Supplementary Materials

A. Derivation of Doubly Robust Policy Gradient Estimator

In this section, we introduce how to derive the doubly robust policy gradient $G_{\text{DR}}(w)$ in eq. (4)

Consider the setting of off-policy sampling specified in Section 2. Note that $J(w)$ has the following alternative form:

$$J(w) = (1-\gamma)\mathbb{E}_{\mu_0}\big[V_{\pi_w}(s_0)\big] + \mathbb{E}_d\big[\rho_{\pi_w}(s,a)(r(s,a,s') - Q_{\pi_w}(s,a) + \gamma\mathbb{E}\big[V_{\pi_w}(s')|s,a\big])\big], \tag{16}$$

where $\rho_{\pi_w}(s,a) = \nu_{\pi_w}(s,a)/d(s,a)$ denotes the *distribution correction ratio*. With a sample $(s,a,r,s',a') \sim \mathcal{D}_d \cdot \pi_w(\cdot)$ and a sample $(s_0, a_0) \sim \mu_0 \cdot \pi_w(\cdot)$, we can formulate the following stochastic estimator of $J(w)$:

$$\hat{J}(w) = \underbrace{(1-\gamma)V_{\pi_w}(s_0)}_{\text{unbiased estimator}} + \underbrace{\rho_{\pi_w}(s,a)(r(s,a,s') - Q_{\pi_w}(s,a) + \gamma V_{\pi_w}(s'))}_{\text{baseline}}. \tag{17}$$

Note that the first term in eq. (17) is an unbiased estimator of $J(w)$ and the second term in eq. (17) is the baseline that can help to reduce the variance (Jiang & Li, 2016; Huang & Jiang, 2020). Note that if we replace the value functions V_{π_w}, Q_{π_w} and the density ratio ρ_{π_w} with their estimators $\hat{V}_{\pi_w}$ $\hat{Q}_{\pi_w}$, and $\hat{\rho}_{\pi_w}$, respectively, we can obtain a doubly robust bias reduced value function estimator (Tang et al., 2019). Next, we take the derivative of $\hat{J}(w)$ to obtain an unbiased estimator of $\nabla J(w)$ which takes the following form:

$$\begin{aligned}
\nabla_w \hat{J}(w) &= (1-\gamma)d^v_{\pi_w}(s_0) + d^\rho_{\pi_w}(s,a)(r(s,a,s') - Q_{\pi_w}(s,a) + \gamma V_{\pi_w}(s')) \\
&\quad + \rho_{\pi_w}(s,a)(-d^q_{\pi_w}(s,a) + \gamma d^v_{\pi_w}(s')) \\
&= (1-\gamma)\mathbb{E}_{\pi_w}[Q_{\pi_w}(s_0,a_0)\nabla_w \log \pi_w(s_0,a_0) + d^q_{\pi_w}(s_0,a_0)] \\
&\quad + d^\rho_{\pi_w}(s,a)(r(s,a,s') - Q_{\pi_w}(s,a) + \gamma\mathbb{E}_{\pi_w}[Q_{\pi_w}(s',a')]) \\
&\quad + \rho_{\pi_w}(s,a)(-d^q_{\pi_w}(s,a) + \gamma\mathbb{E}_{\pi_w}[Q_{\pi_w}(s'_i,a'_i)\nabla_w \log \pi_w(s'_i,a'_i) + d^q_{\pi_w}(s'_i,a'_i)]),
\end{aligned} \tag{18}$$

where $d^v_{\pi_w}$, $d^q_{\pi_w}$, $d^\rho_{\pi_w}$ denote $\nabla_w V_{\pi_w}$, $\nabla_w Q_{\pi_w}$, $\nabla_w \rho_{\pi_w}$, respectively. Given samples $s_0 \sim \mu_0(\cdot)$, $a_0 \sim \pi_w(\cdot|s_0)$ and $(s,a,r,s') \sim \mathcal{D}_d$, $a' \sim \pi_w(\cdot|s')$, and replace Q_{π_w}, ρ_{π_w}, $d^\rho_{\pi_w}$ and $d^q_{\pi_w}$ with estimators $\hat{Q}_{\pi_w}$, $\hat{\rho}_{\pi_w}$, $\hat{d}^\rho_{\pi_w}$ and $\hat{d}^q_{\pi_w}$, respectively, we can obtain the following doubly robust estimator $G_{\text{DR}}(w)$:

$$\begin{aligned}
G_{\text{DR}}(w) &= (1-\gamma)\Big(\hat{Q}_{\pi_w}(s_0,a_0)\nabla_w \log \pi_w(a_0|s_0) + \hat{d}^q_{\pi_w}(s_0,a_0)\Big) + \hat{d}^\rho_{\pi_w}(s,a)\,\Big(r(s,a,s') - \hat{Q}_{\pi_w}(s,a) + \gamma\hat{Q}_{\pi_w}(s',a')\Big) \\
&\quad + \hat{\rho}_{\pi_w}(s,a)\Big[-\hat{d}^q_{\pi_w}(s,a) + \gamma\Big(\hat{Q}_{\pi_w}(s',a')\nabla_w \log \pi_w(a'|s') + \hat{d}^q_{\pi_w}(s',a')\Big)\Big].
\end{aligned} \tag{19}$$

Connection with other off-policy gradient estimators: Our doubly robust estimator G_{DR} can recover a number of existing off-policy policy gradient estimators as special cases by deactivating certain estimators, i.e., letting those estimators be zero.

(1) Deactivating $\hat{d}^q_{\pi_w}$ and $\hat{d}^\rho_{\pi_w}$: In this case, $G_{\text{DR}}(w)$ takes the following form

$$\begin{aligned}
G^I_{\text{DR}}(w) &= (1-\gamma)\hat{Q}_{\pi_w}(s_0,a_0)\nabla_w \log \pi_w(a_0|s_0) + \gamma\hat{\rho}_{\pi_w}(s,a)\Big(\hat{Q}_{\pi_w}(s',a')\nabla_w \log \pi_w(a'|s') + \hat{d}^q_{\pi_w}(s',a')\Big) \\
&= \hat{\rho}_{\pi_w}(s,a)\mathbb{E}_{s'\sim\tilde{\mathsf{P}}(\cdot|s,a),\tilde{a}'\sim\pi_w(\cdot|\tilde{s}')}\Big[\hat{Q}_{\pi_w}(\tilde{s}',\tilde{a}')\nabla_w \log \pi_w(\tilde{a}'|\tilde{s}')\Big],
\end{aligned} \tag{20}$$

where $(\tilde{s}', \tilde{a}')$ is generated using the method in the discussion of Critic III in Section 3.2. Note that the policy gradient $\nabla_w J(w)$ has the following equivalent form

$$\nabla_w J(w) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\Big[\rho_{\pi_w}(s,a)\mathbb{E}_{s'\sim\tilde{\mathsf{P}}(\cdot|s,a),\tilde{a}'\sim\pi_w(\cdot|\tilde{s}')}\big[Q_{\pi_w}(\tilde{s}',\tilde{a}')\nabla_w \log \pi_w(\tilde{a}'|\tilde{s}')\big|(s,a)\big]\Big].$$

Thus, $G^I_{\text{DR}}(w)$ can be viewed as an off-policy policy gradient estimator with only approximations of ρ_{π_w} and Q_{π_w}. Such an estimator has been adopted in the previous studies of provably convergent off-policy actor-critic (Zhang et al., 2019c; Liu et al., 2019), which is referred as AC-DC in our experiment in Section 5. Such an estimator has also been adopted by many off-policy actor-critic algorithms such as ACE (Imani et al., 2018), Geoff-PAC (Zhang et al., 2019b), OPPOSD (Liu et al., 2019) and COF-PAC (Zhang et al., 2019c).

(2) Deactivating ρ_{π_w} and $\hat{d}^{\rho}_{\pi_w}$: In this case, $G_{\text{DR}}(w)$ has the following form

$$G^{II}_{\text{DR}}(w) = (1-\gamma)\Big(\hat{Q}_{\pi_w}(s_0,a_0)\nabla_w \log \pi_w(a_0|s_0) + \hat{d}^{q}_{\pi_w}(s_0,a_0)\Big). \tag{21}$$

Such an estimator $G^{II}_{\text{DR}}(w)$ can be viewed as the one adopted by off-policy DPG/DDPG (Silver et al., 2014; Lillicrap et al., 2016) when the policy π_w converges to a deterministic policy.

(3) Deactivting $\hat{Q}_{\pi_w}$ and $\hat{d}^{q}_{\pi_w}$: In this case, $G_{\text{DR}}(w)$ has the following form

$$G^{III}_{\text{DR}}(w) = \hat{d}^{\rho}_{\pi_w}(s,a) r(s,a,s'). \tag{22}$$

This off-policy policy gradient estimator has been adopted in (Morimura et al., 2010) for averaged MDP setting.

B. Proof of Theorem 1

Without specification, the expectation is taken with respect to the randomness of samples $(s,a,r(s,a,s'),s')$ and s_0, in which $(s,a)\sim d(\cdot)$, $s'\sim \mathsf{P}(\cdot|s,a)$ and $s_0\sim\mu_0(\cdot)$. If a' or a_0 appears, then the expectation is taken with respect to the the policy i.e., $a'\sim\pi_w(\cdot|s')$ and $a_0\sim\pi_w(\cdot|s_0)$. We compute the bias of $G_{\text{DR}}(w)$ as follows.

$$\begin{aligned}
&\mathbb{E}[G_{\text{DR}}(w)] - \nabla_w J(w) \\
&= (1-\gamma)\mathbb{E}[\hat{d}^{v}_{\pi_w}(s_0)] + \mathbb{E}[\hat{d}^{\rho}_{\pi_w}(s,a)(r(s,a,s') - \hat{Q}_{\pi_w}(s,a) + \gamma\hat{V}_{\pi_w}(s'))] + \mathbb{E}[\hat{\rho}_{\pi_w}(s,a)(-\hat{d}^{q}_{\pi_w}(s,a) + \gamma\hat{d}^{v}_{\pi_w}(s'))] \\
&\quad - (1-\gamma)\mathbb{E}[d^{v}_{\pi_w}(s_0)] - \mathbb{E}[d^{\rho}_{\pi_w}(s,a)(r(s,a,s') - Q_{\pi_w}(s,a) + \gamma V_{\pi_w}(s'))] - \mathbb{E}[\rho_{\pi_w}(s,a)(-d^{q}_{\pi_w}(s,a) + \gamma d^{v}_{\pi_w}(s'))] \\
&= \mathbb{E}[(\hat{\rho}_{\pi_w}(s,a) - \rho_{\pi_w}(s,a))(-\hat{d}^{q}_{\pi_w}(s,a) + d^{q}_{\pi_w}(s,a))] + \mathbb{E}[(-\hat{d}^{\rho}_{\pi_w}(s,a) + d^{\rho}_{\pi_w}(s,a))(-\hat{Q}_{\pi_w}(s,a) + Q_{\pi_w}(s,a))] \\
&\quad + \gamma\mathbb{E}[(\hat{\rho}_{\pi_w}(s,a) - \rho_{\pi_w}(s,a))(\hat{d}^{v}_{\pi_w}(s) - d^{v}_{\pi_w}(s))] + \gamma\mathbb{E}[(\hat{d}^{\rho}_{\pi_w}(s,a) - d^{\rho}_{\pi_w}(s,a))(\hat{V}_{\pi_w}(s') - V_{\pi_w}(s'))] \\
&\quad + \mathbb{E}[d^{\rho}_{\pi_w}(s,a)(-\hat{Q}_{\pi_w}(s,a) + Q_{\pi_w}(s,a))] + \mathbb{E}[\rho_{\pi_w}(s,a)(-\hat{d}^{q}_{\pi_w}(s,a) + d^{q}_{\pi_w}(s,a))] \\
&\quad + \gamma\mathbb{E}[d^{\rho}_{\pi_w}(s,a)(\hat{V}_{\pi_w}(s') - V_{\pi_w}(s'))] + \gamma\mathbb{E}[\rho_{\pi_w}(s,a)(\hat{d}^{v}_{\pi_w}(s') - d^{v}_{\pi_w}(s'))] + (1-\gamma)\mathbb{E}[\hat{d}^{v}_{\pi_w}(s_0) - d^{v}_{\pi_w}(s_0)] \\
&\quad + \mathbb{E}[(\hat{\rho}_{\pi_w}(s,a) - \rho_{\pi_w}(s,a))(-d^{q}_{\pi_w}(s,a) + \gamma d^{v}_{\pi_w}(s'))] \\
&\quad + \mathbb{E}[(\hat{d}^{\rho}_{\pi_w}(s,a) - d^{\rho}_{\pi_w}(s,a))(r(s,a,s') - Q_{\pi_w}(s,a) + \gamma V_{\pi_w}(s'))] \\
&= -\mathbb{E}[\varepsilon_\rho(s,a)\varepsilon_{d^q}(s,a)] - \mathbb{E}[\varepsilon_{d^\rho}(s,a)\varepsilon_q(s,a)] + \gamma\mathbb{E}[\varepsilon_\rho(s,a)\varepsilon_{d^v}(s')] + \gamma\mathbb{E}[\varepsilon_\rho(s,a)\varepsilon_v(s')] \\
&\quad + S_1 + S_2 + S_3,
\end{aligned} \tag{23}$$

where

$$\begin{aligned}
S_1 &= \mathbb{E}[d^{\rho}_{\pi_w}(s,a)(-\hat{Q}_{\pi_w}(s,a) + Q_{\pi_w}(s,a))] + \mathbb{E}[\rho_{\pi_w}(s,a)(-\hat{d}^{q}_{\pi_w}(s,a) + d^{q}_{\pi_w}(s,a))] \\
&\quad + \gamma\mathbb{E}[d^{\rho}_{\pi_w}(s,a)(\hat{V}_{\pi_w}(s') - V_{\pi_w}(s'))] + \gamma\mathbb{E}[\rho_{\pi_w}(s,a)(\hat{d}^{v}_{\pi_w}(s') - d^{v}_{\pi_w}(s'))] + (1-\gamma)\mathbb{E}[\hat{d}^{v}_{\pi_w}(s_0) - d^{v}_{\pi_w}(s_0)], \\
S_2 &= \mathbb{E}[(\hat{\rho}_{\pi_w}(s,a) - \rho_{\pi_w}(s,a))(-d^{q}_{\pi_w}(s,a) + \gamma d^{v}_{\pi_w}(s'))], \\
S_3 &= \mathbb{E}[(\hat{d}^{\rho}_{\pi_w}(s,a) - d^{\rho}_{\pi_w}(s,a))(r(s,a,s') - Q_{\pi_w}(s,a) + \gamma V_{\pi_w}(s'))].
\end{aligned}$$

We then proceed to show that $S_1 = S_2 = S_3 = 0$. First consider S_1. Following from the definitions of $\hat{d}^{v}_{\pi_w}$ and $d^{v}_{\pi_w}$, we have

$$\begin{aligned}
S_1 &= \mathbb{E}[d^{\rho}_{\pi_w}(s,a)(-\hat{Q}_{\pi_w}(s,a) + Q_{\pi_w}(s,a))] + \mathbb{E}[\rho_{\pi_w}(s,a)(-\hat{d}^{q}_{\pi_w}(s,a) + d^{q}_{\pi_w}(s,a))] + \gamma\mathbb{E}[d^{\rho}_{\pi_w}(s,a)(\hat{V}_{\pi_w}(s') - V_{\pi_w}(s'))] \\
&\quad + \gamma\mathbb{E}[\rho_{\pi_w}(s,a)(\mathbb{E}[\hat{Q}_{\pi_w}(s',a')\nabla_w \log\pi_w(s',a') + \hat{d}^{q}_{\pi_w}(s')|s'] - \mathbb{E}[Q_{\pi_w}(s',a')\nabla_w \log\pi_w(s',a') + d^{q}_{\pi_w}(s')|s'])] \\
&\quad + (1-\gamma)\mathbb{E}[\mathbb{E}[\hat{Q}_{\pi_w}(s_0,a_0)\nabla_w \log\pi_w(s_0,a_0) + \hat{d}^{q}_{\pi_w}(s_0)|s_0] - \mathbb{E}[Q_{\pi_w}(s_0,a_0)\nabla_w \log\pi_w(s_0,a_0) + d^{q}_{\pi_w}(s_0)|s_0]] \\
&= \mathbb{E}[\rho_{\pi_w}(s,a)(d^{q}_{\pi_w}(s,a) - \hat{d}^{q}_{\pi_w}(s,a))] - \gamma\mathbb{E}[\mathbb{E}[d^{q}_{\pi_w}(s',a') - \hat{d}^{q}_{\pi_w}(s',a')|(s,a)\sim\nu_{\pi_w}(s,a)]] \\
&\quad - (1-\gamma)\mathbb{E}[d^{q}_{\pi_w}(s_0,a_0) - \hat{d}^{q}_{\pi_w}(s_0,a_0)] \\
&\quad + \mathbb{E}[d^{\rho}_{\pi_w}(s,a)(Q_{\pi_w}(s,a) - \hat{Q}_{\pi_w}(s,a))] - \gamma\mathbb{E}[d^{\rho}_{\pi_w}(s,a)(V_{\pi_w}(s') - \hat{V}_{\pi_w}(s'))] \\
&\quad - \gamma\mathbb{E}[\rho_{\pi_w}(s,a)\mathbb{E}[(Q_{\pi_w}(s',a') - \hat{Q}_{\pi_w}(s',a'))\nabla_w \log\pi_w(s',a')|s,a]]
\end{aligned}$$

$$- (1-\gamma)\mathbb{E}[(Q_{\pi_w}(s_0,a_0) - \hat{Q}_{\pi_w}(s_0,a_0))\nabla_w \log \pi_w(s_0,a_0)]. \tag{24}$$

