

## Supplementary Materials

### A EXPERIMENTAL SETTING

In our constrained Cartpole environment, the cart is restricted in the area  $[-2.4, 2.4]$ . Each episode length is no longer than 200 and terminated when the angle of the pole is larger than 12 degree. During the training, the agent receives a reward +1 for every step taken, but is penalized with cost +1 if (1) entering the area  $[-2.4, -2.2]$ ,  $[-1.3, -1.1]$ ,  $[-0.1, 0.1]$ ,  $[1.1, 1.3]$ , and  $[2.2, 2.4]$ ; or (2) having of the angle of pole larger than 6 degree.

In our constrained Acrobot environment, each episode has length 500. During the training, the agent receives a reward +1 when the end-effector is at a height of 0.5, but is penalized with cost +1 when (1) a torque with value +1 is applied when the first pendulum swings at a anticlockwise direction; or (2) a torque with value +1 is applied when the second pendulum swings at a anticlockwise direction with respect to the first pendulum.

For details about the update of PDO please refer to (Achiam et al., 2017, 10.3.3). The performance of PDO is very sensitive to the stepsize of dual variable update. If the stepsize is too small, then the dual variable won't update quickly to enforce the constraints. If the stepsize is too large, then the algorithm will behave conservatively and have low return reward. To appropriately select the stepsize for dual variable, we experiment with learning rate  $\{0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05\}$  for both tasks. The learning rate 0.005 performs the best in the first task, and the learning rate 0.0005 performs the best in the second task. Our reported result of Cartpole is with stepsize 0.005 and our reported result of Acrobot is with stepsize 0.0005.

### B TECHNICAL PROOF OF CRPO IN TABULAR SETTING

#### B.1 SUPPORTING LEMMAS FOR THEOREM 1

**Lemma 3** (The performance difference lemma Kakade & Langford (2002)). *For all policies  $\pi, \pi'$  and initial distribution  $\rho$ , we have*

$$J_i^\rho(\pi) - J_i^\rho(\pi') = \frac{1}{1-\gamma} \mathbb{E}_{s \sim \nu_\rho} \mathbb{E}_{a \sim \pi(\cdot|s)} [A_{\pi'}^i(s, a)]$$

where  $J_i^\rho(\pi)$  and  $\nu_\rho$  denote the accumulated reward (cost) function and visitation distribution under policy  $\pi$  when the initial state distribution is  $\rho$ .

**Lemma 4** (Lemma 5.6. Agarwal et al. (2019)). *Considering the approximated NPG update in line 7 of Algorithm 1 in the tabular setting and  $i = 0$ , the NPG update take the form:*

$$w_{t+1} = w_t + \frac{\alpha}{1-\gamma} \bar{Q}_t^i, \quad \text{and} \quad \pi_{w_{t+1}}(a|s) = \pi_{w_t}(a|s) \frac{\exp(\alpha \bar{Q}_t^i(s, a)/(1-\gamma))}{Z_t(s)},$$

where

$$Z_t(s) = \sum_{a \in \mathcal{A}} \pi_{w_t}(a|s) \exp\left(\frac{\alpha \bar{Q}_t^i(s, a)}{1-\gamma}\right).$$

Note that when we follow the update in line 10 of Algorithm 1, we can obtain similar results for the case  $i \in \{1, \dots, p\}$  as stated in Lemma 4.

**Lemma 5** (Policy gradient property of softmax parameterization). *Considering the softmax policy in the tabular setting (eq. (3)). For any initial state distribution  $\rho$ , we have*

$$\nabla_w J_i^\rho(w) = \mathbb{E}_{s \sim \nu_\rho} \mathbb{E}_{a \sim \pi_w(\cdot|s)} \left[ \left( \mathbb{1}_{as} - \sum_{a' \in \mathcal{A}} \pi_w(a'|s) \mathbb{1}_{a's} \right) Q_{\pi_w}^i(s, a) \right],$$

and

$$\|\nabla_w J_i^\rho(w)\|_2 \leq \frac{2c_{\max}}{1-\gamma},$$

where  $\mathbb{1}_{as}$  is a  $|S| \times |\mathcal{A}|$ -dimension vector, with  $(a, s)$ -th element being one, and the rest elements being zero.

*Proof.* The first result can be obtained directly from Lemma C.1 in Agarwal et al. (2019), we now proceed to prove the second result.

$$\begin{aligned}
\|\nabla_w J_i^\rho(w)\|_2 &= \left\| \mathbb{E} \left[ \left( \mathbb{1}_{as} - \sum_{a' \in \mathcal{A}} \pi_w(a'|s) \mathbb{1}_{a's} \right) Q_{\pi_w}^i(s, a) \right] \right\|_2 \\
&\leq \mathbb{E} \left[ \left\| \left( \mathbb{1}_{as} - \sum_{a' \in \mathcal{A}} \pi_w(a'|s) \mathbb{1}_{a's} \right) Q_{\pi_w}^i(s, a) \right\|_2 \right] \\
&\leq \mathbb{E} \left[ \left\| \mathbb{1}_{as} - \sum_{a' \in \mathcal{A}} \pi_w(a'|s) \mathbb{1}_{a's} \right\|_2 \left\| Q_{\pi_w}^i(s, a) \right\|_2 \right] \\
&\leq 2 \mathbb{E} [Q_{\pi_w}^i(s, a)] \leq \frac{2c_{\max}}{1-\gamma}.
\end{aligned}$$

□

**Lemma 6** (Improvement lower bound for approximated NPG). *For the iterates  $\pi_{w_t}$  generated by the approximated NPG updates in line 7 of Algorithm 1 in the tabular setting, we have for all initial state distribution  $\rho$  and when  $i = 0$ , the following holds*

$$\begin{aligned}
&J_0^\rho(w_{t+1}) - J_0^\rho(w_t) \\
&\geq \frac{1-\gamma}{\alpha} \mathbb{E}_{s \sim \nu_\rho} \left( \log Z_t(s) - \frac{\alpha}{1-\gamma} V_{\pi_{w_t}}^0(s) + \frac{\alpha}{1-\gamma} \sum_{a \in \mathcal{A}} \pi_{w_t}(a|s) \left| \bar{Q}_t^0(s, a) - Q_{\pi_{w_t}}^0(s, a) \right| \right) \\
&\quad - \frac{1}{1-\gamma} \mathbb{E}_{s \sim \nu_\rho} \sum_{a \in \mathcal{A}} \pi_{w_t}(a|s) \left| \bar{Q}_t^0(s, a) - Q_{\pi_{w_t}}^0(s, a) \right| \\
&\quad - \frac{1}{1-\gamma} \mathbb{E}_{s \sim \nu_\rho} \sum_{a \in \mathcal{A}} \pi_{w_{t+1}}(a|s) \left| Q_{\pi_{w_t}}^0(s, a) - \bar{Q}_t^0(s, a) \right|.
\end{aligned}$$

*Proof.* We first provide the following lower bound.

$$\begin{aligned}
&\log Z_t(s) - \frac{\alpha}{1-\gamma} V_{\pi_{w_t}}^i(s) \\
&= \log \sum_{a \in \mathcal{A}} \pi_{w_t}(a|s) \exp \left( \frac{\alpha \bar{Q}_t^i(s, a)}{1-\gamma} \right) - \frac{\alpha}{1-\gamma} V_{\pi_{w_t}}^i(s) \\
&\geq \sum_{a \in \mathcal{A}} \pi_{w_t}(a|s) \log \exp \left( \frac{\alpha \bar{Q}_t^i(s, a)}{1-\gamma} \right) - \frac{\alpha}{1-\gamma} V_{\pi_{w_t}}^i(s) \\
&= \frac{\alpha}{1-\gamma} \sum_{a \in \mathcal{A}} \pi_{w_t}(a|s) (\bar{Q}_t^i(s, a) - Q_{\pi_{w_t}}^i(s, a)) + \frac{\alpha}{1-\gamma} \sum_{a \in \mathcal{A}} \pi_{w_t}(a|s) Q_{\pi_{w_t}}^i(s, a) - \frac{\alpha}{1-\gamma} V_{\pi_{w_t}}^i(s) \\
&= \frac{\alpha}{1-\gamma} \sum_{a \in \mathcal{A}} \pi_{w_t}(a|s) (\bar{Q}_t^i(s, a) - Q_{\pi_{w_t}}^i(s, a)) \\
&\geq \frac{-\alpha}{1-\gamma} \sum_{a \in \mathcal{A}} \pi_{w_t}(a|s) \left| \bar{Q}_t^i(s, a) - Q_{\pi_{w_t}}^i(s, a) \right|.
\end{aligned}$$

Thus we conclude that

$$\log Z_t(s) - \frac{\alpha}{1-\gamma} V_{\pi_{w_t}}^i(s) + \frac{\alpha}{1-\gamma} \sum_{a \in \mathcal{A}} \pi_{w_t}(a|s) \left| \bar{Q}_t^i(s, a) - Q_{\pi_{w_t}}^i(s, a) \right| \geq 0.$$

We then proceed to prove Lemma 6. The performance difference lemma (Lemma 3) implies:

$$\begin{aligned}
&J_i^\rho(w_{t+1}) - J_i^\rho(w_t) \\
&= \frac{1}{1-\gamma} \mathbb{E}_{s \sim \nu_\rho} \sum_{a \in \mathcal{A}} \pi_{w_{t+1}}(a|s) A_{\pi_{w_t}}^i(s, a)
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{1-\gamma} \mathbb{E}_{s \sim \nu_\rho} \sum_{a \in \mathcal{A}} \pi_{w_{t+1}}(a|s) Q_{\pi_{w_t}}^i(s, a) - \frac{1}{1-\gamma} \mathbb{E}_{s \sim \nu_\rho} V_{\pi_{w_t}}^i(s) \\
&= \frac{1}{1-\gamma} \mathbb{E}_{s \sim \nu_\rho} \sum_{a \in \mathcal{A}} \pi_{w_{t+1}}(a|s) \bar{Q}_t^i(s, a) + \frac{1}{1-\gamma} \mathbb{E}_{s \sim \nu_\rho} \sum_{a \in \mathcal{A}} \pi_{w_{t+1}}(a|s) (Q_{\pi_{w_t}}^i(s, a) - \bar{Q}_t^i(s, a)) \\
&\quad - \frac{1}{1-\gamma} \mathbb{E}_{s \sim \nu_\rho} V_{\pi_{w_t}}^i(s) \\
&\stackrel{(i)}{=} \frac{1}{\alpha} \mathbb{E}_{s \sim \nu_\rho} \sum_{a \in \mathcal{A}} \pi_{w_{t+1}}(a|s) \log \left( \frac{\pi_{w_{t+1}}(a|s) Z_t(s)}{\pi_{w_t}(a|s)} \right) \\
&\quad + \frac{1}{1-\gamma} \mathbb{E}_{s \sim \nu_\rho} \sum_{a \in \mathcal{A}} \pi_{w_{t+1}}(a|s) (Q_{\pi_{w_t}}^i(s, a) - \bar{Q}_t^i(s, a)) - \frac{1}{1-\gamma} \mathbb{E}_{s \sim \nu_\rho} V_{\pi_{w_t}}^i(s) \\
&= \frac{1}{\alpha} \mathbb{E}_{s \sim \nu_\rho} D_{\text{KL}}(\pi_{w_{t+1}} || \pi_{w_t}) + \frac{1}{\alpha} \mathbb{E}_{s \sim \nu_\rho} \log Z_t(s) \\
&\quad + \frac{1}{1-\gamma} \mathbb{E}_{s \sim \nu_\rho} \sum_{a \in \mathcal{A}} \pi_{w_{t+1}}(a|s) (Q_{\pi_{w_t}}^i(s, a) - \bar{Q}_t^i(s, a)) - \frac{1}{1-\gamma} \mathbb{E}_{s \sim \nu_\rho} V_{\pi_{w_t}}^i(s) \\
&\geq \frac{1}{\alpha} \mathbb{E}_{s \sim \nu_\rho} \left( \log Z_t(s) - \frac{\alpha}{1-\gamma} V_{\pi_{w_t}}^i(s) + \frac{\alpha}{1-\gamma} \sum_{a \in \mathcal{A}} \pi_{w_t}(a|s) \left| \bar{Q}_t^i(s, a) - Q_{\pi_{w_t}}^i(s, a) \right| \right) \\
&\quad - \frac{1}{1-\gamma} \mathbb{E}_{s \sim \nu_\rho} \sum_{a \in \mathcal{A}} \pi_{w_t}(a|s) \left| \bar{Q}_t^i(s, a) - Q_{\pi_{w_t}}^i(s, a) \right| \\
&\quad - \frac{1}{1-\gamma} \mathbb{E}_{s \sim \nu_\rho} \sum_{a \in \mathcal{A}} \pi_{w_{t+1}}(a|s) \left| Q_{\pi_{w_t}}^i(s, a) - \bar{Q}_t^i(s, a) \right| \\
&\stackrel{(ii)}{\geq} \frac{1-\gamma}{\alpha} \mathbb{E}_{s \sim \nu_\rho} \left( \log Z_t(s) - \frac{\alpha}{1-\gamma} V_{\pi_{w_t}}^i(s) + \frac{\alpha}{1-\gamma} \sum_{a \in \mathcal{A}} \pi_{w_t}(a|s) \left| \bar{Q}_t^i(s, a) - Q_{\pi_{w_t}}^i(s, a) \right| \right) \\
&\quad - \frac{1}{1-\gamma} \mathbb{E}_{s \sim \nu_\rho} \sum_{a \in \mathcal{A}} \pi_{w_t}(a|s) \left| \bar{Q}_t^i(s, a) - Q_{\pi_{w_t}}^i(s, a) \right| \\
&\quad - \frac{1}{1-\gamma} \mathbb{E}_{s \sim \nu_\rho} \sum_{a \in \mathcal{A}} \pi_{w_{t+1}}(a|s) \left| Q_{\pi_{w_t}}^i(s, a) - \bar{Q}_t^i(s, a) \right|
\end{aligned}$$

where (i) follows from the update rule in Lemma 4 and (ii) follows from the facts that  $\|\nu_\rho/\rho\|_\infty \geq 1-\gamma$  and  $\log Z_t(s) - \frac{\alpha}{1-\gamma} V_{\pi_{w_t}}^i(s) + \frac{\alpha}{1-\gamma} \sum_{a \in \mathcal{A}} \pi_{w_t}(a|s) \left| \bar{Q}_t^i(s, a) - Q_{\pi_{w_t}}^i(s, a) \right| \geq 0$ .  $\square$

Note that when we follow the update in line 10 of Algorithm 1, we can obtain similar results for the case  $i \in \{1, \dots, p\}$  as stated in Lemma 6.

