Supplementary Materials

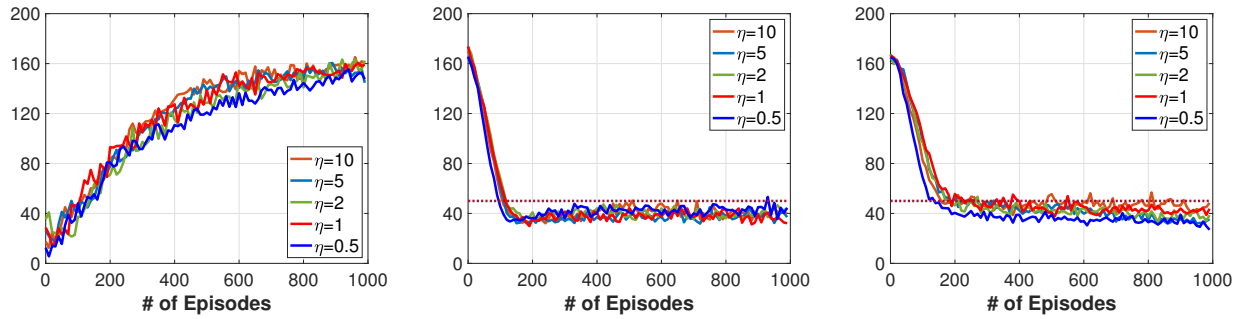
A. Experimental Setting

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 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 along an anticlockwise direction; or (2) a torque with value $+1$ is applied when the second pendulum swings along an anticlockwise direction with respect to the first pendulum.

For details about the update of PD, please refer to (Achiam et al., 2017)[Section 10.3.3]. The performance of PD is very sensitive to the stepsize of the dual variable's update. If the stepsize is too small, then the dual variable will not 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 the dual variable, we conduct the experiments with the learning rates $\{0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05\}$ 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. Thus, our reported result of Cartpole is with the stepsize 0.005 and our reported result of Acrobot is with the stepsize 0.0005.

Next, we investigate the robustness of CRPO with respect to the tolerance parameter η. We conduct the experiments under the following values of η $\{10, 5, 2, 1, 0.5\}$ for the Acrobot environment. It can be seen from Figure 4 that the learning curves of CRPO with the tolerance parameter η taking different values are almost the same, which indicates that the convergence performance of CRPO is robust to the value of η over a wide range. Thus, the tolerance parameter η does not cause much parameter tuning cost for CRPO.



(a) Return Reward (b) Constraint Value I (c) Constraint Value II

Figure 4. Comparison of CRPO in Acrobot with tolerance parameter η taking different values.

B. Proof of Theorem 1: Tabular Setting

B.1. Supporting Lemmas for Poof of Theorem 1

The following lemma characterizes the convergence rate of TD learning in the tabular setting.

Lemma 2 ((Dalal et al., 2019)). *Consider the iteration given in eq. (4) with arbitrary initialization θ_0^i. Assume that the stationary distribution μ_{π_w} is not degenerate for all $w \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$. Let stepsize $\beta_k = \Theta(\frac{1}{t^\sigma})$ $(0 < \sigma < 1)$. Then, with probability at least $1-\delta$, we have*

$$\left\|\theta_K^i - \theta_*^i(\pi_w)\right\|_2 = \mathcal{O}\left(\frac{\log(|\mathcal{S}|^2 |\mathcal{A}|^2 K^2/\delta)}{(1-\gamma)K^{\sigma/2}}\right).$$

Note that σ can be arbitrarily close to 1. Lemma 2 implies that we can obtain an approximation $\bar{Q}_t^i$ such that $\left\|\bar{Q}_t^i - Q_{\pi_w}^i\right\|_2 =$

$\mathcal{O}(1/\sqrt{K_{\text{in}}})$ with high probability.

Lemma 3 (Performance difference lemma (Kakade & Langford, 2002))**.** *For all policies π, π' and initial distribution ρ, we have*

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

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

Lemma 4 (Lemma 5.1. (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 takes the form:*

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

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 if we follow the update in line 10 of Algorithm 1, we can obtain similar results for the case $i \in \{1,\cdots,p\}$ 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 ρ, we have*

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

and

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

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

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

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

□

Lemma 6 (Performance improvement bound for approximated NPG)**.** *For the iterates π_{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 ρ and when $i = 0$, the following holds*

$$J_0^\rho(w_{t+1}) - J_0^\rho(w_t)$$

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

Proof. We first provide the following lower bound.

