terms of $T_4(K)$, we obtain

$$T_4(K) \leq 2T_6(K) + 4(1+\gamma)^2 C_\psi^2 T_5(K), \quad (63)$$

which implies

$$\begin{aligned} & \frac{1}{K - K_0 + 1} \sum_{k=K_0}^K \mathbb{E} \|\nabla J(\theta_k)\|_2^2 \\ & = \mathcal{O}\left(\frac{1}{K^{1-\sigma_1}}\right) + \mathcal{O}\left(\frac{K_0}{K^{\sigma_1}}\right) + \mathcal{O}\left(\frac{K_0^2}{K^{2\sigma_2}}\right) + \mathcal{O}\left(\frac{1}{K - K_0 + 1} \sum_{k=K_0}^K \mathbb{E} \|\omega_k - \omega_k^*\|_2^2\right) + \mathcal{O}(\epsilon_{app}) \end{aligned}$$

We further have

$$\begin{aligned} \frac{1}{K} \sum_{k=1}^K \mathbb{E} \|\nabla J(\theta_k)\|_2^2 & \leq \frac{1}{K} \left(\sum_{k=1}^{K_0-1} L_V^2 + \sum_{k=K_0}^K \mathbb{E} \|\nabla J(\theta_k)\|_2^2 \right) \\ & = \frac{K_0-1}{K} L_V^2 + \frac{K - K_0 + 1}{K} \frac{1}{K - K_0 + 1} \sum_{k=K_0}^K \mathbb{E} \|\nabla J(\theta_k)\|_2^2 \\ & = \mathcal{O}\left(\frac{K_0}{K}\right) + \mathcal{O}\left(\frac{1}{K - K_0 + 1} \sum_{k=K_0}^K \mathbb{E} \|\nabla J(\theta_k)\|_2^2\right) \\ & = \mathcal{O}\left(\frac{1}{K - K_0 + 1} \sum_{k=K_0}^K \mathbb{E} \|\nabla J(\theta_k)\|_2^2\right) \end{aligned} \quad (64)$$

which completes the proof. \square

B.3 PROOF OF THEOREM 3

Given the definition in section B.1, we now give the convergence proof of critic update in Algorithm 1 with linear function approximation and Markovian sampling.

By following the derivation of (38), we have

$$\begin{aligned} \mathbb{E} \|\omega_{k+1} - \omega_{k+1}^*\|_2^2 &\leq (1 - 2\lambda\beta_k) \mathbb{E} \|\omega_k - \omega_k^*\|_2^2 + 2\beta_k \left(C_1 \frac{\alpha_k}{\beta_k} + C_2 K_0 \beta_{k-K_0} \right) \mathbb{E} \|\omega_k - \omega_k^*\|_2 \\ &\quad + 2\beta_k \mathbb{E} \langle \omega_k - \omega_k^*, g(x_{(k)}, \omega_k) - \bar{g}(\theta_k, \omega_k) \rangle + C_q \beta_k^2, \end{aligned} \quad (65)$$

where $C_1 := C_p L_\omega$, $C_2 := C_\delta(1 + \gamma)$ and $C_q := 2C_\delta^2 + 2L_\omega^2 C_p^2 \max_{(k)} \frac{\alpha_k^2}{\beta_k^2} = 2C_\delta^2 + 2L_\omega^2 C_p^2 \frac{c_1^2}{c_2^2}$.

Now we consider the third item in the last inequality. For some $m \in \mathbb{N}^+$, we define $M := (K_0 + 1)m + K_0$. Following Lemma 4 (to be presented in Sec. C.1), for some $d_m \leq M$ and positive constants C_4, C_5, C_6, C_7 , we have

$$\begin{aligned} &\mathbb{E} \langle \omega_k - \omega_k^*, g(x_{(k)}, \omega_k) - \bar{g}(\theta_k, \omega_k) \rangle \\ &\leq C_4 \mathbb{E} \|\theta_k - \theta_{k-d_m}\|_2 + C_5 \sum_{i=\tau_k}^{d_m} \mathbb{E} \|\theta_{k-i} - \theta_{k-d_m}\|_2 + C_6 \mathbb{E} \|\omega_k - \omega_{k-d_m}\|_2 + C_7 \kappa \rho^{m-1} \\ &\leq C_4 \sum_{i=k-d_m}^{k-1} \mathbb{E} \|\theta_{i+1} - \theta_i\|_2 + C_5 \sum_{i=\tau_k}^{d_m-1} \sum_{j=k-d_m}^{k-i-1} \mathbb{E} \|\theta_{j+1} - \theta_j\|_2 + C_6 \sum_{i=k-d_m}^{k-1} \mathbb{E} \|\omega_{i+1} - \omega_i\|_2 + C_7 \kappa \rho^{m-1} \\ &\leq C_4 \sum_{i=k-d_m}^{k-1} \alpha_i C_p + C_5 \sum_{i=\tau_k}^{d_m-1} \sum_{j=k-d_m}^{k-i-1} \alpha_j C_p + C_6 \sum_{i=k-d_m}^{k-1} \beta_i C_\delta + C_7 \kappa \rho^{m-1} \\ &\leq C_4 \alpha_{k-d_m} \sum_{i=k-d_m}^{k-1} C_p + C_5 \alpha_{k-d_m} \sum_{i=\tau_k}^{d_m-1} \sum_{j=k-d_m}^{k-i-1} C_p + C_6 \beta_{k-d_m} \sum_{i=k-d_m}^{k-1} C_\delta + C_7 \kappa \rho^{m-1} \\ &\leq C_4 d_m C_p \alpha_{k-d_m} + C_5 (d_m - \tau_k)^2 C_p \alpha_{k-d_m} + C_6 d_m C_\delta \beta_{k-d_m} + C_7 \kappa \rho^{m-1} \\ &\leq (C_4 M + C_5 M^2) C_p \alpha_{k-M} + C_6 M C_\delta \beta_{k-M} + C_7 \kappa \rho^{m-1}, \end{aligned} \quad (66)$$

where the third last inequality is due to the monotonicity of step size, and the last inequality is due to $\tau_k \geq 0$ and $d_m \leq M$.

Further letting $m = m_K$ which is defined in (20) yields

$$\begin{aligned} &\mathbb{E} \langle \omega_k - \omega_k^*, g(x_{(k)}, \omega_k) - \bar{g}(\theta_k, \omega_k) \rangle \\ &= (C_4 M_K + C_5 M_K^2) C_p \alpha_{k-M_K} + C_6 C_\delta M_K \beta_{k-M_K} + C_7 \kappa \rho^{m_K-1} \\ &\leq (C_4 M_K + C_5 M_K^2) C_p \alpha_{k-M_K} + C_6 C_\delta M_K \beta_{k-M_K} + C_7 \alpha_K, \end{aligned} \quad (67)$$

where $M_K = (K_0 + 1)m_K + K_0$, and the last inequality follows the definition of m_K .

Substituting (67) into (65), then rearranging and summing up both sides over $k = M_K, \dots, K$ yield

$$\begin{aligned} 2\lambda \sum_{k=M_K}^K \mathbb{E} \|\omega_k - \omega_k^*\|_2^2 &\leq \underbrace{\sum_{k=M_K}^K \frac{1}{\beta_k} \left(\mathbb{E} \|\omega_k - \omega_k^*\|_2^2 - \mathbb{E} \|\omega_{k+1} - \omega_{k+1}^*\|_2^2 \right)}_{I_1} + C_q \underbrace{\sum_{k=M_K}^K \beta_k}_{I_2} \\ &\quad + 2 \underbrace{\sum_{k=M_K}^K \left((C_4 M_K + C_5 M_K^2) C_p \alpha_{k-M_K} + C_6 C_\delta M_K \beta_{k-M_K} + C_7 \alpha_K \right)}_{I_3} \\ &\quad + 2 \underbrace{\sum_{k=M_K}^K \left(C_1 \frac{\alpha_k}{\beta_k} + C_2 K_0 \beta_{k-K_0} \right) \mathbb{E} \|\omega_k - \omega_k^*\|_2}_{I_4}. \end{aligned} \quad (68)$$

where the order of I_1 , I_2 and I_4 have already been given by (43), (44) and (48) respectively. We bound I_3 as

$$\begin{aligned} I_3 &= (C_4 M_K + C_5 M_K^2) C_p \sum_{k=M_K}^K \alpha_k + C_6 C_\delta M_K \sum_{k=M_K}^K \beta_k + C_7 \alpha_K \sum_{k=M_K}^K 1 \\ &\leq (C_4 M_K + C_5 M_K^2) C_p c_1 \frac{K^{1-\sigma_1}}{1-\sigma_1} + C_6 C_\delta M_K c_2 \frac{K^{1-\sigma_2}}{1-\sigma_2} + C_7 c_1 K(1+K)^{-\sigma_1} \\ &= \mathcal{O}((K_0^2 \log^2 K) K^{1-\sigma_1}) + \mathcal{O}((K_0 \log K) K^{1-\sigma_2}), \end{aligned} \quad (69)$$

where the last inequality follows from the integration rule $\sum_{k=a}^b k^{-\sigma} \leq \frac{b^{1-\sigma}}{1-\sigma}$, and the last equality is due to $\mathcal{O}(M_K) = \mathcal{O}(K_0 m_K) = \mathcal{O}(K_0 \log K)$.

Collecting the upper bounds of I_1 , I_2 , I_3 and I_4 , and dividing both sides of (68) by $K - M_K + 1$ yield

$$\begin{aligned} &2\lambda \frac{1}{K - M_K + 1} \sum_{k=M_K}^K \mathbb{E} \|\omega_k - \omega_k^*\|_2^2 \\ &\leq \frac{\sqrt{\mathcal{O}(K^{2(\sigma_2-\sigma_1)+1}) + \mathcal{O}(K_0 K^{-\sigma_1+1}) + \mathcal{O}(K_0^2 K^{1-2\sigma_2})}}{K - M_K + 1} \sqrt{\sum_{k=M_K}^K \mathbb{E} \|\omega_k - \omega_k^*\|_2^2} \\ &\quad + \mathcal{O}\left(\frac{1}{K^{1-\sigma_2}}\right) + \mathcal{O}\left(\frac{K_0^2 \log^2 K}{K^{\sigma_1}}\right) + \mathcal{O}\left(\frac{K_0 \log K}{K^{\sigma_2}}\right). \end{aligned} \quad (70)$$

Similar to the derivation of (50), (70) implies

$$\begin{aligned} &\frac{1}{K - M_K + 1} \sum_{k=M_K}^K \mathbb{E} \|\omega_k - \omega_k^*\|_2^2 \\ &= \mathcal{O}\left(\frac{1}{K^{1-\sigma_2}}\right) + \mathcal{O}\left(\frac{1}{K^{2(\sigma_1-\sigma_2)}}\right) + \mathcal{O}\left(\frac{K_0^2}{K^{2\sigma_2}}\right) + \mathcal{O}\left(\frac{K_0^2 \log^2 K}{K^{\sigma_1}}\right) + \mathcal{O}\left(\frac{K_0 \log K}{K^{\sigma_2}}\right). \end{aligned}$$

Similar to (51), we have

$$\begin{aligned} \frac{1}{K} \sum_{k=1}^K \mathbb{E} \|\omega_k - \omega_k^*\|_2^2 &= \mathcal{O}\left(\frac{K_0 \log K}{K}\right) + \mathcal{O}\left(\frac{1}{K - M_K + 1} \sum_{k=M_K}^K \mathbb{E} \|\omega_k - \omega_k^*\|_2^2\right) \\ &= \mathcal{O}\left(\frac{1}{K - M_K + 1} \sum_{k=M_K}^K \mathbb{E} \|\omega_k - \omega_k^*\|_2^2\right) \end{aligned} \quad (71)$$

which completes the proof. \square

B.4 PROOF OF THEOREM 4

Given the definition in section B.2, we now give the convergence proof of actor update in Algorithm 1 with linear value function approximation and Markovian sampling method.