For the first three terms in eq. (24), we have

$$\begin{aligned}
&\mathbb{E}[\rho_{\pi_w}(s,a)(d^q_{\pi_w}(s,a) - \hat{d}^q_{\pi_w}(s,a))] - \gamma\mathbb{E}[\mathbb{E}[d^q_{\pi_w}(s',a') - \hat{d}^q_{\pi_w}(s',a')|(s,a)\sim\nu_{\pi_w}(s,a)]] \\
&\qquad\qquad - (1-\gamma)\mathbb{E}[d^q_{\pi_w}(s_0,a_0) - \hat{d}^q_{\pi_w}(s_0,a_0)] \\
&\quad = \mathbb{E}_{(s,a)\sim\nu_{\pi_w}}[d^q_{\pi_w}(s,a) - \hat{d}^q_{\pi_w}(s,a)] - \gamma\mathbb{E}_{(s,a)\sim\nu_{\pi_w}}[\mathbb{E}[d^q_{\pi_w}(s',a') - \hat{d}^q_{\pi_w}(s',a')|s,a]] \\
&\quad\quad - (1-\gamma)\mathbb{E}[d^q_{\pi_w}(s_0,a_0) - \hat{d}^q_{\pi_w}(s_0,a_0)] \\
&\quad \overset{(i)}{=} \mathbb{E}_{(s,a)\sim\nu_{\pi_w}}[d^q_{\pi_w}(s,a) - \hat{d}^q_{\pi_w}(s,a)] - \mathbb{E}_{(s,a)\sim\nu_{\pi_w}}[\mathbb{E}[d^q_{\pi_w}(s',a') - \hat{d}^q_{\pi_w}(s',a')|s'\sim\tilde{\mathsf{P}}(\cdot|s,a), a'\sim\pi_w(\cdot|s')]] \\
&\quad \overset{(ii)}{=} \mathbb{E}_{(s,a)\sim\nu_{\pi_w}}[d^q_{\pi_w}(s,a) - \hat{d}^q_{\pi_w}(s,a)] - \mathbb{E}_{(s',a')\sim\nu_{\pi_w}}[d^q_{\pi_w}(s',a') - \hat{d}^q_{\pi_w}(s',a')] \\
&\quad = 0,
\end{aligned}$$

where (i) follows from the definition $\tilde{\mathsf{P}}(\cdot|s,a) = \gamma\mathsf{P}(\cdot|s,a) + (1-\gamma)\mu_0(\cdot)$, and (ii) follows from the fact that ν_{π_w} is the stationary distribution of MDP with the transition kernel $\tilde{\mathsf{P}}(\cdot|s,a)$ and policy π_w, i.e., $\pi_w(a'|s')\sum_{(s,a)}\nu_{\pi_w}(s,a)\tilde{\mathsf{P}}(s'|s,a) = \nu_{\pi_w}(s',a')$. For the last four terms in eq. (24), note that for any function $f(s,a)$, we have the following holds

$$\begin{aligned}
&\mathbb{E}[d^\rho_{\pi_w}(s,a)f(s,a)] - \gamma\mathbb{E}[d^\rho_{\pi_w}(s,a)\mathbb{E}[f(s',a')|s']] - \gamma\mathbb{E}[\rho_{\pi_w}(s,a)\mathbb{E}[f(s',a')\nabla_w\log\pi_w(s',a')|s,a]] \\
&\qquad - (1-\gamma)\mathbb{E}[f(s_0,a_0)\nabla_w\log\pi_w(s_0,a_0)] \\
&\quad = \nabla_w\mathbb{E}[\rho_{\pi_w}(s,a)f(s,a)] - \gamma\nabla_w\mathbb{E}[\rho_{\pi_w}(s,a)\mathbb{E}[f(s',a')|s']] - (1-\gamma)\nabla_w\mathbb{E}[\mathbb{E}[f(s_0,a_0)|s_0]] \\
&\quad = \nabla_w(\mathbb{E}[\rho_{\pi_w}(s,a)f(s,a)] - \gamma\mathbb{E}[\rho_{\pi_w}(s,a)\mathbb{E}[f(s',a')|s']] - (1-\gamma)\nabla_w\mathbb{E}[\mathbb{E}[f(s_0,a_0)|s_0]]) \\
&\quad \overset{(i)}{=} \nabla_w(\mathbb{E}_{(s,a)\sim\nu_{\pi_w}}[f(s,a)] - \mathbb{E}_{(s',a')\sim\nu_{\pi_w}}[f(s',a')]) \\
&\quad = 0,
\end{aligned}$$

where (i) follows from the reasons similar to how we proceed eq. (24). Letting $f(s,a) = Q_{\pi_w}(s,a) - \hat{Q}_{\pi_w}(s,a)$, we can then conclude that the summation of the last four terms in eq. (24) is 0, which implies $S_1 = 0$.

We then consider the term S_2. Note that for any function $f(s,a)$, we have

$$\begin{aligned}
&\mathbb{E}[f(s,a)(-d^q_{\pi_w}(s,a) + \gamma d^v_{\pi_w}(s'))] \\
&\quad = \nabla_w\mathbb{E}[f(s,a)(r(s,a,s') + \gamma V_{\pi_w}(s') - Q_{\pi_w}(s,a))] \\
&\quad = \nabla_w\mathbb{E}[f(s,a)(\mathbb{E}[r(s,a,s')|s] + \gamma\mathbb{E}[V_{\pi_w}(s')|s,a] - Q_{\pi_w}(s,a))] \\
&\quad = 0.
\end{aligned} \tag{25}$$

Letting $f(s,a) = \hat{\rho}_{\pi_w}(s,a) - \rho_{\pi_w}(s,a)$, we can then conclude that $S_2 = 0$. To consider S_3, we proceed as follows:

$$\begin{aligned}
S_3 &= \mathbb{E}[(\hat{d}^\rho_{\pi_w}(s,a) - d^\rho_{\pi_w}(s,a))(r(s,a,s') - Q_{\pi_w}(s,a) + \gamma V_{\pi_w}(s'))] \\
&= \mathbb{E}[(\hat{d}^\rho_{\pi_w}(s,a) - d^\rho_{\pi_w}(s,a))(\mathbb{E}[r(s,a,s')|s,a] - Q_{\pi_w}(s,a) + \gamma\mathbb{E}[V_{\pi_w}(s')|s,a])] \\
&= 0.
\end{aligned}$$

Since we have shown that $S_1 = S_2 = S_3 = 0$, eq. (23) becomes

$$\begin{aligned}
&\mathbb{E}[G_{\text{DR}}(w)] - \nabla_w J(w) \\
&\quad = -\mathbb{E}[\varepsilon_\rho(s,a)\varepsilon_{d^q}(s,a)] - \mathbb{E}[\varepsilon_{d^\rho}(s,a)\varepsilon_q(s,a)] + \gamma\mathbb{E}[\varepsilon_\rho(s,a)\varepsilon_{d^v}(s')] + \gamma\mathbb{E}[\varepsilon_\rho(s,a)\varepsilon_v(s')],
\end{aligned}$$

which completes the proof.

C. Supporting Lemmas for Theorem 2

In order to develop the property for Critic I's update in eq. (6), we first introduce the following definitions.

Given a sample from mini-batch (s_i, a_i, r_i, s_i'), $a_i' \sim \pi_{w_t}(\cdot|s_i')$ and a sample $s_{0,i} \sim \mu_0$, we define the following matrix $M_{i,t} \in \mathbb{R}^{(2d_1+1)\times(2d_1+1)}$ and vector $m_{i,t} \in \mathbb{R}^{2d_1+1\times 1}$

$$M_{i,t} = \begin{bmatrix} -\phi_i^\top \phi_i & -(\phi_i - \gamma\phi_i')\phi_i^\top & 0 \\ \phi_i(\phi_i^\top - \gamma\phi_i'^\top) & 0 & -\phi_i \\ 0 & \phi_i^\top & -1 \end{bmatrix}, \qquad m_{i,t} = \begin{bmatrix} (1-\gamma)\phi_{0,i} \\ -r_i\phi_i \\ -1 \end{bmatrix}. \tag{26}$$

Moreover, consider the matrix $M_{i,t}$. We have the following holds

$$\begin{aligned} \|M_{i,t}\|_F^2 &= \left\|\phi_i^\top\phi_i\right\|_F^2 + 2\left\|(\phi_i - \gamma\phi_i')\phi_i^\top\right\|_F^2 + 2\|\phi_i\|_2^2 + 1 \\ &\le C_\phi^4 + 2(1+\gamma)^2 C_\phi^4 + 2C_\phi^2 + 1, \end{aligned} \tag{27}$$

which implies $\|M_{i,t}\|_F \le C_M$, where $C_M = \sqrt{9C_\phi^4 + 2C_\phi^2 + 1}$. For the vector $m_{i,t}$, we have

$$\|m_{i,t}\|_2^2 \le (1-\gamma)^2\|\phi_i\|_2^2 + r_{\max}^2\|\phi_i\|_2^2 + 1 \le [(1-\gamma)^2 + r_{\max}^2]C_\phi^2 + 1, \tag{28}$$

which implies $\|m_{i,t}\|_2 \le C_m$, where $C_m = \sqrt{(1 + r_{\max}^2)C_\phi^2 + 1}$.

We define the semi-stochastic-gradient as $g_{i,t}(\kappa) = M_{i,t}\kappa + m_{i,t}$. Then the iteration in eq. (6) can be rewritten as

$$\kappa_{t+1} = \kappa_t + \beta_1\hat{g}_t(\kappa_t), \tag{29}$$

where $\hat{g}_t(\kappa_t) = \frac{1}{N}\sum_i g_{i,t}(\kappa_t)$. We also define $M_t = \mathbb{E}_i[M_{i,t}]$ and $m_t = \mathbb{E}_i[m_{i,t}]$, i.e.,

$$M_t = \begin{bmatrix} -\mathbb{E}_{\mathcal{D}_d\cdot\pi_{w_t}}[\phi^\top\phi] & -\mathbb{E}_{\mathcal{D}_d\cdot\pi_{w_t}}[(\phi - \gamma\phi')\phi^\top] & 0 \\ \mathbb{E}_{\mathcal{D}_d\cdot\pi_{w_t}}[\phi(\phi^\top - \gamma\phi'^\top)] & 0 & -\mathbb{E}_{\mathcal{D}_d\cdot\pi_{w_t}}[\phi] \\ 0 & \mathbb{E}_{\mathcal{D}_d\cdot\pi_{w_t}}[\phi^\top] & -1 \end{bmatrix}, \qquad m_t = \begin{bmatrix} (1-\gamma)\mathbb{E}_{\mu_0\cdot\pi_{w_t}}[\phi] \\ -\mathbb{E}_{\mathcal{D}}[r\phi] \\ -1 \end{bmatrix},$$

and semi-gradient $g_t(\kappa) = M_t\kappa + m_t$. We further define the fixed point of the iteration eq. (29) as

$$\kappa_t^* = M_t^{-1}m_t = [\theta_{q,t}^{*\top}, \theta_{\rho,t}^{*\top}, \eta_t^*]^\top. \tag{30}$$

Lemma 1. *Consider one step update in eq. (6). Define $\kappa_t = [\theta_{q,t}^\top, \theta_{\rho,t}^\top, \eta_t]^\top$ and κ_t^* in eq. (30). Let $\beta_1 \le \min\{1/\lambda_M, \lambda_m/52C_M^2\}$, we have*

$$\mathbb{E}\left[\|\kappa_{t+1} - \kappa_t^*\|_2^2 \,|\mathcal{F}_t\right] \le \left(1 - \frac{1}{2}\beta_1\lambda_M\right)\|\kappa_t - \kappa_t^*\|_2^2 + \frac{C_1}{N},$$

where $C_1 = 24\beta_1^2(C_M^2R_\kappa^2 + C_m^2)$.

Proof. It has been shown in (Zhang et al., 2020b) that M_t is a Hurwitz matrix which satisfies $(\kappa_t - \kappa_t^*)^\top M_t(\kappa_t - \kappa_t^*) \le -\lambda_M\|\kappa_t - \kappa_t^*\|_2^2$, where $\lambda_M > 0$ is a constant. We then proceed as follows:

$$\begin{aligned} \|\kappa_{t+1} - \kappa_t^*\|_2^2 &= \|\kappa_t + \beta_1\hat{g}_t(\kappa_t) - \kappa_t^*\|_2^2 \\ &= \|\kappa_t - \kappa_t^*\|_2^2 + \beta_1\hat{g}_t(\kappa_t)^\top(\kappa_t - \kappa_t^*) + \beta_1^2\|\hat{g}_t(\kappa_t)\|_2^2 \\ &= \|\kappa_t - \kappa_t^*\|_2^2 + \beta_1 g_t(\kappa_t)^\top(\kappa_t - \kappa_t^*) + \beta_1(\hat{g}_t(\kappa_t) - g_t(\kappa_t))^\top(\kappa_t - \kappa_t^*) + \beta_1^2\|\hat{g}_t(\kappa_t) - g_t(\kappa_t) + g_t(\kappa_t)\|_2^2 \\ &\overset{(i)}{\le} (1 - \beta_1\lambda_M)\|\kappa_t - \kappa_t^*\|_2^2 + \beta_1(\hat{g}_t(\kappa_t) - g_t(\kappa_t))^\top(\kappa_t - \kappa_t^*) + 2\beta_1^2\|\hat{g}_t(\kappa_t) - g_t(\kappa_t)\|_2^2 + 2\beta_1^2\|g_t(\kappa_t)\|_2^2 \\ &\overset{(ii)}{\le} (1 - \beta_1\lambda_M + 2\beta_1^2C_M^2)\|\kappa_t - \kappa_t^*\|_2^2 + \beta_1(\hat{g}_t(\kappa_t) - g_t(\kappa_t))^\top(\kappa_t - \kappa_t^*) + 2\beta_1^2\|\hat{g}_t(\kappa_t) - g_t(\kappa_t)\|_2^2, \end{aligned} \tag{31}$$

where (i) follows because $g_t(\kappa_t) = M_t(\kappa_t - \kappa_t^*)$ and (ii) follows because $\|M\|_F \le C_M$. Taking the expectation on both sides of eq. (31) conditional on $\mathcal{F}_t$ yields

$$\mathbb{E}\left[\|\kappa_{t+1} - \kappa_t^*\|_2^2 \,|\mathcal{F}_t\right] \le (1 - \beta_1\lambda_M + \beta_1^2C_M^2)\|\kappa_t - \kappa_t^*\|_2^2 + 2\beta_1^2\mathbb{E}\left[\|\hat{g}_t(\kappa_t) - g_t(\kappa_t)\|_2^2 \,|\mathcal{F}_t\right]. \tag{32}$$

Next we bound the term $\mathbb{E}\left[\|\hat{g}_t(\kappa_t) - g_t(\kappa_t)\|_2^2 \,|\mathcal{F}_t\right]$ as follows

$$\begin{aligned}
&\mathbb{E}\left[\|\hat{g}_t(\kappa_t) - g_t(\kappa_t)\|_2^2 \,|\mathcal{F}_t\right] \\
&= \mathbb{E}\left[\left\|(\hat{M}_t - M_t)\kappa_t + (\hat{m}_t - m_t)\right\|_2^2 |\mathcal{F}_t\right] \\
&= \mathbb{E}\left[\left\|(\hat{M}_t - M_t)(\kappa_t - \kappa_t^*) + (\hat{M}_t - M_t)\kappa_t^* + (\hat{m}_t - m_t)\right\|_2^2 |\mathcal{F}_t\right] \\
&\leq 3\mathbb{E}\left[\left\|\hat{M}_t - M_t\right\|_2^2 |\mathcal{F}_t\right] \|\kappa_t - \kappa_t^*\|_2^2 + 3R_\kappa^2 \mathbb{E}\left[\left\|\hat{M}_t - M_t\right\|_2^2 |\mathcal{F}_t\right] + 3\mathbb{E}\left[\|\hat{m}_t - m_t\|_2^2 \,|\mathcal{F}_t\right]. \qquad (33)
\end{aligned}$$

Recall that $\|M_{i,t}\|_F \leq C_M$ and $\|m_{i,t}\|_2 \leq C_m$. We then have

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{M}_t - M_t\right\|_2^2 |\mathcal{F}_t\right] \leq \mathbb{E}\left[\left\|\hat{M}_t - M_t\right\|_F^2 |\mathcal{F}_t\right] &= \mathbb{E}\left[\left\|\frac{1}{N}\sum_i M_{i,t} - M_t\right\|_F^2 |\mathcal{F}_t\right] \\
&\leq \frac{1}{N^2}\sum_i\sum_j \mathbb{E}\left[\langle M_{i,t} - M_t, M_{j,t} - M_t\rangle|\mathcal{F}_t\right] \\
&= \frac{1}{N^2}\sum_i \mathbb{E}\left[\|M_{i,t} - M_t\|_2^2 \,|\mathcal{F}_t\right] \leq \frac{4C_M^2}{N}. \qquad (34)
\end{aligned}$$

Similarly, we can obtain

$$\mathbb{E}\left[\|\hat{m}_t - m_t\|_2^2 \,|\mathcal{F}_t\right] \leq \frac{4C_m^2}{N}. \qquad (35)$$

Substituting eq. (34) and eq. (35) into eq. (33) yields

$$\mathbb{E}\left[\|\hat{g}_t(\kappa_t) - g_t(\kappa_t)\|_2^2 \,|\mathcal{F}_t\right] \leq \frac{12C_M^2}{N}\|\kappa_t - \kappa_t^*\|_2^2 + \frac{12(C_M^2 R_\kappa^2 + C_m^2)}{N}. \qquad (36)$$

Substituting eq. (36) into eq. (32) yields

$$\mathbb{E}\left[\|\kappa_{t+1} - \kappa_t^*\|_2^2 \,|\mathcal{F}_t\right] \leq (1 - \beta_1\lambda_M + 26\beta_1^2 C_M^2)\|\kappa_t - \kappa_t^*\|_2^2 + \frac{24\beta_1^2(C_M^2 R_\kappa^2 + C_m^2)}{N}. \qquad (37)$$

Letting $\beta_1 \leq \frac{\lambda_m}{52C_M^2}$, we have

$$\mathbb{E}\left[\|\kappa_{t+1} - \kappa_t^*\|_2^2 \,|\mathcal{F}_t\right] \leq \left(1 - \frac{1}{2}\beta_1\lambda_M\right)\|\kappa_t - \kappa_t^*\|_2^2 + \frac{24\beta_1^2(C_M^2 R_\kappa^2 + C_m^2)}{N},$$

which completes the proof. □

We next develop the property for Critic II's update in eq. (9). We first introduce the following definitions.

Given a sample from mini-batch (s_i, a_i, r_i, s_i'), $a_i' \sim \pi_{w_t}(\cdot|s_i')$, we define the following matrix $U_{i,t} \in \mathbb{R}^{2d_3\times 2d_3}$ and vector $u_{i,t} \in \mathbb{R}^{2d_3\times 1}$ as

$$U_{i,t} = \begin{bmatrix} 0 & (\gamma x_i' - x_i)x_i^\top \\ x_i(\gamma x_i' - x_i)^\top & -I \end{bmatrix}, \qquad u_{i,t}(\theta_{q,t}) = \begin{bmatrix} 0 \\ \phi_i'^\top \theta_{q,t} x_i \nabla_w \log \pi_{w_t}(a_i'|s_i') \end{bmatrix}. \qquad (38)$$

Moreover, consider the matrix U_t and the vector u_t. Following the steps similar to those in eq. (27) and eq. (28), we obtain $\|U_{i,t}\|_F \leq C_U$ and $\|u_{i,t}\|_2 \leq C_u$, where $C_U = \sqrt{8C_x^4 + d_3}$ and $C_u = C_\phi R_q C_{sc}$.