**Lemma 7** (Global improvement upper bound for approximated NPG). *Considering the approximated NPG updates in line 7 of Algorithm 1 in the tabular setting when  $i = 0$ , we have*

$$\begin{aligned}
&J_0(\pi^*) - J_0(\pi_{w_t}) \\
&\leq \frac{1}{\alpha} \mathbb{E}_{s \sim \nu^*} (D_{\text{KL}}(\pi^* || \pi_{w_t}) - D_{\text{KL}}(\pi^* || \pi_{w_{t+1}})) + \frac{2\alpha c_{\max}^2 |\mathcal{S}| |\mathcal{A}|}{(1-\gamma)^3} + \frac{3(1+\alpha c_{\max})}{(1-\gamma)^2} \left\| Q_{\pi_{w_t}}^0 - \bar{Q}_t^0 \right\|_2.
\end{aligned}$$

*Proof.* By the performance difference lemma (Lemma 3), we have

$$\begin{aligned}
&J_i(\pi^*) - J_i(\pi_{w_t}) \\
&= \frac{1}{1-\gamma} \mathbb{E}_{s \sim \nu^*} \sum_{a \in \mathcal{A}} \pi^*(a|s) A_{\pi_{w_t}}^i(s, a) \\
&= \frac{1}{1-\gamma} \mathbb{E}_{s \sim \nu^*} \sum_{a \in \mathcal{A}} \pi^*(a|s) Q_{\pi_{w_t}}^i(s, a) - \frac{1}{1-\gamma} \mathbb{E}_{s \sim \nu^*} V_{\pi_{w_t}}^i(s)
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{1-\gamma} \mathbb{E}_{s \sim \nu^*} \sum_{a \in \mathcal{A}} \pi^*(a|s) \bar{Q}_t^i(s, a) + \frac{1}{1-\gamma} \mathbb{E}_{s \sim \nu^*} \sum_{a \in \mathcal{A}} \pi^*(a|s) (Q_{\pi_{w_t}}^i(s, a) - \bar{Q}_t^i(s, a)) \\
&\quad - \frac{1}{1-\gamma} \mathbb{E}_{s \sim \nu^*} V_{\pi_{w_t}}^i(s) \\
&\stackrel{(i)}{=} \frac{1}{1-\gamma} \mathbb{E}_{s \sim \nu^*} \sum_{a \in \mathcal{A}} \pi^*(a|s) \log \frac{\pi_{w_{t+1}}(a|s) Z_t(s)}{\pi_{w_t}(a|s)} + \frac{1}{1-\gamma} \mathbb{E}_{s \sim \nu^*} \sum_{a \in \mathcal{A}} \pi^*(a|s) (Q_{\pi_{w_t}}^i(s, a) - \bar{Q}_t^i(s, a)) \\
&\quad - \frac{1}{1-\gamma} \mathbb{E}_{s \sim \nu^*} V_{\pi_{w_t}}^i(s) \\
&= \frac{1}{\alpha} \mathbb{E}_{s \sim \nu^*} (D_{\text{KL}}(\pi^* || \pi_{w_t}) - D_{\text{KL}}(\pi^* || \pi_{w_{t+1}})) + \frac{1}{\alpha} \mathbb{E}_{s \sim \nu^*} \left( \log Z_t(s) - \frac{\alpha}{1-\gamma} V_{\pi_{w_t}}^i(s) \right) \\
&\quad + \frac{1}{1-\gamma} \mathbb{E}_{s \sim \nu^*} \sum_{a \in \mathcal{A}} \pi^*(a|s) (Q_{\pi_{w_t}}^i(s, a) - \bar{Q}_t^i(s, a)) \\
&\leq \frac{1}{\alpha} \mathbb{E}_{s \sim \nu^*} (D_{\text{KL}}(\pi^* || \pi_{w_t}) - D_{\text{KL}}(\pi^* || \pi_{w_{t+1}})) \\
&\quad + \frac{1}{\alpha} \mathbb{E}_{s \sim \nu^*} \left( \log Z_t(s) - \frac{\alpha}{1-\gamma} V_{\pi_{w_t}}^i(s) + \frac{\alpha}{1-\gamma} \sum_{a \in \mathcal{A}} \pi_{w_t}(a|s) \left| \bar{Q}_t^i(s, a) - Q_{\pi_{w_t}}^i(s, a) \right| \right) \\
&\quad + \frac{1}{1-\gamma} \mathbb{E}_{s \sim \nu^*} \sum_{a \in \mathcal{A}} \pi^*(a|s) (Q_{\pi_{w_t}}^i(s, a) - \bar{Q}_t^i(s, a)) \\
&\stackrel{(ii)}{\leq} \frac{1}{\alpha} \mathbb{E}_{s \sim \nu^*} (D_{\text{KL}}(\pi^* || \pi_{w_t}) - D_{\text{KL}}(\pi^* || \pi_{w_{t+1}})) \\
&\quad + \frac{1}{1-\gamma} (J_i^{\nu^*}(w_{t+1}) - J_i^{\nu^*}(w_t)) + \frac{1}{(1-\gamma)^2} \mathbb{E}_{s \sim \nu^*} \sum_{a \in \mathcal{A}} \pi_{w_{t+1}}(a|s) \left| Q_{\pi_{w_t}}^i(s, a) - \bar{Q}_t^i(s, a) \right| \\
&\quad + \frac{1}{(1-\gamma)^2} \mathbb{E}_{s \sim \nu^*} \sum_{a \in \mathcal{A}} \pi_{w_t}(a|s) \left| Q_{\pi_{w_t}}^i(s, a) - \bar{Q}_t^i(s, a) \right| \\
&\quad + \frac{1}{1-\gamma} \mathbb{E}_{s \sim \nu^*} \sum_{a \in \mathcal{A}} \pi^*(a|s) \left| Q_{\pi_{w_t}}^i(s, a) - \bar{Q}_t^i(s, a) \right| \\
&\stackrel{(iii)}{\leq} \frac{1}{\alpha} \mathbb{E}_{s \sim \nu^*} (D_{\text{KL}}(\pi^* || \pi_{w_t}) - D_{\text{KL}}(\pi^* || \pi_{w_{t+1}})) + \frac{2c_{\max}}{(1-\gamma)^2} \|w_{t+1} - w_t\|_2 + \frac{3}{(1-\gamma)^2} \|Q_{\pi_{w_t}}^i - \bar{Q}_t^i\|_2 \\
&= \frac{1}{\alpha} \mathbb{E}_{s \sim \nu^*} (D_{\text{KL}}(\pi^* || \pi_{w_t}) - D_{\text{KL}}(\pi^* || \pi_{w_{t+1}})) + \frac{2\alpha c_{\max}}{(1-\gamma)^2} \|\bar{Q}_t^i\|_2 + \frac{3}{(1-\gamma)^2} \|Q_{\pi_{w_t}}^i - \bar{Q}_t^i\|_2 \\
&\leq \frac{1}{\alpha} \mathbb{E}_{s \sim \nu^*} (D_{\text{KL}}(\pi^* || \pi_{w_t}) - D_{\text{KL}}(\pi^* || \pi_{w_{t+1}})) + \frac{2\alpha c_{\max}}{(1-\gamma)^2} \|Q_{\pi_{w_t}}^i\|_2 + \frac{3(1+\alpha c_{\max})}{(1-\gamma)^2} \|Q_{\pi_{w_t}}^i - \bar{Q}_t^i\|_2 \\
&\leq \frac{1}{\alpha} \mathbb{E}_{s \sim \nu^*} (D_{\text{KL}}(\pi^* || \pi_{w_t}) - D_{\text{KL}}(\pi^* || \pi_{w_{t+1}})) + \frac{2\alpha c_{\max}^2 |\mathcal{S}| |\mathcal{A}|}{(1-\gamma)^3} + \frac{3(1+\alpha c_{\max})}{(1-\gamma)^2} \|Q_{\pi_{w_t}}^i - \bar{Q}_t^i\|_2,
\end{aligned}$$

where (i) follows from Lemma 4, (ii) follows from Lemma 6 and (iii) follows from the Lipschitz property of  $J_i^{\nu^*}(w)$  such that  $J_i^{\nu^*}(w_{t+1}) - J_i^{\nu^*}(w_t) \leq \frac{2c_{\max}}{1-\gamma} \|w_{t+1} - w_t\|_2$ , which is proved by Proposition 1 in Xu et al. (2020b).  $\square$

Note that when we follow the update in line 10 of Algorithm 1, we can obtain similar results for the case  $i \in \{1, \dots, p\}$  as stated in Lemma 7:

$$\begin{aligned}
&J_i(\pi_{w_t}) - J_i(\pi^*) \\
&\leq \frac{1}{\alpha} \mathbb{E}_{s \sim \nu^*} (D_{\text{KL}}(\pi^* || \pi_{w_t}) - D_{\text{KL}}(\pi^* || \pi_{w_{t+1}})) + \frac{2\alpha c_{\max}^2 |\mathcal{S}| |\mathcal{A}|}{(1-\gamma)^3} + \frac{3(1+\alpha c_{\max})}{(1-\gamma)^2} \|Q_{\pi_{w_t}}^i - \bar{Q}_t^i\|_2.
\end{aligned}$$

## B.2 PROOF OF THEOREM 1

We first prove the following lemmas.

**Lemma 8.** *Considering the CRPO in Algorithm 1 in the tabular setting. Let  $K_{in} = \Theta(T^{1/\sigma} \log^{2/\sigma}(|\mathcal{S}|^2 |\mathcal{A}|^2 T^{1+2/\sigma}/\delta))$ . With probability at least  $1 - \delta$ , we have*

$$\begin{aligned} & \sum_{t \in \mathcal{N}_0} (J_0(\pi^*) - J_0(\pi_{w_t})) + \alpha \eta \sum_{i=1}^p |\mathcal{N}_i| \\ & \leq \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0}) + \frac{2\alpha^2 c_{\max}^2 |\mathcal{S}| |\mathcal{A}| T}{(1-\gamma)^3} + \frac{\alpha \sqrt{T} (2 + (1-\gamma)^2 + 2\alpha c_{\max})}{(1-\gamma)^2}, \end{aligned}$$

*Proof.* We define the  $\mathcal{N}_i$  as the set of steps that CRPO algorithm chooses to minimize the  $i$ -th constraint. If  $t \in \mathcal{N}_0$ , by Lemma 7 we have

$$\begin{aligned} & \alpha (J_0(\pi^*) - J_0(\pi_{w_t})) \\ & \leq \mathbb{E}_{s \sim \nu^*} (D_{\text{KL}}(\pi^* || \pi_{w_t}) - D_{\text{KL}}(\pi^* || \pi_{w_{t+1}})) + \frac{2\alpha^2 c_{\max}^2 |\mathcal{S}| |\mathcal{A}|}{(1-\gamma)^3} + \frac{3\alpha(1 + \alpha c_{\max})}{(1-\gamma)^2} \|Q_{\pi_{w_t}}^0 - \bar{Q}_t^0\|_2. \end{aligned} \quad (13)$$

If  $t \in \mathcal{N}_i$ , similarly we can obtain

$$\begin{aligned} & \alpha (J_i(\pi_{w_t}) - J_i(\pi^*)) \\ & \leq \mathbb{E}_{s \sim \nu^*} (D_{\text{KL}}(\pi^* || \pi_{w_t}) - D_{\text{KL}}(\pi^* || \pi_{w_{t+1}})) + \frac{2\alpha^2 c_{\max}^2 |\mathcal{S}| |\mathcal{A}|}{(1-\gamma)^3} + \frac{3\alpha(1 + \alpha c_{\max})}{(1-\gamma)^2} \|Q_{\pi_{w_t}}^i - \bar{Q}_t^i\|_2. \end{aligned} \quad (14)$$

Summing eq. (13) and eq. (14) from  $t = 0$  to  $T - 1$  yields

$$\begin{aligned} & \alpha \sum_{t \in \mathcal{N}_0} (J_0(\pi^*) - J_0(\pi_{w_t})) + \alpha \sum_{i=1}^p \sum_{t \in \mathcal{N}_i} (J_i(\pi_{w_t}) - J_i(\pi^*)) \\ & \leq \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0}) + \frac{2\alpha^2 c_{\max}^2 |\mathcal{S}| |\mathcal{A}| T}{(1-\gamma)^3} + \frac{3\alpha(1 + \alpha c_{\max})}{(1-\gamma)^2} \sum_{i=0}^p \sum_{t \in \mathcal{N}_i} \|Q_{\pi_{w_t}}^i - \bar{Q}_t^i\|_2. \end{aligned} \quad (15)$$

Note that when  $t \in \mathcal{N}_i$  ( $i \neq 0$ ), we have  $\bar{J}_i(\theta_t^i) > d_i + \eta$  (line 11 in Algorithm 1), which implies that

$$\begin{aligned} J_i(\pi_{w_t}) - J_i(\pi^*) & \geq \bar{J}_i(\theta_t^i) - J_i(\pi^*) - |\bar{J}_i(\theta_t^i) - J_i(\pi_{w_t})| \\ & \geq d_i + \eta - J_i(\pi^*) - |\bar{J}_i(\theta_t^i) - J_i(\pi_{w_t})| \\ & \geq \eta - \|Q_{\pi_w}^i - \bar{Q}_t^i\|_2. \end{aligned} \quad (16)$$

Substituting eq. (16) into eq. (15) yields

$$\begin{aligned} & \alpha \sum_{t \in \mathcal{N}_0} (J_0(\pi^*) - J_0(\pi_{w_t})) + \alpha \eta \sum_{i=1}^p |\mathcal{N}_i| - \alpha \sum_{i=1}^p \sum_{t \in \mathcal{N}_i} \|Q_{\pi_{w_t}}^i - \bar{Q}_t^i\|_2 \\ & \leq \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0}) + \frac{2\alpha^2 c_{\max}^2 |\mathcal{S}| |\mathcal{A}| T}{(1-\gamma)^3} + \frac{3\alpha(1 + \alpha c_{\max})}{(1-\gamma)^2} \sum_{i=0}^p \sum_{t \in \mathcal{N}_i} \|Q_{\pi_{w_t}}^i - \bar{Q}_t^i\|_2, \end{aligned}$$

which implies

$$\begin{aligned} & \alpha \sum_{t \in \mathcal{N}_0} (J_0(\pi^*) - J_0(\pi_{w_t})) + \alpha \eta \sum_{i=1}^p |\mathcal{N}_i| \\ & \leq \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0}) + \frac{2\alpha^2 c_{\max}^2 |\mathcal{S}| |\mathcal{A}| T}{(1-\gamma)^3} + \frac{\alpha(2 + (1-\gamma)^2 + 3\alpha c_{\max})}{(1-\gamma)^2} \sum_{i=0}^p \sum_{t \in \mathcal{N}_i} \|Q_{\pi_{w_t}}^i - \bar{Q}_t^i\|_2. \end{aligned} \quad (17)$$

By Lemma 1, we have with probability at least  $1 - \delta$  the following holds

$$\|Q_{\pi_{w_t}}^i - \bar{Q}_t^i\|_2 = \mathcal{O} \left( \frac{\log(|\mathcal{S}|^2 |\mathcal{A}|^2 K_{in}^2 / \delta)}{(1-\gamma) K_{in}^{\sigma/2}} \right).$$

Thus, if we let

$$K_{\text{in}} = \Theta \left( \left( \frac{T}{(1-\gamma)^2 |\mathcal{S}| |\mathcal{A}|} \right)^{\frac{1}{\sigma}} \log^{\frac{2}{\sigma}} \left( \frac{T^{\frac{2}{\sigma}+1}}{\delta(1-\gamma)^{\frac{2}{\sigma}} |\mathcal{S}|^{\frac{2}{\sigma}-2} |\mathcal{A}|^{\frac{2}{\sigma}-2}} \right) \right),$$

then with probability at least  $1 - \delta/T$ , we have

$$\|Q_{\pi_{w_t}}^i - \bar{Q}_t^i\|_2 \leq \frac{\sqrt{(1-\gamma) |\mathcal{S}| |\mathcal{A}|}}{\sqrt{T}}. \quad (18)$$

Applying union bound to eq. (18) from  $t = 0$  to  $T - 1$ , we have with probability at least  $1 - \delta$  the following holds

$$\sum_{i=0}^p \sum_{t \in \mathcal{N}_i} \|Q_{\pi_{w_t}}^i - \bar{Q}_t^i\|_2 \leq \sqrt{(1-\gamma) |\mathcal{S}| |\mathcal{A}| T}, \quad (19)$$

which further implies that, with probability at least  $1 - \delta$ , we have

$$\begin{aligned} & \alpha \sum_{t \in \mathcal{N}_0} (J_0(\pi^*) - J_0(\pi_{w_t})) + \alpha \eta \sum_{i=1}^p |\mathcal{N}_i| \\ & \leq \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0}) + \frac{2\alpha^2 c_{\max}^2 |\mathcal{S}| |\mathcal{A}| T}{(1-\gamma)^3} + \frac{\alpha \sqrt{|\mathcal{S}| |\mathcal{A}| T} (2 + (1-\gamma)^2 + 3\alpha c_{\max})}{(1-\gamma)^{1.5}}, \end{aligned}$$

which complete the proof.  $\square$

**Lemma 9.** *If*

$$\frac{1}{2} \alpha \eta T \geq \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0}) + \frac{2\alpha^2 c_{\max}^2 |\mathcal{S}| |\mathcal{A}| T}{(1-\gamma)^3} + \frac{\alpha \sqrt{|\mathcal{S}| |\mathcal{A}| T} (2 + (1-\gamma)^2 + 3\alpha c_{\max})}{(1-\gamma)^{1.5}}, \quad (20)$$

*then with probability at least  $1 - \delta$ , we have the following holds*

1.  $\mathcal{N}_0 \neq \emptyset$ , i.e.,  $w_{\text{out}}$  is well-defined,
2. One of the following two statements must holds,
  - (a)  $|\mathcal{N}_0| \geq T/2$ ,
  - (b)  $\sum_{t \in \mathcal{G}} (J_0(\pi^*) - J_0(w_t)) \leq 0$ .

*Proof.* We consider to prove Lemma 9 in the following event given in eq. (19), which happens with probability at least  $1 - \delta$ :

$$\sum_{i=0}^p \sum_{t \in \mathcal{N}_i} \|Q_{\pi_{w_t}}^i - \bar{Q}_t^i\|_2 \leq \sqrt{(1-\gamma) |\mathcal{S}| |\mathcal{A}| T}.$$

In this event, we have the following inequality holds, which is also the result of Lemma 8.

$$\begin{aligned} & \alpha \sum_{t \in \mathcal{N}_0} (J_0(\pi^*) - J_0(\pi_{w_t})) + \alpha \eta \sum_{i=1}^p |\mathcal{N}_i| \\ & \leq \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0}) + \frac{2\alpha^2 c_{\max}^2 |\mathcal{S}| |\mathcal{A}| T}{(1-\gamma)^3} + \frac{\alpha \sqrt{|\mathcal{S}| |\mathcal{A}| T} (2 + (1-\gamma)^2 + 2\alpha c_{\max})}{(1-\gamma)^{1.5}}. \end{aligned} \quad (21)$$

We first verify item 1. If  $\mathcal{N}_0 = \emptyset$ , then  $\sum_{i=1}^p |\mathcal{N}_i| = T$ , eq. (21) implies that

$$\alpha \eta T \leq \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0}) + \frac{2\alpha^2 c_{\max}^2 |\mathcal{S}| |\mathcal{A}| T}{(1-\gamma)^3} + \frac{\alpha \sqrt{|\mathcal{S}| |\mathcal{A}| T} (2 + (1-\gamma)^2 + 2\alpha c_{\max})}{(1-\gamma)^{1.5}},$$

which contradicts with eq. (20). Thus, we must have  $\mathcal{N}_0 \neq \emptyset$ .