By following the derivation of (53), we have

$$\begin{aligned} \mathbb{E}[J(\theta_{k+1})] &\geq \mathbb{E}[J(\theta_k)] + \underbrace{\alpha_k \mathbb{E} \left\langle \nabla J(\theta_k), \left(\hat{\delta}(x(k), \omega_{k-\tau_k}) - \hat{\delta}(x(k), \omega_k^*) \right) \psi_{\theta_{k-\tau_k}}(s(k), a(k)) \right\rangle}_{I_1} \\ &\quad + \underbrace{\alpha_k \mathbb{E} \left\langle \nabla J(\theta_k), \hat{\delta}(x(k), \omega_k^*) \psi_{\theta_{k-\tau_k}}(s(k), a(k)) \right\rangle}_{I_2} - \frac{L_J}{2} C_p^2 \alpha_k^2. \end{aligned} \quad (72)$$

The item I_1 can be bounded by following (54) as

$$I_1 \geq -\sqrt{2}(1+\gamma) C_\psi \sqrt{\mathbb{E} \|\nabla J(\theta_k)\|_2^2} \sqrt{C_\delta^2 K_0^2 \beta_{k-1}^2 + \mathbb{E} \|\omega_k - \omega_k^*\|_2^2}. \quad (73)$$

Next we consider I_2 . We first decompose it as

$$\begin{aligned}
I_2 &= \mathbb{E} \left\langle \nabla J(\theta_k), \hat{\delta}(x_{(k)}, \omega_k^*) \psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) \right\rangle \\
&= \mathbb{E} \left\langle \nabla J(\theta_k), \underbrace{\left(\hat{\delta}(x_{(k)}, \omega_k^*) - \delta(x_{(k)}, \theta_k) \right)}_{I_2^{(1)}} \psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) \right\rangle \\
&\quad + \mathbb{E} \left\langle \nabla J(\theta_k), \underbrace{\delta(x_{(k)}, \theta_k) \psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) - \nabla J(\theta_k)}_{I_2^{(2)}} \right\rangle + \mathbb{E} \|\nabla J(\theta_k)\|_2^2. \tag{74}
\end{aligned}$$

For some $m \in \mathbb{N}^+$, define $M := (K_0 + 1)m + K_0$. Following Lemma 5, for some $d_m \leq M$ and positive constants $D_2, D_3, D_4, D_5, I_2^{(1)}$ can be bounded as

$$\begin{aligned}
I_2^{(1)} &= \mathbb{E} \left\langle \nabla J(\theta_k), \left(\hat{\delta}(x_{(k)}, \omega_k^*) - \delta(x_{(k)}, \theta_k) \right) \psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) \right\rangle \\
&\geq -D_2 \mathbb{E} \|\theta_{k-\tau_k} - \theta_{k-d_m}\|_2 - D_3 \mathbb{E} \|\theta_k - \theta_{k-d_m}\|_2 - D_4 \sum_{i=k-d_m}^{k-\tau_k} \mathbb{E} \|\theta_i - \theta_{k-d_m}\|_2 \\
&\quad - D_5 \kappa \rho^{m-1} - L_V C_\psi (1 + \gamma) \epsilon_{app} \\
&\geq -D_2 (d_m - \tau_k) C_p \alpha_{k-d_m} - D_3 d_m C_p \alpha_{k-d_m} - D_4 (d_m - \tau_k)^2 C_p \alpha_{k-d_m} \\
&\quad - D_5 \kappa \rho^{m-1} - (1 + \gamma) L_V C_\psi \epsilon_{app}, \tag{75}
\end{aligned}$$

where the derivation of the last inequality is similar to that of (66). By setting $m = m_K$ in (75), and following the fact that $d_{m_K} \leq M_K$ and $\tau_k \geq 0$, we have

$$\begin{aligned}
I_2^{(1)} &\geq -D_2 M_K C_p \alpha_{k-M_K} - D_3 M_K C_p \alpha_{k-M_K} - D_4 M_K^2 C_p \alpha_{k-M_K} - D_5 \kappa \rho^{m_K-1} - (1 + \gamma) L_V C_\psi \epsilon_{app} \\
&= -((D_2 + D_3) C_p M_K + D_4 C_p M_K^2) \alpha_{k-M_K} - D_5 \kappa \rho^{m_K-1} - (1 + \gamma) L_V C_\psi \epsilon_{app} \\
&\geq -((D_2 + D_3) C_p M_K + D_4 C_p M_K^2) \alpha_{k-M_K} - D_5 \alpha_K - (1 + \gamma) L_V C_\psi \epsilon_{app}, \tag{76}
\end{aligned}$$

where the last inequality is due to the definition of m_K .

Following Lemma 6, for some positive constants D_6, D_7 and D_8 , we bound $I_2^{(2)}$ as

$$\begin{aligned}
I_2^{(2)} &= \mathbb{E} \left\langle \nabla J(\theta_k), \delta(x_{(k)}, \theta_k) \psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) - \nabla J(\theta_k) \right\rangle \\
&\geq -D_6 \mathbb{E} \|\theta_{k-\tau_k} - \theta_{k-d_m}\|_2 - D_7 \mathbb{E} \|\theta_k - \theta_{k-d_m}\|_2 - D_8 \sum_{i=\tau_k}^{d_m} \mathbb{E} \|\theta_{k-i} - \theta_{k-d_m}\|_2 - D_9 \kappa \rho^{m-1}.
\end{aligned}$$

Similar to the derivation of (76), we have

$$I_2^{(2)} \geq - (D_6 C_p M_K + D_7 C_p M_K + D_8 C_p M_K^2) \alpha_{k-M_K} - D_9 \alpha_K. \tag{77}$$

Collecting the lower bounds of $I_2^{(1)}$ and $I_2^{(2)}$ yields

$$I_2 \geq -D_K \alpha_{k-M_K} - (D_5 + D_9) \alpha_K - (1 + \gamma) L_V C_\psi \epsilon_{app} + \mathbb{E} \|\nabla J(\theta_k)\|_2^2, \tag{78}$$

where we define $D_K := C_p (D_4 + D_8) M_K^2 + C_p (D_2 + D_3 + D_6 + D_7) M_K$ for brevity.

Substituting (73) and (78) into (72) yields

$$\begin{aligned}
\mathbb{E}[J(\theta_{k+1})] &\geq \mathbb{E}[J(\theta_k)] - \alpha_k \sqrt{2} (1 + \gamma) C_\psi \sqrt{\mathbb{E} \|\nabla J(\theta_k)\|_2^2} \sqrt{C_\delta^2 K_0^2 \beta_{k-1}^2 + \mathbb{E} \|\omega_k - \omega_k^*\|_2^2} \\
&\quad - \alpha_k (D_K \alpha_{k-M_K} + (D_5 + D_9) \alpha_K) - \alpha_k (1 + \gamma) L_V C_\psi \epsilon_{app} + \alpha_k \mathbb{E} \|\nabla J(\theta_k)\|_2^2 - \frac{L_J}{2} C_p^2 \alpha_k^2.
\end{aligned}$$

Rearranging and dividing both sides by α_k yield

$$\begin{aligned}
\mathbb{E} \|\nabla J(\theta_k)\|_2^2 &\leq \frac{1}{\alpha_k} (\mathbb{E}[J(\theta_{k+1})] - \mathbb{E}[J(\theta_k)]) + D_K \alpha_{k-M_K} + (D_5 + D_9) \alpha_K + \frac{L_J}{2} C_p^2 \alpha_k \\
&\quad + \sqrt{2} (1 + \gamma) C_\psi \sqrt{\mathbb{E} \|\nabla J(\theta_k)\|_2^2} \sqrt{C_\delta^2 K_0^2 \beta_{k-1}^2 + \mathbb{E} \|\omega_k - \omega_k^*\|_2^2} + (1 + \gamma) L_V C_\psi \epsilon_{app}.
\end{aligned}$$

Taking summation gives

$$\begin{aligned}
\sum_{k=M_K}^K \mathbb{E} \|\nabla J(\theta_k)\|_2^2 &\leq \underbrace{\sum_{k=M_K}^K \frac{1}{\alpha_k} (\mathbb{E}[J(\theta_{k+1})] - \mathbb{E}[J(\theta_k)])}_{I_3} \\
&\quad + \underbrace{\sum_{k=M_K}^K \left(D_K \alpha_{k-M_K} + \frac{L_J}{2} C_p^2 \alpha_k + (D_5 + D_9) \alpha_K \right)}_{I_4} \\
&\quad + \sqrt{2}(1+\gamma)C_\psi \underbrace{\sum_{k=M_K}^K \sqrt{\mathbb{E} \|\nabla J(\theta_k)\|_2^2} \sqrt{C_\delta^2 K_0^2 \beta_{k-1}^2 + \mathbb{E} \|\omega_k - \omega_k^*\|_2^2}}_{I_5} \\
&\quad + (K - M_K + 1)(1 + \gamma)L_V C_\psi \epsilon_{app}. \tag{79}
\end{aligned}$$

in which the upper bounds of I_3 and I_5 have already been given by (59) and (61) respectively.