We also define the semi-stochastic-gradient as $\ell_{i,t}(\zeta, \theta_q) = U_{i,t}\xi + u_{i,t}(\theta_q)$. Then the iteration in eq. (14) can be rewritten as

$$\zeta_{t+1} = \zeta_t + \beta_3 \hat{\ell}_t(\zeta_t, \theta_{q,t}), \qquad (39)$$

where $\hat{\ell}_t(\zeta_t, \theta_{q,t}) = \frac{1}{N}\sum_i \ell_{i,t}(\zeta_t, \theta_{q_t})$. We define $U_t = \mathbb{E}_i[U_{i,t}]$ and $u_t(\theta^*_{q,t}) = \mathbb{E}_i[u_{i,t}(\theta^*_{q,t})]$, i.e.,

$$U_t = \begin{bmatrix} 0 & \mathbb{E}_{\mathcal{D}_d\cdot\pi_{w_t}}[(\gamma x' - x)x^\top] \\ \mathbb{E}_{\mathcal{D}_d\cdot\pi_{w_t}}[x(\gamma x' - x)^\top] & -I \end{bmatrix}, \qquad u_t(\theta^*_{q,t}) = \begin{bmatrix} 0 \\ \mathbb{E}_{\mathcal{D}_d\cdot\pi_{w_t}}[\phi'^\top\theta^*_{q,t} x \nabla_w \log\pi_{w_t}(a'|s')] \end{bmatrix}.$$

We define the semi-gradient as $\ell_t(\zeta_t, \theta^*_{q,t}) = U_t\zeta_t + u_t(\theta^*_{q,t})$, and the fixed point of the iteration eq. (45) as

$$\zeta^*_t = U_t^{-1}u_t(\theta^*_{q,t}) = [\theta^{*\top}_{d_q,t}, 0^\top]^\top. \tag{40}$$

where $\theta^{*\top}_{d_q,w_t} = A^{d_q-1}_{w_t} b^{d_q}_{w_t}$, with $A^{d_q}_w = \mathbb{E}_{\mathcal{D}_d\cdot\pi_w}[(\gamma x' - x)x'^\top]$ and $b^{d_q}_w = \mathbb{E}_{\mathcal{D}_d\cdot\pi_w}[\phi'^\top\theta^*_{q,w} x\nabla_w \log\pi_w(a'|s')]$.

Lemma 2. *Consider one step update in eq. (9). Define $\zeta_t = [\theta^\top_{d_q,t}, w^\top_{d_q.t}]^\top$ and ζ^*_t in eq. (40). Let $\beta_3 \le \min\{1/\lambda_U, \lambda_U/16C_U^2\}$ and $N \ge \frac{192C_U^2}{\lambda_U}\left(\frac{2}{\lambda_U} + 2\beta_3\right)$, we have*

$$\mathbb{E}\left[\|\zeta_{t+1} - \zeta^*_t\|_2^2 \,|\mathcal{F}_t\right] \le \left(1 - \frac{1}{4}\beta_3\lambda_U\right)\|\zeta_t - \zeta^*_t\|_2^2 + C_3\beta_3\left\|\theta_{q,t} - \theta^*_{q,t}\right\|_2^2 + \frac{C_4}{N},$$

where $C_3 = \left(\frac{4}{\lambda_U} + 4\beta_3\right)C_\phi^2 C_x^2 C_\pi^2$ and $C_4 = \left(\frac{48\beta_3}{\lambda_U} + 48\beta_3^2\right)(C_U^2 R_\zeta^2 + C_\phi^2 R^2_{\theta_q} C_x^2 C_\pi^2)$.

Proof. Following the steps similar to those in the proof of Theorem 1 in Chapter 5 of (Maei, 2011), we can show that U_t is a Hurwitz matrix which $(\zeta_t - \zeta^*_t)^\top U_t(\zeta_t - \zeta^*_t) \le -\lambda_U\|\zeta_t - \zeta^*_t\|_2^2$, where $\lambda_U > 0$ is a constant. Following steps similar to those in the proof of Theorem 4 in (Xu et al., 2020b), we can obtain

$$\begin{aligned}\mathbb{E}\left[\|\zeta_{t+1} - \zeta^*_t\|_2^2\,|\mathcal{F}_t\right] &= \left\|\zeta_t + \beta_3\hat{\ell}_t(\zeta_t,\theta_{q,t}) - \zeta^*_t\right\|_2^2 \\ &\le \left(1 - \frac{1}{2}\beta_3\lambda_U + 2C_U^2\beta_3^2\right)\|\zeta_t - \zeta^*_t\|_2^2 + \left(\frac{2\beta_3}{\lambda_U} + 2\beta_3^2\right)\mathbb{E}\left[\left\|\hat{\ell}_t(\zeta_t,\theta_{q,t}) - \ell_t(\zeta_t,\theta^*_{q,t})\right\|_2^2\Big|\mathcal{F}_t\right].\end{aligned} \tag{41}$$

Next we bound the term $\mathbb{E}\left[\left\|\hat{\ell}_t(\zeta_t,\theta_{q,t}) - \ell_t(\zeta_t,\theta^*_{q,t})\right\|_2^2\Big|\mathcal{F}_t\right]$ as follows:

$$\begin{aligned}&\mathbb{E}\left[\left\|\hat{\ell}_t(\zeta_t,\theta_{q,t}) - \ell_t(\zeta_t,\theta^*_{q,t})\right\|_2^2\Big|\mathcal{F}_t\right] \\ &= \mathbb{E}\left[\left\|\hat{\ell}_t(\zeta_t,\theta_{q,t}) - \hat{\ell}_t(\zeta_t,\theta^*_{q,t}) + \hat{\ell}_t(\zeta_t,\theta^*_{q,t}) - \ell_t(\zeta_t,\theta^*_{q,t})\right\|_2^2\Big|\mathcal{F}_t\right] \\ &= 2\mathbb{E}\left[\left\|\hat{\ell}_t(\zeta_t,\theta_{q,t}) - \hat{\ell}_t(\zeta_t,\theta^*_{q,t})\right\|_2^2\Big|\mathcal{F}_t\right] + 2\mathbb{E}\left[\left\|\hat{\ell}_t(\zeta_t,\theta^*_{q,t}) - \ell_t(\zeta_t,\theta^*_{q,t})\right\|_2^2\Big|\mathcal{F}_t\right] \\ &= 2\mathbb{E}\left[\left\|\frac{1}{N}\sum_i(\ell_{i,t}(\zeta_t,\theta_{q_t}) - \ell_{i,t}(\zeta_t,\theta^*_{q_t}))\right\|_2^2\Big|\mathcal{F}_t\right] + 2\mathbb{E}\left[\left\|\hat{\ell}_t(\zeta_t,\theta^*_{q,t}) - \ell_t(\zeta_t,\theta^*_{q,t})\right\|_2^2\Big|\mathcal{F}_t\right] \\ &\le \frac{2}{N}\sum_i\mathbb{E}\left[\|(\ell_{i,t}(\zeta_t,\theta_{q_t}) - \ell_{i,t}(\zeta_t,\theta^*_{q_t}))\|_2^2\Big|\mathcal{F}_t\right] + 2\mathbb{E}\left[\left\|\hat{\ell}_t(\zeta_t,\theta^*_{q,t}) - \ell_t(\zeta_t,\theta^*_{q,t})\right\|_2^2\Big|\mathcal{F}_t\right] \\ &= \frac{2}{N}\sum_i\mathbb{E}\left[\|\phi_i'^\top(\theta_{q,t} - \theta^*_{q,t})x_i\nabla_w\log\pi_{w_t}(a_i'|s_i')\|_2^2\Big|\mathcal{F}_t\right] + 2\mathbb{E}\left[\left\|\hat{\ell}_t(\zeta_t,\theta^*_{q,t}) - \ell_t(\zeta_t,\theta^*_{q,t})\right\|_2^2\Big|\mathcal{F}_t\right] \\ &= 2C_\phi^2C_x^2C_\pi^2\left\|\theta_{q,t} - \theta^*_{q,t}\right\|_2^2 + 2\mathbb{E}\left[\left\|\hat{\ell}_t(\zeta_t,\theta^*_{q,t}) - \ell_t(\zeta_t,\theta^*_{q,t})\right\|_2^2\Big|\mathcal{F}_t\right].\end{aligned} \tag{42}$$

To bound the term $\mathbb{E}\left[\left\|\hat{\ell}_t(\zeta_t,\theta^*_{q,t}) - \ell_t(\zeta_t,\theta^*_{q,t})\right\|_2^2\Big|\mathcal{F}_t\right]$, we follow the steps similar to those in the proof of bounding the term $\mathbb{E}\left[\|\hat{g}_t(\kappa_t) - g_t(\kappa_t)\|_2^2\,|\mathcal{F}_t\right]$ in Lemma 1 to obtain

$$\mathbb{E}\left[\left\|\hat{\ell}_t(\zeta_t,\theta^*_{q,t}) - \ell_t(\zeta_t,\theta^*_{q,t})\right\|_2^2|\mathcal{F}_t\right] \le \frac{12C_U^2}{N}\|\zeta_t - \zeta^*_t\|_2^2 + \frac{12(C_U^2R_\zeta^2 + C_\phi^2R^2_{\theta_q}C_x^2C_\pi^2)}{N}, \tag{43}$$

Substituting eq. (43) and eq. (42) into eq. (41) yields

$$
\begin{aligned}
&\mathbb{E}\left[\|\zeta_{t+1}-\zeta_t^*\|_2^2\,|\mathcal{F}_t\right] \\
&\leq \left(1-\frac{1}{2}\beta_3\lambda_U+2C_U^2\beta_3^2\right)\|\zeta_t-\zeta_t^*\|_2^2 \\
&\quad+\left(\frac{2\beta_3}{\lambda_U}+2\beta_3^2\right)\left(2C_\phi^2C_x^2C_\pi^2\left\|\theta_{q,t}-\theta_{q,t}^*\right\|_2^2+\frac{24C_U^2}{N}\|\zeta_t-\zeta_t^*\|_2^2+\frac{24(C_U^2R_\zeta^2+C_\phi^2R_{\theta_q}^2C_x^2C_\pi^2)}{N}\right) \\
&=\left[1-\frac{1}{2}\beta_3\lambda_U+2C_U^2\beta_3^2+\left(\frac{2\beta_3}{\lambda_U}+2\beta_3^2\right)\frac{24C_U^2}{N}\right]\|\zeta_t-\zeta_t^*\|_2^2 \\
&\quad+\left(\frac{4\beta_3}{\lambda_U}+4\beta_3^2\right)C_\phi^2C_x^2C_\pi^2\left\|\theta_{q,t}-\theta_{q,t}^*\right\|_2^2+\left(\frac{2\beta_3}{\lambda_U}+2\beta_3^2\right)\frac{24(C_U^2R_\zeta^2+C_\phi^2R_{\theta_q}^2C_x^2C_\pi^2)}{N}.
\end{aligned}
$$

Letting $\beta_3\leq\min\{1/\lambda_U,\lambda_U/16C_U^2\}$ and $N\geq\frac{192C_U^2}{\lambda_U}\left(\frac{2}{\lambda_U}+2\beta_3\right)$, we have

$$
\mathbb{E}\left[\|\zeta_{t+1}-\zeta_t^*\|_2^2\,|\mathcal{F}_t\right]\leq\left(1-\frac{1}{4}\beta_3\lambda_U\right)\|\zeta_t-\zeta_t^*\|_2^2+C_3\beta_3\left\|\theta_{q,t}-\theta_{q,t}^*\right\|_2^2+\frac{C_4}{N},
$$

where $C_3=\left(\frac{4}{\lambda_U}+4\beta_3\right)C_\phi^2C_x^2C_\pi^2$ and $C_4=\left(\frac{48\beta_3}{\lambda_U}+48\beta_3^2\right)(C_U^2R_\zeta^2+C_\phi^2R_{\theta_q}^2C_x^2C_\pi^2)$. □

We next develop the property for Critic III's update. We first introduce the following definitions.

Given a sample $(s_i,a_i,r_i,\tilde{s}_i')$ generated as we discuss in Section 3 and $a_i'\sim\pi_{w_t}(\cdot|s_i')$. We define the following matrix $P_{i,t}\in\mathbb{R}^{2d_1\times 2d_2}$ and vector $p_{i,t}\in\mathbb{R}^{2d_2\times 1}$ as

$$
P_{i,t}=\begin{bmatrix}0 & (\varphi_i-\tilde{\varphi}_i')\tilde{\varphi}_i'^\top\\ \tilde{\varphi}_i'(\varphi_i-\tilde{\varphi}_i')^\top & -I\end{bmatrix},\qquad p_{i,t}=\begin{bmatrix}0\\ \tilde{\varphi}_i'\nabla_w\log\pi_{w_t}(a_i'|\tilde{s}_i')\end{bmatrix}. \tag{44}
$$

Consider the matrix $P_{i,t}$ and the vector $p_{i,t}$. Following the steps similar to those in eq. (27) and eq. (28), we obtain $\|P_{i,t}\|_F\leq C_P$ and $\|p_{i,t}\|_2\leq C_p$, where $C_P=\sqrt{8C_\varphi^4+d_2}$ and $C_p=C_\varphi C_{sc}$.

We also define the semi-stochastic-gradient as $h_{i,t}(\xi)=P_{i,t}\xi+p_{i,t}$. Then the iteration in eq. (14) can be rewritten as

$$
\xi_{t+1}=\xi_t+\beta_2\hat{h}_t(\xi_t), \tag{45}
$$

where $\hat{h}_t(\xi_t)=\frac{1}{N}\sum_i h_{i,t}(\xi_t)$. We define $P_t=\mathbb{E}_i[P_{i,t}]$ and $p_t=\mathbb{E}_i[p_{i,t}]$, i.e.,

$$
P_t=\begin{bmatrix}0 & \mathbb{E}_{\tilde{\mathcal{D}}_d\cdot\pi_{w_t}}[(\varphi-\varphi')\varphi'^\top]\\ \mathbb{E}_{\tilde{\mathcal{D}}_d\cdot\pi_{w_t}}[\varphi'(\varphi-\varphi')^\top] & -I\end{bmatrix},\qquad p_t=\begin{bmatrix}0\\ \mathbb{E}_{\tilde{\mathcal{D}}_d\cdot\pi_{w_t}}[\varphi'\nabla_w\log\pi_{w_t}(a'|s')]\end{bmatrix}.
$$

We further define the fixed point of the iteration eq. (45) as

$$
\xi_t^*=P_t^{-1}p_t=[\theta_{\psi,w_t}^{*\top},0^\top]^\top, \tag{46}
$$

where $\theta_{\psi,w_t}^{*\top}=A_{w_t}^{\xi-1}b_{w_t}^\xi$, with $A_w^\xi=\mathbb{E}_{\tilde{\mathcal{D}}_d\cdot\pi_w}[(\varphi-\varphi')\varphi'^\top]$ and $b_w^\xi=\mathbb{E}_{\tilde{\mathcal{D}}_d\cdot\pi_w}[\varphi'\nabla_w\log\pi_w(a'|\tilde{s}')]$.

Lemma 3. *Consider one step update in eq. (14). Define $\xi_t=[\theta_{\psi,t}^\top,w_{\psi,t}^\top]^\top$ and ξ_t^* in eq. (46). Let $\beta_2\leq\min\{1/\lambda_P,\lambda_P/52C_P^2\}$, we have*

$$
\mathbb{E}\left[\|\xi_{t+1}-\xi_t^*\|_2^2\,|\mathcal{F}_t\right]\leq\left(1-\frac{1}{2}\beta_2\lambda_P\right)\|\xi_t-\xi_t^*\|_2^2+\frac{C_2}{N},
$$

where $C_2=24\beta_2^2(C_P^2R_\xi^2+C_p^2)$.

Proof. Following the steps similar to those in the proof of Theorem 1 in Chapter 5 of (Maei, 2011), we can show that P_t is a Hurwitz matrix that satisfies $(\xi_t - \xi_t^*)^\top P_t(\xi_t - \xi_t^*) \leq -\lambda_P \|\xi_t - \xi_t^*\|_2^2$, where $\lambda_P > 0$ is a constant. Then letting $\beta_2 \leq \min\{\frac{\lambda_p}{52C_P^2}, 1/\lambda_P\}$, following the steps similar to those in the proof of bounding the term $\mathbb{E}\left[\|\hat{g}_t(\kappa_t) - g_t(\kappa_t)\|_2^2 |\mathcal{F}_t\right]$ in Lemma 1, we can obtain

$$\mathbb{E}\left[\|\xi_{t+1} - \xi_t^*\|_2^2 |\mathcal{F}_t\right] \leq \left(1 - \frac{1}{2}\beta_2\lambda_P\right)\|\xi_t - \xi_t^*\|_2^2 + \frac{24\beta_2^2(C_P^2 R_\xi^2 + C_p^2)}{N}.$$

□

Lemma 4. *Consider policy π_{w_1} and π_{w_2}, respectively, with the fixed points $\kappa_1^* = [\theta_{q,w_1}^{*\top}, \theta_{\rho,w_1}^{*\top}, \eta_{w_1}^*]^\top$ and $\kappa_2^* = [\theta_{q,w_1}^{*\top}, \theta_{\rho,w_1}^{*\top}, \eta_{w_1}^*]^\top$ as defined in eq. (46). We have*

$$\begin{aligned}
\left\|\theta_{q,w_1}^{*\top} - \theta_{q,w_2}^{*\top}\right\|_2 &\leq L_q \left\|w_1 - w_2\right\|_2, \\
\left\|\theta_{\rho,w_1}^{*\top} - \theta_{\rho,w_2}^{*\top}\right\|_2 &\leq L_\rho \left\|w_1 - w_2\right\|_2, \\
\left\|\eta_{w_1}^{*} - \eta_{w_2}^{*}\right\|_2 &\leq L_\eta \left\|w_1 - w_2\right\|_2,
\end{aligned}$$

where $L_q = \frac{L_C^\kappa(C_E^\kappa + C_A^\kappa R_\rho) + \lambda_C^\kappa(L_E^\kappa + C_A^\kappa L\rho + L_A^\kappa R_\rho)}{(\lambda_C^\kappa)^2}$, $L_\rho = \frac{C_G^\kappa L_F^\kappa + \lambda_F^\kappa L_G^\kappa}{(\lambda_F^\kappa)^2}$, and $L_\eta = C_D^\kappa L_\rho + L_D^\kappa R_\rho$, which further implies that

$$\|\kappa_1^* - \kappa_2^*\|_2 \leq L_\kappa \left\|w_1 - w_2\right\|_2,$$

where $L_\kappa = \sqrt{L_q^2 + L_p^2 + L_\eta^2}$.