We then proceed to verify the second item. If  $\sum_{t \in \mathcal{N}_0} (J_0(\pi^*) - J_0(w_t)) \leq 0$ , then (b) hold. If  $\sum_{t \in \mathcal{N}_0} (J_0(\pi^*) - J_0(w_t)) \leq 0$ , then Equation (21) implies that

$$\alpha \eta \sum_{i=1}^p |\mathcal{N}_i| \leq \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0}) + \frac{2\alpha^2 c_{\max}^2 |\mathcal{S}| |\mathcal{A}| T}{(1-\gamma)^3} + \frac{\alpha \sqrt{|\mathcal{S}| |\mathcal{A}| T} (2 + (1-\gamma)^2 + 3\alpha c_{\max})}{(1-\gamma)^{1.5}}.$$

Suppose that  $|\mathcal{N}_0| < T/2$ , i.e.,  $\sum_{i=1}^p |\mathcal{N}_i| \geq T/2$ . Then,

$$\begin{aligned} \frac{1}{2} \alpha \eta T &\leq \alpha \eta \sum_{i=1}^p |\mathcal{N}_i| \\ &\leq \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0}) + \frac{2\alpha^2 c_{\max}^2 |\mathcal{S}| |\mathcal{A}| T}{(1-\gamma)^3} + \frac{\alpha \sqrt{|\mathcal{S}| |\mathcal{A}| T} (2 + (1-\gamma)^2 + 3\alpha c_{\max})}{(1-\gamma)^{1.5}}, \end{aligned}$$

which contradicts with eq. (20). Hence, (a) holds.  $\square$

Now, we are ready to prove our main theorem. We restate Theorem 1 as follows to include the specifics of the parameters.

**Theorem 3** (Restatement of Theorem 1). *Consider Algorithm 1 in the tabular setting, let  $\alpha = (1-\gamma)^{1.5} / \sqrt{|\mathcal{S}| |\mathcal{A}| T}$ ,  $\eta = \frac{2\sqrt{|\mathcal{S}| |\mathcal{A}|}}{(1-\gamma)^{1.5} \sqrt{T}} (3 + \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0}) + 3c_{\max} + c_{\max}^2)$ , and*

$$K_{in} = \Theta \left( \left( \frac{T}{(1-\gamma) |\mathcal{S}| |\mathcal{A}|} \right)^{\frac{1}{\sigma}} \log^{\frac{2}{\sigma}} \left( \frac{T^{\frac{2}{\sigma}+1}}{\delta (1-\gamma)^{\frac{2}{\sigma}} |\mathcal{S}|^{\frac{2}{\sigma}-2} |\mathcal{A}|^{\frac{2}{\sigma}-2}} \right) \right).$$

*Suppose the same setting for policy evaluation in Lemma 1 hold. Then, with probability at least  $1 - \delta$ , we have*

$$J_0(\pi^*) - \mathbb{E}[J_0(w_{out})] = \frac{2\sqrt{|\mathcal{S}| |\mathcal{A}|}}{(1-\gamma)^{1.5} \sqrt{T}} (\mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0}) + 3 + 2c_{\max}^2 + 3c_{\max}),$$

*and for all  $i \in \{1, \dots, p\}$ , we have*

$$\mathbb{E}[J_i(\pi_{w_{out}})] - d_i \leq \frac{2\sqrt{|\mathcal{S}| |\mathcal{A}|}}{(1-\gamma)^{1.5} \sqrt{T}} (3 + \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0}) + 3c_{\max} + c_{\max}^2) + \frac{2\sqrt{(1-\gamma) |\mathcal{S}| |\mathcal{A}|}}{\sqrt{T}}.$$

*Proof of Theorem 1.* Similar to the proof of Lemma 9. To prove Theorem 1 (or Theorem 3), we still consider the following event given in eq. (19) that happens with probability at least  $1 - \delta$ :

$$\sum_{i=0}^p \sum_{t \in \mathcal{N}_i} \|Q_{\pi_{w_t}}^i - \bar{Q}_t^i\|_2 \leq \sqrt{(1-\gamma) |\mathcal{S}| |\mathcal{A}| T},$$

which implies

$$\begin{aligned} &\alpha \sum_{t \in \mathcal{N}_0} (J_0(\pi^*) - J_0(\pi_{w_t})) + \alpha \eta \sum_{i=1}^p |\mathcal{N}_i| \\ &\leq \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0}) + \frac{2\alpha^2 c_{\max}^2 |\mathcal{S}| |\mathcal{A}| T}{(1-\gamma)^3} + \frac{\alpha \sqrt{|\mathcal{S}| |\mathcal{A}| T} (2 + (1-\gamma)^2 + 3\alpha c_{\max})}{(1-\gamma)^{1.5}}. \end{aligned}$$

We first consider the convergence rate of the objective function. In the above mentioned event, we have the following holds

$$\begin{aligned} &\alpha \sum_{t \in \mathcal{N}_0} (J_0(\pi^*) - J_0(\pi_{w_t})) \\ &\leq \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0}) + \frac{2\alpha^2 c_{\max}^2 |\mathcal{S}| |\mathcal{A}| T}{(1-\gamma)^3} + \frac{\alpha \sqrt{|\mathcal{S}| |\mathcal{A}| T} (2 + (1-\gamma)^2 + 3\alpha c_{\max})}{(1-\gamma)^{1.5}}. \end{aligned}$$

If  $\sum_{t \in \mathcal{N}_0} (J_0(\pi^*) - J_0(\pi_{w_t})) \leq 0$ , then we have  $J_0(\pi^*) - J_0(\pi_{w_{\text{out}}}) \leq 0$ . If  $\sum_{t \in \mathcal{N}_0} (J_0(\pi^*) - J_0(\pi_{w_t})) \geq 0$ , we have  $|\mathcal{N}_0| \geq T/2$ , which implies the following convergence rate

$$\begin{aligned}
& J_0(\pi^*) - \mathbb{E}[J_0(\pi_{w_{\text{out}}})] \\
&= \frac{1}{|\mathcal{N}_0|} \sum_{t \in \mathcal{N}_0} (J_0(\pi^*) - J_0(\pi_{w_t})) \\
&\leq \frac{2}{\alpha T} \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0}) + \frac{4\alpha c_{\max}^2 |\mathcal{S}| |\mathcal{A}|}{(1-\gamma)^3} + \frac{2\sqrt{|\mathcal{S}| |\mathcal{A}|} (2 + (1-\gamma)^2 + 3\alpha c_{\max})}{(1-\gamma)^{1.5} \sqrt{T}} \\
&\leq \frac{\sqrt{|\mathcal{S}| |\mathcal{A}|}}{(1-\gamma)^{1.5} \sqrt{T}} (2\mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0}) + 6 + 4c_{\max}^2 + 6c_{\max}).
\end{aligned}$$

We then proceed to bound the constraints violation cost. For any  $i \in \{1, \dots, p\}$ , we have

$$\begin{aligned}
\mathbb{E}[J_i(\pi_{w_{\text{out}}})] - d_i &= \frac{1}{|\mathcal{N}_0|} \sum_{t \in \mathcal{N}_0} J_i(\pi_{w_t}) - d_i \\
&\leq \frac{1}{|\mathcal{N}_0|} \sum_{t \in \mathcal{N}_0} (\bar{J}_i(\theta_t^i) - d_i) + \frac{1}{|\mathcal{N}_0|} \sum_{t \in \mathcal{N}_0} |J_i(\pi_{w_t}) - \bar{J}_i(\theta_t^i)| \\
&\leq \eta + \frac{1}{|\mathcal{N}_0|} \sum_{t=0}^{T-1} |J_i(\pi_{w_t}) - \bar{J}_i(\theta_t^i)| \\
&\leq \eta + \frac{1}{|\mathcal{N}_0|} \sum_{i=0}^p \sum_{t \in \mathcal{N}_i} \|Q_{\pi_{w_t}}^i - \bar{Q}_t^i\|_2 \\
&\leq \eta + \frac{2}{T} \sum_{i=0}^p \sum_{t \in \mathcal{N}_i} \|Q_{\pi_{w_t}}^i - \bar{Q}_t^i\|_2.
\end{aligned}$$

In the event defined in eq. (19), we have  $\sum_{i=0}^p \sum_{t \in \mathcal{N}_i} \|Q_{\pi_{w_t}}^i - \bar{Q}_t^i\|_2 \leq \sqrt{(1-\gamma) |\mathcal{S}| |\mathcal{A}| T}$ . Recall the value of the tolerance  $\eta = \frac{2\sqrt{|\mathcal{S}| |\mathcal{A}|}}{(1-\gamma)^{1.5} \sqrt{T}} (3 + \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0}) + 3c_{\max} + c_{\max}^2)$ . With probability at least  $1 - \delta$ , we have

$$\mathbb{E}[J_i(\pi_{w_{\text{out}}})] - d_i \leq \frac{2\sqrt{|\mathcal{S}| |\mathcal{A}|}}{(1-\gamma)^{1.5} \sqrt{T}} (3 + \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0}) + 3c_{\max} + c_{\max}^2) + \frac{2\sqrt{(1-\gamma) |\mathcal{S}| |\mathcal{A}|}}{\sqrt{T}}.$$

□

## C TECHNICAL PROOF OF CRPO IN FUNCTION APPROXIMATION SETTING

For notation simplicity, we denote state action pairs  $(s, a)$  and  $(s', a')$  as  $x$  and  $x'$ , respectively. For notation simplicity, we will write  $\theta_k^i$  as  $\theta_k$  without making any confusing in this subsection. We define

$$f_0(x, \theta) = \frac{1}{\sqrt{m}} \sum_{r=1}^m b_r \mathbf{1}(\theta_{0,r}^\top \psi(x) > 0) \theta_r^\top \psi(x)$$

as the local linearization of  $f(x, \theta)$  at the initial point  $\theta_0$ . We denote the temporal differences as  $\delta_0(x, x'; \theta_k) = f_0((s', a'); \theta_k) - \gamma f_0((s, a); \theta_k) - r(s, a, s')$  and  $\delta_k(x, x'; \theta_k) = f((s', a'); \theta_k) - \gamma f((s, a); \theta_k) - r(s, a, s')$ . We define the stochastic semi-gradient  $g_k(\theta_k) = \delta_k(x_k, x'_k; \theta_k) \nabla_\theta f(x_k, \theta_k)$ , and full semi-gradients  $\bar{g}_0(\theta_k) = \mathbb{E}_{\mu_\pi} [\delta_0(x, x'; \theta_k) \nabla_\theta f_0(x, \theta_k)]$ , and  $\bar{g}_k(\theta_k) = \mathbb{E}_{\mu_\pi} [\delta_k(x, x'; \theta_k) \nabla_\theta f(x, \theta_k)]$ . The approximated stationary point  $\theta^*$  satisfies  $\bar{g}_0(\theta)^\top (\theta - \theta^*) \geq 0$  for any  $\theta \in \mathcal{B}$ . We define the following function spaces

$$\mathcal{F}_{0,m} = \left\{ \frac{1}{\sqrt{m}} \sum_{r=1}^m b_r \mathbf{1}(\theta_{0,r}^\top \psi(x) > 0) \theta_r^\top \psi(x) : \|\theta - \theta_0\|_2 \leq R \right\},$$



and

$$\bar{\mathcal{F}}_{0,m} = \left\{ \frac{1}{\sqrt{m}} \sum_{r=1}^m b_r \mathbb{1}(\theta_{0,r}^\top \psi(s) > 0) \theta_r^\top \psi(x) : \|\theta_r - \theta_{0,r}\|_\infty \leq R/\sqrt{md} \right\},$$

and  $f_0(x, \theta_\pi^*)$  as the projection of  $Q_\pi(x)$  onto the function space  $\mathcal{F}_{0,m}$  in terms of  $\|\cdot\|_{\mu_\pi}$  norm. Without loss of generality, we assume  $0 < \delta < \frac{1}{e}$  in the sequel.

### C.1 SUPPORTING LEMMAS FOR LEMMA 2

We provide the proof of supporting lemmas for Lemma 2. Note that similar properties have been given in Cai et al. (2019). Here we established these properties again under a slightly different initialization scheme compared with that in Cai et al. (2019).

**Lemma 10 (Rahimi & Recht (2009)).** *Let  $f \in \mathcal{F}_{0,\infty}$ , where  $\mathcal{F}_{0,\infty}$  is defined in Assumption 2. For any  $\delta > 0$ , it holds with probability at least  $1 - \delta$  that*

$$\left\| \Pi_{\bar{\mathcal{F}}_{0,m}} f - f \right\|_d^2 \leq \frac{4R^2 \log(\frac{1}{\delta})}{m},$$

where  $d$  could be any distribution over  $\mathcal{S} \times \mathcal{A}$ .

**Lemma 11.** *Suppose Assumption 1 holds. For any policy  $\pi$  and all  $k \geq 0$ , it holds that*

$$\mathbb{E}_{\mu_\pi} \left[ \frac{1}{m} \sum_{r=1}^m |\mathbb{1}(\theta_{k,r}^\top \psi(x) > 0) - \mathbb{1}(\theta_{0,r}^\top \psi(x) > 0)| \right] \leq \frac{C_0 R}{d_1 \sqrt{m}}.$$

*Proof.* Note that  $\mathbb{1}(\theta_{k,r}^\top \psi(x) > 0) \neq \mathbb{1}(\theta_{0,r}^\top \psi(x) > 0)$  implies

$$|\theta_{0,r}^\top \psi(x)| \leq |\theta_{k,r}^\top \psi(x) - \theta_{0,r}^\top \psi(x)| \leq \|\theta_{k,r} - \theta_{0,r}\|_2,$$

which further implies

$$|\mathbb{1}(\theta_{k,r}^\top \psi(x) > 0) - \mathbb{1}(\theta_{0,r}^\top \psi(x) > 0)| \leq \mathbb{1}(|\theta_{0,r}^\top \psi(x)| \leq \|\theta_{k,r} - \theta_{0,r}\|_2). \quad (22)$$

Then, we can derive the following upper bound

$$\begin{aligned} & \mathbb{E}_{\mu_\pi} \left[ \frac{1}{m} \sum_{r=1}^m |\mathbb{1}(\theta_{k,r}^\top \psi(x) > 0) - \mathbb{1}(\theta_{0,r}^\top \psi(x) > 0)| \right] \\ & \leq \mathbb{E}_{\mu_\pi} \left[ \frac{1}{m} \sum_{r=1}^m \mathbb{1}(|\theta_{0,r}^\top \psi(x)| \leq \|\theta_{k,r} - \theta_{0,r}\|_2) \right] \end{aligned} \quad (23)$$

$$\begin{aligned} & = \frac{1}{m} \sum_{r=1}^m \mathbf{P}_{\mu_\pi}(|\theta_{0,r}^\top \psi(x)| \leq \|\theta_{k,r} - \theta_{0,r}\|_2) \\ & \stackrel{(i)}{\leq} \frac{C_0}{m} \sum_{r=1}^m \frac{\|\theta_{k,r} - \theta_{0,r}\|_2}{\|\theta_{0,r}\|_2} \\ & \leq \frac{C_0}{m} \left( \sum_{r=1}^m \|\theta_{k,r} - \theta_{0,r}\|_2^2 \right)^{1/2} \left( \sum_{r=1}^m \frac{1}{\|\theta_{0,r}\|_2^2} \right)^{1/2} \\ & \stackrel{(ii)}{\leq} \frac{C_0 R}{d_1 \sqrt{m}}. \end{aligned} \quad (24)$$

where (i) follows from Assumption 1 and (ii) follows from the fact that  $\|\theta_{0,r}\|_2 \geq d_1$ .  $\square$

**Lemma 12.** *Suppose Assumption 1 holds. For any policy  $\pi$  and all  $k \geq 0$ , it holds that*

$$\mathbb{E}_{\mu_\pi} \left[ |f((s, a); \theta_k) - f_0((s, a); \theta_k)|^2 \right] \leq \frac{4C_0 R^3}{d_1 \sqrt{m}}$$