We bound I_4 as

$$\begin{aligned}
I_4 &= \sum_{k=M_K}^K \left(D_K \alpha_{k-M_K} + \frac{L_J}{2} C_p^2 \alpha_k + (D_5 + D_9) \alpha_K \right) \\
&\leq \sum_{k=M_K}^K \left(D_K \alpha_{k-M_K} + \frac{L_J}{2} C_p^2 \alpha_{k-M_K} + (D_5 + D_9) \alpha_K \right) \\
&= \left(D_K + \frac{L_J}{2} C_p^2 \right) \sum_{k=M_K}^K \alpha_{k-M_K} + (D_5 + D_9)(K - M_K + 1) \alpha_K \\
&= \left(D_K + \frac{L_J}{2} C_p^2 \right) \sum_{k=0}^{K-M_K} \alpha_k + (D_5 + D_9)(K - M_K + 1) \alpha_K \\
&\leq \left(D_K + \frac{L_J}{2} C_p^2 \right) \frac{c_1}{1 - \sigma_1} K^{1-\sigma_1} + c_1 (D_5 + D_9)(K + 1)^{1-\sigma_1} \\
&= \mathcal{O}((K_0^2 \log^2 K) K^{1-\sigma_1}) \tag{80}
\end{aligned}$$

where the last inequality uses $\sum_{k=a}^b k^{-\sigma} \leq \frac{b^{1-\sigma}}{1-\sigma}$, and the last equality is due to the fact that

$$\mathcal{O}(D_K) = \mathcal{O}(M_K^2 + M_K) = \mathcal{O}((K_0 m_K)^2 + K_0 m_K) = \mathcal{O}(K_0^2 \log^2 K).$$

Substituting the upper bounds of I_3 , I_4 and I_5 into (79), and dividing both sides by $K - M_K + 1$ give

$$\begin{aligned}
\frac{1}{K - M_K + 1} \sum_{k=M_K}^K \mathbb{E} \|\nabla J(\theta_k)\|_2^2 &\leq \frac{\sqrt{2}(1+\gamma)C_\psi}{K - M_K + 1} \sqrt{\sum_{k=M_K}^K \mathbb{E} \|\nabla J(\theta_k)\|_2^2} \sqrt{\mathcal{O}(K_0^2 K^{1-2\sigma_2}) + \sum_{k=M_K}^K \mathbb{E} \|\omega_k - \omega_k^*\|_2^2} \\
&\quad + \mathcal{O}\left(\frac{1}{K^{1-\sigma_1}}\right) + \mathcal{O}\left(\frac{K_0^2 \log^2 K}{K^{\sigma_1}}\right) + \mathcal{O}(\epsilon_{app}). \tag{81}
\end{aligned}$$

Following the similar steps of those in (63), (81) essentially implies

$$\begin{aligned}
\frac{1}{K - M_K + 1} \sum_{k=M_K}^K \mathbb{E} \|\nabla J(\theta_k)\|_2^2 &= \mathcal{O}\left(\frac{1}{K^{1-\sigma_1}}\right) + \mathcal{O}\left(\frac{K_0^2 \log^2 K}{K^{\sigma_1}}\right) + \mathcal{O}\left(\frac{K_0^2}{K^{2\sigma_2}}\right) + \mathcal{O}\left(\frac{1}{K - M_K + 1} \sum_{k=M_K}^K \mathbb{E} \|\omega_k - \omega_{\theta_k}^*\|_2^2\right) + \mathcal{O}(\epsilon_{app}).
\end{aligned}$$

Similar to (64), we have

$$\begin{aligned} \frac{1}{K} \sum_{k=1}^K \mathbb{E} \|\nabla J(\theta_k)\|_2^2 &= \mathcal{O}\left(\frac{K_0 \log K}{K}\right) + \mathcal{O}\left(\frac{1}{K - M_K + 1} \sum_{k=M_K}^K \mathbb{E} \|\nabla J(\theta_k)\|_2^2\right) \\ &= \mathcal{O}\left(\frac{1}{K - M_K + 1} \sum_{k=M_K}^K \mathbb{E} \|\nabla J(\theta_k)\|_2^2\right) \end{aligned}$$

which completes the proof. \square

C SUPPORTING LEMMAS

C.1 SUPPORTING LEMMAS FOR THEOREM 3

Lemma 4. *For any $m \geq 1$ and $k \geq (K_0 + 1)m + K_0 + 1$, we have*

$$\begin{aligned} \mathbb{E} \langle \omega_k - \omega_{\theta_k}^*, g(x_{(k)}, \omega_k) - \bar{g}(\theta_k, \omega_k) \rangle &\leq C_4 \mathbb{E} \|\theta_k - \theta_{k-d_m}\|_2 + C_5 \sum_{i=\tau_k}^{d_m} \mathbb{E} \|\theta_{k-i} - \theta_{k-d_m}\|_2 \\ &\quad + C_6 \mathbb{E} \|\omega_k - \omega_{k-d_m}\|_2 + C_7 \kappa \rho^{m-1}, \end{aligned}$$

where $d_m \leq (K_0 + 1)m + K_0$, and $C_4 := 2C_\delta L_\omega + 4R_\omega C_\delta |\mathcal{A}| L_\pi (1 + \log_\rho \kappa^{-1} + (1 - \rho)^{-1})$, $C_5 := 4R_\omega C_\delta |\mathcal{A}| L_\pi$ and $C_6 := 4(1 + \gamma)R_\omega + 2C_\delta$, $C_7 := 8R_\omega C_\delta$.

Proof. Consider the collection of random samples $\{x_{(k-K_0-1)}, x_{(k-K_0)}, \dots, x_{(k)}\}$. Suppose $x_{(k)}$ is sampled by worker n , then due to Assumption 1, $\{x_{(k-K_0-1)}, x_{(k-K_0)}, \dots, x_{(k-1)}\}$ will contain at least another sample drawn by worker n . Therefore, $\{x_{(k-(K_0+1)m)}, x_{(k-(K_0+1)m+1)}, \dots, x_{(k-1)}\}$ will contain at least m samples from worker n .

Consider the Markov chain formed by $m + 1$ samples in $\{x_{(k-(K_0+1)m)}, x_{(k-(K_0+1)m+1)}, \dots, x_{(k)}\}$:

$$s_{t-m} \xrightarrow{\theta_{k-d_m}} a_{t-m} \xrightarrow{\tilde{\mathcal{P}}} s_{t-m+1} \xrightarrow{\theta_{k-d_m-1}} a_{t-m+1} \cdots s_{t-1} \xrightarrow{\theta_{k-d_1}} a_{t-1} \xrightarrow{\tilde{\mathcal{P}}} s_t \xrightarrow{\theta_{k-d_0}} a_t \xrightarrow{\tilde{\mathcal{P}}} s_{t+1},$$

where $(s_t, a_t, s_{t+1}) = (s_{(k)}, a_{(k)}, s'_{(k)})$, and $\{d_j\}_{j=0}^m$ is some increasing sequence with $d_0 := \tau_k$.

Suppose θ_{k-d_m} was used to do the k_m th update, then we have $x_{t-m} = x_{(k_m)}$. Following Assumption 1, we have $\tau_{k_m} = k_m - (k - d_m) \leq K_0$. Since $x_{(k_m)}$ is in $\{x_{(k-(K_0+1)m)}, \dots, x_{(k)}\}$, we have $k_m \geq k - (K_0 + 1)m$. Combining these two inequalities, we have

$$d_m \leq (K_0 + 1)m + K_0. \quad (82)$$

Given $(s_{t-m}, a_{t-m}, s_{t-m+1})$ and θ_{k-d_m} , we construct an auxiliary Markov chain as that in Lemma 2:

$$s_{t-m} \xrightarrow{\theta_{k-d_m}} a_{t-m} \xrightarrow{\tilde{\mathcal{P}}} s_{t-m+1} \xrightarrow{\theta_{k-d_m}} \tilde{a}_{t-m+1} \cdots \tilde{s}_{t-1} \xrightarrow{\theta_{k-d_m}} \tilde{a}_{t-1} \xrightarrow{\tilde{\mathcal{P}}} \tilde{s}_t \xrightarrow{\theta_{k-d_m}} \tilde{a}_t \xrightarrow{\tilde{\mathcal{P}}} \tilde{s}_{t+1}.$$

For brevity, we define

$$\Delta_1(x, \theta, \omega) := \langle \omega - \omega_\theta^*, g(x, \omega) - \bar{g}(\theta, \omega) \rangle.$$

Throughout this proof, we use $\theta, \theta', \omega, \omega', x$ and \tilde{x} as shorthand notations of $\theta_k, \theta_{k-d_m}, \omega_k, \omega_{k-d_m}, x_t$ and \tilde{x}_t respectively.

First we decompose $\Delta_1(x, \theta, \omega)$ as

$$\begin{aligned} \Delta_1(x, \theta, \omega) &= \underbrace{\Delta_1(x, \theta, \omega) - \Delta_1(x, \theta', \omega)}_{I_1} + \underbrace{\Delta_1(x, \theta', \omega) - \Delta_1(x, \theta', \omega')}_{I_2} \\ &\quad + \underbrace{\Delta_1(x, \theta', \omega') - \Delta_1(\tilde{x}, \theta', \omega')}_{I_3} + \underbrace{\Delta_1(\tilde{x}, \theta', \omega')}_{I_4}. \end{aligned} \quad (83)$$

We bound I_1 in (83) as

$$\begin{aligned} \Delta_1(x, \theta, \omega) - \Delta_1(x, \theta', \omega) &= \langle \omega - \omega_{\theta}^*, g(x, \omega) - \bar{g}(\theta, \omega) \rangle - \langle \omega - \omega_{\theta'}^*, g(x, \omega) - \bar{g}(\theta', \omega) \rangle \\ &\leq |\langle \omega - \omega_{\theta}^*, g(x, \omega) - \bar{g}(\theta, \omega) \rangle - \langle \omega - \omega_{\theta'}^*, g(x, \omega) - \bar{g}(\theta, \omega) \rangle| \\ &\quad + |\langle \omega - \omega_{\theta'}^*, g(x, \omega) - \bar{g}(\theta, \omega) \rangle - \langle \omega - \omega_{\theta'}^*, g(x, \omega) - \bar{g}(\theta', \omega) \rangle|. \end{aligned} \quad (84)$$

For the first term in (84), we have

$$\begin{aligned} |\langle \omega - \omega_{\theta}^*, g(x, \omega) - \bar{g}(\theta, \omega) \rangle - \langle \omega - \omega_{\theta'}^*, g(x, \omega) - \bar{g}(\theta, \omega) \rangle| &= |\langle \omega_{\theta}^* - \omega_{\theta'}^*, g(x, \omega) - \bar{g}(\theta, \omega) \rangle| \\ &\leq \|\omega_{\theta}^* - \omega_{\theta'}^*\|_2 \|g(x, \omega) - \bar{g}(\theta, \omega)\| \\ &\leq 2C_{\delta} \|\omega_{\theta}^* - \omega_{\theta'}^*\|_2 \\ &\leq 2C_{\delta} L_{\omega} \|\theta - \theta'\|_2, \end{aligned}$$

where the last inequality is due to Proposition 2.