Proof. Define $A_w^\kappa = \mathbb{E}_{\mathcal{D}_d\cdot\pi_w}[(\phi - \gamma\phi')\phi^\top]$, $C_w^\kappa = \mathbb{E}_{\mathcal{D}_d\cdot\pi_{w_t}}[\phi^\top\phi]$, $D_w^\kappa = \mathbb{E}_{\mathcal{D}_d\cdot\pi_{w_t}}[\phi]$ and $E_w^\kappa = (1-\gamma)\mathbb{E}_{\mu_0\cdot\pi_{w_t}}[\phi]$. (Zhang et al., 2020b;a) showed that

$$\begin{aligned}
\theta_{\rho,w}^* &= -F_w^{\kappa\,-1} G_w^\kappa, \\
\theta_{q,w}^* &= C_w^{\kappa-1}(E_w^\kappa - A_w^\kappa\theta_{\rho,w}^*), \\
\eta_w^* &= D_w^\kappa\theta_{\rho,w}^* - 1,
\end{aligned}$$

where

$$\begin{aligned}
F_w^\kappa &= A_w^{\kappa\top} C_w^{\kappa\,-1} A_w^\kappa + D_w^\kappa D_w^{\kappa\top}, \\
G_w^\kappa &= A_w^{\kappa\top} C_w^{\kappa\,-1} E_w^\kappa + D_w^\kappa.
\end{aligned}$$

We first develop the Lipschitz property for the matrices A_w^κ, C_w^κ, D_w^κ, and E_w^κ. For A_w^κ, we obtain the following

$$\begin{aligned}
&\left\|A_{w_1}^\kappa - A_{w_2}^\kappa\right\|_2 \\
&= \left\|-\mathbb{E}_{\tilde{\mathcal{D}}_d\cdot\pi_{w_1}}[(\gamma\phi' - \phi)\phi^\top] + \mathbb{E}_{\tilde{\mathcal{D}}_d\cdot\pi_{w_2}}[(\gamma\phi' - \phi)\phi^\top]\right\|_2 \\
&= \left\|\int \gamma\phi(s',a')\phi(s,a)^\top(\pi_{w_2}(da'|s') - \pi_{w_1}(da'|s'))\tilde{\mathsf{P}}(ds'|s,a)\mathcal{D}(ds,da)\right\|_2 \\
&\quad + \left\|\int \phi(s,a)\phi(s,a)^\top(\pi_{w_1}(da'|s') - \pi_{w_2}(da'|s'))\tilde{\mathsf{P}}(ds'|s,a)\mathcal{D}(ds,da)\right\|_2 \\
&\leq \int \left\|\gamma\phi(s',a')\phi(s,a)^\top\right\|_2 |(\pi_{w_1}(da'|s') - \pi_{w_2}(da'|s'))|\,\tilde{\mathsf{P}}(ds'|s,a)\mathcal{D}(ds,da) \\
&\quad + \int \left\|\phi(s,a)\phi(s,a)^\top\right\|_2 |(\pi_{w_2}(da'|s') - \pi_{w_1}(da'|s'))|\,\tilde{\mathsf{P}}(ds'|s,a)\mathcal{D}(ds,da) \\
&\leq 2C_\phi^2 \int |\pi_{w_1}(da'|s') - \pi_{w_2}(da'|s')|\,\tilde{\mathsf{P}}(ds'|s,a)\mathcal{D}(ds,da)
\end{aligned}$$

$$
\begin{aligned}
&\leq 2C_\phi^2 \left\| \pi_{w_1}(\cdot) - \pi_{w_2}(\cdot) \right\|_{TV} \\
&\leq 2C_\phi^2 L_\pi \left\| w_1 - w_2 \right\|_2 .
\end{aligned} \tag{47}
$$

For C_w^κ, D_w^κ, and E_w^κ, following the steps similar to those in eq. (47), we obtain

$$
\begin{aligned}
\left\| C_{w_1}^\kappa - C_{w_2}^\kappa \right\|_2 &\leq C_\phi^2 L_\pi \left\| w_1 - w_2 \right\|_2 = L_C^\kappa \left\| w_1 - w_2 \right\|_2 , \\
\left\| D_{w_1}^\kappa - D_{w_2}^\kappa \right\|_2 &\leq C_\phi L_\pi \left\| w_1 - w_2 \right\|_2 = L_D^\kappa \left\| w_1 - w_2 \right\|_2 , \\
\left\| E_{w_1}^\kappa - E_{w_2}^\kappa \right\|_2 &\leq (1-\gamma) C_\phi L_\pi \left\| w_1 - w_2 \right\|_2 = L_E^\kappa \left\| w_1 - w_2 \right\|_2 ,
\end{aligned} \tag{48}
$$

where $L_C^\kappa = C_\phi^2 L_\pi$, $L_D^\kappa = C_\phi L_\pi$, and $L_E^\kappa = (1-\gamma)C_\phi L_\pi$. We then proceed to bound the two terms $\left\| F_{w_1}^\kappa - F_{w_2}^\kappa \right\|_2$ and $\left\| G_{w_1}^\kappa - G_{w_2}^\kappa \right\|_2$. For $\left\| F_{w_1}^\kappa - F_{w_2}^\kappa \right\|_2$, we obtain the following bound:

$$
\begin{aligned}
&\left\| F_{w_1}^\kappa - F_{w_2}^\kappa \right\|_2 \\
&= \left\| A_{w_1}^{\kappa\top} C_{w_1}^{\kappa\,-1} A_{w_1}^\kappa - A_{w_2}^{\kappa\top} C_{w_2}^{\kappa\,-1} A_{w_2}^\kappa + D_{w_1}^\kappa D_{w_1}^{\kappa\top} - D_{w_2}^\kappa D_{w_2}^{\kappa\top} \right\|_2 \\
&\leq \left\| A_{w_1}^{\kappa\top} C_{w_1}^{\kappa\,-1} A_{w_1}^\kappa - A_{w_2}^{\kappa\top} C_{w_2}^{\kappa\,-1} A_{w_2}^\kappa \right\|_2 + \left\| D_{w_1}^\kappa D_{w_1}^{\kappa\top} - D_{w_2}^\kappa D_{w_2}^{\kappa\top} \right\|_2 \\
&= \left\| A_{w_1}^{\kappa\top} C_{w_1}^{\kappa\,-1} (C_{w_2}^\kappa - C_{w_1}^\kappa) C_{w_2}^{\kappa\,-1} A_{w_1}^\kappa + (A_{w_1}^\kappa - A_{w_2}^\kappa)^\top C_{w_2}^{\kappa\,-1} A_{w_1}^\kappa + A_{w_2}^\kappa C_{w_2}^{\kappa\,-1} (A_{w_1}^\kappa - A_{w_2}^\kappa) \right\|_2 \\
&\quad + \left\| D_{w_1}^\kappa (D_{w_1}^\kappa - D_{w_2}^\kappa)^\top + (D_{w_1}^\kappa - D_{w_2}^\kappa) D_{w_2}^{\kappa\top} \right\|_2 \\
&\leq \left\| A_{w_1}^{\kappa\top} C_{w_1}^{\kappa\,-1} (C_{w_2}^\kappa - C_{w_1}^\kappa) C_{w_2}^{\kappa\,-1} A_{w_1}^\kappa \right\|_2 + \left\| (A_{w_1}^\kappa - A_{w_2}^\kappa)^\top C_{w_2}^{\kappa\,-1} A_{w_1}^\kappa \right\|_2 + \left\| A_{w_2}^\kappa C_{w_2}^{\kappa\,-1} (A_{w_1}^\kappa - A_{w_2}^\kappa) \right\|_2 \\
&\quad + \left\| D_{w_1}^\kappa (D_{w_1}^\kappa - D_{w_2}^\kappa)^\top \right\|_2 + \left\| (D_{w_1}^\kappa - D_{w_2}^\kappa) D_{w_2}^{\kappa\top} \right\|_2 \\
&\leq \left(\frac{L_C^\kappa (C_A^\kappa)^2 + 2 C_A^\kappa \lambda_C^\kappa L_C^\kappa}{(\lambda_C^\kappa)^2} + C_D^\kappa L_D^\kappa \right) \left\| w_1 - w_2 \right\|_2 \\
&= L_F^\kappa \left\| w_1 - w_2 \right\|_2 ,
\end{aligned} \tag{49}
$$

where

$$
L_F^\kappa = \frac{L_C^\kappa (C_A^\kappa)^2 + 2 C_A^\kappa \lambda_C^\kappa L_C^\kappa}{(\lambda_C^\kappa)^2} + C_D^\kappa L_D^\kappa .
$$

For $\left\| G_{w_1}^\kappa - G_{w_2}^\kappa \right\|_2$, we have

$$
\begin{aligned}
&\left\| G_{w_1}^\kappa - G_{w_2}^\kappa \right\|_2 \\
&= \left\| A_{w_1}^{\kappa\top} C_{w_1}^{\kappa\,-1} E_{w_1}^\kappa - A_{w_2}^{\kappa\top} C_{w_2}^{\kappa\,-1} E_{w_2}^\kappa + D_{w_1}^\kappa - D_{w_2}^\kappa \right\|_2 \\
&\leq \left\| A_{w_1}^{\kappa\top} C_{w_1}^{\kappa\,-1} E_{w_1}^\kappa - A_{w_2}^{\kappa\top} C_{w_2}^{\kappa\,-1} E_{w_2}^\kappa \right\|_2 + \left\| D_{w_1}^\kappa - D_{w_2}^\kappa \right\|_2 \\
&= \left\| A_{w_1}^{\kappa\top} C_{w_1}^{\kappa\,-1} (C_{w_2}^\kappa - C_{w_1}^\kappa) C_{w_2}^{\kappa\,-1} E_{w_1}^\kappa + (A_{w_1}^\kappa - A_{w_2}^\kappa)^\top C_{w_2}^{\kappa\,-1} E_{w_1}^\kappa + A_{w_2}^\kappa C_{w_2}^{\kappa\,-1} (E_{w_1}^\kappa - E_{w_2}^\kappa) \right\|_2 + \left\| D_{w_1}^\kappa - D_{w_2}^\kappa \right\|_2 \\
&\leq \left(\frac{L_C^\kappa C_A^\kappa C_E^\kappa + C_E^\kappa \lambda_C^\kappa L_A^\kappa + C_A^\kappa \lambda_C^\kappa L_E^\kappa}{(\lambda_C^\kappa)^2} + L_D^\kappa \right) \left\| w_1 - w_2 \right\|_2 \\
&= L_G^\kappa \left\| w_1 - w_2 \right\|_2 ,
\end{aligned} \tag{50}
$$

where

$$
L_G^\kappa = \frac{L_C^\kappa C_A^\kappa C_E^\kappa + C_E^\kappa \lambda_C^\kappa L_A^\kappa + C_A^\kappa \lambda_C^\kappa L_E^\kappa}{(\lambda_C^\kappa)^2} + L_D^\kappa .
$$

We next prove the Lipschitz property for θ_ρ^*. To bound $\left\| \theta_{\rho,w_1}^{*\top} - \theta_{\rho,w_2}^{*\top} \right\|_2$, we proceed as follows.

$$
\begin{aligned}
\theta_{\rho,w_1}^{*\top} - \theta_{\rho,w_2}^{*\top} &= F_{w_1}^{\kappa-1} G_{w_1}^\kappa - F_{w_2}^{\kappa-1} G_{w_2}^\kappa = (F_{w_1}^{\kappa-1} - F_{w_2}^{\kappa-1}) G_{w_1}^\kappa + F_{w_2}^{\kappa-1} (G_{w_1}^\kappa - G_{w_2}^\kappa) \\
&= (F_{w_1}^{\kappa-1} F_{w_2}^\kappa F_{w_2}^{\kappa-1} - F_{w_1}^{\kappa-1} F_{w_1}^\kappa F_{w_2}^{\kappa-1}) G_{w_1}^\kappa + F_{w_2}^{\kappa-1} (G_{w_1}^\kappa - G_{w_2}^\kappa) \\
&= F_{w_1}^{\kappa-1} (F_{w_2}^\kappa - F_{w_1}^\kappa) F_{w_2}^{\kappa-1} G_{w_1}^\kappa + F_{w_2}^{\kappa-1} (G_{w_1}^\kappa - G_{w_2}^\kappa),
\end{aligned} \tag{51}
$$

which implies

$$
\begin{aligned}
\left\|\theta^{*\top}_{\rho,w_1}-\theta^{*\top}_{\rho,w_2}\right\|_2 &\le \left\|F^{\kappa-1}_{w_1}\right\|_2\left\|F^{\kappa}_{w_2}-F^{\kappa}_{w_1}\right\|_2\left\|F^{\kappa-1}_{w_2}\right\|_2\left\|G^{\kappa}_{w_1}\right\|_2+\left\|F^{\kappa-1}_{w_2}\right\|_2\left\|G^{\kappa}_{w_1}-G^{\kappa}_{w_2}\right\|_2\\
&\le \frac{C^{\kappa}_G}{(\lambda^{\kappa}_F)^2}\left\|F^{\kappa}_{w_1}-F^{\kappa}_{w_2}\right\|_2+\frac{1}{\lambda^{\kappa}_F}\left\|G^{\kappa}_{w_1}-G^{\kappa}_{w_2}\right\|_2\\
&\le \frac{C^{\kappa}_G L^{\kappa}_F+\lambda^{\kappa}_F L^{\kappa}_G}{(\lambda^{\kappa}_F)^2}\left\|w_1-w_2\right\|_2 = L_\rho\left\|w_1-w_2\right\|_2, \qquad (52)
\end{aligned}
$$

where $L_\rho=\frac{C^{\kappa}_G L^{\kappa}_F+\lambda^{\kappa}_F L^{\kappa}_G}{(\lambda^{\kappa}_F)^2}$.

We then consider the Lipschitz property of θ^*_q. To bound $\left\|\theta^*_{q,w_1}-\theta^{*\top}_{q,w_2}\right\|_2$, we proceed as follows.

$$
\begin{aligned}
&\theta^*_{q,w_1}-\theta^*_{q,w_2}\\
&=C^{\kappa-1}_{w_1}(E^{\kappa}_{w_1}-A^{\kappa}_{w_1}\theta^*_{\rho,w_1})-C^{\kappa-1}_{w_2}(E^{\kappa}_{w_2}-A^{\kappa}_{w_2}\theta^*_{\rho,w_2})\\
&=(C^{\kappa-1}_{w_1}-C^{\kappa-1}_{w_2})(E^{\kappa}_{w_1}-A^{\kappa}_{w_1}\theta^*_{\rho,w_1})+C^{\kappa-1}_{w_2}(E^{\kappa}_{w_1}-E^{\kappa}_{w_2}+A^{\kappa}_{w_2}\theta^*_{\rho,w_2}-A^{\kappa}_{w_1}\theta^*_{\rho,w_1})\\
&=C^{\kappa-1}_{w_1}(C^{\kappa}_{w_2}-C^{\kappa}_{w_1})C^{\kappa-1}_{w_2}(E^{\kappa}_{w_1}-A^{\kappa}_{w_1}\theta^*_{\rho,w_1})\\
&\quad+C^{\kappa-1}_{w_2}\left[(E^{\kappa}_{w_1}-E^{\kappa}_{w_2})+A^{\kappa}_{w_2}(\theta^*_{\rho,w_2}-\theta^*_{\rho,w_1})+(A^{\kappa}_{w_2}-A^{\kappa}_{w_1})\theta^*_{\rho,w_1}\right],
\end{aligned}
$$

which implies

$$
\begin{aligned}
&\left\|\theta^*_{q,w_1}-\theta^*_{q,w_2}\right\|_2\\
&\le\left\|C^{\kappa-1}_{w_1}\right\|_2\left\|C^{\kappa}_{w_2}-C^{\kappa}_{w_1}\right\|_2\left\|C^{\kappa-1}_{w_2}\right\|_2\left(\left\|E^{\kappa}_{w_1}\right\|_2+\left\|A^{\kappa}_{w_1}\right\|_2\left\|\theta^*_{\rho,w_1}\right\|_2\right)\\
&\quad+\left\|C^{\kappa-1}_{w_2}\right\|_2\left[\left\|E^{\kappa}_{w_1}-E^{\kappa}_{w_2}\right\|_2+\left\|A^{\kappa}_{w_2}\right\|_2\left\|\theta^*_{\rho,w_2}-\theta^*_{\rho,w_1}\right\|_2+\left\|A^{\kappa}_{w_2}-A^{\kappa}_{w_1}\right\|_2\left\|\theta^*_{\rho,w_1}\right\|_2\right]\\
&\le\left[\frac{L^{\kappa}_C(C^{\kappa}_E+C^{\kappa}_A R_\rho)+\lambda^{\kappa}_C(L^{\kappa}_E+C^{\kappa}_A L\rho+L^{\kappa}_A R_\rho)}{(\lambda^{\kappa}_C)^2}\right]\left\|w_1-w_2\right\|_2=L_q\left\|w_1-w_2\right\|_2,
\end{aligned}
$$

where $L_q=\frac{L^{\kappa}_C(C^{\kappa}_E+C^{\kappa}_A R_\rho)+\lambda^{\kappa}_C(L^{\kappa}_E+C^{\kappa}_A L\rho+L^{\kappa}_A R_\rho)}{(\lambda^{\kappa}_C)^2}$.

Finally, we consider the Lipschitz property of η^*. To bound $\left\|\eta^*_{w_1}-\eta^*_{w_2}\right\|_2$, we proceed as follows

$$
\begin{aligned}
&\left\|\eta^*_{w_1}-\eta^*_{w_2}\right\|_2\\
&=\left\|D^{\kappa}_{w_1}\theta^*_{\rho,w_1}-D^{\kappa}_{w_2}\theta^*_{\rho,w_2}\right\|_2\\
&=\left\|D^{\kappa}_{w_1}(\theta^*_{\rho,w_1}-\theta^*_{\rho,w_2})+(D^{\kappa}_{w_1}-D^{\kappa}_{w_2})\theta^*_{\rho,w_2}\right\|_2\\
&\le\left\|D^{\kappa}_{w_1}(\theta^*_{\rho,w_1}-\theta^*_{\rho,w_2})\right\|_2+\left\|(D^{\kappa}_{w_1}-D^{\kappa}_{w_2})\theta^*_{\rho,w_2}\right\|_2\\
&\le(C^{\kappa}_D L_\rho+L^{\kappa}_D R_\rho)\left\|w_1-w_2\right\|_2=L_\eta\left\|w_1-w_2\right\|_2, \qquad (53)
\end{aligned}
$$

where $L_\eta=C^{\kappa}_D L_\rho+L^{\kappa}_D R_\rho$. □

Lemma 5. *Consider the policies π_{w_1} and π_{w_2}, respectively, with the fixed points ξ^*_1 and ξ^*_2 defined in eq. (46). We have*

$$
\|\xi^*_1-\xi^*_2\|_2\le L_\xi\left\|w_1-w_2\right\|_2,
$$

where $L_\xi=\frac{2C^2_\varphi C^{\xi}_b L_\pi+\lambda^{\xi}_A C_\varphi(C_{sc}L_\pi+L_{sc})}{(\lambda^{\xi}_A)^2}$.