*Proof.* By definition, we have

$$\begin{aligned}
& |f((s, a); \theta_t) - f_0((s, a); \theta_t)| \\
&= \frac{1}{\sqrt{m}} \left| \sum_{r=1}^m (\mathbb{1}(\theta_{k,r}^\top \psi(x) > 0) - \mathbb{1}(\theta_{0,r}^\top \psi(x) > 0)) b_r \theta_{k,r}^\top \psi(x) \right| \\
&\leq \frac{1}{\sqrt{m}} \sum_{r=1}^m |(\mathbb{1}(\theta_{k,r}^\top \psi(x) > 0) - \mathbb{1}(\theta_{0,r}^\top \psi(x) > 0))| |b_r| \|\theta_{k,r}^\top \psi(x)\|_2 \\
&\stackrel{(i)}{\leq} \frac{1}{\sqrt{m}} \sum_{r=1}^m \mathbb{1}(|\theta_{0,r}^\top \psi(x)| \leq \|\theta_{k,r} - \theta_{0,r}\|_2) \|\theta_{k,r}^\top \psi(x)\|_2 \\
&\leq \frac{1}{\sqrt{m}} \sum_{r=1}^m \mathbb{1}(|\theta_{0,r}^\top \psi(x)| \leq \|\theta_{k,r} - \theta_{0,r}\|_2) (\|\theta_{0,r} - \theta_{k,r}\|_2 + \|\theta_{0,r}^\top \psi(x)\|_2) \\
&\leq \frac{2}{\sqrt{m}} \sum_{r=1}^m \mathbb{1}(|\theta_{0,r}^\top \psi(x)| \leq \|\theta_{k,r} - \theta_{0,r}\|_2) \|\theta_{0,r} - \theta_{k,r}\|_2. \tag{25}
\end{aligned}$$

where (i) follows from eq. (22). We can then obtain the following upper bound.

$$\begin{aligned}
& \mathbb{E}_{\mu_\pi} [|f((s, a); \theta_t) - f_0((s, a); \theta_t)|^2] \\
&\leq \frac{4}{m} \mathbb{E}_{\mu_\pi} \left[ \left( \sum_{r=1}^m \mathbb{1}(|\theta_{0,r}^\top \psi(x)| \leq \|\theta_{k,r} - \theta_{0,r}\|_2) \|\theta_{0,r} - \theta_{k,r}\|_2 \right)^2 \right] \\
&\stackrel{(i)}{\leq} \frac{4}{m} \mathbb{E}_{\mu_\pi} \left[ \sum_{r=1}^m \mathbb{1}(|\theta_{0,r}^\top \psi(x)| \leq \|\theta_{k,r} - \theta_{0,r}\|_2) \sum_{r=1}^m \|\theta_{0,r} - \theta_{k,r}\|_2^2 \right] \\
&= \frac{4R^2}{m} \mathbb{E}_{\mu_\pi} \left[ \sum_{r=1}^m \mathbb{1}(|\theta_{0,r}^\top \psi(x)| \leq \|\theta_{k,r} - \theta_{0,r}\|_2) \right] \\
&\stackrel{(ii)}{\leq} \frac{4C_0 R^3}{d_1 \sqrt{m}}. \tag{26}
\end{aligned}$$

where in (i) we apply Holder's inequality, and (ii) follows from the derivation in Lemma 11 after eq. (23).  $\square$

**Lemma 13.** Suppose Assumption 1 holds. For any policy  $\pi$  and all  $k \geq 0$ , with probability at least  $1 - \delta$ , we have

$$\|\bar{g}_k(\theta_k) - \bar{g}_0(\theta_k)\|_2 \leq \Theta \left( \frac{\sqrt{\log(\frac{1}{\delta})}}{(1-\gamma)m^{1/4}} \right).$$

*Proof.* By definition, we have

$$\begin{aligned}
& \|\bar{g}_k(\theta_k) - \bar{g}_0(\theta_k)\|_2 \\
&= \|\mathbb{E}_{\mu_\pi} [\delta_k(x, x'.\theta_k) \nabla_\theta f(x, \theta_k)] - \mathbb{E}_{\mu_\pi} [\delta_0(x, x'.\theta_k) \nabla_\theta f_0(x, \theta_k)]\|_2 \\
&= \|\mathbb{E}_{\mu_\pi} [(\delta_k(x, x'.\theta_k) - \delta_0(x, x'.\theta_k)) \nabla_\theta f(x, \theta_k) + \delta_0(x, x'.\theta_k) (\nabla_\theta f(x, \theta_k) - \nabla_\theta f_0(x, \theta_k))]\|_2 \\
&\leq \mathbb{E}_{\mu_\pi} [|\delta_k(x, x'.\theta_k) - \delta_0(x, x'.\theta_k)| \|\nabla_\theta f(x, \theta_k)\|_2] + |\delta_0(x, x'.\theta_k)| \|\nabla_\theta f(x, \theta_k) - \nabla_\theta f_0(x, \theta_k)\|_2 \\
&\stackrel{(i)}{\leq} \mathbb{E}_{\mu_\pi} [|\delta_k(x, x'.\theta_k) - \delta_0(x, x'.\theta_k)|] + \mathbb{E}_{\mu_\pi} [|\delta_0(x, x'.\theta_k)| \|\nabla_\theta f(x, \theta_k) - \nabla_\theta f_0(x, \theta_k)\|_2], \tag{27}
\end{aligned}$$

where (i) follows from the fact that  $\|\nabla_\theta f(x, \theta_k)\|_2 \leq 1$ . eq. (27) implies that

$$\begin{aligned}
& \|\bar{g}_k(\theta_k) - \bar{g}_0(\theta_k)\|_2^2 \\
&\leq 2\mathbb{E}_{\mu_\pi} [|\delta_k(x, x'.\theta_k) - \delta_0(x, x'.\theta_k)|^2] + 2(\mathbb{E}_{\mu_\pi} [|\delta_0(x, x'.\theta_k)| \|\nabla_\theta f(x, \theta_k) - \nabla_\theta f_0(x, \theta_k)\|_2])^2
\end{aligned}$$

$$\leq 2\mathbb{E}_{\mu_\pi} [|\delta_k(x, x'.\theta_k) - \delta_0(x, x'.\theta_k)|^2] + 2\mathbb{E}_{\mu_\pi} [|\delta_0(x, x'.\theta_k)|^2] \mathbb{E}_{\mu_\pi} [\|\nabla_\theta f(x, \theta_k) - \nabla_\theta f_0(x, \theta_k)\|_2^2]. \quad (28)$$

We first upper bound the term  $\mathbb{E}_{\mu_\pi} [|\delta_k(x, x'.\theta_k) - \delta_0(x, x'.\theta_k)|^2]$ . By definition, we have

$$\begin{aligned} & |\delta_k(x, x'.\theta_k) - \delta_0(x, x'.\theta_k)| \\ &= |f(x, \theta_k) - f_0(x, \theta_k) - \gamma(f(x', \theta_k) - f_0(x', \theta_k))| \\ &\leq |f(x, \theta_k) - f_0(x, \theta_k)| + |f(x', \theta_k) - f_0(x', \theta_k)|, \end{aligned}$$

which implies

$$\begin{aligned} & \mathbb{E}_{\mu_\pi} [|\delta_k(x, x'.\theta_k) - \delta_0(x, x'.\theta_k)|^2] \\ &\leq 2\mathbb{E}_{\mu_\pi} [|f(x, \theta_k) - f_0(x, \theta_k)|^2] + 2\mathbb{E}_{\mu_\pi} [|f(x', \theta_k) - f_0(x', \theta_k)|^2] \\ &= 4\mathbb{E}_{\mu_\pi} [|f(x, \theta_k) - f_0(x, \theta_k)|^2] \\ &\stackrel{(i)}{\leq} \frac{16C_0R^2}{d_1\sqrt{m}}. \end{aligned} \quad (29)$$

We then proceed to bound the term  $\mathbb{E}_{\mu_\pi} [\|\nabla_\theta f(x, \theta_k) - \nabla_\theta f_0(x, \theta_k)\|_2^2]$ . By definition, we have

$$\begin{aligned} & \|\nabla_\theta f(x, \theta_k) - \nabla_\theta f_0(x, \theta_k)\|_2 \\ &= \frac{1}{\sqrt{m}} \left\| \sum_{r=1}^m [\mathbb{1}(\theta_{k,r}^\top \psi(x) > 0) - \mathbb{1}(\theta_{0,r}^\top \psi(x) > 0)] b_r \theta_{0,r}^\top \psi(x) \right\|_2 \\ &\stackrel{(i)}{\leq} \frac{1}{\sqrt{m}} \sum_{r=1}^m |\mathbb{1}(\theta_{k,r}^\top \psi(x) > 0) - \mathbb{1}(\theta_{0,r}^\top \psi(x) > 0)| \|\theta_{0,r}\|_2 \\ &\stackrel{(ii)}{\leq} \frac{1}{\sqrt{m}} \sum_{r=1}^m \mathbb{1}(|\theta_{0,r}^\top \psi(x)| \leq \|\theta_{k,r} - \theta_{0,r}\|_2) \|\theta_{0,r}\|_2, \end{aligned} \quad (30)$$

where (i) follows from  $|b_r| \leq 1$  and  $\|\psi(s)\|_2 \leq 1$ , and (ii) follows from eq. (22). eq. (30) implies that

$$\begin{aligned} & \mathbb{E}_{\mu_\pi} [\|\nabla_\theta f(x, \theta_k) - \nabla_\theta f_0(x, \theta_k)\|_2^2] \\ &\leq \frac{1}{m} \mathbb{E}_{\mu_\pi} \left[ \left( \sum_{r=1}^m \mathbb{1}(|\theta_{0,r}^\top \psi(x)| \leq \|\theta_{k,r} - \theta_{0,r}\|_2) \right) \left( \sum_{r=1}^m \|\theta_{0,r}\|_2^2 \right) \right] \\ &\leq \frac{R^2}{m} \sum_{r=1}^m \mathbb{1}(|\theta_{0,r}^\top \psi(x)| \leq \|\theta_{k,r} - \theta_{0,r}\|_2) \\ &\stackrel{(i)}{\leq} \frac{C_0R^3}{d_1\sqrt{m}}, \end{aligned} \quad (31)$$

where (i) follows from the derivation in Lemma 11 after eq. (23).

Finally, we consider the upper bound of  $\mathbb{E}_{\mu_\pi} [|\delta_0(x, x'.\theta_k)|^2]$ . We proceed as follows.

$$\begin{aligned} & \mathbb{E}_{\mu_\pi} [|\delta_0(x, x'.\theta_k)|^2] \\ &\leq \mathbb{E}_{\mu_\pi} [|f_0(x, \theta_k) - r(x, x') - \gamma f_0(x', \theta_k)|^2] \\ &\leq 3\mathbb{E}_{\mu_\pi} [|f_0(x, \theta_k)|^2] + 3\mathbb{E}_{\mu_\pi} [r^2(x, x')] + 3\gamma^2 \mathbb{E}_{\mu_\pi} [|f_0(x', \theta_k)|^2] \\ &\leq 6\mathbb{E}_{\mu_\pi} [|f_0(x, \theta_k)|^2] + 3c_{\max}^2 \\ &= 6\mathbb{E}_{\mu_\pi} [|f_0(x, \theta_k) - f_0(x, \theta_\pi^*) + f_0(x, \theta_\pi^*) - Q_\pi(x) + Q_\pi(x)|^2] + 3c_{\max}^2 \\ &= 18\mathbb{E}_{\mu_\pi} [|f_0(x, \theta_k) - f_0(x, \theta_\pi^*)|^2] + 18\mathbb{E}_{\mu_\pi} [|f_0(x, \theta_\pi^*) - Q_\pi(x)|^2] + 18\mathbb{E}_{\mu_\pi} [|Q_\pi(x)|^2] + 3c_{\max}^2 \\ &\stackrel{(i)}{\leq} 18R^2 + \frac{21c_{\max}^2}{(1-\gamma)^2} + 18\mathbb{E}_{\mu_\pi} [|f_0(x, \theta_\pi^*) - Q_\pi(x)|^2]. \end{aligned} \quad (32)$$

Since  $\bar{\mathcal{F}}_{0,m} \subset \mathcal{F}_{0,m}$ . Lemma 10 implies that with probability at least  $1 - \delta$ , we have

$$\mathbb{E}_{\mu_\pi} [|f_0(x, \theta_\pi^*) - Q_\pi(x)|^2] \leq \frac{4R^2 \log(\frac{1}{\delta})}{m} \leq 4R^2 \log\left(\frac{1}{\delta}\right). \quad (33)$$

Thus, with probability at least  $1 - \delta$ , we have

$$\mathbb{E}_{\mu_\pi} [|\delta_0(x, x', \theta_k)|^2] \leq 18R^2 + \frac{21c_{\max}^2}{(1-\gamma)^2} + 72R^2 \log\left(\frac{1}{\delta}\right). \quad (34)$$

Combining eq. (29), eq. (31) and eq. (34), we can obtain that, with probability at least  $1 - \delta$ , we have

$$\|\bar{g}_k(\theta_k) - \bar{g}_0(\theta_k)\|_2^2 \leq \Theta\left(\frac{\log(\frac{1}{\delta})}{(1-\gamma)^2 \sqrt{m}}\right),$$

which implies that with probability at least  $1 - \delta$ , we have

$$\|\bar{g}_k(\theta_k) - \bar{g}_0(\theta_k)\|_2 \leq \Theta\left(\frac{\sqrt{\log(\frac{1}{\delta})}}{(1-\gamma)m^{1/4}}\right),$$

which complete the proof.  $\square$

## C.2 PROOF OF LEMMA 2

We consider the convergence of  $\theta_k^i$  for a given  $i$  under a fixed policy  $\pi$ . For the iteration of  $\theta_k$ , we proceed as follows.

$$\begin{aligned} & \|\theta_{k+1} - \theta^*\|_2^2 \\ &= \|\Pi_{\mathbf{B}}(\theta_k - \beta g_k(\theta_k)) - \Pi_{\mathbf{B}}(\theta^* - \beta \bar{g}_0(\theta^*))\|_2^2 \\ &\leq \|(\theta_k - \theta^*) - \beta(g_k(\theta_k) - \bar{g}_0(\theta^*))\|_2^2 \\ &= \|\theta_k - \theta^*\|_2^2 - 2\beta(g_k(\theta_k) - \bar{g}_0(\theta^*))^\top(\theta_k - \theta^*) + \beta^2 \|g_k(\theta_k) - \bar{g}_0(\theta^*)\|_2^2 \\ &= \|\theta_k - \theta^*\|_2^2 - 2\beta(\bar{g}_0(\theta_k) - \bar{g}_0(\theta^*))^\top(\theta_k - \theta^*) + 2\beta(\bar{g}_k(\theta_k) - g_k(\theta_k))^\top(\theta_k - \theta^*) \\ &\quad + 2\beta(\bar{g}_0(\theta_k) - \bar{g}_k(\theta_k))^\top(\theta_k - \theta^*) + \beta^2 \|g_k(\theta_k) - \bar{g}_0(\theta^*)\|_2^2 \\ &\leq \|\theta_k - \theta^*\|_2^2 - 2\beta(\bar{g}_0(\theta_k) - \bar{g}_0(\theta^*))^\top(\theta_k - \theta^*) + 2\beta(\bar{g}_k(\theta_k) - g_k(\theta_k))^\top(\theta_k - \theta^*) \\ &\quad + 2\beta(\bar{g}_0(\theta_k) - \bar{g}_k(\theta_k))^\top(\theta_k - \theta^*) + 3\beta^2 \|g_k(\theta_k) - \bar{g}_k(\theta_k)\|_2^2 + 3\beta^2 \|\bar{g}_k(\theta_k) - \bar{g}_0(\theta_k)\|_2^2 \\ &\quad + 3\beta^2 \|\bar{g}_0(\theta_k) - \bar{g}_0(\theta^*)\|_2^2 \\ &\stackrel{(i)}{\leq} \|\theta_k - \theta^*\|_2^2 - 2(1-\gamma)\beta \mathbb{E}_{\mu_\pi} [(f_0((s, a); \theta_k) - f_0((s, a); \theta^*))^2] \\ &\quad + 2\beta(\bar{g}_k(\theta_k) - g_k(\theta_k))^\top(\theta_k - \theta^*) + 4R\beta \|\bar{g}_0(\theta_k) - \bar{g}_k(\theta_k)\|_2 + 3\beta^2 \|g_k(\theta_k) - \bar{g}_k(\theta_k)\|_2^2 \\ &\quad + 3\beta^2 \|\bar{g}_k(\theta_k) - \bar{g}_0(\theta_k)\|_2^2 + 3\beta^2 \|\bar{g}_0(\theta_k) - \bar{g}_0(\theta^*)\|_2^2 \\ &\stackrel{(ii)}{\leq} \|\theta_k - \theta^*\|_2^2 - [2\beta(1-\gamma) - 12\beta^2] \mathbb{E}_{\mu_\pi} [(f_0((s, a); \theta_k) - f_0((s, a); \theta^*))^2] \\ &\quad + 2\beta(\bar{g}_k(\theta_k) - g_k(\theta_k))^\top(\theta_k - \theta^*) + 4R\beta \|\bar{g}_0(\theta_k) - \bar{g}_k(\theta_k)\|_2 + 3\beta^2 \|g_k(\theta_k) - \bar{g}_k(\theta_k)\|_2^2 \\ &\quad + 3\beta^2 \|\bar{g}_k(\theta_k) - \bar{g}_0(\theta_k)\|_2^2, \end{aligned} \quad (35)$$

where (i) follows from the fact that

$$\begin{aligned} & (\bar{g}_0(\theta_k) - \bar{g}_0(\theta^*))^\top(\theta_k - \theta^*) \\ & \geq (1-\gamma) \mathbb{E}_{\mu_\pi} [(f_0((s, a); \theta_k) - f_0((s, a); \theta^*))^2] - R \|\bar{g}_k(\theta_k) - \bar{g}_0(\theta_k)\|_2, \end{aligned}$$

and (ii) follows from the fact that

$$\|\bar{g}_0(\theta_k) - \bar{g}_0(\theta^*)\|_2^2 \leq 4\mathbb{E}_{\mu_\pi} [(f_0((s, a); \theta_k) - f_0((s, a); \theta^*))^2].$$