We use $x \sim \theta'$ as shorthand notations to represent that $s \sim \mu_{\theta'}$, $a \sim \pi_{\theta'}$, $s' \sim \tilde{\mathcal{P}}$. For the second term in (84), we have

$$\begin{aligned} &|\langle \omega - \omega_{\theta'}^*, g(x, \omega) - \bar{g}(\theta, \omega) \rangle - \langle \omega - \omega_{\theta'}^*, g(x, \omega) - \bar{g}(\theta', \omega) \rangle| \\ &= |\langle \omega - \omega_{\theta'}^*, \bar{g}(\theta', \omega) - \bar{g}(\theta, \omega) \rangle| \\ &\leq \|\omega - \omega_{\theta'}^*\|_2 \|\bar{g}(\theta', \omega) - \bar{g}(\theta, \omega)\|_2 \\ &\leq 2R_{\omega} \|\bar{g}(\theta', \omega) - \bar{g}(\theta, \omega)\|_2 \\ &= 2R_{\omega} \left\| \mathbb{E}_{x \sim \theta'} [g(x, \omega)] - \mathbb{E}_{x \sim \theta} [g(x, \omega)] \right\|_2 \\ &\leq 2R_{\omega} \sup_x \|g(x, \omega)\|_2 \|\mu_{\theta'} \otimes \pi_{\theta'} \otimes \tilde{\mathcal{P}} - \mu_{\theta} \otimes \pi_{\theta} \otimes \tilde{\mathcal{P}}\|_{TV} \\ &\leq 2R_{\omega} C_{\delta} \|\mu_{\theta'} \otimes \pi_{\theta'} \otimes \tilde{\mathcal{P}} - \mu_{\theta} \otimes \pi_{\theta} \otimes \tilde{\mathcal{P}}\|_{TV} \\ &= 4R_{\omega} C_{\delta} d_{TV} \left(\mu_{\theta'} \otimes \pi_{\theta'} \otimes \tilde{\mathcal{P}}, \mu_{\theta} \otimes \pi_{\theta} \otimes \tilde{\mathcal{P}} \right) \\ &\leq 4R_{\omega} C_{\delta} |\mathcal{A}| L_{\pi} (1 + \log_{\rho} \kappa^{-1} + (1 - \rho)^{-1}) \|\theta - \theta'\|_2, \end{aligned}$$

where the third inequality follows the definition of TV norm, the second last inequality follows (30), and the last inequality follows Lemma A.1. in [17].

Collecting the upper bounds of the two terms in (84) yields

$$I_1 \leq [2C_{\delta} L_{\omega} + 4R_{\omega} C_{\delta} |\mathcal{A}| L_{\pi} (1 + \log_{\rho} \kappa^{-1} + (1 - \rho)^{-1})] \|\theta - \theta'\|_2.$$

Next we bound $\mathbb{E}[I_2]$ in (83) as

$$\begin{aligned} \mathbb{E}[I_2] &= \mathbb{E}[\Delta_1(x, \theta', \omega) - \Delta_1(x, \theta', \omega')] \\ &= \mathbb{E} \langle \omega - \omega_{\theta'}^*, g(x, \omega) - \bar{g}(\theta', \omega) \rangle - \langle \omega' - \omega_{\theta'}^*, g(x, \omega') - \bar{g}(\theta', \omega') \rangle \\ &\leq \mathbb{E} |\langle \omega - \omega_{\theta'}^*, g(x, \omega) - \bar{g}(\theta', \omega) \rangle - \langle \omega - \omega_{\theta'}^*, g(x, \omega') - \bar{g}(\theta', \omega') \rangle| \\ &\quad + \mathbb{E} |\langle \omega - \omega_{\theta'}^*, g(x, \omega') - \bar{g}(\theta', \omega') \rangle - \langle \omega' - \omega_{\theta'}^*, g(x, \omega') - \bar{g}(\theta', \omega') \rangle|. \end{aligned} \quad (85)$$

We bound the first term in (85) as

$$\begin{aligned} &\mathbb{E} |\langle \omega - \omega_{\theta'}^*, g(x, \omega) - \bar{g}(\theta', \omega) \rangle - \langle \omega - \omega_{\theta'}^*, g(x, \omega') - \bar{g}(\theta', \omega') \rangle| \\ &= \mathbb{E} |\langle \omega - \omega_{\theta'}^*, g(x, \omega) - g(x, \omega') + \bar{g}(\theta', \omega') - \bar{g}(\theta', \omega) \rangle| \\ &\leq 2R_{\omega} (\mathbb{E} \|g(x, \omega) - g(x, \omega')\|_2 + \mathbb{E} \|\bar{g}(\theta', \omega') - \bar{g}(\theta', \omega)\|_2) \\ &\leq 2R_{\omega} \left(\mathbb{E} \|g(x, \omega) - g(x, \omega')\|_2 + \mathbb{E} \left\| \mathbb{E}_{x \sim \theta'} [g(x, \omega')] - \mathbb{E}_{x \sim \theta'} [g(x, \omega)] \right\|_2 \right) \\ &= 2R_{\omega} \left(\mathbb{E} \|(\gamma\phi(s') - \phi(s))^\top (\omega - \omega')\|_2 + \mathbb{E} \left\| \mathbb{E}_{x \sim \theta'} [(\gamma\phi(s') - \phi(s))^\top] (\omega' - \omega) \right\|_2 \right) \\ &\leq 2R_{\omega} ((1 + \gamma) \mathbb{E} \|\omega - \omega'\|_2 + (1 + \gamma) \mathbb{E} \|\omega - \omega'\|_2) \\ &= 4R_{\omega} (1 + \gamma) \mathbb{E} \|\omega - \omega'\|_2. \end{aligned}$$

We bound the second term in (85) as

$$\begin{aligned} & \mathbb{E} |\langle \omega - \omega_{\theta'}^*, g(x, \omega') - \bar{g}(\theta', \omega') \rangle - \langle \omega' - \omega_{\theta'}^*, g(x, \omega') - \bar{g}(\theta', \omega') \rangle| \\ &= \mathbb{E} |\langle \omega - \omega', g(x, \omega') - \bar{g}(\theta', \omega') \rangle| \\ &\leq 2C_\delta \mathbb{E} \|\omega - \omega'\|_2. \end{aligned}$$

Collecting the upper bounds of the two terms in (85) yields

$$\mathbb{E}[I_2] \leq (4(1 + \gamma)R_\omega + 2C_\delta) \mathbb{E} \|\omega - \omega'\|_2.$$

We first bound I_3 as

$$\begin{aligned} \mathbb{E}[I_3 | \theta', \omega', s_{t-m+1}] &= \mathbb{E} [\Delta_1(x, \theta', \omega') - \Delta_1(\tilde{x}, \theta', \omega') | \theta', \omega', s_{t-m+1}] \\ &\leq |\mathbb{E} [\Delta_1(x, \theta', \omega') | \theta', \omega', s_{t-m+1}] - \mathbb{E} [\Delta_1(\tilde{x}, \theta', \omega') | \theta', \omega', s_{t-m+1}]| \\ &\leq \sup_x |\Delta_1(x, \theta', \omega')| \|\mathbb{P}(x \in \cdot | \theta', \omega', s_{t-m+1}) - \mathbb{P}(\tilde{x} \in \cdot | \theta', \omega', s_{t-m+1})\|_{TV} \\ &\leq 8R_\omega C_\delta d_{TV} (\mathbb{P}(x \in \cdot | \theta', s_{t-m+1}), \mathbb{P}(\tilde{x} \in \cdot | \theta', s_{t-m+1})), \end{aligned} \quad (86)$$

where the second last inequality follows the definition of TV norm, and the last inequality follows the fact that

$$|\Delta_1(x, \theta', \omega')| \leq \|\omega' - \omega_{\theta'}^*\|_2 \|g(x, \omega') - \bar{g}(\theta', \omega')\|_2 \leq 4R_\omega C_\delta.$$

By following (21) in Lemma 2, we have

$$d_{TV} (\mathbb{P}(x \in \cdot | \theta', s_{t-m+1}), \mathbb{P}(\tilde{x} \in \cdot | \theta', s_{t-m+1})) \leq \frac{1}{2} |\mathcal{A}| L_\pi \sum_{i=\tau_k}^{d_m} \mathbb{E} [\|\theta_{k-i} - \theta_{k-d_m}\|_2 | \theta', s_{t-m+1}].$$

Substituting the last inequality into (86), then taking total expectation on both sides yield

$$\mathbb{E}[I_3] \leq 4R_\omega C_\delta |\mathcal{A}| L_\pi \sum_{i=\tau_k}^{d_m} \mathbb{E} \|\theta_{k-i} - \theta_{k-d_m}\|_2.$$

Next we bound I_4 . Define $\bar{x} := (\bar{s}, \bar{a}, \bar{s}')$ where $\bar{s} \sim \mu_{\theta'}$, $\bar{a} \sim \pi_{\theta'}$ and $\bar{s}' \sim \tilde{\mathcal{P}}$. It is immediate that

$$\begin{aligned} \mathbb{E} [\Delta_1(\bar{x}, \theta', \omega') | \theta', \omega', s_{t-m+1}] &= \langle \omega' - \omega_{\theta'}^*, \mathbb{E} [g(\bar{x}, \omega') | \theta', \omega', s_{t-m+1}] - \bar{g}(\theta', \omega') \rangle \\ &= \langle \omega' - \omega_{\theta'}^*, \bar{g}(\theta', \omega') - \bar{g}(\theta', \omega') \rangle = 0. \end{aligned} \quad (87)$$

Then we have

$$\begin{aligned} \mathbb{E}[I_4 | \theta', \omega', s_{t-m+1}] &= \mathbb{E} [\Delta_1(\tilde{x}, \theta', \omega') - \Delta_1(\bar{x}, \theta', \omega') | \theta', \omega', s_{t-m+1}] \\ &\leq |\mathbb{E} [\Delta_1(\tilde{x}, \theta', \omega') | \theta', \omega', s_{t-m+1}] - \mathbb{E} [\Delta_1(\bar{x}, \theta', \omega') | \theta', \omega', s_{t-m+1}]| \\ &\leq \sup_x |\Delta_1(x, \theta', \omega')| \|\mathbb{P}(\tilde{x} \in \cdot | \theta', s_{t-m+1}) - \mathbb{P}(\bar{x} \in \cdot | \theta', s_{t-m+1})\|_{TV} \\ &\leq 8R_\omega C_\delta d_{TV} (\mathbb{P}(\tilde{x} \in \cdot | \theta', s_{t-m+1}), \mathbb{P}(\bar{x} \in \cdot | \theta', s_{t-m+1})) \\ &= 8R_\omega C_\delta d_{TV} (\mathbb{P}(\tilde{x} \in \cdot | \theta', s_{t-m+1}), \mu_{\theta'} \otimes \pi_{\theta'} \otimes \tilde{\mathcal{P}}), \end{aligned} \quad (88)$$

where the second inequality follows the definition of TV norm, and the third inequality follows (87).