Proof. Recall that for $k=1$ or 2, we have $\xi^*_k=P^{-1}_k p_k=[\theta^{*\top}_{\psi,w_k},0^\top]^\top$, where $\theta^{*\top}_{\psi,w_k}=A^{\xi-1}_{w_k}b^{\xi}_{w_k}$, with $A^{\xi}_w=\mathbb{E}_{\tilde{\mathcal{D}}_d\cdot\pi_w}[(\varphi-\varphi')\varphi'^\top]$ and $b^{\xi}_w=\mathbb{E}_{\tilde{\mathcal{D}}_d\cdot\pi_w}[\varphi'\nabla_w\log\pi_w(a'|s')]$, which implies that

$$
\|\xi^*_1-\xi^*_2\|_2=\left\|\theta^{*\top}_{\psi,w_1}-\theta^{*\top}_{\psi,w_2}\right\|_2.
$$

To bound $\left\|\theta^{*\top}_{\psi,w_1} - \theta^{*\top}_{\psi,w_2}\right\|_2$, following the steps similar to those in eq. (51) and eq. (52), we obtain

$$\left\|\theta^{*\top}_{\psi,w_1} - \theta^{*\top}_{\psi,w_2}\right\|_2 \leq \frac{C_b^\xi}{(\lambda_A^\xi)^2} \left\|A^\xi_{w_1} - A^\xi_{w_2}\right\|_2 + \frac{1}{\lambda_A^\xi} \left\|b^\xi_{w_1} - b^\xi_{w_2}\right\|_2 . \tag{54}$$

We first bound the term $\left\|A^\xi_{w_2} - A^\xi_{w_1}\right\|_2$. Following the steps similar to those in eq. (47), we obtain

$$\left\|A^\xi_{w_2} - A^\xi_{w_1}\right\|_2 \leq 2C_\varphi^2 L_\pi \left\|w_1 - w_2\right\|_2 . \tag{55}$$

We then bound the term $\left\|b^\xi_{w_1} - b^\xi_{w_2}\right\|_2$. Based on the definition of b^ξ_w, we have

$$\begin{aligned}
&\left\|b^\xi_{w_1} - b^\xi_{w_2}\right\|_2 \\
&= \left\|\mathbb{E}_{\tilde{\mathcal{D}}_d\cdot\pi_{w_1}}[\varphi'\nabla_w \log\pi_{w_1}(a'|s')] - \mathbb{E}_{\tilde{\mathcal{D}}_d\cdot\pi_{w_2}}[\varphi'\nabla_w \log\pi_{w_2}(a'|s')]\right\|_2 \\
&\leq \left\|\mathbb{E}_{\tilde{\mathcal{D}}_d\cdot\pi_{w_1}}[\varphi'\nabla_w \log\pi_{w_1}(a'|s')] - \mathbb{E}_{\tilde{\mathcal{D}}_d\cdot\pi_{w_2}}[\varphi'\nabla_w \log\pi_{w_1}(a'|s')]\right\|_2 \\
&\quad + \left\|\mathbb{E}_{\tilde{\mathcal{D}}_d\cdot\pi_{w_2}}[\varphi'\nabla_w \log\pi_{w_1}(a'|s')] - \mathbb{E}_{\tilde{\mathcal{D}}_d\cdot\pi_{w_2}}[\varphi'\nabla_w \log\pi_{w_2}(a'|s')]\right\|_2 \\
&= \left\|\int \varphi(s',a')\nabla_w \log\pi_{w_1}(a'|s')(\pi_{w_1}(da'|s') - \pi_{w_2}(da'|s'))\tilde{\mathsf{P}}(ds'|s,a)\mathcal{D}(ds,da)\right\|_2 \\
&\quad + \left\|\int \varphi(s',a')(\nabla_w \log\pi_{w_1}(a'|s') - \nabla_w \log\pi_{w_2}(a'|s'))\pi_{w_2}(da'|s')\tilde{\mathsf{P}}(ds'|s,a)\mathcal{D}(ds,da)\right\|_2 \\
&= \int \left\|\varphi(s',a')\nabla_w \log\pi_{w_1}(a'|s')\right\|_2 \left|\pi_{w_1}(da'|s') - \pi_{w_2}(da'|s')\right| \tilde{\mathsf{P}}(ds'|s,a)\mathcal{D}(ds,da) \\
&\quad + \int \left\|\varphi(s',a')\right\|_2 \left\|\nabla_w \log\pi_{w_1}(a'|s') - \nabla_w \log\pi_{w_2}(a'|s')\right\|_2 \pi_{w_2}(da'|s')\tilde{\mathsf{P}}(ds'|s,a)\mathcal{D}(ds,da) \\
&\leq C_\varphi C_{sc} \int \left|\pi_{w_1}(da'|s') - \pi_{w_2}(da'|s')\right| \tilde{\mathsf{P}}(ds'|s,a)\mathcal{D}(ds,da) + C_\varphi L_{sc}\left\|w_1 - w_2\right\|_2 \\
&\leq C_\varphi(C_{sc}L_\pi + L_{sc})\left\|w_1 - w_2\right\|_2 .
\end{aligned} \tag{56}$$

Substituting eq. (55) and eq. (56) into eq. (52) yields

$$\left\|\theta^{*\top}_{\psi,w_1} - \theta^{*\top}_{\psi,w_2}\right\|_2 \leq \frac{2C_\varphi^2 C_b^\xi L_\pi + \lambda_A^\xi C_\varphi(C_{sc}L_\pi + L_{sc})}{(\lambda_A^\xi)^2}\left\|w_1 - w_2\right\|_2 = L_\xi \left\|w_1 - w_2\right\|_2 .$$

Thus, we have $\left\|\xi_1^* - \xi_2^*\right\|_2 \leq L_\xi \left\|w_1 - w_2\right\|_2$, which completes the proof. □

Lemma 6. *Consider the policies π_{w_1} and π_{w_2}, respectively, with the fixed points ζ_1^* and ζ_2^* defined in eq. (40). Then, we have*

$$\left\|\zeta_1^* - \zeta_2^*\right\|_2 \leq L_\zeta \left\|w_1 - w_2\right\|_2 ,$$

where $L_\zeta = \frac{2C_x^2 L_\pi C_b^\zeta + \lambda_A^\zeta C_\phi C_x(R_q L_{sc} + L_q C_{sc}) + \lambda_A^\zeta C_\phi R_q C_x C_{sc} L_\pi}{(\lambda_A^\zeta)^2}$.

Proof. Recall that for $k = 1$ or 2, we have $\zeta_k^* = U_k^{-1}u_k(\theta^*_{q,k}) = [\theta^{*\top}_{d_q,k}, 0^\top]^\top$, where $\theta^{*\top}_{d_q,w_k} = A^{\zeta-1}_{w_k} b^\zeta_{w_k}$, with $A^\zeta_w = \mathbb{E}_{\mathcal{D}_d\cdot\pi_w}[(\gamma x(s',a') - x(s,a))x(s,a)^\top]$ and $b^\zeta_w = \mathbb{E}_{\mathcal{D}_d\cdot\pi_w}[\phi(s',a')^\top\theta^*_{q,w}x(s,a)\nabla_w \log\pi_w(a'|s')]$, which implies that

$$\left\|\zeta_1^* - \zeta_2^*\right\|_2 = \left\|\theta^{*\top}_{d_q,1} - \theta^{*\top}_{d_q,2}\right\|_2 .$$

To bound $\left\|\theta^{*\top}_{d_q,1} - \theta^{*\top}_{d_q,2}\right\|_2$, following the steps similar to those in eq. (51) and eq. (52), we obtain

$$\left\|\theta^{*\top}_{d_q,1} - \theta^{*\top}_{d_q,2}\right\|_2 \leq \frac{C_b^\zeta}{(\lambda_A^\zeta)^2} \left\|A^\zeta_{w_1} - A^\zeta_{w_2}\right\|_2 + \frac{1}{\lambda_A^\zeta} \left\|b^\zeta_{w_1} - b^\zeta_{w_2}\right\|_2 . \tag{57}$$

We first bound the term $\left\|A^{\zeta}_{w_1} - A^{\zeta}_{w_2}\right\|_2$. Following the steps similar to those in eq. (47), we obtain

$$\left\|A^{\zeta}_{w_1} - A^{\zeta}_{w_2}\right\|_2 \leq 2C_x^2 L_\pi \left\|w_1 - w_2\right\|_2. \tag{58}$$

We then bound the term $\left\|b^{\zeta}_{w_1} - b^{\zeta}_{w_2}\right\|_2$. Based on to the definition of b^{ζ}_w, we have

$$\begin{aligned}
&\left\|b^{\zeta}_{w_1} - b^{\zeta}_{w_2}\right\|_2 \\
&= \left\|\mathbb{E}_{\mathcal{D}_d\cdot\pi_{w_1}}[\phi(s',a')^\top \theta^*_{q,w_1} x(s,a)\nabla_w \log \pi_{w_1}(a'|s')] - \mathbb{E}_{\mathcal{D}_d\cdot\pi_{w_2}}[\phi(s',a')^\top \theta^*_{q,w_2} x(s,a)\nabla_w \log \pi_{w_2}(a'|s')]\right\|_2 \\
&\leq \left\|\mathbb{E}_{\mathcal{D}_d\cdot\pi_{w_1}}[\phi(s',a')^\top \theta^*_{q,w_1} x(s,a)\nabla_w \log \pi_{w_1}(a'|s')] - \mathbb{E}_{\mathcal{D}_d\cdot\pi_{w_2}}[\phi(s',a')^\top \theta^*_{q,w_1} x(s,a)\nabla_w \log \pi_{w_1}(a'|s')]\right\|_2 \\
&\quad + \left\|\mathbb{E}_{\mathcal{D}_d\cdot\pi_{w_2}}[\phi(s',a')^\top \theta^*_{q,w_1} x(s,a)\nabla_w \log \pi_{w_1}(a'|s')] - \mathbb{E}_{\mathcal{D}_d\cdot\pi_{w_2}}[\phi(s',a')^\top \theta^*_{q,w_2} x(s,a)\nabla_w \log \pi_{w_2}(a'|s')]\right\|_2 \\
&\leq \left\|\mathbb{E}_{\mathcal{D}_d\cdot\pi_{w_2}}[\phi(s',a')^\top \theta^*_{q,w_1} x(s,a)\nabla_w \log \pi_{w_1}(a'|s')] - \mathbb{E}_{\mathcal{D}_d\cdot\pi_{w_2}}[\phi(s',a')^\top \theta^*_{q,w_2} x(s,a)\nabla_w \log \pi_{w_2}(a'|s')]\right\|_2 \\
&\quad + C_\phi R_q C_x C_{sc} L_\pi \left\|w_1 - w_2\right\|_2.
\end{aligned} \tag{59}$$

Consider the term

$$\left\|\mathbb{E}_{\mathcal{D}_d\cdot\pi_{w_2}}[\phi(s',a')^\top \theta^*_{q,w_1} x(s,a)\nabla_w \log \pi_{w_1}(a'|s')] - \mathbb{E}_{\mathcal{D}_d\cdot\pi_{w_2}}[\phi(s',a')^\top \theta^*_{q,w_2} x(s,a)\nabla_w \log \pi_{w_2}(a'|s')]\right\|_2,$$

we have

$$\begin{aligned}
&\left\|\mathbb{E}_{\mathcal{D}_d\cdot\pi_{w_2}}[\phi(s',a')^\top \theta^*_{q,w_1} x(s,a)\nabla_w \log \pi_{w_1}(a'|s')] - \mathbb{E}_{\mathcal{D}_d\cdot\pi_{w_2}}[\phi(s',a')^\top \theta^*_{q,w_2} x(s,a)\nabla_w \log \pi_{w_2}(a'|s')]\right\|_2 \\
&\leq \left\|\mathbb{E}_{\mathcal{D}_d\cdot\pi_{w_2}}[\phi(s',a')^\top \theta^*_{q,w_1} x(s,a)(\nabla_w \log \pi_{w_1}(a'|s') - \nabla_w \log \pi_{w_2}(a'|s'))]\right\|_2 \\
&\quad + \left\|\mathbb{E}_{\mathcal{D}_d\cdot\pi_{w_2}}[\phi(s',a')^\top (\theta^*_{q,w_1} - \theta^*_{q,w_2}) x(s,a)\nabla_w \log \pi_{w_2}(a'|s')]\right\|_2 \\
&\overset{(i)}{\leq} C_\phi C_x (R_q L_{sc} + L_q C_{sc}) \left\|w_1 - w_2\right\|_2,
\end{aligned} \tag{60}$$

where (i) follows from the Lipschitz property of $\theta^*_{q,w}$ given in Lemma 4. Combining eq. (47), eq. (59), eq. (60) and eq. (57) yields

$$\left\|\theta^{*\top}_{d_q,w_1} - \theta^{*\top}_{d_q,w_2}\right\|_2 \leq \frac{2C_x^2 L_\pi C_b^\zeta + \lambda_A^\zeta C_\phi C_x (R_q L_{sc} + L_q C_{sc}) + \lambda_A^\zeta C_\phi R_q C_x C_{sc} L_\pi}{(\lambda_A^\zeta)^2} \left\|w_1 - w_2\right\|_2 = L_\zeta \left\|w_1 - w_2\right\|_2.$$

Thus we have $\left\|\zeta_1^* - \zeta_2^*\right\|_2 \leq L_\zeta \left\|w_1 - w_2\right\|_2$, which completes the proof. □

Lemma 7. *Define $\Delta_t = \|\kappa_t - \kappa_t^*\|_2^2 + \|\xi_t - \xi_t^*\|_2^2 + \|\zeta_t - \zeta_t^*\|_2^2$. Then, we have*

$$\mathbb{E}[\Delta_{t+1}|\mathcal{F}_t] \leq \left(1 - \frac{1}{2}\varrho\right)\Delta_t + \frac{2}{\varrho}(L_\kappa^2 + L_\xi^2 + L_\zeta^2)\mathbb{E}[\|w_{t+1} - w_t\|_2^2 |\mathcal{F}_t] + \frac{C_5}{(1-\varrho)N},$$

where $\varrho = \frac{1}{4}\min\{\beta_1\lambda_M, \beta_2\lambda_P, \beta_3\lambda_U\}$ and $C_5 = C_1 + C_2 + C_4$, where C_1, C_2 and C_4 are defined in Lemma 1, Lemma 2, and Lemma 3.

Proof. Following from the iteration property developed in Lemma 1, Lemma 2, and Lemma 3, we have

$$\mathbb{E}\left[\|\kappa_{t+1} - \kappa_t^*\|_2^2 |\mathcal{F}_t\right] \leq \left(1 - \frac{1}{2}\beta_1\lambda_M\right)\|\kappa_t - \kappa_t^*\|_2^2 + \frac{C_1}{N}, \tag{61}$$

$$\mathbb{E}\left[\|\xi_{t+1} - \xi_t^*\|_2^2 |\mathcal{F}_t\right] \leq \left(1 - \frac{1}{2}\beta_2\lambda_P\right)\|\xi_t - \xi_t^*\|_2^2 + \frac{C_2}{N}, \tag{62}$$

$$\begin{aligned}
\mathbb{E}\left[\|\zeta_{t+1} - \zeta_t^*\|_2^2 |\mathcal{F}_t\right] &\leq \left(1 - \frac{1}{4}\beta_3\lambda_U\right)\|\zeta_t - \zeta_t^*\|_2^2 + C_3\beta_3\left\|\theta_{q,t} - \theta^*_{q,t}\right\|_2^2 + \frac{C_4}{N} \\
&\leq \left(1 - \frac{1}{4}\beta_3\lambda_U\right)\|\zeta_t - \zeta_t^*\|_2^2 + C_3\beta_3\|\kappa_t - \kappa_t^*\|_2^2 + \frac{C_4}{N}.
\end{aligned} \tag{63}$$

Summarizing eq. (61), eq. (62) and eq. (63), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\|\kappa_{t+1}-\kappa_t^*\|_2^2\,|\mathcal{F}_t\right]+\mathbb{E}\left[\|\xi_{t+1}-\xi_t^*\|_2^2\,|\mathcal{F}_t\right]+\mathbb{E}\left[\|\zeta_{t+1}-\zeta_t^*\|_2^2\,|\mathcal{F}_t\right]\\
&\leq\left(1-\frac{1}{2}\beta_1\lambda_M+C_3\beta_3\right)\|\kappa_t-\kappa_t^*\|_2^2+\left(1-\frac{1}{2}\beta_2\lambda_P\right)\|\xi_t-\xi_t^*\|_2^2+\left(1-\frac{1}{4}\beta_3\lambda_U\right)\|\zeta_t-\zeta_t^*\|_2^2+\frac{C_1+C_2+C_4}{N}.
\end{aligned}$$

Let $\beta_3\leq\frac{\beta_1\lambda_M}{4C_3}$ and define $\varrho=\frac{1}{4}\min\{\beta_1\lambda_M,\beta_2\lambda_P,\beta_3\lambda_U\}$, $C_5=C_1+C_2+C_4$. Then, we have

$$\begin{aligned}
&\mathbb{E}\left[\|\kappa_{t+1}-\kappa_t^*\|_2^2\,|\mathcal{F}_t\right]+\mathbb{E}\left[\|\xi_{t+1}-\xi_t^*\|_2^2\,|\mathcal{F}_t\right]+\mathbb{E}\left[\|\zeta_{t+1}-\zeta_t^*\|_2^2\,|\mathcal{F}_t\right]\\
&\leq(1-\varrho)\left(\|\kappa_t-\kappa_t^*\|_2^2+\|\xi_t-\xi_t^*\|_2^2+\|\zeta_t-\zeta_t^*\|_2^2\right)+\frac{C_5}{N}.
\end{aligned}\tag{64}$$

We then proceed to investigate the iteration of $\mathit{\Delta}_t$. By Young's inequality, we have

$$\begin{aligned}
\mathit{\Delta}_{t+1}&\leq\left(\frac{2-\varrho}{2-2\varrho}\right)\left(\mathbb{E}\left[\|\kappa_{t+1}-\kappa_t^*\|_2^2\,|\mathcal{F}_t\right]+\mathbb{E}\left[\|\xi_{t+1}-\xi_t^*\|_2^2\,|\mathcal{F}_t\right]+\mathbb{E}\left[\|\zeta_{t+1}-\zeta_t^*\|_2^2\,|\mathcal{F}_t\right]\right)\\
&\quad+\left(\frac{2-\varrho}{\varrho}\right)\left(\mathbb{E}[\left\|\kappa_{t+1}^*-\kappa_t^*\right\|_2^2|\mathcal{F}_t]+\mathbb{E}[\left\|\xi_{t+1}^*-\xi_t^*\right\|_2^2|\mathcal{F}_t]+\mathbb{E}[\left\|\zeta_{t+1}^*-\zeta_t^*\right\|_2^2|\mathcal{F}_t]\right)\\
&\overset{(i)}{\leq}\left(1-\frac{1}{2}\varrho\right)\mathit{\Delta}_t+\frac{2}{\varrho}\left(\mathbb{E}[\left\|\kappa_{t+1}^*-\kappa_t^*\right\|_2^2|\mathcal{F}_t]+\mathbb{E}[\left\|\xi_{t+1}^*-\xi_t^*\right\|_2^2|\mathcal{F}_t]+\mathbb{E}[\left\|\zeta_{t+1}^*-\zeta_t^*\right\|_2^2|\mathcal{F}_t]\right)+\frac{C_5}{(1-\varrho)N}\\
&\leq\left(1-\frac{1}{2}\varrho\right)\mathit{\Delta}_t+\frac{2}{\varrho}(L_\kappa^2+L_\xi^2+L_\zeta^2)\mathbb{E}[\|w_{t+1}-w_t\|_2^2\,|\mathcal{F}_t]+\frac{C_5}{(1-\varrho)N}.
\end{aligned}\tag{65}$$

where (i) follows from eq. (64). □

D. Proof of Theorem 2

In order to prove the convergence for the DR-Off-PAC algorithm, we first introduce the following Lipschitz property of $J(w)$, which was established in (Xu et al., 2020b).