Rearrange eq. (35) yields

$$\begin{aligned} & [2\beta(1-\gamma) - 12\beta^2] \mathbb{E}_{\mu_\pi} [(f_0((s, a); \theta_k) - f_0((s, a); \theta^*))^2] \\ & \leq \|\theta_k - \theta^*\|_2^2 - \|\theta_{k+1} - \theta^*\|_2^2 + 2\beta(\bar{g}_k(\theta_k) - g_k(\theta_k))^\top (\theta_k - \theta^*) + 4R\beta \|\bar{g}_0(\theta_k) - \bar{g}_k(\theta_k)\|_2 \\ & \quad + 3\beta^2 \|g_k(\theta_k) - \bar{g}_k(\theta_k)\|_2^2 + 3\beta^2 \|\bar{g}_k(\theta_k) - \bar{g}_0(\theta_k)\|_2^2. \end{aligned} \quad (36)$$

Summing eq. (36) over  $t = 0$  to  $K - 1$  yields

$$\begin{aligned} & [2\beta(1-\gamma) - 12\beta^2] \sum_{t=0}^{K-1} \mathbb{E}_{\mu_\pi} [(f_0((s, a); \theta_k) - f_0((s, a); \theta^*))^2] \\ & \leq \|\theta_0 - \theta^*\|_2^2 - \|\theta_K - \theta^*\|_2^2 + 2\beta \sum_{t=0}^{K-1} (\bar{g}_k(\theta_k) - g_k(\theta_k))^\top (\theta_k - \theta^*) + 4R\beta \sum_{t=0}^{K-1} \|\bar{g}_0(\theta_k) - \bar{g}_k(\theta_k)\|_2 \\ & \quad + 3\beta^2 \sum_{t=0}^{K-1} \|g_k(\theta_k) - \bar{g}_k(\theta_k)\|_2^2 + 3\beta^2 \sum_{t=0}^{K-1} \|\bar{g}_k(\theta_k) - \bar{g}_0(\theta_k)\|_2^2 \\ & \stackrel{(i)}{\leq} R^2 + 2\beta \sum_{t=0}^{K-1} \zeta_k(\theta_k)^\top (\theta_k - \theta^*) + 3\beta^2 \sum_{t=0}^{K-1} \|\zeta_k(\theta_k)\|_2^2 + 4R\beta \sum_{t=0}^{K-1} \|\xi_k(\theta_k)\|_2 + 3\beta^2 \sum_{t=0}^{K-1} \|\xi_k(\theta_k)\|_2^2, \end{aligned}$$

where in (i) we define  $\zeta_k(\theta_k) = \bar{g}_k(\theta_k) - g_k(\theta_k)$  and  $\xi_k(\theta_k) = \bar{g}_k(\theta_k) - \bar{g}_0(\theta_k)$ .

We first consider the term  $\sum_{t=0}^{K-1} \|\zeta_k(\theta_k)\|_2^2$ . We proceed as follows.

$$\begin{aligned} & P_{\mu_\pi} \left( \sum_{t=0}^{K-1} \|\zeta_k(\theta_k)\|_2^2 \geq (1 + \Lambda) C_\zeta^2 K \right) \\ & = P_{\mu_\pi} \left( \frac{\sum_{t=0}^{K-1} \|\zeta_k(\theta_k)\|_2^2}{C_\zeta^2 K} \geq 1 + \Lambda \right) \\ & = P_{\mu_\pi} \left( \exp \left( \frac{\sum_{t=0}^{K-1} \|\zeta_k(\theta_k)\|_2^2}{C_\zeta^2 K} \right) \geq \exp(1 + \Lambda) \right) \\ & \leq P_{\mu_\pi} \left( \frac{1}{K} \sum_{t=0}^{K-1} \exp \left( \frac{\|\zeta_k(\theta_k)\|_2^2}{C_\zeta^2} \right) \geq \exp(1 + \Lambda) \right) \\ & \stackrel{(i)}{\leq} \frac{1}{K} \sum_{t=0}^{K-1} \mathbb{E}_{\mu_\pi} \left[ \exp \left( \frac{\|\zeta_k(\theta_k)\|_2^2}{C_\zeta^2} \right) \right] / \exp(1 + \Lambda) \\ & \stackrel{(ii)}{\leq} \exp(-\Lambda), \end{aligned} \quad (37)$$

where (i) follows from Markov's inequality, (ii) follows from Assumption 4. eq. (37) implies that with probability at least  $1 - \delta_1$ , we have

$$\sum_{t=0}^{K-1} \|\zeta_k(\theta_k)\|_2^2 \leq \left( 1 + \log \left( \frac{1}{\delta_1} \right) \right) C_\zeta^2 K \leq 2 \log \left( \frac{1}{\delta_1} \right) C_\zeta^2 K. \quad (38)$$

We then consider the term  $\sum_{t=0}^{K-1} \zeta_k(\theta_k)^\top (\theta_k - \theta^*)$ . Note that for any  $0 \leq k \leq K - 1$ , we have

$$|\zeta_k(\theta_k)^\top (\theta_k - \theta^*)|^2 \leq \|\zeta_k(\theta_k)\|_2^2 \|\theta_k - \theta^*\|_2^2 \leq R^2 \|\zeta_k(\theta_k)\|_2^2,$$

which implies

$$\mathbb{E}_{\mu_\pi} \left[ \exp \left( \frac{|\zeta_k(\theta_k)^\top (\theta_k - \theta^*)|^2}{B^2 C_\zeta^2} \right) \right] \leq \mathbb{E}_{\mu_\pi} \left[ \exp \left( \frac{\|\zeta_k(\theta_k)\|_2^2}{C_\zeta^2} \right) \right] \leq \exp(1).$$

Applying Bernstein's inequality for martingale (Ghadimi & Lan, 2013, Lemma 2.3), we can obtain

$$P_{\mu_\pi} \left( \left| \sum_{t=0}^{K-1} \zeta_k(\theta_k)^\top (\theta_k - \theta^*) \right| \geq \sqrt{2}(1 + \Lambda) C_\zeta \sqrt{K} \right) \leq \exp(-\Lambda^2/3),$$

which implies with probability at least  $1 - \delta_2$ , we have

$$\left| \sum_{t=0}^{K-1} \zeta_k(\theta_k)^\top (\theta_k - \theta^*) \right| \leq \sqrt{2} \left( 1 + \sqrt{3 \log \left( \frac{1}{\delta_2} \right)} \right) C_\zeta \sqrt{K} \leq 5C_\zeta \sqrt{\log \left( \frac{1}{\delta_2} \right)} \sqrt{K}. \quad (39)$$

We then consider the term  $\sum_{t=0}^{K-1} \|\xi_k(\theta_k)\|_2$  and  $\sum_{t=0}^{K-1} \|\xi_k(\theta_k)\|_2^2$ . Lemma 13 implies that with probability at least  $1 - \delta_3/K$ , we have

$$\|\xi_k(\theta_k)\|_2 \leq \Theta \left( \frac{\sqrt{\log(\frac{K}{\delta_3})}}{(1-\gamma)m^{1/4}} \right).$$

Applying union bound we can obtain that with probability at least  $1 - \delta_3$ , we have

$$\sum_{t=0}^{K-1} \|\xi_k(\theta_k)\|_2 \leq \Theta \left( \frac{K \sqrt{\log(\frac{K}{\delta_3})}}{(1-\gamma)m^{1/4}} \right). \quad (40)$$

Similarly, we can obtain that with probability at least  $1 - \delta_3$ , we have

$$\sum_{t=0}^{K-1} \|\xi_k(\theta_k)\|_2^2 \leq \Theta \left( \frac{K \log(\frac{K}{\delta_3})}{(1-\gamma)^2 m^{1/2}} \right). \quad (41)$$

Combining eq. (38), eq. (39), eq. (40) and eq. (41) and applying union bound, we can obtain that with probability at least  $1 - (\delta_1 + \delta_2 + \delta_3 + \delta_4)$ , we have

$$\begin{aligned} & [2\beta(1-\gamma) - 12\beta^2] \sum_{t=0}^{K-1} \mathbb{E}_{\mu_\pi} [(f_0((s, a); \theta_k) - f_0((s, a); \theta^*))^2] \\ & \leq R^2 + 10\beta C_\zeta \sqrt{\log \left( \frac{1}{\delta_2} \right)} \sqrt{K} + 6\beta^2 \log \left( \frac{1}{\delta_1} \right) C_\zeta^2 K + \beta K \Theta \left( \frac{\sqrt{\log(\frac{K}{\delta_3})}}{(1-\gamma)m^{1/4}} \right) \\ & \quad + \beta^2 K \Theta \left( \frac{\log(\frac{K}{\delta_3})}{(1-\gamma)^2 m^{1/2}} \right). \end{aligned} \quad (42)$$

Divide both sides of eq. (42) by  $[2\beta(1-\gamma) - 12\beta^2]K$ . Recalling that the stepsize  $\beta = \min\{1/\sqrt{K}, (1-\gamma)/12\}$ , which implies that  $\frac{1}{\sqrt{K}[2\beta(1-\gamma) - 12\beta^2]} \leq \frac{12}{(1-\gamma)^2}$ . Then, with probability at least  $1 - (\delta_1 + \delta_2 + \delta_3 + \delta_4)$ , we have

$$\begin{aligned} & \|f_0((s, a); \bar{\theta}_K) - f_0((s, a); \theta^*)\|_{\mu_\pi}^2 \leq \frac{1}{K} \sum_{t=0}^{K-1} \mathbb{E}_{\mu_\pi} [(f_0((s, a); \theta_k) - f_0((s, a); \theta^*))^2] \\ & \leq \frac{R^2}{[2\beta(1-\gamma) - 12\beta^2]K} + \frac{10\beta C_\zeta \sqrt{\log \left( \frac{1}{\delta_2} \right)}}{[2\beta(1-\gamma) - 12\beta^2]\sqrt{K}} + \frac{6\beta \log \left( \frac{1}{\delta_1} \right) C_\zeta^2}{[2\beta(1-\gamma) - 12\beta^2]\sqrt{K}} \\ & \quad + \Theta \left( \frac{\sqrt{\log(\frac{K}{\delta_3})}}{(1-\gamma)m^{1/4}} \right) \frac{1}{[2\beta(1-\gamma) - 12\beta^2]\sqrt{K}} \\ & \quad + \Theta \left( \frac{\log(\frac{K}{\delta_3})}{(1-\gamma)^2 m^{1/2}} \right) \frac{1}{[2\beta(1-\gamma) - 12\beta^2]\sqrt{K}} \\ & \leq \Theta \left( \frac{1}{(1-\gamma)^2 \sqrt{K}} \right) + \Theta \left( \frac{1}{(1-\gamma)^2 \sqrt{K}} \sqrt{\log \left( \frac{1}{\delta_1} \right)} \right) + \Theta \left( \frac{1}{(1-\gamma)^2 \sqrt{K}} \sqrt{\log \left( \frac{1}{\delta_2} \right)} \right) \\ & \quad + \Theta \left( \frac{\sqrt{\log(\frac{K}{\delta_3})}}{(1-\gamma)^3 m^{1/4}} \right) + \Theta \left( \frac{\sqrt{\log(\frac{K}{\delta_4})}}{(1-\gamma)^3 m^{1/4}} \right) \end{aligned}$$

$$\begin{aligned}
&= \Theta \left( \frac{1}{(1-\gamma)^2 \sqrt{K}} \left( \sqrt{\log \left( \frac{1}{\delta_1} \right)} + \sqrt{\log \left( \frac{1}{\delta_1} \right)} \right) \right) \\
&\quad + \Theta \left( \frac{1}{(1-\gamma)^3 m^{1/4}} \left( \sqrt{\log \left( \frac{K}{\delta_3} \right)} + \sqrt{\log \left( \frac{K}{\delta_4} \right)} \right) \right). \tag{43}
\end{aligned}$$

Finally, we consider the upper bound of  $\|f((s, a); \bar{\theta}_K) - Q_\pi(s, a)\|_{\mu_\pi}^2$ . We proceed as follows

$$\begin{aligned}
&\|f((s, a); \bar{\theta}_K) - Q_\pi(s, a)\|_{\mu_\pi}^2 \\
&\leq 3 \|f((s, a); \bar{\theta}_K) - f_0((s, a); \bar{\theta}_K)\|_{\mu_\pi}^2 + 3 \|f_0((s, a); \bar{\theta}_K) - f_0((s, a); \theta^*)\|_{\mu_\pi}^2 \\
&\quad + 3 \|f_0((s, a); \theta^*) - Q_\pi(s, a)\|_{\mu_\pi}^2 \\
&\stackrel{(i)}{\leq} \Theta \left( \frac{1}{\sqrt{m}} \right) + 3 \|f_0((s, a); \bar{\theta}_K) - f_0((s, a); \theta^*)\|_{\mu_\pi}^2 + \frac{3}{1-\gamma} \|f_0((s, a); \theta^*) - Q_\pi(s, a)\|_{\mu_\pi}^2, \tag{44}
\end{aligned}$$

where (i) follows from Lemma 12 and the fact that

$$\|f_0((s, a); \theta^*) - Q_\pi(s, a)\|_{\mu_\pi}^2 \leq \frac{1}{1-\gamma} \|f_0((s, a); \theta^*) - Q_\pi(s, a)\|_{\mu_\pi}^2,$$

which is given in Cai et al. (2019). eq. (33) implies that, with probability at least  $\delta_5$ , we have

$$\|f_0((s, a); \theta^*) - Q_\pi(s, a)\|_{\mu_\pi}^2 \leq \frac{4R^2 \log \left( \frac{1}{\delta_5} \right)}{m}. \tag{45}$$

Substituting eq. (43) and eq. (45) into eq. (44), we have with probability at least  $1 - (\delta_1 + \delta_2 + \delta_3 + \delta_4 + \delta_5)$ , the following holds:

$$\begin{aligned}
&\|f((s, a); \bar{\theta}_K) - Q_\pi(s, a)\|_{\mu_\pi}^2 \\
&\leq \Theta \left( \frac{1}{(1-\gamma)^2 \sqrt{K}} \left( \sqrt{\log \left( \frac{1}{\delta_1} \right)} + \sqrt{\log \left( \frac{1}{\delta_1} \right)} \right) \right) \\
&\quad + \Theta \left( \frac{1}{(1-\gamma)^3 m^{1/4}} \left( \sqrt{\log \left( \frac{K}{\delta_3} \right)} + \sqrt{\log \left( \frac{K}{\delta_4} \right)} \right) \right) \\
&\quad + \Theta \left( \frac{1}{(1-\gamma)m} \log \left( \frac{1}{\delta_5} \right) \right).
\end{aligned}$$

Letting  $\delta_1 = \delta_2 = \delta_3 = \delta_4 = \delta_5 = \frac{\delta}{5}$ , we have with probability at least  $1 - \delta$ , the following holds:

$$\begin{aligned}
&\|f((s, a); \bar{\theta}_K) - Q_\pi(s, a)\|_{\mu_\pi}^2 \\
&\leq \Theta \left( \frac{1}{(1-\gamma)^2 \sqrt{K}} \sqrt{\log \left( \frac{1}{\delta} \right)} \right) + \Theta \left( \frac{1}{(1-\gamma)^3 m^{1/4}} \sqrt{\log \left( \frac{K}{\delta} \right)} \right),
\end{aligned}$$

which complete the proof.