The auxiliary Markov chain with policy $\pi_{\theta'}$ starts from initial state s_{t-m+1} , and \tilde{s}_t is the $(m-1)$ th state on the chain. Following Lemma 1, we have:

$$\begin{aligned} & d_{TV} (\mathbb{P}(\tilde{x} \in \cdot | \theta', s_{t-m+1}), \mu_{\theta'} \otimes \pi_{\theta'} \otimes \tilde{\mathcal{P}}) \\ &= d_{TV} (\mathbb{P}((\tilde{s}_t, \tilde{a}_t, \tilde{s}_{t+1}) \in \cdot | \theta', s_{t-m+1}), \mu_{\theta'} \otimes \pi_{\theta'} \otimes \tilde{\mathcal{P}}) \leq \kappa \rho^{m-1}. \end{aligned}$$

Substituting the last inequality into (88) and taking total expectation on both sides yield

$$\mathbb{E}[I_4] \leq 8R_\omega C_\delta \kappa \rho^{m-1}.$$

Taking total expectation on (83) and collecting bounds of I_1, I_2, I_3, I_4 yield

$$\begin{aligned} \mathbb{E} [\Delta_1(x, \theta, \omega)] &\leq C_4 \mathbb{E} \|\theta_k - \theta_{k-d_m}\|_2 + C_5 \sum_{i=\tau_k}^{d_m} \mathbb{E} \|\theta_{k-i} - \theta_{k-d_m}\|_2 \\ &\quad + C_6 \mathbb{E} \|\omega_k - \omega_{k-d_m}\|_2 + C_7 \kappa \rho^{m-1}, \end{aligned}$$

where $C_4 := 2C_\delta L_\omega + 4R_\omega C_\delta |\mathcal{A}| L_\pi (1 + \log_\rho \kappa^{-1} + (1 - \rho)^{-1})$, $C_5 := 4R_\omega C_\delta |\mathcal{A}| L_\pi$, $C_6 := 4(1 + \gamma)R_\omega + 2C_\delta$ and $C_7 := 8R_\omega C_\delta$. \square

C.2 SUPPORTING LEMMAS FOR THEOREM 4

Lemma 5. For any $m \geq 1$ and $k \geq (K_0 + 1)m + K_0 + 1$, we have

$$\begin{aligned} \mathbb{E} \left\langle \nabla J(\theta_k), \left(\hat{\delta}(x_{(k)}, \omega_k^*) - \delta(x_{(k)}, \theta_k) \right) \psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) \right\rangle &\geq -D_2 \mathbb{E} \|\theta_{k-\tau_k} - \theta_{k-d_m}\|_2 \\ &\quad - D_3 \mathbb{E} \|\theta_k - \theta_{k-d_m}\|_2 - D_4 \sum_{i=\tau_k}^{d_m} \mathbb{E} \|\theta_{k-i} - \theta_{k-d_m}\|_2 - D_5 \kappa \rho^{m-1} - L_V C_\psi (1 + \gamma) \epsilon_{app}, \end{aligned}$$

where $D_2 := 2L_V L_\psi C_\delta$, $D_3 := (2C_\delta C_\psi L_J + L_V C_\psi (L_\omega + L_V)(1 + \gamma))$, $D_4 := 2L_V C_\psi C_\delta |A| L_\pi$ and $D_5 := 4L_V C_\psi C_\delta$.

Proof. For the worker that contributes to the k th update, we construct its Markov chain:

$$s_{t-m} \xrightarrow{\theta_{k-d_m}} a_{t-m} \xrightarrow{\tilde{\mathcal{P}}} s_{t-m+1} \xrightarrow{\theta_{k-d_m-1}} a_{t-m+1} \cdots s_{t-1} \xrightarrow{\theta_{k-d_1}} a_{t-1} \xrightarrow{\tilde{\mathcal{P}}} s_t \xrightarrow{\theta_{k-d_0}} a_t \xrightarrow{\tilde{\mathcal{P}}} s_{t+1},$$

where $(s_t, a_t, s_{t+1}) = (s_{(k)}, a_{(k)}, s'_{(k)})$, and $\{d_j\}_{j=0}^m$ is some increasing sequence with $d_0 := \tau_k$. By (82) in Lemma 4, we have $d_m \leq (K_0 + 1)m + K_0$.

Given $(s_{t-m}, a_{t-m}, s_{t-m+1})$ and θ_{k-d_m} , we construct an auxiliary Markov chain:

$$s_{t-m} \xrightarrow{\theta_{k-d_m}} a_{t-m} \xrightarrow{\tilde{\mathcal{P}}} s_{t-m+1} \xrightarrow{\theta_{k-d_m}} \tilde{a}_{t-m+1} \cdots \tilde{s}_{t-1} \xrightarrow{\theta_{k-d_m}} \tilde{a}_{t-1} \xrightarrow{\tilde{\mathcal{P}}} \tilde{s}_t \xrightarrow{\theta_{k-d_m}} \tilde{a}_t \xrightarrow{\tilde{\mathcal{P}}} \tilde{s}_{t+1}.$$

First we have

$$\begin{aligned} &\left\langle \nabla J(\theta_k), \left(\hat{\delta}(x_{(k)}, \omega_k^*) - \delta(x_{(k)}, \theta_k) \right) \psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) \right\rangle \\ &= \left\langle \nabla J(\theta_k), \left(\hat{\delta}(x_{(k)}, \omega_k^*) - \delta(x_{(k)}, \theta_k) \right) \left(\psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) - \psi_{\theta_{k-d_m}}(s_{(k)}, a_{(k)}) \right) \right\rangle \\ &\quad + \left\langle \nabla J(\theta_k), \left(\hat{\delta}(x_{(k)}, \omega_k^*) - \delta(x_{(k)}, \theta_k) \right) \psi_{\theta_{k-d_m}}(s_{(k)}, a_{(k)}) \right\rangle. \end{aligned} \quad (89)$$

We first bound the first term in (89) as

$$\begin{aligned} &\left\langle \nabla J(\theta_k), \left(\hat{\delta}(x_{(k)}, \omega_k^*) - \delta(x_{(k)}, \theta_k) \right) \left(\psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) - \psi_{\theta_{k-d_m}}(s_{(k)}, a_{(k)}) \right) \right\rangle \\ &\geq -\|J(\theta_k)\|_2 \|\hat{\delta}(x_{(k)}, \omega_k^*) - \delta(x_{(k)}, \theta_k)\| \|\psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) - \psi_{\theta_{k-d_m}}(s_{(k)}, a_{(k)})\|_2 \\ &\geq -\|J(\theta_k)\|_2 \left(|\hat{\delta}(x_{(k)}, \omega_k^*)| + |\delta(x_{(k)}, \theta_k)| \right) \|\psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) - \psi_{\theta_{k-d_m}}(s_{(k)}, a_{(k)})\|_2 \\ &\geq -L_V \left(|\hat{\delta}(x_{(k)}, \omega_k^*)| + |\delta(x_{(k)}, \theta_k)| \right) \|\psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) - \psi_{\theta_{k-d_m}}(s_{(k)}, a_{(k)})\|_2 \\ &\geq -2L_V C_\delta \|\psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) - \psi_{\theta_{k-d_m}}(s_{(k)}, a_{(k)})\|_2 \\ &\geq -2L_V L_\psi C_\delta \|\theta_{k-\tau_k} - \theta_{k-d_m}\|_2, \end{aligned} \quad (90)$$

where the last inequality follows Assumption 3 and second last inequality follows

$$\begin{aligned} |\hat{\delta}(x, \omega_\theta^*)| &\leq |r(x)| + \gamma \|\phi(s')\|_2 \|\omega_\theta^*\|_2 + \|\phi(s)\|_2 \|\omega_\theta^*\|_2 \leq r_{\max} + (1 + \gamma) R_\omega \leq C_\delta, \\ |\delta(x, \theta)| &\leq |r(x)| + \gamma |V_{\pi_\theta}(s')| + |V_{\pi_\theta}(s)| \leq r_{\max} + (1 + \gamma) \frac{r_{\max}}{1 - \gamma} \leq C_\delta. \end{aligned}$$

Substituting (90) into (89) gives

$$\begin{aligned} &\left\langle \nabla J(\theta_k), \left(\hat{\delta}(x_{(k)}, \omega_k^*) - \delta(x_{(k)}, \theta_k) \right) \psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) \right\rangle \\ &\geq -2L_V L_\psi C_\delta \|\theta_{k-\tau_k} - \theta_{k-d_m}\|_2 + \left\langle \nabla J(\theta_k), \left(\hat{\delta}(x_{(k)}, \omega_k^*) - \delta(x_{(k)}, \theta_k) \right) \psi_{\theta_{k-d_m}}(s_{(k)}, a_{(k)}) \right\rangle. \end{aligned} \quad (91)$$

Then we start to bound the second term in (91). For brevity, we define

$$\Delta_2(x, \theta) := \left\langle \nabla J(\theta), \left(\hat{\delta}(x, \omega_\theta^*) - \delta(x, \theta) \right) \psi_{\theta_{k-d_m}}(s, a) \right\rangle.$$

In the following proof, we use $\theta, \theta', \omega_\theta^*, \omega_{\theta'}^*, x$ and \tilde{x} as shorthand notations for $\theta_k, \theta_{k-d_m}, \omega_k^*, \omega_{k-d_m}^*, x_t$ and \tilde{x}_t respectively. We also define $\bar{x} := (\bar{s}, \bar{a}, \bar{s}')$, where $\bar{s} \sim \mu_{\theta'}$, $\bar{a} \sim \pi_{\theta'}$ and $\bar{s}' \sim \tilde{\mathcal{P}}$.