Proposition 1. *Suppose Assumptions 2 and 3 hold. For any $w,w'\in\mathbb{R}^d$, we have $\|\nabla_w J(w)-\nabla_w J(w')\|_2\leq L_J\|w-w'\|_2$, for all $w,w'\in\mathbb{R}^d$, where $L_J=\Theta(1/(1-\gamma))$.*

The Lipschitz property established in Proposition 1 is important to establish the local convergence of the gradient-based algorithm.

To proceed the proof of Theorem 2, consider the update in Algorithm 1. For brevity, we define $G_{\text{DR}}(w_t,\mathcal{M}_t)=\frac{1}{N}\sum_i G_{\text{DR}}^i(w_t)$, where $\mathcal{M}_t$ represents the sample set $\mathcal{B}_t\cup\mathcal{B}_{t,0}$. Following the L_J-Lipschitz property of objective $J(w)$, we have

$$\begin{aligned}
J(w_{t+1})&\geq J(w_t)+\langle\nabla_w J(w_t),w_{t+1}-w_t\rangle-\frac{L_J}{2}\|w_{t+1}-w_t\|_2^2\\
&=J(w_t)+\alpha\langle\nabla_w J(w_t),G_{\text{DR}}(w_t,\mathcal{M}_t)-\nabla_w J(w_t)+\nabla_w J(w_t)\rangle-\frac{L_J\alpha^2}{2}\|G_{\text{DR}}(w_t,\mathcal{M}_t)\|_2^2\\
&=J(w_t)+\alpha\|\nabla_w J(w_t)\|_2^2+\alpha\langle\nabla_w J(w_t),G_{\text{DR}}(w_t,\mathcal{M}_t)-\nabla_w J(w_t)\rangle\\
&\quad-\frac{L_J\alpha^2}{2}\|G_{\text{DR}}(w_t,\mathcal{M}_t)-\nabla_w J(w_t)+\nabla_w J(w_t)\|_2^2\\
&\overset{(i)}{\geq}J(w_t)+\left(\frac{1}{2}\alpha-L_J\alpha^2\right)\|\nabla_w J(w_t)\|_2^2-\left(\frac{1}{2}\alpha+L_J\alpha^2\right)\|G_{\text{DR}}(w_t,\mathcal{M}_t)-\nabla_w J(w_t)\|_2^2,
\end{aligned}\tag{66}$$

where (i) follows because

$$\langle\nabla_w J(w_t),G_{\text{DR}}(w_t,\mathcal{M}_t)-\nabla_w J(w_t)\rangle\geq-\frac{1}{2}\|\nabla_w J(w_t)\|_2^2-\frac{1}{2}\|G_{\text{DR}}(w_t,\mathcal{M}_t)-\nabla_w J(w_t)\|_2^2,$$

and

$$\|G_{\rm DR}(w_t,\mathcal{M}_t)-\nabla_w J(w_t)+\nabla_w J(w_t)\|_2^2\le 2\|G_{\rm DR}(w_t,\mathcal{M}_t)-\nabla_w J(w_t)\|_2^2+2\|\nabla_w J(w_t)\|_2^2.$$

Taking the expectation on both sides of eq. (66) conditioned on $\mathcal{F}_t$ and rearranging eq. (66) yield

$$\begin{aligned}&\Big(\frac{1}{2}\alpha-L_J\alpha^2\Big)\mathbb{E}[\|\nabla_w J(w_t)\|_2^2|\mathcal{F}_t]\\&\quad\le\mathbb{E}[J(w_{t+1})|\mathcal{F}_t]-J(w_t)+\Big(\frac{1}{2}\alpha+L_J\alpha^2\Big)\mathbb{E}[\|G_{\rm DR}(w_t,\mathcal{M}_t)-\nabla_w J(w_t)\|_2^2|\mathcal{F}_t].\end{aligned}\tag{67}$$

Then, we upper-bound the term $\mathbb{E}[\|G_{\rm DR}(w_t,\mathcal{M}_t)-\nabla_w J(w_t)\|_2^2|\mathcal{F}_t]$ as follows. By definition, we have

$$\begin{aligned}&\|G_{\rm DR}(w_t,\mathcal{M}_t)-\nabla_w J(w_t)\|_2^2\\&=\left\|G_{\rm DR}(w_t,\theta_{q,t},\theta_{\rho,t},\theta_{\psi,t},\theta_{d_q,t},\mathcal{M}_t)-\nabla_w J(w_t)\right\|_2^2\\&\le 6\left\|G_{\rm DR}(w_t,\theta_{q,t},\theta_{\rho,t},\theta_{\psi,t},\theta_{d_q,t},\mathcal{M}_t)-G_{\rm DR}(w_t,\theta_{q,t},\theta_{\rho,t},\theta_{\psi,t},\theta^*_{d_q,t},\mathcal{M}_t)\right\|_2^2\\&\quad+6\left\|G_{\rm DR}(w_t,\theta_{q,t},\theta_{\rho,t},\theta_{\psi,t},\theta^*_{d_q,t},\mathcal{M}_t)-G_{\rm DR}(w_t,\theta_{q,t},\theta_{\rho,t},\theta^*_{\psi,t},\theta^*_{d_q,t},\mathcal{M}_t)\right\|_2^2\\&\quad+6\left\|G_{\rm DR}(w_t,\theta_{q,t},\theta_{\rho,t},\theta^*_{\psi,t},\theta^*_{d_q,t},\mathcal{M}_t)-G_{\rm DR}(w_t,\theta^*_{q,t},\theta_{\rho,t},\theta^*_{\psi,t},\theta^*_{d_q,t},\mathcal{M}_t)\right\|_2^2\\&\quad+6\left\|G_{\rm DR}(w_t,\theta^*_{q,t},\theta_{\rho,t},\theta^*_{\psi,t},\theta^*_{d_q,t},\mathcal{M}_t)-G_{\rm DR}(w_t,\theta^*_{q,t},\theta^*_{\rho,t},\theta^*_{\psi,t},\theta^*_{d_q,t},\mathcal{M}_t)\right\|_2^2\\&\quad+6\left\|G_{\rm DR}(w_t,\theta^*_{q,t},\theta^*_{\rho,t},\theta^*_{\psi,t},\theta^*_{d_q,t},\mathcal{M}_t)-\mathbb{E}[G^i_{\rm DR}(w_t,\theta^*_{q,t},\theta^*_{\rho,t},\theta^*_{\psi,t},\theta^*_{d_q,t})]\right\|_2^2\\&\quad+6\left\|\mathbb{E}[G^i_{\rm DR}(w_t,\theta^*_{q,t},\theta^*_{\rho,t},\theta^*_{\psi,t},\theta^*_{d_q,t})]-\nabla_w J(w_t)\right\|_2^2\\&\le\frac{6}{N}\sum_i\left\|G^i_{\rm DR}(w_t,\theta_{q,t},\theta_{\rho,t},\theta_{\psi,t},\theta_{d_q,t})-G^i_{\rm DR}(w_t,\theta_{q,t},\theta_{\rho,t},\theta_{\psi,t},\theta^*_{d_q,t})\right\|_2^2\\&\quad+\frac{6}{N}\sum_i\left\|G^i_{\rm DR}(w_t,\theta_{q,t},\theta_{\rho,t},\theta_{\psi,t},\theta^*_{d_q,t})-G^i_{\rm DR}(w_t,\theta_{q,t},\theta_{\rho,t},\theta^*_{\psi,t},\theta^*_{d_q,t}\right\|_2^2\\&\quad+\frac{6}{N}\sum_i\left\|G^i_{\rm DR}(w_t,\theta_{q,t},\theta_{\rho,t},\theta^*_{\psi,t},\theta^*_{d_q,t})-G^i_{\rm DR}(w_t,\theta^*_{q,t},\theta_{\rho,t},\theta^*_{\psi,t},\theta^*_{d_q,t})\right\|_2^2\\&\quad+\frac{6}{N}\sum_i\left\|G^i_{\rm DR}(w_t,\theta^*_{q,t},\theta_{\rho,t},\theta^*_{\psi,t},\theta^*_{d_q,t})-G^i_{\rm DR}(w_t,\theta^*_{q,t},\theta^*_{\rho,t},\theta^*_{\psi,t},\theta^*_{d_q,t})\right\|_2^2\\&\quad+6\left\|G_{\rm DR}(w_t,\theta^*_{q,t},\theta^*_{\rho,t},\theta^*_{\psi,t},\theta^*_{d_q,t},\mathcal{M}_t)-\mathbb{E}[G^i_{\rm DR}(w_t,\theta^*_{q,t},\theta^*_{\rho,t},\theta^*_{\psi,t},\theta^*_{d_q,t})]\right\|_2^2\\&\quad+6\left\|\mathbb{E}[G^i_{\rm DR}(w_t,\theta^*_{q,t},\theta^*_{\rho,t},\theta^*_{\psi,t},\theta^*_{d_q,t})]-\nabla_w J(w_t)\right\|_2^2.\end{aligned}\tag{68}$$

We next bound each term in eq. (68). For $\left\|G^i_{\rm DR}(w_t,\theta_{q,t},\theta_{\rho,t},\theta_{\psi,t},\theta_{d_q,t})-G^i_{\rm DR}(w_t,\theta_{q,t},\theta_{\rho,t},\theta_{\psi,t},\theta^*_{d_q,t})\right\|_2$, we proceed as follows:

$$\begin{aligned}&\left\|G^i_{\rm DR}(w_t,\theta_{q,t},\theta_{\rho,t},\theta_{\psi,t},\theta_{d_q,t})-G^i_{\rm DR}(w_t,\theta_{q,t},\theta_{\rho,t},\theta_{\psi,t},\theta^*_{d_q,t})\right\|_2\\&\le(1-\gamma)\left\|\mathbb{E}_{\pi_{w_t}}[x(s_{0,i},a_{0,i})]^\top(\theta_{d_q,t}-\theta^*_{d_q,t})\right\|_2+\left\|\hat\rho_{\pi_{w_t}}(s_i,a_i)x(s_i,a_i)(\theta^*_{d_q,t}-\theta_{d_q,t})\right\|_2\\&\quad+\gamma\left\|\hat\rho_{\pi_{w_t}}(s_i,a_i)\mathbb{E}_{\pi_{w_t}}[x(s_i',a_i')]^\top(\theta_{d_q,t}-\theta^*_{d_q,t})\right\|_2\\&\le(1+2R_\rho C_\phi)C_x\left\|\theta^*_{d_q,t}-\theta_{d_q,t}\right\|_2\\&=C_6\left\|\theta^*_{d_q,t}-\theta_{d_q,t}\right\|_2,\end{aligned}\tag{69}$$

where $C_6 = (1 + 2R_\rho C_\phi)$.

For $\left\|G^i_{\text{DR}}(w_t, \theta_{q,t}, \theta_{\rho,t}, \theta_{\psi,t}, \theta^*_{d_q,t}) - G^i_{\text{DR}}(w_t, \theta_{q,t}, \theta_{\rho,t}, \theta^*_{\psi,t}, \theta^*_{d_q,t}\right\|_2$, we proceed as follows

$$\begin{aligned}
&\left\|G^i_{\text{DR}}(w_t, \theta_{q,t}, \theta_{\rho,t}, \theta_{\psi,t}, \theta^*_{d_q,t}) - G^i_{\text{DR}}(w_t, \theta_{q,t}, \theta_{\rho,t}, \theta^*_{\psi,t}, \theta^*_{d_q,t}\right\|_2 \\
&\leq \left\|\hat{\rho}_{\pi_{w_t}}(s_i, a_i)\varphi(s_i, a_i)^\top(\theta_{\psi,t} - \theta^*_{\psi,t})\left(r(s_i, a_i, s'_i) + \hat{Q}_{\pi_{w_t}}(s_i, a_i) - \gamma\mathbb{E}_{\pi_{w_t}}[\hat{Q}_{\pi_{w_t}}(s'_i, a'_i)]\right)\right\|_2 \\
&\leq R_\rho C_\varphi(r_{\max} + 2R_q C_\phi)\left\|\theta_{\psi,t} - \theta^*_{\psi,t}\right\|_2 \\
&= C_7\left\|\theta_{\psi,t} - \theta^*_{\psi,t}\right\|_2,
\end{aligned} \tag{70}$$

where $C_7 = R_\rho C_\varphi(r_{\max} + 2R_q C_\phi)$.

For $\left\|G^i_{\text{DR}}(w_t, \theta_{q,t}, \theta_{\rho,t}, \theta^*_{\psi,t}, \theta^*_{d_q,t}) - G^i_{\text{DR}}(w_t, \theta^*_{q,t}, \theta_{\rho,t}, \theta^*_{\psi,t}, \theta^*_{d_q,t})\right\|_2$, we proceed as follows

$$\begin{aligned}
&\left\|G^i_{\text{DR}}(w_t, \theta_{q,t}, \theta_{\rho,t}, \theta^*_{\psi,t}, \theta^*_{d_q,t}) - G^i_{\text{DR}}(w_t, \theta^*_{q,t}, \theta_{\rho,t}, \theta^*_{\psi,t}, \theta^*_{d_q,t})\right\|_2 \\
&\leq (1-\gamma)\left\|\mathbb{E}_{\pi_{w_t}}\left[\phi(s_{0,i}, a_{0,i})^\top(\theta_{q,t} - \theta^*_{q,t})\nabla_w \log\pi_{w_t}(s_{0,i}, a_{0,i})\right]\right\|_2 \\
&\quad + \left\|\hat{\rho}_{\pi_{w_t}}(s_i, a_i)\varphi(s_i, a_i)^\top\theta^*_{\psi,t}\phi(s_i, a_i)^\top(\theta_{q,t} - \theta^*_{q,t})\right\|_2 \\
&\quad + \gamma\left\|\hat{\rho}_{\pi_{w_t}}(s_i, a_i)\varphi(s_i, a_i)^\top\theta^*_{\psi,t}\mathbb{E}_{\pi_{w_t}}[\phi(s'_i, a'_i)]^\top(\theta_{q,t} - \theta^*_{q,t})\right\|_2 \\
&\leq \left[(1-\gamma)C_\phi C_{sc} + (1+\gamma)R_\rho C^2_\phi C_\varphi R_\psi\right]\left\|\theta_{q,t} - \theta^*_{q,t}\right\|_2 \\
&= C_8\left\|\theta_{q,t} - \theta^*_{q,t}\right\|_2,
\end{aligned} \tag{71}$$

where $C_8 = (1-\gamma)C_\phi C_{sc} + (1+\gamma)R_\rho C^2_\phi C_\varphi R_\psi$.

For $\left\|G^i_{\text{DR}}(w_t, \theta^*_{q,t}, \theta_{\rho,t}, \theta^*_{\psi,t}, \theta^*_{d_q,t}) - G^i_{\text{DR}}(w_t, \theta^*_{q,t}, \theta^*_{\rho,t}, \theta^*_{\psi,t}, \theta^*_{d_q,t})\right\|_2$, we proceed as follows

$$\begin{aligned}
&\left\|G^i_{\text{DR}}(w_t, \theta^*_{q,t}, \theta_{\rho,t}, \theta^*_{\psi,t}, \theta^*_{d_q,t}) - G^i_{\text{DR}}(w_t, \theta^*_{q,t}, \theta^*_{\rho,t}, \theta^*_{\psi,t}, \theta^*_{d_q,t})\right\|_2 \\
&\leq \left\|\phi(s_i, a_i)^\top(\theta_{\rho,t} - \theta^*_{\rho,t})\varphi(s_i, a_i)^\top\theta^*_{\psi,t}\left(r(s_i, a_i, s'_i) - \phi(s_i, a_i)^\top\theta^*_{q,t} + \gamma\mathbb{E}[\phi(s'_i, a'_i)]^\top\theta^*_{q,t}\right)\right\|_2 \\
&\quad + \left\|\phi(s_i, a_i)^\top(\theta_{\rho,t} - \theta^*_{\rho,t})\left(-x(s_i, a_i)^\top\theta^*_{d_q,t} + \gamma\mathbb{E}_{\pi_{w_t}}[\phi(s'_i, a'_i)^\top\theta^*_{q,t}\nabla_w\log\pi_{w_t}(s'_i, a'_i) + x(s'_i, a'_i)^\top\theta^*_{d_q,t}]\right)\right\|_2 \\
&\leq \left[C_\phi C_\varphi R_\psi(r_{\max} + (1+\gamma)C_\phi R_q) + C_\phi(C_x R_{d_q} + \gamma(C_\phi R_q C_{sc} + C_x R_{d_q}))\right]\left\|\theta_{\rho,t} - \theta^*_{\rho,t}\right\|_2 \\
&= C_9\left\|\theta_{\rho,t} - \theta^*_{\rho,t}\right\|_2,
\end{aligned} \tag{72}$$

where $C_9 = C_\phi C_\varphi R_\psi(r_{\max} + (1+\gamma)C_\phi R_q) + C_\phi(C_x R_{d_q} + \gamma(C_\phi R_q C_{sc} + C_x R_{d_q}))$.