### C.3 SUPPORTING LEMMAS FOR THEOREM 2

For the two-layer neural network defined in eq. (6), we have the following property:  $\tau \cdot f(x, W) = f(x, \tau W)$ . Thus, in the sequel, we write  $\pi_W^\tau(a|s) = \pi_{\tau W}(a|s)$ . In the technical proof, we consider the following policy class:

$$\pi_W(a|s) := \frac{\exp(f((s, a); W))}{\sum_{a' \in \mathcal{A}} \exp(f((s, a'); W))}, \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}, \tag{46}$$

and  $J_i(W)$  as the accumulated cost with policy  $\pi_W$ . We denote  $\phi_W^i(s, a) = \nabla_W f_i((s, a), W)$ . We define the diameter of  $\mathcal{B}_W$  as  $R_W$ . When performing each NPG update, we will need to solve the linear regression problem specified in eq. (11). As shown in Wang et al. (2019), when the neural network for the policy parametrization and value function approximation share the same initialization,  $\bar{\theta}_t$  is an approximated solution of the problem eq. (11). Thus, instead of solving the problem eq. (11) directly, here we simply use  $\theta_t$  as the approximated NPG update at each iteration:

$$\tau_{t+1} \cdot W_{t+1} = \tau_t \cdot W_t + \frac{\alpha}{1-\gamma} \bar{\theta}_t.$$

**Lemma 14.** For any  $\theta, \theta' \in \mathcal{B}$  and  $\pi$ , we have

$$\|\phi_\theta(s, a)^\top \theta' - \phi_{\theta_0}(s, a)^\top \theta'\|_{\mu_\pi}^2 \leq \frac{4C_0 R^3}{d_1 \sqrt{m}}.$$

*Proof.* By definition, we have

$$\begin{aligned} & \phi_\theta(s, a)^\top \theta' - \phi_{\theta_0}(s, a)^\top \theta' \\ &= \frac{1}{\sqrt{m}} \left| \sum_{r=1}^m (\mathbb{1}(\theta_r^\top \psi(x) > 0) - \mathbb{1}(\theta_{0,r}^\top \psi(x) > 0)) b_r \theta_r'^\top \psi(x) \right| \\ &\leq \frac{1}{\sqrt{m}} \sum_{r=1}^m |(\mathbb{1}(\theta_r^\top \psi(x) > 0) - \mathbb{1}(\theta_{0,r}^\top \psi(x) > 0))| |b_r| \|\theta_r'^\top \psi(x)\|_2 \\ &\stackrel{(i)}{\leq} \frac{1}{\sqrt{m}} \sum_{r=1}^m \mathbb{1}(|\theta_{0,r}^\top \psi(x)| \leq \|\theta_r - \theta_{0,r}\|_2) \|\theta_r'^\top \psi(x)\|_2 \\ &\leq \frac{1}{\sqrt{m}} \sum_{r=1}^m \mathbb{1}(|\theta_{0,r}^\top \psi(x)| \leq \|\theta_r - \theta_{0,r}\|_2) \left( \|\theta_r'^\top \psi(x) - \theta_{0,r}^\top \psi(x)\|_2 + \|\theta_{0,r}^\top \psi(x)\|_2 \right) \\ &\leq \frac{1}{\sqrt{m}} \sum_{r=1}^m \mathbb{1}(|\theta_{0,r}^\top \psi(x)| \leq \|\theta_r - \theta_{0,r}\|_2) \left( \|\theta_r' - \theta_{0,r}\|_2 + \|\theta_{0,r}^\top \psi(x)\|_2 \right) \\ &\leq \frac{1}{\sqrt{m}} \sum_{r=1}^m \mathbb{1}(|\theta_{0,r}^\top \psi(x)| \leq \|\theta_r - \theta_{0,r}\|_2) (\|\theta_r' - \theta_{0,r}\|_2 + \|\theta_r - \theta_{0,r}\|_2), \end{aligned} \quad (47)$$

where (i) follows from eq. (22). Following from Holder's inequality, we obtain from eq. (47) that

$$\begin{aligned} & |\phi_\theta(s, a)^\top \theta' - \phi_{\theta_0}(s, a)^\top \theta'|^2 \\ &\leq \frac{1}{m} \left[ \sum_{r=1}^m \mathbb{1}(|\theta_{0,r}^\top \psi(x)| \leq \|\theta_r - \theta_{0,r}\|_2) \right] \left[ \sum_{r=1}^m (\|\theta_r' - \theta_{0,r}\|_2 + \|\theta_r - \theta_{0,r}\|_2)^2 \right] \\ &\leq \frac{2}{m} \left[ \sum_{r=1}^m \mathbb{1}(|\theta_{0,r}^\top \psi(x)| \leq \|\theta_r - \theta_{0,r}\|_2) \right] \left[ \sum_{r=1}^m \|\theta_r' - \theta_{0,r}\|_2^2 + \sum_{r=1}^m \|\theta_r - \theta_{0,r}\|_2^2 \right] \\ &\leq \frac{4R^2}{m} \sum_{r=1}^m \mathbb{1}(|\theta_{0,r}^\top \psi(x)| \leq \|\theta_r - \theta_{0,r}\|_2), \end{aligned}$$

which implies

$$\|\phi_\theta(s, a)^\top \theta' - \phi_{\theta_0}(s, a)^\top \theta'\|_{\mu_\pi}^2 = \mathbb{E}_{\mu_\pi} [|\phi_\theta(s, a)^\top \theta' - \phi_{\theta_0}(s, a)^\top \theta'|^2] \leq \frac{4C_0 R^3}{d_1 \sqrt{m}}, \quad (48)$$

where (i) follows from the derivation in Lemma 11 after eq. (23).  $\square$

**Lemma 15** (Global improvement upper bound for neural NPG). *Considering the approximated NPG updates in the neural network approximation setting, we have*

$$\begin{aligned} & \alpha(1-\gamma)(J_0(\pi^*) - J_0(\pi_{\tau_t W_t})) \\ &\leq \mathbb{E}_{\nu^*} [D_{KL}(\pi^* || \pi_{\tau_t W_t})] - \mathbb{E}_{\nu^*} [D_{KL}(\pi^* || \pi_{\tau_{t+1} W_{t+1}})] + \frac{8\alpha C_{RN} \sqrt{C_0} R^{1.5}}{\sqrt{d_1} m^{1/4}} + \alpha^2 L_f (R^2 + m d_2^2) \\ &\quad + 2\alpha C_{RN} \|f((s, a), \bar{\theta}_t) - Q_{\pi_{\tau_t W_t}}(s, a)\|_{\mu_{\pi_{\tau_t W_t}}}. \end{aligned}$$



*Proof.* It has been verified that the feature mapping  $\phi_W^r(s, a)$  is bounded Wang et al. (2019); Cai et al. (2019). By following similar argument in (Agarwal et al., 2019, Example 6.3), we can show that  $\log(\pi_w(a|s))$  is  $L_f$ -Lipschitz. Applying the Lipschitz property of  $\log(\pi_w(a|s))$ , we can obtain the following.

$$\begin{aligned}
& \mathbb{E}_{\nu^*} [D_{\text{KL}}(\pi^* || \pi_{\tau_t W_t})] - \mathbb{E}_{\nu^*} [D_{\text{KL}}(\pi^* || \pi_{\tau_{t+1} W_{t+1}})] \\
&= \mathbb{E}_{\nu^*} [\log(\pi_{\tau_{t+1} W_{t+1}}(a|s)) - \log(\pi_{\tau_t W_t}(a|s))] \\
&\stackrel{(i)}{\geq} \mathbb{E}_{\nu^*} [\nabla_W \log(\pi_{\tau_t W_t}(a|s))]^\top (\tau_{t+1} W_{t+1} - \tau_t W_t) - \frac{L_f}{2} \|\tau_{t+1} W_{t+1} - \tau_t W_t\|_2^2 \\
&= \alpha \mathbb{E}_{\nu^*} [\nabla_W \log(\pi_{\tau_t W_t}(a|s))]^\top \bar{\theta}_t - \frac{\alpha^2 L_f}{2} \|\bar{\theta}_t\|_2^2 \\
&= \alpha \mathbb{E}_{\nu^*} [\phi_{W_t}(s, a) - \mathbb{E}_{\pi_{\tau_t W_t}} [\phi_{W_t}(s, a')]]^\top \bar{\theta}_t - \frac{\alpha^2 L_f}{2} \|\bar{\theta}_t\|_2^2 \\
&= \alpha \mathbb{E}_{\nu^*} [Q_{\pi_{\tau_t W_t}}(s, a) - \mathbb{E}_{\pi_{\tau_t W_t}} [Q_{\pi_{\tau_t W_t}}(s, a')]] + \alpha \mathbb{E}_{\nu^*} [\phi_{W_t}(s, a)^\top \bar{\theta}_t - Q_{\pi_{\tau_t W_t}}(s, a)] \\
&\quad + \alpha \mathbb{E}_{\nu^*} \mathbb{E}_{\pi_{\tau_t W_t}} [Q_{\pi_{\tau_t W_t}}(s, a') - \phi_{W_t}(s, a')^\top \bar{\theta}_t] - \frac{\alpha^2 L_f}{2} \|\bar{\theta}_t\|_2^2 \\
&= \alpha(1 - \gamma)(J_0(\pi^*) - J_0(\pi_{\tau_t W_t})) + \alpha \mathbb{E}_{\nu^*} [\phi_{W_t}(s, a)^\top \bar{\theta}_t - f((s, a), \bar{\theta}_t)] \\
&\quad + \alpha \mathbb{E}_{\nu^*} [f((s, a), \bar{\theta}_t) - Q_{\pi_{\tau_t W_t}}(s, a)] + \alpha \mathbb{E}_{\nu^*} \mathbb{E}_{\pi_{\tau_t W_t}} [Q_{\pi_{\tau_t W_t}}(s, a') - f((s, a'), \bar{\theta}_t)] \\
&\quad + \alpha \mathbb{E}_{\nu^*} \mathbb{E}_{\pi_{\tau_t W_t}} [f((s, a'), \bar{\theta}_t) - \phi_{W_t}(s, a')^\top \bar{\theta}_t] - \frac{\alpha^2 L_f}{2} \|\bar{\theta}_t\|_2^2 \\
&= \alpha(1 - \gamma)(J_0(\pi^*) - J_0(\pi_{\tau_t W_t})) + \alpha \mathbb{E}_{\nu^*} [\phi_{W_t}(s, a)^\top \bar{\theta}_t - f((s, a), \bar{\theta}_t)] \\
&\quad + \alpha \mathbb{E}_{\nu^*} [f((s, a), \bar{\theta}_t) - Q_{\pi_{\tau_t W_t}}(s, a)] + \alpha \mathbb{E}_{\nu^*} \mathbb{E}_{\pi_{\tau_t W_t}} [Q_{\pi_{\tau_t W_t}}(s, a') - f((s, a'), \bar{\theta}_t)] \\
&\quad + \alpha \mathbb{E}_{\nu^*} \mathbb{E}_{\pi_{\tau_t W_t}} [f((s, a'), \bar{\theta}_t) - \phi_{W_t}(s, a')^\top \bar{\theta}_t] - \frac{\alpha^2 L_f}{2} \|\bar{\theta}_t\|_2^2 \\
&= \alpha(1 - \gamma)(J_0(\pi^*) - J_0(\pi_{\tau_t W_t})) + \alpha \mathbb{E}_{\nu^*} \mathbb{E}_{\bar{\theta}_t} [\phi_{W_t}(s, a)^\top \bar{\theta}_t - f((s, a), \bar{\theta}_t)] \\
&\quad + \alpha \mathbb{E}_{\nu^*} [f((s, a), \bar{\theta}_t) - Q_{\pi_{\tau_t W_t}}(s, a)] + \alpha \mathbb{E}_{\nu^*} \mathbb{E}_{\pi_{\tau_t W_t}} [Q_{\pi_{\tau_t W_t}}(s, a') - f((s, a'), \bar{\theta}_t)] \\
&\quad + \alpha \mathbb{E}_{\nu^*} \mathbb{E}_{\pi_{\tau_t W_t}} [f((s, a'), \bar{\theta}_t) - \phi_{W_t}(s, a')^\top \bar{\theta}_t] - \frac{\alpha^2 L_f}{2} \|\bar{\theta}_t\|_2^2 \\
&\geq \alpha(1 - \gamma)(J_0(\pi^*) - J_0(\pi_{\tau_t W_t})) - \alpha \sqrt{\mathbb{E}_{\nu^*} [(\phi_{W_t}(s, a)^\top \bar{\theta}_t - f((s, a), \bar{\theta}_t))^2]} \\
&\quad - \alpha \sqrt{\mathbb{E}_{\nu^*} [(f((s, a), \bar{\theta}_t) - Q_{\pi_{\tau_t W_t}}(s, a))^2]} - \alpha \sqrt{\mathbb{E}_{\nu^*} \mathbb{E}_{\pi_{\tau_t W_t}} [(Q_{\pi_{\tau_t W_t}}(s, a') - f((s, a'), \bar{\theta}_t))^2]} \\
&\quad - \alpha \sqrt{\mathbb{E}_{\nu^*} \mathbb{E}_{\pi_{\tau_t W_t}} [(f((s, a'), \bar{\theta}_t) - \phi_{W_t}(s, a')^\top \bar{\theta}_t)^2]} - \frac{\alpha^2 L_f}{2} \|\bar{\theta}_t\|_2^2. \tag{49}
\end{aligned}$$

Note that for any  $x \sim \nu_{\pi_W}$ , and any function  $h(x)$ , we have

$$\begin{aligned}
\int_x h(x) d\nu^*(x) &= \int_x h(x) \frac{d\nu^*(x)}{d\mu_{\pi_W}(x)} d\mu_{\pi_W}(x) \\
&\stackrel{(i)}{\leq} \sqrt{\int_x h^2(x) d\mu_{\pi_W}(x)} \sqrt{\int_x \left( \frac{d\nu^*(x)}{d\mu_{\pi_W}(x)} \right)^2 d\mu_{\pi_W}(x)} \\
&\stackrel{(ii)}{\leq} C_{RN}^2 \|h(x)\|_{\mu_{\pi_W}}, \tag{50}
\end{aligned}$$

where (i) follows from Holder's inequality, and (ii) follows from eq. (10). Similarly, we can obtain

$$\int_x h(x) d(\nu^* \pi_W)(x) \leq C_{RN}^2 \|h(x)\|_{\mu_{\pi_W}}. \tag{51}$$

Substituting eq. (50) and eq. (51) into eq. (49) and using the fact that  $\|\bar{\theta}_t\|_2 \leq R + \sqrt{m}d_2$  yield

$$\mathbb{E}_{\nu^*} [D_{\text{KL}}(\pi^* || \pi_{\tau_t W_t})] - \mathbb{E}_{\nu^*} [D_{\text{KL}}(\pi^* || \pi_{\tau_{t+1} W_{t+1}})]$$

$$\begin{aligned}
&\geq \alpha(1-\gamma)(J_0(\pi^*) - J_0(\pi_{\tau_t W_t})) - \alpha C_{RN} \sqrt{\mathbb{E}_{\nu_{\pi_{\tau_t W_t}}} \left[ (\phi_{W_t}(s, a)^\top \bar{\theta}_t - f((s, a), \bar{\theta}_t))^2 \right]} \\
&\quad - \alpha C_{RN} \sqrt{\mathbb{E}_{\mu_{\pi_{\tau_t W_t}}} \left[ (f((s, a), \bar{\theta}_t) - Q_{\pi_{\tau_t W_t}}(s, a))^2 \right]} \\
&\quad - \alpha C_{RN} \sqrt{\mathbb{E}_{\mu_{\pi_{\tau_t W_t}}} \left[ (Q_{\pi_{\tau_t W_t}}(s, a') - f((s, a'), \bar{\theta}_t))^2 \right]} \\
&\quad - \alpha C_{RN} \sqrt{\mathbb{E}_{\mu_{\pi_{\tau_t W_t}}} \left[ (f((s, a'), \bar{\theta}_t) - \phi_{W_t}(s, a')^\top \bar{\theta}_t)^2 \right]} - \alpha^2 L_f (R^2 + m d_2^2) \\
&= \alpha(1-\gamma)(J_0(\pi^*) - J_0(\pi_{\tau_t W_t})) - 2\alpha C_{RN} \sqrt{\mathbb{E}_{\mu_{\pi_{\tau_t W_t}}} \left[ (\phi_{W_t}(s, a)^\top \bar{\theta}_t - f((s, a), \bar{\theta}_t))^2 \right]} \\
&\quad - 2\alpha C_{RN} \sqrt{\mathbb{E}_{\mu_{\pi_{\tau_t W_t}}} \left[ (f((s, a), \bar{\theta}_t) - Q_{\pi_{\tau_t W_t}}(s, a))^2 \right]} - \alpha^2 L_f (R^2 + m d_2^2) \\
&= \alpha(1-\gamma)(J_0(\pi^*) - J_0(\pi_{\tau_t W_t})) - 2\alpha C_{RN} \left\| \phi_{W_t}(s, a)^\top \bar{\theta}_t - f((s, a), \bar{\theta}_t) \right\|_{\mu_{\pi_{\tau_t W_t}}} \\
&\quad - 2\alpha C_{RN} \left\| f((s, a), \bar{\theta}_t) - Q_{\pi_{\tau_t W_t}}(s, a) \right\|_{\mu_{\pi_{\tau_t W_t}}} - \alpha^2 L_f (R^2 + m d_2^2). \tag{52}
\end{aligned}$$

We then proceed to upper bound the term  $\left\| \phi_{W_t}(s, a)^\top \bar{\theta}_t - f((s, a), \bar{\theta}_t) \right\|_{\mu_{\pi_{\tau_t W_t}}}^2$ .