We decompose the second term in (91) as

$$\Delta_2(x, \theta) = \underbrace{\Delta_2(x, \theta) - \Delta_2(x, \theta')}_{I_1} + \underbrace{\Delta_2(x, \theta') - \Delta_2(\tilde{x}, \theta')}_{I_2} + \underbrace{\Delta_2(\tilde{x}, \theta') - \Delta_2(\bar{x}, \theta')}_{I_3} + \underbrace{\Delta_2(\bar{x}, \theta')}_{I_4}.$$

We bound the term I_1 as

$$\begin{aligned} I_1 &= \left\langle \nabla J(\theta), \left(\hat{\delta}(x, \omega_\theta^*) - \delta(x, \theta) \right) \psi_{\theta'}(s, a) \right\rangle - \left\langle \nabla J(\theta'), \left(\hat{\delta}(x, \omega_{\theta'}^*) - \delta(x, \theta') \right) \psi_{\theta'}(s, a) \right\rangle \\ &= \left\langle \nabla J(\theta), \left(\hat{\delta}(x, \omega_\theta^*) - \delta(x, \theta) \right) \psi_{\theta'}(s, a) \right\rangle - \left\langle \nabla J(\theta'), \left(\hat{\delta}(x, \omega_\theta^*) - \delta(x, \theta) \right) \psi_{\theta'}(s, a) \right\rangle \\ &\quad + \left\langle \nabla J(\theta'), \left(\hat{\delta}(x, \omega_\theta^*) - \delta(x, \theta) \right) \psi_{\theta'}(s, a) \right\rangle - \left\langle \nabla J(\theta'), \left(\hat{\delta}(x, \omega_{\theta'}^*) - \delta(x, \theta') \right) \psi_{\theta'}(s, a) \right\rangle. \end{aligned}$$

For the first term in I_1 , we have

$$\begin{aligned} &\left\langle \nabla J(\theta), \left(\hat{\delta}(x, \omega_\theta^*) - \delta(x, \theta) \right) \psi_{\theta'}(s, a) \right\rangle - \left\langle \nabla J(\theta'), \left(\hat{\delta}(x, \omega_\theta^*) - \delta(x, \theta) \right) \psi_{\theta'}(s, a) \right\rangle \\ &= \left\langle \nabla J(\theta) - \nabla J(\theta'), \left(\hat{\delta}(x, \omega_\theta^*) - \delta(x, \theta) \right) \psi_{\theta'}(s, a) \right\rangle \\ &\geq -\|\nabla J(\theta) - \nabla J(\theta')\|_2 \|\hat{\delta}(x, \omega_\theta^*) - \delta(x, \theta)\|_2 \|\psi_{\theta'}(s, a)\|_2 \\ &\geq -2C_\delta C_\psi \|\nabla J(\theta) - \nabla J(\theta')\|_2 \\ &\geq -2C_\delta C_\psi L_J \|\theta - \theta'\|_2, \end{aligned}$$

where the last inequality is due to the L_J -Lipschitz of policy gradient shown in Proposition 1.

For the second term in I_1 , we have

$$\begin{aligned} &\left\langle \nabla J(\theta'), \left(\hat{\delta}(x, \omega_\theta^*) - \delta(x, \theta) \right) \psi_{\theta'}(s, a) \right\rangle - \left\langle \nabla J(\theta'), \left(\hat{\delta}(x, \omega_{\theta'}^*) - \delta(x, \theta') \right) \psi_{\theta'}(s, a) \right\rangle \\ &= \left\langle \nabla J(\theta'), \left(\hat{\delta}(x, \omega_\theta^*) - \hat{\delta}(x, \omega_{\theta'}^*) + \delta(x, \theta') - \delta(x, \theta) \right) \psi_{\theta'}(s, a) \right\rangle \\ &\geq -L_V C_\psi \left| \hat{\delta}(x, \omega_\theta^*) - \hat{\delta}(x, \omega_{\theta'}^*) + \delta(x, \theta') - \delta(x, \theta) \right| \\ &\geq -L_V C_\psi \left| \gamma \phi(s')^\top (\omega_\theta^* - \omega_{\theta'}^*) + \phi(s)^\top (\omega_\theta^* - \omega_{\theta'}^*) + \gamma V_{\pi_{\theta'}}(s') - \gamma V_{\pi_\theta}(s') + V_{\pi_\theta}(s) - V_{\pi_{\theta'}}(s) \right| \\ &\geq -L_V C_\psi (\gamma \|\omega_\theta^* - \omega_{\theta'}^*\|_2 + \|\omega_{\theta'}^* - \omega_\theta^*\|_2 + \gamma |V_{\pi_{\theta'}}(s') - V_{\pi_\theta}(s')| + |V_{\pi_\theta}(s) - V_{\pi_{\theta'}}(s)|) \\ &\geq -L_V C_\psi (\gamma L_\omega \|\theta - \theta'\|_2 + L_\omega \|\theta - \theta'\|_2 + \gamma L_V \|\theta - \theta'\|_2 + L_V \|\theta - \theta'\|_2) \\ &= -L_V C_\psi (L_\omega + L_V)(1 + \gamma) \|\theta - \theta'\|_2, \end{aligned}$$

where the last inequality is due to the L_ω -Lipschitz continuity of ω_θ^* shown in Proposition 2 and L_V -Lipschitz continuity of $V_{\pi_\theta}(s)$ shown in Lemma 3. Collecting the upper bounds of I_1 yields

$$I_1 \geq -(2C_\delta C_\psi L_J + L_V C_\psi (L_\omega + L_V)(1 + \gamma)) \|\theta - \theta'\|_2.$$

First we bound I_2 as

$$\begin{aligned} \mathbb{E}[I_2 | \theta', s_{t-m+1}] &= \mathbb{E}[\Delta_2(x, \theta') - \Delta_2(\tilde{x}, \theta') | \theta', s_{t-m+1}] \\ &\geq -|\mathbb{E}[\Delta_2(x, \theta') | \theta', s_{t-m+1}] - \mathbb{E}[\Delta_2(\tilde{x}, \theta') | \theta', s_{t-m+1}]| \\ &\geq -\sup_x |\Delta_2(x, \theta')| \|\mathbb{P}(x \in \cdot | \theta', s_{t-m+1}) - \mathbb{P}(\tilde{x} \in \cdot | \theta', s_{t-m+1})\|_{TV} \\ &\geq -4L_V C_\psi C_\delta d_{TV} (\mathbb{P}(x \in \cdot | \theta', s_{t-m+1}), \mathbb{P}(\tilde{x} \in \cdot | \theta', s_{t-m+1})) \\ &\geq -2L_V C_\psi C_\delta |\mathcal{A}| L_\pi \sum_{i=\tau_k}^{d_m} \mathbb{E}[\|\theta_{k-i} - \theta_{k-d_m}\|_2 | \theta', s_{t-m+1}], \end{aligned} \quad (92)$$

where the second inequality is due to the definition of TV norm, the last inequality follows (21) in Lemma 2, and the second last inequality follows the fact that

$$|\Delta_2(x, \theta')| \leq \|\nabla J(\theta')\|_2 \|\hat{\delta}(x, \omega_{\theta'}^*) - \delta(x, \theta')\| \|\psi_{\theta'}(s, a)\|_2 \leq 2L_V C_\delta C_\psi. \quad (93)$$

Taking total expectation on both sides of (92) yields

$$\mathbb{E}[I_2] \geq -2L_V C_\psi C_\delta |\mathcal{A}| L_\pi \sum_{i=\tau_k}^{d_m} \mathbb{E} \|\theta_{k-i} - \theta_{k-d_m}\|_2.$$

Next we bound I_3 as

$$\begin{aligned} \mathbb{E}[I_3|\theta', s_{t-m+1}] &= \mathbb{E} [\Delta_2(\tilde{x}, \theta') - \Delta_2(\bar{x}, \theta') | \theta', s_{t-m+1}] \\ &\geq -|\mathbb{E} [\Delta_2(\tilde{x}, \theta') | \theta', s_{t-m+1}] - \mathbb{E} [\Delta_2(\bar{x}, \theta') | \theta', s_{t-m+1}]| \\ &\geq -\sup_x |\Delta_2(x, \theta')| \|\mathbb{P}(\tilde{x} \in \cdot | \theta', s_{t-m+1}) - \mathbb{P}(\bar{x} \in \cdot | \theta', s_{t-m+1})\|_{TV} \\ &\geq -4L_V C_\psi C_\delta d_{TV} \left(\mathbb{P}(\tilde{x} \in \cdot | \theta', s_{t-m+1}), \mu_{\theta'} \otimes \pi_{\theta'} \otimes \tilde{\mathcal{P}} \right), \end{aligned} \quad (94)$$

where the second inequality is due to the definition of TV norm, and the last inequality follows (93).

The auxiliary Markov chain with policy $\pi_{\theta'}$ starts from initial state s_{t-m+1} , and \tilde{s}_t is the $(m-1)$ th state on the chain. Following Lemma 1, we have:

$$\begin{aligned} d_{TV} \left(\mathbb{P}(\tilde{x} \in \cdot | \theta', s_{t-m+1}), \mu_{\theta'} \otimes \pi_{\theta'} \otimes \tilde{\mathcal{P}} \right) &= d_{TV} \left(\mathbb{P}((\tilde{s}_t, \tilde{a}_t, \tilde{s}_{t+1}) \in \cdot | \theta', s_{t-m+1}), \mu_{\theta'} \otimes \pi_{\theta'} \otimes \tilde{\mathcal{P}} \right) \\ &\leq \kappa \rho^{m-1}. \end{aligned}$$

Substituting the last inequality into (94) and taking total expectation on both sides yield

$$\mathbb{E}[I_3] \geq -4L_V C_\psi C_\delta \kappa \rho^{m-1}$$

We bound I_4 as

$$\begin{aligned} \mathbb{E}[I_4|\theta'] &= \mathbb{E} \left[\left\langle \nabla J(\theta'), \left(\hat{\delta}(\bar{x}, \omega_{\theta'}^*) - \delta(\bar{x}, \theta') \right) \psi_{\theta'}(s, a) \right\rangle \middle| \theta' \right] \\ &\geq -L_V C_\psi \mathbb{E} \left[\left| \hat{\delta}(\bar{x}, \omega_{\theta'}^*) - \delta(\bar{x}, \theta') \right| \middle| \theta' \right] \\ &= -L_V C_\psi \mathbb{E} \left[\left| \gamma \left(\phi(\bar{s}')^\top \omega_{\theta'}^* - V_{\pi_{\theta'}}(\bar{s}') \right) + V_{\pi_{\theta'}}(\bar{s}) - \phi(\bar{s})^\top \omega_{\theta'}^* \right| \middle| \theta' \right] \\ &\geq -L_V C_\psi \left(\gamma \mathbb{E} \left[\left| \phi(\bar{s}')^\top \omega_{\theta'}^* - V_{\pi_{\theta'}}(\bar{s}') \right| \middle| \theta' \right] + \mathbb{E} \left[\left| V_{\pi_{\theta'}}(\bar{s}) - \phi(\bar{s})^\top \omega_{\theta'}^* \right| \middle| \theta' \right] \right) \\ &\geq -L_V C_\psi \left(\gamma \sqrt{\mathbb{E} \left[\left| \phi(\bar{s}')^\top \omega_{\theta'}^* - V_{\pi_{\theta'}}(\bar{s}') \right|^2 \middle| \theta' \right]} + \sqrt{\mathbb{E} \left[\left| V_{\pi_{\theta'}}(\bar{s}) - \phi(\bar{s})^\top \omega_{\theta'}^* \right|^2 \middle| \theta' \right]} \right) \\ &= -L_V C_\psi \left(\gamma \sqrt{\mathbb{E}_{\bar{s}' \sim \mu_{\theta'}} \left[\left| \phi(\bar{s}')^\top \omega_{\theta'}^* - V_{\pi_{\theta'}}(\bar{s}') \right|^2 \right]} + \sqrt{\mathbb{E}_{\bar{s} \sim \mu_{\theta'}} \left[\left| V_{\pi_{\theta'}}(\bar{s}) - \phi(\bar{s})^\top \omega_{\theta'}^* \right|^2 \right]} \right) \\ &\geq -L_V C_\psi (1 + \gamma) \epsilon_{app} \end{aligned}$$

where the second last inequality follows Jensen's inequality.