For $\left\|G_{\text{DR}}(w_t, \theta^*_{q,t}, \theta^*_{\rho,t}, \theta^*_{\psi,t}, \theta^*_{d_q,t}, \mathcal{M}_t) - \mathbb{E}[G^i_{\text{DR}}(w_t, \theta^*_{q,t}, \theta^*_{\rho,t}, \theta^*_{\psi,t}, \theta^*_{d_q,t})]\right\|^2_2$, note that for all i, we have

$$\begin{aligned}
&\left\|G^i_{\text{DR}}(w_t, \theta^*_{q,t}, \theta^*_{\rho,t}, \theta^*_{\psi,t}, \theta^*_{d_q,t})\right\|_2 \\
&\leq (1-\gamma)\left\|\mathbb{E}_{\pi_w}[\phi^\top_{0,i}\theta^*_{q,t}\nabla_w\log\pi_w(s_{0,i}, a_{0,i}) + x^\top_{0,i}\theta^*_{d_q,t}]\right\|_2 + \left\|\psi^\top_i\theta^*_{\psi,t}\left(r(s_i, a_i, s'_i) - \phi^\top_i\theta^*_{q,t} + \gamma\mathbb{E}_{\pi_{w_t}}[\phi'^\top_i\theta^*_{q,t}]\right)\right\|_2 \\
&\quad + \left\|\phi^\top_i\theta^*_{\rho,t}(-x^\top_i\theta^*_{d_q,t} + \gamma\mathbb{E}_{\pi_w}[\phi^\top_i\theta^*_{q,t}\nabla_w\log\pi_w(s_{t,i}, a_{t,i}) + x^\top_i\theta^*_{d_q,t}])\right\|_2 \\
&\leq (1-\gamma)(C_\phi R_q C_{sc} + C_x R_{d_q}) + C_\psi R_\psi(r_{\max} + (1+\gamma)C_\phi R_q) + C_\phi R_\rho(C_x R_{d_q} + \gamma C_\phi R_q C_{sc} + \gamma C_x R_{d-q}).
\end{aligned}$$

Let $C_{10} = (1-\gamma)(C_\phi R_q C_{sc} + C_x R_{d_q}) + C_\psi R_\psi(r_{\max} + (1+\gamma)C_\phi R_q) + C_\phi R_\rho(C_x R_{d_q} + \gamma C_\phi R_q C_{sc} + \gamma C_x R_{d-q})$. Then following the steps similar to those in eq. (34) and eq. (35), we obtain

$$\mathbb{E}\left[\left\|G_{\text{DR}}(w_t, \theta^*_{q,t}, \theta^*_{\rho,t}, \theta^*_{\psi,t}, \theta^*_{d_q,t}, \mathcal{M}_t) - \mathbb{E}[G^i_{\text{DR}}(w_t, \theta^*_{q,t}, \theta^*_{\rho,t}, \theta^*_{\psi,t}, \theta^*_{d_q,t})]\right\|^2_2 \middle| \mathcal{F}_t\right] \leq \frac{4C^2_{10}}{N}. \tag{73}$$

Finally, consider the term $\left\|\mathbb{E}[G^i_{\text{DR}}(w_t,\theta^*_{q,t},\theta^*_{\rho,t},\theta^*_{\psi,t},\theta^*_{d_q,t})]-\nabla_w J(w_t)\right\|_2^2$. Following the steps similar to those in proving Theorem 1, we obtain

$$\begin{aligned}
&\mathbb{E}[G^i_{\text{DR}}(w_t,\theta^*_{q,t},\theta^*_{\rho,t},\theta^*_{\psi,t},\theta^*_{d_q,t})]-\nabla_w J(w_t)\\
&=\mathbb{E}_{\mathcal{D}}[(\hat{\rho}_{\pi_{w_t}}(s,a,\theta^*_{\rho,t})-\rho_{\pi_{w_t}}(s,a))(-\hat{d}^q_{\pi_{w_t}}(s,a,\theta^*_{d_q,t})+d^q_{\pi_{w_t}}(s,a))]\\
&\quad+\mathbb{E}_{\mathcal{D}}[(-\hat{d}^\rho_{\pi_{w_t}}(s,a,\theta^*_{\rho,t},\theta^*_{\psi,t})+d^\rho_{\pi_{w_t}}(s,a))(-\hat{Q}_{\pi_{w_t}}(s,a,\theta^*_{q,t})+Q_{\pi_{w_t}}(s,a))]\\
&\quad+\gamma\mathbb{E}_{\mathcal{D}}[(\hat{\rho}_{\pi_{w_t}}(s,a,\theta^*_{\rho,t})-\rho_{\pi_{w_t}}(s,a))\mathbb{E}_{\pi_{w_t}}[\hat{d}^q_{\pi_{w_t}}(s',a',\theta^*_{d_q,t})-d^q_{\pi_{w_t}}(s',a')]]\\
&\quad+\gamma\mathbb{E}_{\mathcal{D}}[(\hat{\rho}_{\pi_{w_t}}(s,a,\theta^*_{\rho,t})-\rho_{\pi_{w_t}}(s,a))\mathbb{E}_{\pi_{w_t}}[(\hat{Q}_{\pi_{w_t}}(s',a',\theta^*_{q,t})-Q_{\pi_{w_t}}(s',a'))\nabla_w\log\pi_{w_t}(s',a')]]\\
&\quad+\gamma\mathbb{E}_{\mathcal{D}}[(\hat{d}^\rho_{\pi_{w_t}}(s,a,\theta^*_{\rho,t},\theta^*_{\psi,t})-d^\rho_{\pi_{w_t}}(s,a))\mathbb{E}_{\pi_{w_t}}[\hat{Q}_{\pi_{w_t}}(s',a',\theta^*_{q,t})-Q_{\pi_{w_t}}(s',a')]]\\
&\le\sqrt{\mathbb{E}_{\mathcal{D}}[(\hat{\rho}_{\pi_{w_t}}(s,a,\theta^*_{\rho,t})-\rho_{\pi_{w_t}}(s,a))^2]}\sqrt{\mathbb{E}_{\mathcal{D}}[(\hat{d}^q_{\pi_{w_t}}(s,a,\theta^*_{d_q,t})-d^q_{\pi_{w_t}}(s,a))^2]}\\
&\quad+\sqrt{\mathbb{E}_{\mathcal{D}}[(\hat{d}^\rho_{\pi_{w_t}}(s,a,\theta^*_{\rho,t},\theta^*_{\psi,t})-d^\rho_{\pi_{w_t}}(s,a))^2]}\sqrt{\mathbb{E}_{\mathcal{D}}[(\hat{Q}_{\pi_{w_t}}(s,a,\theta^*_{q,t})-Q_{\pi_{w_t}}(s,a))^2]}\\
&\quad+\sqrt{\mathbb{E}_{\mathcal{D}}[(\hat{\rho}_{\pi_{w_t}}(s,a,\theta^*_{q,t})-\rho_{\pi_{w_t}}(s,a))^2]}\sqrt{\mathbb{E}_{\mathcal{D}_d\cdot\pi_{w_t}}[(\hat{d}^q_{\pi_{w_t}}(s',a',\theta^*_{d_q,t})-d^q_{\pi_{w_t}}(s',a'))^2]}\\
&\quad+C_{sc}\sqrt{\mathbb{E}_{\mathcal{D}}[(\hat{\rho}_{\pi_{w_t}}(s,a,\theta^*_{q,t})-\rho_{\pi_{w_t}}(s,a))^2]}\sqrt{\mathbb{E}_{\mathcal{D}_d\cdot\pi_{w_t}}[(\hat{Q}_{\pi_{w_t}}(s',a',\theta^*_{q,t})-Q_{\pi_{w_t}}(s',a'))^2]}\\
&\quad+\sqrt{\mathbb{E}_{\mathcal{D}}[(\hat{d}^\rho_{\pi_{w_t}}(s,a,\theta^*_{\rho,t},\theta^*_{\psi,t})-d^\rho_{\pi_{w_t}}(s,a))^2]}\sqrt{\mathbb{E}_{\mathcal{D}_d\cdot\pi_{w_t}}[(\hat{Q}_{\pi_{w_t}}(s',a',\theta^*_{q,t})-Q_{\pi_{w_t}}(s',a'))^2]}\\
&\le 2\epsilon_\rho\epsilon_{d_q}+2\epsilon_{d_\rho}\epsilon_q+C_{sc}\epsilon_\rho\epsilon_q.
\end{aligned}\tag{74}$$

Recall that we define

$$\begin{aligned}
\epsilon_\rho&=\max_w\sqrt{\mathbb{E}_{\mathcal{D}}[(\hat{\rho}_{\pi_w}(s,a,\theta^*_{\rho,w})-\rho_{\pi_w}(s,a))^2]}\\
\epsilon_{d_\rho}&=\max_w\sqrt{\mathbb{E}_{\mathcal{D}}[(\hat{d}^\rho_{\pi_w}(s,a,\theta^*_{\rho,w},\theta^*_{\psi,w})-d^\rho_{\pi_w}(s,a))^2]}\\
\epsilon_q&=\max\left\{\max_w\sqrt{\mathbb{E}_{\mathcal{D}}[(\hat{Q}_{\pi_w}(s,a,\theta^*_{q,w})-Q_{\pi_w}(s,a))^2]},\max_w\sqrt{\mathbb{E}_{\mathcal{D}_d\cdot\pi_w}[(\hat{Q}_{\pi_w}(s',a',\theta^*_{q,w})-Q_{\pi_w}(s',a'))^2]}\right\}\\
\epsilon_{d_q}&=\max\left\{\max_w\sqrt{\mathbb{E}_{\mathcal{D}}[(\hat{d}^q_{\pi_w}(s,a,\theta^*_{d_q,w})-d^q_{\pi_w}(s,a))^2]},\max_w\sqrt{\mathbb{E}_{\mathcal{D}_d\cdot\pi_w}[(\hat{d}^q_{\pi_w}(s',a',\theta^*_{d_q,w})-d^q_{\pi_w}(s',a'))^2]}\right\},
\end{aligned}$$

where

$$\begin{aligned}
\hat{\rho}_{\pi_w}(s,a,\theta^*_{q,w})&=\phi(s,a)^\top\theta^*_{\rho,w},\\
\hat{d}^\rho_{\pi_w}(s,a,\theta^*_{\rho,w},\theta^*_{\psi,w})&=\phi(s,a)^\top\theta^*_{\rho,w}\varphi(s,a)^\top\theta^*_{\psi,w},\\
\hat{Q}_{\pi_w}(s,a,\theta^*_{q,w})&=\phi(s,a)^\top\theta^*_{q,w},\\
\hat{d}^q_{\pi_w}(s',a',\theta^*_{d_q,w})&=x(s,a)^\top\theta^*_{d_q,w}.
\end{aligned}$$

Then, eq. (74) implies that

$$\left\|\mathbb{E}[G^i_{\text{DR}}(w_t,\theta^*_{q,t},\theta^*_{\rho,t},\theta^*_{\psi,t},\theta^*_{d_q,t})]-\nabla_w J(w_t)\right\|_2^2\le 6\epsilon_\rho^2\epsilon_{d_q}^2+6\epsilon_{d_\rho}^2\epsilon_q^2+3C_{sc}\epsilon_\rho^2\epsilon_q^2.\tag{75}$$

Substituting eq. (69), eq. (70), eq. (71), eq. (72), eq. (73) and eq. (75) into eq. (68) yields

$$\begin{aligned}
&\mathbb{E}\left[\|G_{\text{DR}}(w_t,\mathcal{M}_t)-\nabla_w J(w_t)\|_2^2\,|\mathcal{F}_t\right]\\
&\le 6C_6^2\left\|\theta^*_{d_q,t}-\theta_{d_q,t}\right\|_2^2+6C_7^2\left\|\theta_{\psi,t}-\theta^*_{\psi,t}\right\|_2^2+6C_8^2\left\|\theta_{q,t}-\theta^*_{q,t}\right\|_2+6C_9^2\left\|\theta_{\rho,t}-\theta^*_{\rho,t}\right\|_2^2+\frac{24C_{10}^2}{N}\\
&\quad+36\epsilon_\rho^2\epsilon_{d_q}^2+36\epsilon_{d_\rho}^2\epsilon_q^2+18C_{sc}\epsilon_\rho^2\epsilon_q^2\\
&\le C_{11}\left(\left\|\theta^*_{d_q,t}-\theta_{d_q,t}\right\|_2^2+\left\|\theta_{\psi,t}-\theta^*_{\psi,t}\right\|_2^2+\left\|\theta_{q,t}-\theta^*_{q,t}\right\|_2+\left\|\theta_{\rho,t}-\theta^*_{\rho,t}\right\|_2^2\right)+\frac{24C_{10}^2}{N}
\end{aligned}$$

$$
\begin{aligned}
&+ C_{12}(\epsilon_\rho^2\epsilon_{d_q}^2 + \epsilon_{d_\rho}^2\epsilon_q^2 + \epsilon_\rho^2\epsilon_q^2) \\
&\le C_{11}\varDelta_t + \frac{24C_{10}^2}{N} + C_{12}(\epsilon_\rho^2\epsilon_{d_q}^2 + \epsilon_{d_\rho}^2\epsilon_q^2 + \epsilon_\rho^2\epsilon_q^2),
\end{aligned} \tag{76}
$$

where $C_{11} = 6\max\{C_6^2, C_7^2, C_8^2, C_9^2\}$, and $C_{12} = \max\{36, 18C_{sc}\}$. Substituting eq. (76) into eq. (67) yields

$$
\begin{aligned}
&\Big(\frac{1}{2}\alpha - L_J\alpha^2\Big)\mathbb{E}[\|\nabla_w J(w_t)\|_2^2\,|\mathcal{F}_t] \\
&\quad\le \mathbb{E}[J(w_{t+1})|\mathcal{F}_t] - J(w_t) + \Big(\frac{1}{2}\alpha + L_J\alpha^2\Big)C_{11}\varDelta_t + \Big(\frac{1}{2}\alpha + L_J\alpha^2\Big)\frac{24C_{10}^2}{N} \\
&\quad\quad + \Big(\frac{1}{2}\alpha + L_J\alpha^2\Big)C_{12}(\epsilon_\rho^2\epsilon_{d_q}^2 + \epsilon_{d_\rho}^2\epsilon_q^2 + \epsilon_\rho^2\epsilon_q^2).
\end{aligned} \tag{77}
$$

Taking the expectation on both sides of eq. (77) and taking the summation over $t = 0\cdots T-1$ yield

$$
\begin{aligned}
&\Big(\frac{1}{2}\alpha - L_J\alpha^2\Big)\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla_w J(w_t)\|_2^2] \\
&\quad\le \mathbb{E}[J(w_T)] - J(w_0) + \Big(\frac{1}{2}\alpha + L_J\alpha^2\Big)C_{11}\sum_{t=0}^{T-1}\mathbb{E}[\varDelta_t] + \Big(\frac{1}{2}\alpha + L_J\alpha^2\Big)\frac{24C_{10}^2 T}{N} \\
&\quad\quad + \Big(\frac{1}{2}\alpha + L_J\alpha^2\Big)C_{12}(\epsilon_\rho^2\epsilon_{d_q}^2 + \epsilon_{d_\rho}^2\epsilon_q^2 + \epsilon_\rho^2\epsilon_q^2)T.
\end{aligned} \tag{78}
$$

We then proceed to bound the term $\sum_{t=0}^{T-1}\mathbb{E}[\varDelta_t]$. Lemma 7 implies that

$$
\begin{aligned}
&\mathbb{E}[\varDelta_{t+1}|\mathcal{F}_t] \\
&\le \left(1 - \frac{1}{2}\varrho\right)\varDelta_t + \frac{2L_{fix}^2}{\varrho}\mathbb{E}[\|w_{t+1} - w_t\|_2^2\,|\mathcal{F}_t] + \frac{C_5}{(1-\varrho)N} \\
&= \left(1 - \frac{1}{2}\varrho\right)\varDelta_t + \frac{2L_{fix}^2}{\varrho}\alpha^2\mathbb{E}[\|G_{\text{DR}}(w_t, \mathcal{M}_t)\|_2^2\,|\mathcal{F}_t] + \frac{C_5}{(1-\varrho)N} \\
&\le \left(1 - \frac{1}{2}\varrho\right)\varDelta_t + \frac{4L_{fix}^2}{\varrho}\alpha^2\mathbb{E}[\|G_{\text{DR}}(w_t, \mathcal{M}_t) - \nabla_w J(w_t)\|_2^2\,|\mathcal{F}_t] + \frac{4L_{fix}^2}{\varrho}\alpha^2\mathbb{E}[\|\nabla_w J(w_t)\|_2^2\,|\mathcal{F}_t] + \frac{C_5}{(1-\varrho)N} \\
&\overset{(i)}{\le} \left(1 - \frac{1}{2}\varrho + \frac{4C_{11}L_{fix}^2\alpha^2}{\varrho}\right)\varDelta_t + \frac{4L_{fix}^2}{\varrho}\alpha^2\mathbb{E}[\|\nabla_w J(w_t)\|_2^2\,|\mathcal{F}_t] + \left[\frac{96L_{fix}^2C_{10}^2\alpha^2}{\varrho} + \frac{C_5}{1-\varrho}\right]\frac{1}{N} \\
&\quad + \frac{4C_{12}L_{fix}^2\alpha^2}{\varrho}(\epsilon_\rho^2\epsilon_{d_q}^2 + \epsilon_{d_\rho}^2\epsilon_q^2 + \epsilon_\rho^2\epsilon_q^2) \\
&\overset{(ii)}{\le} \left(1 - \frac{1}{4}\varrho\right)\varDelta_t + C_{12}\alpha^2\mathbb{E}[\|\nabla_w J(w_t)\|_2^2\,|\mathcal{F}_t] + \frac{C_{13}}{N} + C_{14}(\epsilon_\rho^2\epsilon_{d_q}^2 + \epsilon_{d_\rho}^2\epsilon_q^2 + \epsilon_\rho^2\epsilon_q^2),
\end{aligned} \tag{79}
$$

where $L_{fix}^2 = (L_\kappa^2 + L_\xi^2 + L_\zeta^2)$, (i) follows from eq. (76), (ii) follows from the fact that $C_{12} = \frac{4L_{fix}^2}{\varrho}$, $C_{13} = \frac{96L_{fix}^2C_{10}^2\alpha^2}{\varrho} + \frac{C_5}{1-\varrho}$ and $C_{14} = \frac{4C_{12}L_{fix}^2\alpha^2}{\varrho}$, and for small enough α, we have $\alpha \le \frac{\varrho}{4\sqrt{C_{11}}L_{fix}}$. Taking the expectation on both sides of eq. (79) and applying it iteratively yield

$$
\begin{aligned}
\mathbb{E}[\varDelta_t] &\le \left(1 - \frac{1}{4}\varrho\right)^t\varDelta_0 + C_{12}\alpha^2\sum_{i=0}^{t-1}\left(1 - \frac{1}{4}\varrho\right)^{t-1-i}\mathbb{E}[\|\nabla_w J(w_i)\|_2^2] + \frac{C_{13}}{N}\sum_{i=0}^{t-1}\left(1 - \frac{1}{4}\varrho\right)^{t-1-i} \\
&\quad + C_{14}(\epsilon_\rho^2\epsilon_{d_q}^2 + \epsilon_{d_\rho}^2\epsilon_q^2 + \epsilon_\rho^2\epsilon_q^2)\sum_{i=0}^{t-1}\left(1 - \frac{1}{4}\varrho\right)^{t-1-i}.
\end{aligned} \tag{80}
$$