$$\begin{aligned}
&\left\| \phi_{W_t}(s, a)^\top \bar{\theta}_t - f((s, a), \bar{\theta}_t) \right\|_{\mu_{\pi_{\tau_t W_t}}}^2 \\
&= \left\| \phi_{W_t}(s, a)^\top \bar{\theta}_t - \phi_{W_0}(s, a)^\top \bar{\theta}_t + \phi_{W_0}(s, a)^\top \bar{\theta}_t - f((s, a), \bar{\theta}_t) \right\|_{\mu_{\pi_{\tau_t W_t}}}^2 \\
&\leq 2 \left\| \phi_{W_t}(s, a)^\top \bar{\theta}_t - \phi_{W_0}(s, a)^\top \bar{\theta}_t \right\|_{\mu_{\pi_{\tau_t W_t}}}^2 + 2 \left\| \phi_{W_0}(s, a)^\top \bar{\theta}_t - f((s, a), \bar{\theta}_t) \right\|_{\mu_{\pi_{\tau_t W_t}}}^2 \\
&\stackrel{(i)}{\leq} \frac{16C_0 R^3}{d_1 \sqrt{m}}, \tag{53}
\end{aligned}$$

where (i) follows from Lemma 12 and Lemma 14. Substituting eq. (53) into eq. (52) yields

$$\begin{aligned}
&\mathbb{E}_{\nu^*} [D_{KL}(\pi^* || \pi_{\tau_t W_t})] - \mathbb{E}_{\nu^*} [D_{KL}(\pi^* || \pi_{\tau_{t+1} W_{t+1}})] \\
&\leq \alpha(1-\gamma)(J_0(\pi^*) - J_0(\pi_{\tau_t W_t})) - \frac{8\alpha C_{RN} \sqrt{C_0} R^{1.5}}{\sqrt{d_1} m^{1/4}} - \alpha^2 L_f (R^2 + m d_2^2) \\
&\quad - 2\alpha C_{RN} \left\| f((s, a), \bar{\theta}_t) - Q_{\pi_{\tau_t W_t}}(s, a) \right\|_{\mu_{\pi_{\tau_t W_t}}}.
\end{aligned}$$

Rearranging the above inequality yields the desired result.  $\square$

Note that when we follow the update in line 10 of Algorithm 1, we can obtain similar results for the case  $i \in \{1, \dots, p\}$  as stated in Lemma 15:

$$\begin{aligned}
&\alpha(1-\gamma)(J_i(\pi_{\tau_t W_t}) - J_i(\pi^*)) \\
&\leq \mathbb{E}_{\nu^*} [D_{KL}(\pi^* || \pi_{\tau_t W_t})] - \mathbb{E}_{\nu^*} [D_{KL}(\pi^* || \pi_{\tau_{t+1} W_{t+1}})] + \frac{8\alpha C_{RN} \sqrt{C_0} R^{1.5}}{\sqrt{d_1} m^{1/4}} + \alpha^2 L_f (R^2 + m d_2^2) \\
&\quad + 2\alpha C_{RN} \left\| f((s, a), \bar{\theta}_t) - Q_{\pi_{\tau_t W_t}}(s, a) \right\|_{\mu_{\pi_{\tau_t W_t}}}.
\end{aligned}$$

#### C.4 PROOF OF THEOREM 2

We first prove the following lemmas.

**Lemma 16.** *Considering the CRPO update in Algorithm 1 in the neural network approximation setting. Let  $K_{in} = C_1((1-\gamma)^2 \sqrt{m})$  and  $N = T \log(2T/\delta)$ . With probability at least  $1 - \delta$ , we have*

$$\begin{aligned}
&\alpha(1-\gamma) \sum_{t \in \mathcal{N}_0} (J_0(\pi^*) - J_0(\pi_{w_t})) + \alpha(1-\gamma) \eta \sum_{i=1}^p |\mathcal{N}_i| \\
&\leq \mathbb{E}_{s \sim \nu^*} D_{KL}(\pi^* || \pi_{w_0}) + C_3 \left( \frac{\alpha T}{m^{1/4}} \right) + C_4(\alpha^2 m T)
\end{aligned}$$

$$+ C_5 \left( \frac{\alpha T}{(1-\gamma)^{1.5} m^{1/8}} \log^{\frac{1}{4}} \left( \frac{T^3}{\delta} \right) \right) + C_6 \left( \alpha(1-\gamma)\sqrt{T} \right).$$

where  $C_3 = \frac{8C_{RN}\sqrt{C_0}R^{1.5}}{\sqrt{d_1}}$ ,  $C_4 = L_f(R^2 + d_2^2)$ ,  $C_5 = 3\alpha C_2 C_{RN}$ ,  $C_6 = 2C_f$  and  $C_2$  is a positive constant depend on  $C_1$ .

*Proof.* We define the  $\mathcal{N}_i$  as the set of steps that CRPO algorithm chooses to minimize the  $i$ -th constraint. If  $t \in \mathcal{N}_0$ , by Lemma 15 we have

$$\begin{aligned} & \alpha(1-\gamma)(J_0(\pi^*) - J_0(\pi_{\tau_t W_t})) \\ & \leq \mathbb{E}_{\nu^*} [D_{\text{KL}}(\pi^* || \pi_{\tau_t W_t})] - \mathbb{E}_{\nu^*} [D_{\text{KL}}(\pi^* || \pi_{\tau_{t+1} W_{t+1}})] + \frac{8\alpha C_{RN}\sqrt{C_0}R^{1.5}}{\sqrt{d_1}m^{1/4}} + \alpha^2 L_f(R^2 + md_2^2) \\ & \quad + 2\alpha C_{RN} \left\| f_0((s, a), \bar{\theta}_t) - Q_{\pi_{\tau_t W_t}}^0(s, a) \right\|_{\mu_{\pi_{\tau_t W_t}}}. \end{aligned} \quad (54)$$

If  $t \in \mathcal{N}_i$ , similarly we can obtain

$$\begin{aligned} & \alpha(1-\gamma)(J_i(\pi_{\tau_t W_t}) - J_i(\pi^*)) \\ & \leq \mathbb{E}_{\nu^*} [D_{\text{KL}}(\pi^* || \pi_{\tau_t W_t})] - \mathbb{E}_{\nu^*} [D_{\text{KL}}(\pi^* || \pi_{\tau_{t+1} W_{t+1}})] + \frac{8\alpha C_{RN}\sqrt{C_0}R^{1.5}}{\sqrt{d_1}m^{1/4}} + \alpha^2 L_f(R^2 + md_2^2) \\ & \quad + 2\alpha C_{RN} \left\| f_i((s, a), \bar{\theta}_t) - Q_{\pi_{\tau_t W_t}}^i(s, a) \right\|_{\mu_{\pi_{\tau_t W_t}}}. \end{aligned} \quad (55)$$

Summing eq. (13) and eq. (14) from  $t = 0$  to  $T - 1$  yields

$$\begin{aligned} & \alpha(1-\gamma) \sum_{t \in \mathcal{N}_0} (J_0(\pi^*) - J_0(\pi_{w_t})) + \alpha(1-\gamma) \sum_{i=1}^p \sum_{t \in \mathcal{N}_i} (J_i(\pi_{w_t}) - J_i(\pi^*)) \\ & \leq \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0}) + \frac{8\alpha C_{RN}\sqrt{C_0}R^{1.5}T}{\sqrt{d_1}m^{1/4}} + \alpha^2 L_f(R^2 + md_2^2)T \\ & \quad + 2\alpha C_{RN} \sum_{i=0}^p \sum_{t \in \mathcal{N}_i} \left\| f_i((s, a), \bar{\theta}_t) - Q_{\pi_{\tau_t W_t}}^i(s, a) \right\|_{\mu_{\pi_{\tau_t W_t}}}. \end{aligned} \quad (56)$$

Note that when  $t \in \mathcal{N}_i$  ( $i \neq 0$ ), we have  $\bar{J}_i(\theta_t^i) > d_i + \eta$  (line 11 in Algorithm 1), which implies that

$$\begin{aligned} J_i(\pi_{\tau_t W_t}) - J_i(\pi^*) & \geq \bar{J}_i(\theta_t^i) - J_i(\pi^*) - |\bar{J}_i(\theta_t^i) - J_i(\pi_{\tau_t W_t})| \\ & \geq d_i + \eta - J_i(\pi^*) - |\bar{J}_i(\theta_t^i) - J_i(\pi_{\tau_t W_t})| \\ & \geq \eta - |\bar{J}_i(\theta_t^i) - J_i(\pi_{\tau_t W_t})|. \end{aligned} \quad (57)$$

To bound the term  $|\bar{J}_i(\theta_t^i) - J_i(\pi_{\tau_t W_t})|$ , we proceed as follows

$$\begin{aligned} & |\bar{J}_i(\theta_t^i) - J_i(\pi_{\tau_t W_t})| \\ & = \left| \bar{J}_i(\theta_t^i) - \mathbb{E}_{\nu_{\pi_{\tau_t W_t}}} [f_i((s, a), \bar{\theta}_t)] + \mathbb{E}_{\nu_{\pi_{\tau_t W_t}}} [f_i((s, a), \bar{\theta}_t)] - J_i(\pi_{\tau_t W_t}) \right| \\ & \leq \left| \bar{J}_i(\theta_t^i) - \mathbb{E}_{\nu_{\pi_{\tau_t W_t}}} [f_i((s, a), \bar{\theta}_t)] \right| + \left\| f_i((s, a), \bar{\theta}_t) - Q_{\pi_{\tau_t W_t}}^i(s, a) \right\|_{\nu_{\pi_{\tau_t W_t}}} \\ & \stackrel{(i)}{\leq} \left| \bar{J}_i(\theta_t^i) - \mathbb{E}_{\nu_{\pi_{\tau_t W_t}}} [f_i((s, a), \bar{\theta}_t)] \right| + C_{RN} \left\| f_i((s, a), \bar{\theta}_t) - Q_{\pi_{\tau_t W_t}}^i(s, a) \right\|_{\mu_{\pi_{\tau_t W_t}}}, \end{aligned} \quad (58)$$

where (i) can be obtained by following similar steps in eq. (50). Substituting eq. (58) into eq. (57) yields

$$\begin{aligned} & J_i(\pi_{\tau_t W_t}) - J_i(\pi^*) \\ & \geq \eta - \left( \left| \bar{J}_i(\theta_t^i) - \mathbb{E}_{\nu_{\pi_{\tau_t W_t}}} [f_i((s, a), \bar{\theta}_t)] \right| + C_{RN} \left\| f_i((s, a), \bar{\theta}_t) - Q_{\pi_{\tau_t W_t}}^i(s, a) \right\|_{\mu_{\pi_{\tau_t W_t}}} \right). \end{aligned} \quad (59)$$

Then, substituting eq. (59) into eq. (56) yields

$$\begin{aligned}
& \alpha(1-\gamma) \sum_{t \in \mathcal{N}_0} (J_0(\pi^*) - J_0(\pi_{w_t})) + \alpha(1-\gamma)\eta \sum_{i=1}^p |\mathcal{N}_i| \\
& \leq \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0}) + \frac{8\alpha C_{RN} \sqrt{C_0} R^{1.5} T}{\sqrt{d_1} m^{1/4}} + \alpha^2 L_f (R^2 + m d_2^2) T \\
& \quad + 3\alpha C_{RN} \sum_{i=0}^p \sum_{t \in \mathcal{N}_i} \left\| f_i((s, a), \bar{\theta}_t) - Q_{\pi_{\tau_t} w_t}^i(s, a) \right\|_{\mu_{\pi_{\tau_t} w_t}} \\
& \quad + \alpha(1-\gamma) \sum_{i=1}^p \sum_{t \in \mathcal{N}_i} \left| \bar{J}_i(\theta_t^i) - \mathbb{E}_{\nu_{\pi_{\tau_t} w_t}} [f_i((s, a), \bar{\theta}_t)] \right| \\
& \leq \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0}) + \frac{8\alpha C_{RN} \sqrt{C_0} R^{1.5} T}{\sqrt{d_1} m^{1/4}} + \alpha^2 L_f (R^2 + m d_2^2) T \\
& \quad + 3\alpha C_{RN} \sum_{t=0}^{T-1} \left\| f_i((s, a), \bar{\theta}_t) - Q_{\pi_{\tau_t} w_t}^i(s, a) \right\|_{\mu_{\pi_{\tau_t} w_t}} \\
& \quad + \alpha(1-\gamma) \sum_{t=0}^{T-1} \left| \bar{J}_i(\theta_t^i) - \mathbb{E}_{\nu_{\pi_{\tau_t} w_t}} [f_i((s, a), \bar{\theta}_t)] \right|. \tag{60}
\end{aligned}$$

We then consider upper bound the term  $\sum_{t=0}^{T-1} \left\| f_i((s, a), \bar{\theta}_t) - Q_{\pi_{\tau_t} w_t}^i(s, a) \right\|_{\mu_{\pi_{\tau_t} w_t}}$ . Lemma 2 implies that if we let  $K_{\text{in}} = C_1((1-\gamma)^2 \sqrt{m})$ , then with probability at least  $1 - \delta_1/T$ , we have

$$\left\| f((s, a); \bar{\theta}_K) - Q_{\pi}(s, a) \right\|_{\mu_{\pi}} \leq C_2 \left( \frac{1}{(1-\gamma)^{1.5} m^{1/8}} \log^{\frac{1}{4}} \left( \frac{(1-\gamma)^2 T \sqrt{m}}{\delta_1} \right) \right),$$

where  $C_1$  and  $C_2$  are positive constant. Applying union bound, we have with probability at least  $1 - \delta_1$ ,

$$\sum_{t=0}^{T-1} \left\| f_i((s, a), \bar{\theta}_t) - Q_{\pi_{\tau_t} w_t}^i(s, a) \right\|_{\mu_{\pi_{\tau_t} w_t}} \leq C_2 \left( \frac{T}{(1-\gamma)^{1.5} m^{1/8}} \log^{\frac{1}{4}} \left( \frac{(1-\gamma)^2 T \sqrt{m}}{\delta_1} \right) \right). \tag{61}$$

We then consider bound the term  $\sum_{t=0}^{T-1} \left| \bar{J}_i(\theta_t^i) - \mathbb{E}_{\nu_{\pi_{\tau_t} w_t}} [f_i((s, a), \bar{\theta}_t)] \right|$ . For simplicity, we denote  $J'_i(\bar{\theta}_t) = \mathbb{E}_{\xi, \mu_{\pi_{\tau_t} w_t}} [f_i((s, a), \bar{\theta}_t)]$ . Recall that  $\bar{J}_i(\theta_t^i) = \frac{1}{N} \sum_{j=1}^N f_i((s_j, a_j), \bar{\theta}_t)$ . For each  $t \geq 0$ , consider the error  $\bar{J}_i(\theta_t^i) - J'_i(\bar{\theta}_t)$ , we have

$$\begin{aligned}
& \mathbf{P} \left( \left( \frac{1}{N} \sum_{j=1}^N f_i((s_j, a_j), \bar{\theta}_t) - J'_i(\bar{\theta}_t) \right)^2 \geq \frac{(1+\Lambda) C_f^2}{N} \right) \\
& \leq \mathbf{P} \left( \frac{1}{N} \sum_{j=1}^N \frac{[f_i((s_j, a_j), \bar{\theta}_t) - J'_i(\bar{\theta}_t)]^2}{C_f^2} \geq 1 + \Lambda \right) \\
& = \mathbf{P} \left( \exp \left( \frac{1}{N} \sum_{j=1}^N \frac{[f_i((s_j, a_j), \bar{\theta}_t) - J'_i(\bar{\theta}_t)]^2}{C_f^2} \right) \geq 1 + \Lambda \right) \\
& \leq \mathbf{P} \left( \frac{1}{N} \sum_{j=1}^N \exp \left( \frac{[f_i((s_j, a_j), \bar{\theta}_t) - J'_i(\bar{\theta}_t)]^2}{C_f^2} \right) \geq 1 + \Lambda \right) \\
& \stackrel{(i)}{\leq} \frac{1}{N} \sum_{j=1}^N \mathbb{E} \left[ \exp \left( \frac{[f_i((s_j, a_j), \bar{\theta}_t) - J'_i(\bar{\theta}_t)]^2}{C_f^2} \right) \right] / \exp(1 + \Lambda)
\end{aligned}$$

$$\leq \exp(-\Lambda), \quad (62)$$

where (i) follows from Markov's inequality. eq. (62) implies that with probability at least  $1 - \delta_2/T$ , we have

$$\left| \frac{1}{N} \sum_{j=1}^N f_i((s_j, a_j), \bar{\theta}_t) - J'_i(\bar{\theta}_t) \right| \leq \frac{C_f}{\sqrt{N}} \left( 1 + \sqrt{\log \left( \frac{T}{\delta_2} \right)} \right).$$

Applying union bound, we have with probability at least  $1 - \delta_2$ ,

$$\sum_{t=0}^{T-1} \left| \bar{J}_i(\theta_t^i) - \mathbb{E}_{\nu_{\pi_{\tau_t} w_t}} [f_i((s, a), \bar{\theta}_t)] \right| \leq \frac{C_f T}{\sqrt{N}} \left( 1 + \sqrt{\log \left( \frac{T}{\delta_2} \right)} \right). \quad (63)$$