Taking total expectation on both sides of (91), and collecting lower bounds of I_1, I_2, I_3 and I_4 yield

$$\begin{aligned} &\mathbb{E} \left\langle \nabla J(\theta_k), \left(\hat{\delta}(x_{(k)}, \omega_k^*) - \delta(x_{(k)}, \theta_k) \right) \psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) \right\rangle \\ &\geq -D_2 \mathbb{E} \|\theta_{k-\tau_k} - \theta_{k-d_m}\|_2 - D_3 \mathbb{E} \|\theta_k - \theta_{k-d_m}\|_2 - D_4 \sum_{i=\tau_k}^{d_m} \mathbb{E} \|\theta_{k-i} - \theta_{k-d_m}\|_2 \\ &\quad - D_5 \kappa \rho^{m-1} - L_V C_\psi (1 + \gamma) \epsilon_{app}, \end{aligned}$$

where $D_2 := 2L_V L_\psi C_\delta$, $D_3 := (2C_\delta C_\psi L_J + L_V C_\psi (L_\omega + L_V)(1 + \gamma))$, $D_4 := 2L_V C_\psi C_\delta |\mathcal{A}| L_\pi$ and $D_5 := 4L_V C_\psi C_\delta$. \square

Lemma 6. For any $m \geq 1$ and $k \geq (K_0 + 1)m + K_0 + 1$, we have

$$\begin{aligned} &\mathbb{E} \left\langle \nabla J(\theta_k), \delta(x_{(k)}, \theta_k) \psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) - \nabla J(\theta_k) \right\rangle \\ &\geq -D_6 \mathbb{E} \|\theta_{k-\tau_k} - \theta_{k-d_m}\|_2 - D_7 \mathbb{E} \|\theta_k - \theta_{k-d_m}\|_2 - D_8 \sum_{i=\tau_k}^{d_m} \mathbb{E} \|\theta_{k-i} - \theta_{k-d_m}\|_2 - D_9 \kappa \rho^{m-1}, \end{aligned}$$

where $d_m \leq (K_0 + 1)m + K_0$, $D_6 := L_V C_\delta L_\psi$, $D_7 := C_p L_J + (1 + \gamma) L_V^2 C_\psi + 2L_V L_J$, $D_8 := L_V C_p |\mathcal{A}| L_\pi$ and $D_9 := 2L_V C_p$.

Proof. For the worker that contributes to the k th update, we construct its Markov chain:

$$s_{t-m} \xrightarrow{\theta_{k-d_m}} a_{t-m} \xrightarrow{\tilde{\mathcal{P}}} s_{t-m+1} \xrightarrow{\theta_{k-d_{m-1}}} a_{t-m+1} \cdots s_{t-1} \xrightarrow{\theta_{k-d_1}} a_{t-1} \xrightarrow{\tilde{\mathcal{P}}} s_t \xrightarrow{\theta_{k-d_0}} a_t \xrightarrow{\tilde{\mathcal{P}}} s_{t+1},$$

where $(s_t, a_t, s_{t+1}) = (s^{(k)}, a^{(k)}, s'^{(k)})$, and $\{d_j\}_{j=0}^m$ is some increasing sequence with $d_0 := \tau_k$. By (82) in Lemma 4, we have $d_m \leq (K_0 + 1)m + K_0$.

Given $(s_{t-m}, a_{t-m}, s_{t-m+1})$ and θ_{k-d_m} , we construct an auxiliary Markov chain:

$$s_{t-m} \xrightarrow{\theta_{k-d_m}} a_{t-m} \xrightarrow{\tilde{\mathcal{P}}} s_{t-m+1} \xrightarrow{\theta_{k-d_m}} \tilde{a}_{t-m+1} \cdots \tilde{s}_{t-1} \xrightarrow{\theta_{k-d_m}} \tilde{a}_{t-1} \xrightarrow{\tilde{\mathcal{P}}} \tilde{s}_t \xrightarrow{\theta_{k-d_m}} \tilde{a}_t \xrightarrow{\tilde{\mathcal{P}}} \tilde{s}_{t+1}.$$

First we have

$$\begin{aligned} & \left\langle \nabla J(\theta_k), \delta(x^{(k)}, \theta_k) \psi_{\theta_{k-\tau_k}}(s^{(k)}, a^{(k)}) - \nabla J(\theta_k) \right\rangle \\ &= \left\langle \nabla J(\theta_k), \delta(x^{(k)}, \theta_k) \left(\psi_{\theta_{k-\tau_k}}(s^{(k)}, a^{(k)}) - \psi_{\theta_{k-d_m}}(s^{(k)}, a^{(k)}) \right) \right\rangle \\ & \quad + \left\langle \nabla J(\theta_k), \delta(x^{(k)}, \theta_k) \psi_{\theta_{k-d_m}}(s^{(k)}, a^{(k)}) - \nabla J(\theta_k) \right\rangle. \end{aligned} \quad (95)$$

We bound the first term in (95) as

$$\begin{aligned} & \left\langle \nabla J(\theta_k), \delta(x^{(k)}, \theta_k) \left(\psi_{\theta_{k-\tau_k}}(s^{(k)}, a^{(k)}) - \psi_{\theta_{k-d_m}}(s^{(k)}, a^{(k)}) \right) \right\rangle \\ & \geq -\|\nabla J(\theta_k)\|_2 \|\delta(x^{(k)}, \theta_k)\|_2 \|\psi_{\theta_{k-\tau_k}}(s^{(k)}, a^{(k)}) - \psi_{\theta_{k-d_m}}(s^{(k)}, a^{(k)})\|_2 \\ & \geq -L_V \|\delta(x^{(k)}, \theta_k)\|_2 \|\psi_{\theta_{k-\tau_k}}(s^{(k)}, a^{(k)}) - \psi_{\theta_{k-d_m}}(s^{(k)}, a^{(k)})\|_2 \\ & \geq -L_V C_\delta \|\psi_{\theta_{k-\tau_k}}(s^{(k)}, a^{(k)}) - \psi_{\theta_{k-d_m}}(s^{(k)}, a^{(k)})\|_2 \\ & \geq -L_V C_\delta L_\psi \|\theta_{k-\tau_k} - \theta_{k-d_m}\|_2, \end{aligned} \quad (96)$$

where the last inequality follows Assumption 3, and the second last inequality follows the fact that

$$|\delta(x, \theta)| \leq |r(x)| + \gamma |V_{\pi_\theta}(s')| + |V_{\pi_\theta}(s)| \leq r_{\max} + (1 + \gamma) \frac{r_{\max}}{1 - \gamma} \leq C_\delta.$$

Substituting (96) into (95) gives

$$\begin{aligned} & \left\langle \nabla J(\theta_k), \delta(x^{(k)}, \theta_k) \psi_{\theta_{k-\tau_k}}(s^{(k)}, a^{(k)}) - \nabla J(\theta_k) \right\rangle \\ & \geq -L_V C_\delta L_\psi \|\theta_{k-\tau_k} - \theta_{k-d_m}\|_2 + \left\langle \nabla J(\theta_k), \delta(x^{(k)}, \theta_k) \psi_{\theta_{k-d_m}}(s^{(k)}, a^{(k)}) - \nabla J(\theta_k) \right\rangle. \end{aligned} \quad (97)$$

Then we start to bound the second term in (97). For brevity, we define

$$\Delta_3(x, \theta) := \left\langle \nabla J(\theta), \delta(x, \theta) \psi_{\theta_{k-d_m}}(s, a) - \nabla J(\theta) \right\rangle.$$

Throughout the following proof, we use θ, θ', x and \tilde{x} as shorthand notations of $\theta_k, \theta_{k-d_m}, x_t$ and \tilde{x}_t respectively.

We decompose $\Delta_3(x, \theta)$ as

$$\Delta_3(x, \theta) = \underbrace{\Delta_3(x, \theta) - \Delta_3(x, \theta')}_{I_1} + \underbrace{\Delta_3(x, \theta') - \Delta_3(\tilde{x}, \theta')}_{I_2} + \underbrace{\Delta_3(\tilde{x}, \theta')}_{I_3}.$$

We first bound I_1 as

$$\begin{aligned} |I_1| &= |\Delta_3(x, \theta) - \Delta_3(x, \theta')| \\ &= \left| \langle \nabla J(\theta), \delta(x, \theta) \psi_{\theta'}(s, a) \rangle - \|\nabla J(\theta)\|_2^2 - \langle \nabla J(\theta'), \delta(x, \theta') \psi_{\theta'}(s, a) \rangle + \|\nabla J(\theta')\|_2^2 \right| \\ &\leq \left| \langle \nabla J(\theta), \delta(x, \theta) \psi_{\theta'}(s, a) \rangle - \langle \nabla J(\theta'), \delta(x, \theta') \psi_{\theta'}(s, a) \rangle \right| + \left| \|\nabla J(\theta')\|_2^2 - \|\nabla J(\theta)\|_2^2 \right| \\ &\leq \left| \langle \nabla J(\theta), \delta(x, \theta) \psi_{\theta'}(s, a) \rangle - \langle \nabla J(\theta'), \delta(x, \theta') \psi_{\theta'}(s, a) \rangle \right| + \|\nabla J(\theta') + \nabla J(\theta)\|_2 \|\nabla J(\theta') - \nabla J(\theta)\|_2 \\ &\leq \left| \langle \nabla J(\theta), \delta(x, \theta) \psi_{\theta'}(s, a) \rangle - \langle \nabla J(\theta'), \delta(x, \theta') \psi_{\theta'}(s, a) \rangle \right| + 2L_V L_J \|\theta - \theta'\|_2, \end{aligned} \quad (98)$$

where the last equality is due to L_V -Lipschitz of value function and L_J -Lipschitz of policy gradient. We bound the first term in (98) as

$$\begin{aligned}
& |\langle \nabla J(\theta), \delta(x, \theta) \psi_{\theta'}(s, a) \rangle - \langle \nabla J(\theta'), \delta(x, \theta') \psi_{\theta'}(s, a) \rangle| \\
& \leq |\langle \nabla J(\theta), \delta(x, \theta) \psi_{\theta'}(s, a) \rangle - \langle \nabla J(\theta), \delta(x, \theta') \psi_{\theta'}(s, a) \rangle| \\
& \quad + |\langle \nabla J(\theta), \delta(x, \theta') \psi_{\theta'}(s, a) \rangle - \langle \nabla J(\theta'), \delta(x, \theta') \psi_{\theta'}(s, a) \rangle| \\
& = |\langle \nabla J(\theta), (\delta(x, \theta) - \delta(x, \theta')) \psi_{\theta'}(s, a) \rangle| + |\langle \nabla J(\theta) - \nabla J(\theta'), \delta(x, \theta') \psi_{\theta'}(s, a) \rangle| \\
& \leq L_V C_\psi |\delta(x, \theta) - \delta(x, \theta')| + C_p \|\nabla J(\theta) - \nabla J(\theta')\|_2 \\
& = L_V C_\psi |\gamma(V_{\pi_\theta}(s') - V_{\pi_{\theta'}}(s')) + V_{\pi_{\theta'}}(s) - V_{\pi_\theta}(s)| + C_p \|\nabla J(\theta) - \nabla J(\theta')\|_2 \\
& \leq L_V C_\psi (\gamma |V_{\pi_\theta}(s') - V_{\pi_{\theta'}}(s')| + |V_{\pi_{\theta'}}(s) - V_{\pi_\theta}(s)|) + C_p \|\nabla J(\theta) - \nabla J(\theta')\|_2 \\
& \leq L_V C_\psi (\gamma L_V \|\theta - \theta'\|_2 + L_V \|\theta' - \theta\|) + C_p L_J \|\theta - \theta'\|_2 \\
& = (C_p L_J + (1 + \gamma) L_V^2 C_\psi) \|\theta - \theta'\|_2.
\end{aligned}$$