Taking the summation on eq. (80) over $t = 0, \cdots, T-1$ yields

$$
\sum_{t=0}^{T-1}\mathbb{E}[\varDelta_t] \le \varDelta_0\sum_{t=0}^{T-1}\left(1 - \frac{1}{4}\varrho\right)^t + C_{12}\alpha^2\sum_{t=0}^{T-1}\sum_{i=0}^{t-1}\left(1 - \frac{1}{4}\varrho\right)^{t-1-i}\mathbb{E}[\|\nabla_w J(w_i)\|_2^2] + \frac{C_{13}}{N}\sum_{t=0}^{T-1}\sum_{i=0}^{t-1}\left(1 - \frac{1}{4}\varrho\right)^{t-1-i}
$$

$$
\begin{aligned}
&+ C_{14}(\epsilon_\rho^2\epsilon_{d_q}^2 + \epsilon_{d_\rho}^2\epsilon_q^2 + \epsilon_\rho^2\epsilon_q^2)\sum_{t=0}^{T-1}\sum_{i=0}^{t-1}\left(1-\frac{1}{4}\varrho\right)^{t-1-i} \\
&\le \frac{4}{\varrho}\Delta_0 + \frac{4C_{12}\alpha^2}{\varrho}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla_w J(w_t)\|_2^2] + \frac{4C_{13}T}{\varrho N} + \frac{4C_{14}T}{\varrho}(\epsilon_\rho^2\epsilon_{d_q}^2 + \epsilon_{d_\rho}^2\epsilon_q^2 + \epsilon_\rho^2\epsilon_q^2). \qquad (81)
\end{aligned}
$$

Substituting eq. (81) into eq. (78) yields

$$
\begin{aligned}
&\left(\frac{1}{2}\alpha - L_J\alpha^2\right)\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla_w J(w_t)\|_2^2] \\
&\le \mathbb{E}[J(w_T)] - J(w_0) + \left(\frac{1}{2}\alpha + L_J\alpha^2\right)\frac{4C_{11}}{\varrho}\Delta_0 + \frac{4C_{12}\alpha^2}{\varrho}\left(\frac{1}{2}\alpha + L_J\alpha^2\right)\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla_w J(w_i)\|_2^2] \\
&+ \left(\frac{1}{2}\alpha + L_J\alpha^2\right)\frac{24\varrho C_{10}^2 T + 4C_{13}T}{\varrho N} + \left(\frac{1}{2}\alpha + L_J\alpha^2\right)\left(C_{12} + \frac{4C_{14}T}{\varrho}\right)(\epsilon_\rho^2\epsilon_{d_q}^2 + \epsilon_{d_\rho}^2\epsilon_q^2 + \epsilon_\rho^2\epsilon_q^2)T, \qquad (82)
\end{aligned}
$$

which implies

$$
\begin{aligned}
&\left(\frac{1}{2}\alpha - L_J\alpha^2 - \frac{4C_{12}\alpha^3}{\varrho}\left(\frac{1}{2} + L_J\alpha\right)\right)\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla_w J(w_t)\|_2^2] \\
&\le \mathbb{E}[J(w_T)] - J(w_0) + \left(\frac{1}{2}\alpha + L_J\alpha^2\right)\frac{4C_{11}}{\varrho}\Delta_0 + \left(\frac{1}{2}\alpha + L_J\alpha^2\right)\frac{24\varrho C_{10}^2 T + 4C_{13}T}{\varrho N} \\
&+ \left(\frac{1}{2}\alpha + L_J\alpha^2\right)\left(C_{12} + \frac{4C_{14}T}{\varrho}\right)(\epsilon_\rho^2\epsilon_{d_q}^2 + \epsilon_{d_\rho}^2\epsilon_q^2 + \epsilon_\rho^2\epsilon_q^2)T. \qquad (83)
\end{aligned}
$$

For small enough α, we can guarantee that $\frac{1}{2}\alpha - L_J\alpha^2 - \frac{4C_{12}\alpha^3}{\varrho}\left(\frac{1}{2} + L_J\alpha\right) > 0$. Dividing both sides of eq. (83) by $T\left(\frac{1}{2}\alpha - L_J\alpha^2 - \frac{4C_{12}\alpha^3}{\varrho}\left(\frac{1}{2} + L_J\alpha\right)\right)$, we obtain

$$
\mathbb{E}[\left\|\nabla_w J(w_{\hat{T}})\right\|_2^2] = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla_w J(w_t)\|_2^2] \le \Theta\left(\frac{1}{T}\right) + \Theta\left(\frac{1}{N}\right) + \Theta(\epsilon_\rho^2\epsilon_{d_q}^2 + \epsilon_{d_\rho}^2\epsilon_q^2 + \epsilon_\rho^2\epsilon_q^2).
$$

Note that $\mathbb{E}[\left\|\nabla_w J(w_{\hat{T}})\right\|_2] \le \sqrt{\mathbb{E}[\left\|\nabla_w J(w_{\hat{T}})\right\|_2^2]}$ and $\sqrt{\sum_i a_i} \le \sum_i \sqrt{a_i}$ for $a_i \ge 0$. We obtain

$$
\mathbb{E}[\left\|\nabla_w J(w_{\hat{T}})\right\|_2] \le \Theta\left(\frac{1}{\sqrt{T}}\right) + \Theta\left(\frac{1}{\sqrt{N}}\right) + \Theta(\epsilon_\rho\epsilon_{d_q} + \epsilon_{d_\rho}\epsilon_q + \epsilon_\rho\epsilon_q),
$$

which completes the proof.

E. Proof of Theorem 3

The following proposition can be directly obtained from Corollary 6.10. in (Agarwal et al., 2019).

Proposition 2. *Consider the DR-Off-PAC update given in Algorithm 1. Suppose Assumption 5 holds. Let $w_t^* = F^{-1}(w_t)\nabla_w J(w_t)$ be the exact NPG update direction at w_t. Then, we have*

$$
\begin{aligned}
&J(\pi^*) - J(w_{\hat{T}}) \\
&\le \frac{\epsilon_{approx}}{1-\gamma} + \frac{1}{\alpha T}\mathbb{E}_{\nu_{\pi^*}}[KL(\pi^*(\cdot|s)||\pi_{w_0}(\cdot|s))] + \frac{L_{sc}\alpha}{2T}\sum_{t=0}^{T-1}\|G_{DR}(w_t,\mathcal{M}_t)\|_2^2 + \frac{C_{sc}}{T}\sum_{t=0}^{T-1}\|G_{DR}(w_t,\mathcal{M}_t) - w_t^*\|_2. \qquad (84)
\end{aligned}
$$

Proposition 2 indicates that, as long as the DR-Off-PAC update is close enough to the exact NPG update, then DP-Off-PAC is guaranteed to converge to a neighbourhood of the global optimal $J(\pi^*)$ with a $\left(\frac{\epsilon_{approx}}{1-\gamma}\right)$-level gap. We then proceed to prove Theorem 3.

Proof. We start with eq. (84) as follows:

$$
\begin{aligned}
&J(\pi^*) - \mathbb{E}[J(w_{\hat{T}})] \\
&\le \frac{\epsilon_{approx}}{1-\gamma} + \frac{1}{\alpha T}\mathbb{E}_{\nu_{\pi^*}}[\mathrm{KL}(\pi^*(\cdot|s)||\pi_{w_0}(\cdot|s))] + \frac{L_{sc}\alpha}{2T}\sum_{t=0}^{T-1}\mathbb{E}[\|G_{\mathrm{DR}}(w_t,\mathcal{M}_t)\|_2^2] + \frac{C_{sc}}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|G_{\mathrm{DR}}(w_t,\mathcal{M}_t) - w_t^*\|_2\right] \\
&\le \frac{\epsilon_{approx}}{1-\gamma} + \frac{1}{\alpha T}\mathbb{E}_{\nu_{\pi^*}}[\mathrm{KL}(\pi^*(\cdot|s)||\pi_{w_0}(\cdot|s))] + \frac{L_{sc}\alpha}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|G_{\mathrm{DR}}(w_t,\mathcal{M}_t) - \nabla_w J(w_t)\|_2^2] + \frac{L_{sc}\alpha}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla_w J(w_t)\|_2^2\right] \\
&\quad + \frac{C_{sc}}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|G_{\mathrm{DR}}(w_t,\mathcal{M}_t) - \nabla_w J(w_t)\|_2] + \frac{C_{sc}}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla_w J(w_t) - w_t^*\|_2]. \qquad (85)
\end{aligned}
$$

We then bound the error terms on the right-hand side of eq. (85) separately.

First consider the term $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla_w J(w_t)\|_2^2\right]$. We have

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla_w J(w_t)\|_2^2\right] &= \mathbb{E}\left[\left\|\nabla_w J(w_{\hat{T}})\right\|_2^2\right] \\
&\overset{(i)}{\le} \Theta\left(\frac{1}{T}\right) + \Theta\left(\frac{1}{N}\right) + \Theta(\epsilon_\rho^2\epsilon_{d_q}^2 + \epsilon_{d_\rho}^2\epsilon_q^2 + \epsilon_\rho^2\epsilon_q^2), \qquad (86)
\end{aligned}
$$

where (i) follows from Theorem 2.

Then we consider the term $\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla_w J(w_t) - w_t^*\|_2$. We proceed as follows.

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla_w J(w_t) - w_t^*\|_2\right] \\
&\quad = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|(I - F^{-1}(w_t))\nabla_w J(w_t)\right\|_2\right] \le \left(1 + \frac{1}{\lambda_F}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla_w J(w_t)\|_2\right] \\
&\quad = \left(1 + \frac{1}{\lambda_F}\right)\mathbb{E}\left[\left\|\nabla_w J(w_{\hat{T}})\right\|_2\right] = \left(1 + \frac{1}{\lambda_F}\right)\sqrt{\mathbb{E}\left[\left\|\nabla_w J(w_{\hat{T}})\right\|_2^2\right]} \\
&\quad \overset{(i)}{\le} \Theta\left(\frac{1}{\sqrt{T}}\right) + \Theta\left(\frac{1}{\sqrt{N}}\right) + \Theta(\epsilon_\rho\epsilon_{d_q} + \epsilon_{d_\rho}\epsilon_q + \epsilon_\rho\epsilon_q), \qquad (87)
\end{aligned}
$$

where (i) follows from eq. (86) and the fact that $\sqrt{\sum_i a_i} \le \sum_i \sqrt{a_i}$ for $a_i \ge 0$.

We then consider the term $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|G_{\mathrm{DR}}(w_t,\mathcal{M}_t) - \nabla_w J(w_t)\|_2]$. Recalling eq. (76), we have

$$
\mathbb{E}\left[\|G_{\mathrm{DR}}(w_t,\mathcal{M}_t) - \nabla_w J(w_t)\|_2^2\,|\mathcal{F}_t\right] \le C_{11}\varDelta_t + \frac{24C_{10}^2}{N} + C_{12}(\epsilon_\rho^2\epsilon_{d_q}^2 + \epsilon_{d_\rho}^2\epsilon_q^2 + \epsilon_\rho^2\epsilon_q^2),
$$

which implies

$$
\begin{aligned}
\mathbb{E}[\|G_{\mathrm{DR}}(w_t,\mathcal{M}_t) - \nabla_w J(w_t)\|_2\,|\mathcal{F}_t] &\le \sqrt{\mathbb{E}\left[\|G_{\mathrm{DR}}(w_t,\mathcal{M}_t) - \nabla_w J(w_t)\|_2^2\,|\mathcal{F}_t\right]} \\
&\le \sqrt{C_{11}}\sqrt{\varDelta_t} + \frac{5C_{10}}{\sqrt{N}} + \sqrt{C_{12}}(\epsilon_\rho\epsilon_{d_q} + \epsilon_{d_\rho}\epsilon_q + \epsilon_\rho\epsilon_q). \qquad (88)
\end{aligned}
$$

Taking the expectation on both sides of eq. (88) and taking the summation over $t = 1\cdots T-1$ yield

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|G_{\mathrm{DR}}(w_t,\mathcal{M}_t) - \nabla_w J(w_t)\|_2] \le \frac{\sqrt{C_{11}}}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\sqrt{\varDelta_t}\right] + \frac{5C_{10}}{\sqrt{N}} + \sqrt{C_{12}}(\epsilon_\rho\epsilon_{d_q} + \epsilon_{d_\rho}\epsilon_q + \epsilon_\rho\epsilon_q). \qquad (89)
$$

We then bound the term $\sum_{t=0}^{T-1}\mathbb{E}\left[\sqrt{\Delta_t}\right]$. Note that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\sqrt{\Delta_t}] = \mathbb{E}[\sqrt{\Delta_{\hat{T}}}] \leq \sqrt{\mathbb{E}[\Delta_{\hat{T}}]} = \sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\Delta_t]}. \tag{90}$$

Recalling eq. (81), we have

$$\sum_{t=0}^{T-1}\mathbb{E}[\Delta_t] \leq \frac{4}{\varrho}\Delta_0 + \frac{4C_{12}\alpha^2}{\varrho}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla_w J(w_t)\|_2^2] + \frac{4C_{13}T}{\varrho N} + \frac{4C_{14}T}{\varrho}(\epsilon_\rho^2\epsilon_{d_q}^2 + \epsilon_{d_\rho}^2\epsilon_q^2 + \epsilon_\rho^2\epsilon_q^2),$$

which implies

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\sqrt{\Delta_t}] &\leq \sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\Delta_t]}\\ &\leq \sqrt{\frac{4\Delta_0}{\varrho T}} + \frac{2\sqrt{C_{12}}\alpha}{\sqrt{\varrho}}\sqrt{\mathbb{E}[\left\|\nabla_w J(w_{\hat{T}})\right\|_2^2]} + \sqrt{\frac{4C_{13}}{\varrho N}} + \sqrt{\frac{4C_{14}}{\varrho}}(\epsilon_\rho\epsilon_{d_q} + \epsilon_{d_\rho}\epsilon_q + \epsilon_\rho\epsilon_q)\\ &\overset{(i)}{\leq} \Theta\left(\frac{1}{\sqrt{T}}\right) + \Theta\left(\frac{1}{\sqrt{N}}\right) + \Theta(\epsilon_\rho\epsilon_{d_q} + \epsilon_{d_\rho}\epsilon_q + \epsilon_\rho\epsilon_q),\end{aligned} \tag{91}$$

where (i) follows from eq. (86). Substituting eq. (91) into eq. (89) yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|G_{\text{DR}}(w_t,\mathcal{M}_t) - \nabla_w J(w_t)\|_2] \leq \Theta\left(\frac{1}{\sqrt{T}}\right) + \Theta\left(\frac{1}{\sqrt{N}}\right) + \Theta(\epsilon_\rho\epsilon_{d_q} + \epsilon_{d_\rho}\epsilon_q + \epsilon_\rho\epsilon_q). \tag{92}$$

Finally, we consider the term $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|G_{\text{DR}}(w_t,\mathcal{M}_t) - \nabla_w J(w_t)\|_2^2]$. Taking the expectation on both sides of eq. (76) and taking the summation over $t = 0,\cdots,T-1$ yield

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|G_{\text{DR}}(w_t,\mathcal{M}_t) - \nabla_w J(w_t)\|_2^2\right] \leq \frac{C_{11}}{T}\sum_{t=0}^{T-1}\mathbb{E}[\Delta_t] + \frac{24C_{10}^2}{N} + C_{12}(\epsilon_\rho^2\epsilon_{d_q}^2 + \epsilon_{d_\rho}^2\epsilon_q^2 + \epsilon_\rho^2\epsilon_q^2)\\ &\overset{(i)}{\leq} \frac{4C_{11}}{\varrho T}\Delta_0 + \frac{4C_{11}C_{12}\alpha^2}{\varrho T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla_w J(w_t)\|_2^2] + \frac{4C_{11}C_{13}}{\varrho N} + \frac{4C_{11}C_{14}}{\varrho}(\epsilon_\rho^2\epsilon_{d_q}^2 + \epsilon_{d_\rho}^2\epsilon_q^2 + \epsilon_\rho^2\epsilon_q^2)\\ &\quad + \frac{24C_{10}^2}{N} + C_{12}(\epsilon_\rho^2\epsilon_{d_q}^2 + \epsilon_{d_\rho}^2\epsilon_q^2 + \epsilon_\rho^2\epsilon_q^2)\\ &\overset{(ii)}{\leq} \Theta\left(\frac{1}{T}\right) + \Theta\left(\frac{1}{N}\right) + \Theta(\epsilon_\rho^2\epsilon_{d_q}^2 + \epsilon_{d_\rho}^2\epsilon_q^2 + \epsilon_\rho^2\epsilon_q^2),\end{aligned} \tag{93}$$

where (i) follows from eq. (81), and (ii) follows from Theorem 2.

Substituting eq. (86), eq. (87), eq. (92) and eq. (93) into eq. (85) yields

$$J(\pi^*) - J(w_{\hat{T}}) \leq \frac{\epsilon_{approx}}{1-\gamma} + \Theta\left(\frac{1}{\sqrt{T}}\right) + \Theta\left(\frac{1}{\sqrt{N}}\right) + \Theta(\epsilon_\rho\epsilon_{d_q} + \epsilon_{d_\rho}\epsilon_q + \epsilon_\rho\epsilon_q),$$

which completes the proof.

$\square$