Letting  $\delta_1 = \delta_2 = \frac{\delta}{2}$ ,  $N = T \log(2T/\delta)$ , and combining eq. (61) and eq. (63), we have with probability at least  $1 - \delta$

$$\begin{aligned} & \alpha(1 - \gamma) \sum_{t \in \mathcal{N}_0} (J_0(\pi^*) - J_0(\pi_{w_t})) + \alpha(1 - \gamma) \eta \sum_{i=1}^p |\mathcal{N}_i| \\ & \leq \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* \| \pi_{w_0}) + C_3 \left( \frac{\alpha T}{m^{1/4}} \right) + C_4(\alpha^2 m T) \\ & \quad + C_5 \left( \frac{\alpha T}{(1 - \gamma)^{1.5} m^{1/8}} \log^{\frac{1}{4}} \left( \frac{(1 - \gamma)^2 T \sqrt{m}}{\delta} \right) \right) + C_6 \left( \alpha(1 - \gamma) \sqrt{T} \right), \end{aligned}$$

where  $C_3 = \frac{8C_{RN}\sqrt{C_0}R^{1.5}}{\sqrt{d_1}}$ ,  $C_4 = L_f(R^2 + d_2^2)$ ,  $C_5 = 3\alpha C_2 C_{RN}$ , and  $C_6 = 2C_f$  are positive constants.  $\square$

**Lemma 17.** Let  $K_{in} = C_1((1 - \gamma)^2 \sqrt{m})$  and  $N = T \log(2T/\delta)$ , if we let

$$\begin{aligned} \frac{1}{2} \alpha(1 - \gamma) \eta T & \geq \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* \| \pi_{w_0}) + C_3 \left( \frac{\alpha T}{m^{1/4}} \right) + C_4(\alpha^2 m T) \\ & \quad + C_5 \left( \frac{\alpha T}{(1 - \gamma)^{1.5} m^{1/8}} \log^{\frac{1}{4}} \left( \frac{(1 - \gamma)^2 T \sqrt{m}}{\delta} \right) \right) + C_6 \left( \alpha(1 - \gamma) \sqrt{T} \right), \end{aligned} \quad (64)$$

then with probability at least  $1 - \delta$ , we have the following holds

1.  $\mathcal{N}_0 \neq \emptyset$ , i.e.,  $w_{out}$  is well-defined,
2. One of the following two statements must holds,
  - (a)  $|\mathcal{N}_0| \geq T/2$ ,
  - (b)  $\sum_{t \in \mathcal{G}} (J_0(\pi^*) - J_0(w_t)) \leq 0$ .

*Proof.* We consider the event given in Lemma 16, which happens with probability at least  $1 - \delta$ , we have

$$\begin{aligned} & \alpha(1 - \gamma) \sum_{t \in \mathcal{N}_0} (J_0(\pi^*) - J_0(\pi_{w_t})) + \alpha(1 - \gamma) \eta \sum_{i=1}^p |\mathcal{N}_i| \\ & \leq \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* \| \pi_{w_0}) + C_3 \left( \frac{\alpha T}{m^{1/4}} \right) + C_4(\alpha^2 m T) \\ & \quad + C_5 \left( \frac{\alpha T}{(1 - \gamma)^{1.5} m^{1/8}} \log^{\frac{1}{4}} \left( \frac{(1 - \gamma)^2 T \sqrt{m}}{\delta} \right) \right) + C_6 \left( \alpha(1 - \gamma) \sqrt{T} \right). \end{aligned} \quad (65)$$

We first verify item 1. If  $\mathcal{N}_0 = \emptyset$ , then  $\sum_{i=1}^p |\mathcal{N}_i| = T$ , Lemma 16 implies that

$$\begin{aligned} \alpha(1 - \gamma) \eta T & \leq \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* \| \pi_{w_0}) + C_3 \left( \frac{\alpha T}{m^{1/4}} \right) + C_4(\alpha^2 m T) \\ & \quad + C_5 \left( \frac{\alpha T}{(1 - \gamma)^{1.5} m^{1/8}} \log^{\frac{1}{4}} \left( \frac{(1 - \gamma)^2 T \sqrt{m}}{\delta} \right) \right) + C_6 \left( \alpha(1 - \gamma) \sqrt{T} \right), \end{aligned}$$

which contradicts with eq. (64). Thus, we must have  $\mathcal{N}_0 \neq \emptyset$ .

We then proceed to verify the second item. If  $\sum_{t \in \mathcal{G}} (J_0(\pi^*) - J_0(w_t)) \leq 0$ , then (b) hold. If  $\sum_{t \in \mathcal{G}} (J_0(\pi^*) - J_0(w_t)) \leq 0$ , then Equation (65) implies that

$$\begin{aligned} \alpha(1-\gamma)\eta \sum_{i=1}^p |\mathcal{N}_i| &\leq \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0}) + C_3 \left( \frac{\alpha T}{m^{1/4}} \right) + C_4(\alpha^2 m T) \\ &\quad + C_5 \left( \frac{\alpha T}{(1-\gamma)^{1.5} m^{1/8}} \log^{\frac{1}{4}} \left( \frac{(1-\gamma)^2 T \sqrt{m}}{\delta} \right) \right) + C_6 \left( \alpha(1-\gamma) \sqrt{T} \right). \end{aligned}$$

Suppose that  $|\mathcal{N}_0| < T/2$ , i.e.,  $\sum_{i=1}^p |\mathcal{N}_i| \geq T/2$ . Then,

$$\begin{aligned} \frac{1}{2} \alpha(1-\gamma)\eta T &\leq \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0}) + C_3 \left( \frac{\alpha T}{m^{1/4}} \right) + C_4(\alpha^2 m T) \\ &\quad + C_5 \left( \frac{\alpha T}{(1-\gamma)^{1.5} m^{1/8}} \log^{\frac{1}{4}} \left( \frac{(1-\gamma)^2 T \sqrt{m}}{\delta} \right) \right) + C_6 \left( \alpha(1-\gamma) \sqrt{T} \right), \end{aligned}$$

which contradicts with eq. (64). Hence, (a) holds.  $\square$

Now, we are ready to prove our main theorem. We restate Theorem 1 as follows to include the specifics of the parameters.

**Theorem 4** (Restatement of Theorem 2). *Consider Algorithm 1 in the neural network approximation setting. Suppose Assumption 1-5 hold. Let  $\alpha = \frac{1}{2C_4\sqrt{T}}$  and*

$$\begin{aligned} \eta &= \frac{4C_4 \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0})}{(1-\gamma)\sqrt{T}} + \frac{2C_3}{(1-\gamma)m^{1/4}} + \frac{m}{(1-\gamma)\sqrt{T}} \\ &\quad + 2C_5 \left( \frac{\alpha T}{(1-\gamma)^{2.5} m^{1/8}} \log^{\frac{1}{4}} \left( \frac{(1-\gamma)^2 T \sqrt{m}}{\delta} \right) \right) + \frac{2C_6}{\sqrt{T}}. \end{aligned}$$

*Suppose performing neural TD with  $K_{in} = C_1(1-\gamma)^2\sqrt{m}$  iterations at each iteration of CRPO. Then, with probability at least  $1 - \delta$ , we have*

$$J_0(\pi^*) - \mathbb{E}[J_0(\pi_{w_{out}})] \leq \frac{C_7 m}{(1-\gamma)\sqrt{T}} + \frac{C_8}{(1-\gamma)^{2.5} m^{1/8}} \log^{\frac{1}{4}} \left( \frac{(1-\gamma)^2 T \sqrt{m}}{\delta} \right),$$

where

$$C_7 = \frac{4C_4 D_{\text{KL}}(\pi^* || \pi_{w_0})}{m} + \frac{2(1-\gamma)C_6}{m} + 1,$$

and

$$C_8 = 2C_5 + \frac{2C_3(1-\gamma)^{1.5}}{m^{1/8}}.$$

For all  $i \in \{1, \dots, p\}$ , we have

$$\begin{aligned} \mathbb{E}[J_i(\pi_{w_{out}})] - d_i &\leq \frac{4C_4 \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0})}{(1-\gamma)\sqrt{T}} + \frac{2C_3}{(1-\gamma)m^{1/4}} + \frac{m}{(1-\gamma)\sqrt{T}} \\ &\quad + 2(C_2 + C_5) \left( \frac{\alpha T}{(1-\gamma)^{2.5} m^{1/8}} \log^{\frac{1}{4}} \left( \frac{(1-\gamma)^2 T \sqrt{m}}{\delta} \right) \right) + \frac{4C_6}{\sqrt{T}}. \end{aligned}$$

*Proof of Theorem 2.* We consider the event given in Lemma 16, which happens with probability at least  $1 - \delta$ :

$$\begin{aligned} &\alpha(1-\gamma) \sum_{t \in \mathcal{N}_0} (J_0(\pi^*) - J_0(\pi_{w_t})) + \alpha(1-\gamma)\eta \sum_{i=1}^p |\mathcal{N}_i| \\ &\leq \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0}) + C_3 \left( \frac{\alpha T}{m^{1/4}} \right) + C_4(\alpha^2 m T) \end{aligned}$$

$$+ C_5 \left( \frac{\alpha T}{(1-\gamma)^{1.5} m^{1/8}} \log^{\frac{1}{4}} \left( \frac{(1-\gamma)^2 T \sqrt{m}}{\delta} \right) \right) + C_6 \left( \alpha(1-\gamma) \sqrt{T} \right). \quad (66)$$

We first consider the convergence rate of the objective function. In the aforementioned event, we have the following holds:

$$\begin{aligned} & \alpha(1-\gamma) \sum_{t \in \mathcal{N}_0} (J_0(\pi^*) - J_0(\pi_{w_t})) \\ & \leq \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* \| \pi_{w_0}) + C_3 \left( \frac{\alpha T}{m^{1/4}} \right) + C_4(\alpha^2 m T) \\ & \quad + C_5 \left( \frac{\alpha T}{(1-\gamma)^{1.5} m^{1/8}} \log^{\frac{1}{4}} \left( \frac{(1-\gamma)^2 T \sqrt{m}}{\delta} \right) \right) + C_6 \left( \alpha(1-\gamma) \sqrt{T} \right). \end{aligned}$$

If  $\sum_{t \in \mathcal{N}_0} (J_0(\pi^*) - J_0(\pi_{w_t})) \leq 0$ , then we have  $J_0(\pi^*) - J_0(\pi_{w_{\text{out}}}) \leq 0$ . If  $\sum_{t \in \mathcal{N}_0} (J_0(\pi^*) - J_0(\pi_{w_t})) \geq 0$ , we have  $|\mathcal{N}_0| \geq T/2$ , which implies the following convergence rate

$$\begin{aligned} J_0(\pi^*) - \mathbb{E}[J_0(\pi_{w_{\text{out}}})] &= \frac{1}{|\mathcal{N}_0|} \sum_{t \in \mathcal{N}_0} (J_0(\pi^*) - J_0(\pi_{w_t})) \\ &\leq \frac{2\mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* \| \pi_{w_0})}{\alpha(1-\gamma)T} + \frac{2C_3}{(1-\gamma)m^{1/4}} + \frac{2C_4\alpha m}{1-\gamma} \\ &\quad + \frac{2C_5}{(1-\gamma)^{2.5} m^{1/8}} \log^{\frac{1}{4}} \left( \frac{(1-\gamma)^2 T \sqrt{m}}{\delta} \right) + \frac{2C_6}{\sqrt{T}}. \end{aligned}$$

Letting  $\alpha = \frac{1}{2C_4\sqrt{T}}$ , we can obtain the following convergence rate

$$J_0(\pi^*) - \mathbb{E}[J_0(\pi_{w_{\text{out}}})] \leq \frac{C_7 m}{(1-\gamma)\sqrt{T}} + \frac{C_8}{(1-\gamma)^{2.5} m^{1/8}} \log^{\frac{1}{4}} \left( \frac{(1-\gamma)^2 T \sqrt{m}}{\delta} \right),$$

where

$$C_7 = \frac{4C_4 D_{\text{KL}}(\pi^* \| \pi_{w_0})}{m} + \frac{2(1-\gamma)C_6}{m} + 1,$$

and

$$C_8 = 2C_5 + \frac{2C_3(1-\gamma)^{1.5}}{m^{1/8}}.$$

We then proceed to bound the constraints violation cost. For any  $i \in \{1, \dots, p\}$ , we have

$$\begin{aligned} \mathbb{E}[J_i(\pi_{w_{\text{out}}})] - d_i &= \frac{1}{|\mathcal{N}_0|} \sum_{t \in \mathcal{N}_0} J_i(\pi_{w_t}) - d_i \\ &\leq \frac{1}{|\mathcal{N}_0|} \sum_{t \in \mathcal{N}_0} (\bar{J}_i(\theta_t^i) - d_i) + \frac{1}{|\mathcal{N}_0|} \sum_{t \in \mathcal{N}_0} |J_i(\pi_{w_t}) - \bar{J}_i(\theta_t^i)| \\ &\leq \eta + \frac{1}{|\mathcal{N}_0|} \sum_{t=0}^{T-1} |J_i(\pi_{w_t}) - \bar{J}_i(\theta_t^i)| \\ &\leq \eta + \frac{1}{|\mathcal{N}_0|} \sum_{t=0}^{T-1} \left| \bar{J}_i(\theta_t^i) - \mathbb{E}_{\nu_{\pi_{\tau_t} w_t}} [f_i((s, a), \bar{\theta}_t)] \right| \\ &\quad + \frac{C_{RN}}{|\mathcal{N}_0|} \sum_{t=0}^{T-1} \left\| f_i((s, a), \bar{\theta}_t) - Q_{\pi_{\tau_t} w_t}^i(s, a) \right\|_{\mu_{\pi_{\tau_t} w_t}}. \end{aligned}$$

Recall eq. (61) and eq. (63), in the event defined in eq. (66), we have

$$\sum_{t=0}^{T-1} \left| \bar{J}_i(\theta_t^i) - \mathbb{E}_{\nu_{\pi_{\tau_t} w_t}} [f_i((s, a), \bar{\theta}_t)] \right| \leq C_6 \sqrt{T}, \quad (67)$$

and

$$\begin{aligned} & \sum_{t=0}^{T-1} \left\| f_i((s, a), \bar{\theta}_t) - Q_{\pi_{\tau_t} w_t}^i(s, a) \right\|_{\mu_{\pi_{\tau_t} w_t}} \\ & \leq C_2 \left( \frac{T}{(1-\gamma)^{1.5} m^{1/8}} \log^{\frac{1}{4}} \left( \frac{2(1-\gamma)^2 T \sqrt{m}}{\delta} \right) \right). \end{aligned} \quad (68)$$

Let the value of the tolerance  $\eta$  be

$$\begin{aligned} \eta &= \frac{4C_4 \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0})}{(1-\gamma)\sqrt{T}} + \frac{2C_3}{(1-\gamma)m^{1/4}} + \frac{m}{(1-\gamma)\sqrt{T}} \\ &+ 2C_5 \left( \frac{\alpha T}{(1-\gamma)^{2.5} m^{1/8}} \log^{\frac{1}{4}} \left( \frac{(1-\gamma)^2 T \sqrt{m}}{\delta} \right) \right) + \frac{2C_6}{\sqrt{T}}, \end{aligned} \quad (69)$$

we have

$$\begin{aligned} \frac{1}{2} \alpha (1-\gamma) \eta T &\geq \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0}) + C_3 \left( \frac{\alpha T}{m^{1/4}} \right) + C_4 (\alpha^2 m T) \\ &+ C_5 \left( \frac{\alpha T}{(1-\gamma)^{1.5} m^{1/8}} \log^{\frac{1}{4}} \left( \frac{(1-\gamma)^2 T \sqrt{m}}{\delta} \right) \right) + C_6 \left( \alpha (1-\gamma) \sqrt{T} \right), \end{aligned}$$

which satisfies the requirement specified in Lemma 17. Combining eq. (67), eq. (68) and eq. (69), according to Lemma 17, we have with probability at least  $1 - \delta$  at least one of the following holds:

$$\mathbb{E}[J_i(\pi_{w_{\text{out}}})] - d_i \leq 0,$$

or  $|\mathcal{N}_0| \geq T/2$ , which further implies

$$\begin{aligned} \mathbb{E}[J_i(\pi_{w_{\text{out}}})] - d_i &\leq \frac{4C_4 \mathbb{E}_{s \sim \nu^*} D_{\text{KL}}(\pi^* || \pi_{w_0})}{(1-\gamma)\sqrt{T}} + \frac{2C_3}{(1-\gamma)m^{1/4}} + \frac{m}{(1-\gamma)\sqrt{T}} \\ &+ 2(C_2 + C_5) \left( \frac{\alpha T}{(1-\gamma)^{2.5} m^{1/8}} \log^{\frac{1}{4}} \left( \frac{(1-\gamma)^2 T \sqrt{m}}{\delta} \right) \right) + \frac{4C_6}{\sqrt{T}}. \end{aligned}$$

□