Substituting the above inequality into (98) gives the lower bound of I_1 :

$$I_1 \geq - (C_p L_J + (1 + \gamma) L_V^2 C_\psi + 2L_V L_J) \|\theta - \theta'\|_2.$$

First we bound I_2 as

$$\begin{aligned}
\mathbb{E}[I_2 | \theta', s_{t-m+1}] &= \mathbb{E}[\Delta_3(x, \theta') - \Delta_3(\tilde{x}, \theta') | \theta', s_{t-m+1}] \\
&\geq - |\mathbb{E}[\Delta_3(x, \theta') | \theta', s_{t-m+1}] - \mathbb{E}[\Delta_3(\tilde{x}, \theta') | \theta', s_{t-m+1}]| \\
&\geq - \sup_x |\Delta_3(x, \theta')| \|\mathbb{P}(x \in \cdot | \theta', s_{t-m+1}) - \mathbb{P}(\tilde{x} \in \cdot | \theta', s_{t-m+1})\|_{TV} \\
&\geq -2L_V (C_p + L_V) d_{TV}(\mathbb{P}(x \in \cdot | \theta', s_{t-m+1}), \mathbb{P}(\tilde{x} \in \cdot | \theta', s_{t-m+1})) \\
&\geq -L_V (C_p + L_V) |\mathcal{A}| L_\pi \sum_{i=\tau_k}^{d_m} \mathbb{E}[\|\theta_{k-i} - \theta_{k-d_m}\|_2 | \theta', s_{t-m+1}], \quad (99)
\end{aligned}$$

where the second inequality is due to the definition of TV norm, the last inequality is due to (21) in Lemma 2, and thesecond last inequality follows the fact that

$$|\Delta_3(x, \theta')| \leq \|\nabla J(\theta)\|_2 (\|\delta(x, \theta) \psi_{\theta_{k-d_m}}(s, a)\|_2 + \|\nabla J(\theta)\|_2) \leq L_V (C_p + L_V). \quad (100)$$

Taking total expectation on both sides of (99) yields

$$\mathbb{E}[I_2] \geq -L_V (C_p + L_V) |\mathcal{A}| L_\pi \sum_{i=\tau_k}^{d_m} \mathbb{E}[\|\theta_{k-i} - \theta_{k-d_m}\|_2].$$

Define $\bar{x} := (\bar{s}, \bar{a}, \bar{s}')$, where $\bar{s} \sim d_{\theta'}$, $\bar{a} \sim \pi_{\theta'}$ and $\bar{s}' \sim \tilde{\mathcal{P}}$. Then we have

$$\begin{aligned}
\mathbb{E}[\Delta_3(\bar{x}, \theta') | \theta', s_{t-m+1}] &= \mathbb{E}[\langle \nabla J(\theta'), \delta(\bar{x}, \theta') \psi_{\theta'}(\bar{s}, \bar{a}) - \nabla J(\theta') \rangle | \theta', s_{t-m+1}] \\
&= \langle \nabla J(\theta'), \mathbb{E}[\delta(\bar{x}, \theta') \psi_{\theta'}(\bar{s}, \bar{a}) | \theta', s_{t-m+1}] - \nabla J(\theta') \rangle \\
&= \langle \nabla J(\theta'), \nabla J(\theta') - \nabla J(\theta') \rangle = 0.
\end{aligned}$$

Therefore, we bound I_3 as

$$\begin{aligned}
\mathbb{E}[I_3 | \theta', s_{t-m+1}] &= \mathbb{E}[\Delta_3(\tilde{x}, \theta') - \Delta_3(\bar{x}, \theta') | \theta', s_{t-m+1}] \\
&\geq - |\mathbb{E}[\Delta_3(\tilde{x}, \theta') | \theta', s_{t-m+1}] - \mathbb{E}[\Delta_3(\bar{x}, \theta') | \theta', s_{t-m+1}]| \\
&\geq - \sup_x |\Delta_3(x, \theta')| \|\mathbb{P}(\tilde{x} \in \cdot | \theta', s_{t-m+1}) - \mathbb{P}(\bar{x} \in \cdot | \theta', s_{t-m+1})\|_{TV} \\
&\geq -2L_V (C_p + L_V) d_{TV}(\mathbb{P}(\tilde{x} \in \cdot | \theta', s_{t-m+1}), \mathbb{P}(\bar{x} \in \cdot | \theta', s_{t-m+1})) \\
&= -2L_V (C_p + L_V) d_{TV}(\mathbb{P}(\tilde{x} \in \cdot | \theta', s_{t-m+1}), d_{\theta'} \otimes \pi_{\theta'} \otimes \tilde{\mathcal{P}}) \\
&= -2L_V (C_p + L_V) d_{TV}(\mathbb{P}(\tilde{x} \in \cdot | \theta', s_{t-m+1}), \mu_{\theta'} \otimes \pi_{\theta'} \otimes \tilde{\mathcal{P}}) \quad (101)
\end{aligned}$$

where the second inequality follows the definition of total variation norm, and the third inequality follows (100). The last equality is due to the fact shown by [6] that $\mu_{\theta'}(\cdot) = d_{\theta'}(\cdot)$, where $\mu_{\theta'}$ is the stationary distribution of an artificial MDP with transition kernel $\tilde{\mathcal{P}}(\cdot | s, a)$ and policy $\pi_{\theta'}$.

The auxiliary Markov chain with policy $\pi_{\theta'}$ starts from initial state s_{t-m+1} , and \tilde{s}_t is the $(m-1)$ th state on the chain. Following Lemma 1, we have:

$$d_{TV} \left(\mathbb{P}(\tilde{x} \in \cdot | \theta', s_{t-m+1}), \mu_{\theta'} \otimes \pi_{\theta'} \otimes \tilde{\mathcal{P}} \right) = d_{TV} \left(\mathbb{P}((\tilde{s}_t, \tilde{a}_t, \tilde{s}_{t+1}) \in \cdot | \theta', s_{t-m+1}), \mu_{\theta'} \otimes \pi_{\theta'} \otimes \tilde{\mathcal{P}} \right) \leq \kappa \rho^{m-1}.$$

Substituting the last inequality into (101) and taking total expectation on both sides yield

$$\mathbb{E}[I_3] \geq -2L_V(C_p + L_V)\kappa\rho^{m-1}$$

Taking total expectation on $\Delta_3(x, \theta)$ and collecting lower bounds of I_1, I_2, I_3 yield

$$\begin{aligned} \mathbb{E}[\Delta_3(x, \theta)] &\geq - (C_p L_J + (1 + \gamma)L_V^2 C_\psi + 2L_V L_J) \mathbb{E} \|\theta_k - \theta_{k-d_m}\|_2 \\ &\quad - L_V(C_p + L_V)|\mathcal{A}|L_\pi \sum_{i=\tau_k}^{d_m} \mathbb{E} \|\theta_{k-i} - \theta_{k-d_m}\|_2 - 2L_V(C_p + L_V)\kappa\rho^{m-1} \end{aligned}$$

Taking total expectation on (97) and substituting the above inequality into it yield

$$\begin{aligned} &\mathbb{E} \left\langle \nabla J(\theta_k), \delta(x_{(k)}, \theta_k) \psi_{\theta_{k-\tau_k}}(s_{(k)}, a_{(k)}) - \nabla J(\theta_k) \right\rangle \\ &\geq -D_6 \mathbb{E} \|\theta_{k-\tau_k} - \theta_{k-d_m}\|_2 - D_7 \mathbb{E} \|\theta_k - \theta_{k-d_m}\|_2 - D_8 \sum_{i=\tau_k}^{d_m} \mathbb{E} \|\theta_{k-i} - \theta_{k-d_m}\|_2 - D_9 \kappa \rho^{m-1}, \end{aligned}$$

where $D_6 := L_V C_\delta L_\psi$, $D_7 := C_p L_J + (1 + \gamma)L_V^2 C_\psi + 2L_V L_J$, $D_8 := L_V(C_p + L_V)|\mathcal{A}|L_\pi$, $D_9 := 2L_V(C_p + L_V)$. \square

D EXPERIMENT DETAILS

Hardware device. The tests on synthetic environment and CartPole was performed in a 16-core CPU computer. The test on Atari game was run in a 4 GPU computer.

Parameterization. For the synthetic environment, we used linear value function approximation and tabular softmax policy [34]. For CartPole, we used a 3-layer MLP with 128 neurons and sigmoid activation function in each layer. The first two layers are shared for both actor and critic network. For the Atari sequest game, we used a convolution-LSTM network. For network details, see [39].

Hyper-parameters	Value
Number of workers	16
Optimizer	Adam
Step size	0.00015
Batch size	20
Discount factor	0.99
Entropy coefficient	0.01
Frame size	80 × 80
Frame skip rate	4
Grayscaleing	Yes
Training reward clipping	[-1,1]

Table 1: Hyper-parameters of A3C-TD(0) in the Atari sequest game.

Hyper-parameters. For the synthetic environment tests, we run Algorithm 1 with actor step size $\alpha_k = \frac{0.05}{(1+k)^{0.6}}$ and critic step size $\beta_k = \frac{0.05}{(1+k)^{0.4}}$. In tests of CartPole, we run Algorithm 1 with a minibatch of 20 samples. We update the actor network with a step size of $\alpha_k = \frac{0.01}{(1+k)^{0.6}}$ and critic network with a step size of $\beta_k = \frac{0.01}{(1+k)^{0.4}}$. See Table 1 for hyper-parameters to generate the Atari game results in Figure 4.