$$\log Z_t(s) - \frac{\alpha}{1-\gamma}V^i_{\pi_{w_t}}(s)$$
$$= \log\sum_{a\in\mathcal{A}}\pi_{w_t}(a|s)\exp\left(\frac{\alpha\bar{Q}^i_t(s,a)}{1-\gamma}\right) - \frac{\alpha}{1-\gamma}V^i_{\pi_{w_t}}(s)$$
$$\geq \sum_{a\in\mathcal{A}}\pi_{w_t}(a|s)\log\exp\left(\frac{\alpha\bar{Q}^i_t(s,a)}{1-\gamma}\right) - \frac{\alpha}{1-\gamma}V^i_{\pi_{w_t}}(s)$$
$$= \frac{\alpha}{1-\gamma}\sum_{a\in\mathcal{A}}\pi_{w_t}(a|s)(\bar{Q}^i_t(s,a) - Q^i_{\pi_{w_t}}(s,a)) + \frac{\alpha}{1-\gamma}\sum_{a\in\mathcal{A}}\pi_{w_t}(a|s)Q^i_{\pi_{w_t}}(s,a) - \frac{\alpha}{1-\gamma}V^i_{\pi_{w_t}}(s)$$
$$= \frac{\alpha}{1-\gamma}\sum_{a\in\mathcal{A}}\pi_{w_t}(a|s)(\bar{Q}^i_t(s,a) - Q^i_{\pi_{w_t}}(s,a))$$
$$\geq \frac{-\alpha}{1-\gamma}\sum_{a\in\mathcal{A}}\pi_{w_t}(a|s)\left|\bar{Q}^i_t(s,a) - Q^i_{\pi_{w_t}}(s,a)\right|.$$

Thus, we conclude that

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

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

$$J^\rho_i(w_{t+1}) - J^\rho_i(w_t)$$
$$= \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu_\rho}\sum_{a\in\mathcal{A}}\pi_{w_{t+1}}(a|s)A^i_{\pi_{w_t}}(s,a)$$
$$= \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu_\rho}\sum_{a\in\mathcal{A}}\pi_{w_{t+1}}(a|s)Q^i_{\pi_{w_t}}(s,a) - \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu_\rho}V^i_{\pi_{w_t}}(s)$$
$$= \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu_\rho}\sum_{a\in\mathcal{A}}\pi_{w_{t+1}}(a|s)\bar{Q}^i_t(s,a) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu_\rho}\sum_{a\in\mathcal{A}}\pi_{w_{t+1}}(a|s)(Q^i_{\pi_{w_t}}(s,a) - \bar{Q}^i_t(s,a))$$
$$- \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu_\rho}V^i_{\pi_{w_t}}(s)$$
$$\overset{(i)}{=} \frac{1}{\alpha}\mathbb{E}_{s\sim\nu_\rho}\sum_{a\in\mathcal{A}}\pi_{w_{t+1}}(a|s)\log\left(\frac{\pi_{w_{t+1}}(a|s)Z_t(s)}{\pi_{w_t}(a|s)}\right)$$
$$+ \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu_\rho}\sum_{a\in\mathcal{A}}\pi_{w_{t+1}}(a|s)(Q^i_{\pi_{w_t}}(s,a) - \bar{Q}^i_t(s,a)) - \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu_\rho}V^i_{\pi_{w_t}}(s)$$
$$= \frac{1}{\alpha}\mathbb{E}_{s\sim\nu_\rho}D_{\mathrm{KL}}(\pi_{w_{t+1}}||\pi_{w_t}) + \frac{1}{\alpha}\mathbb{E}_{s\sim\nu_\rho}\log Z_t(s)$$
$$+ \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu_\rho}\sum_{a\in\mathcal{A}}\pi_{w_{t+1}}(a|s)(Q^i_{\pi_{w_t}}(s,a) - \bar{Q}^i_t(s,a)) - \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu_\rho}V^i_{\pi_{w_t}}(s)$$
$$\geq \frac{1}{\alpha}\mathbb{E}_{s\sim\nu_\rho}\left(\log Z_t(s) - \frac{\alpha}{1-\gamma}V^i_{\pi_{w_t}}(s) + \frac{\alpha}{1-\gamma}\sum_{a\in\mathcal{A}}\pi_{w_t}(a|s)\left|\bar{Q}^i_t(s,a) - Q^i_{\pi_{w_t}}(s,a)\right|\right)$$

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

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

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

Lemma 7 (Upper bound on optimality gap for approximated NPG)**.** *Consider the approximated NPG updates in line 7 of Algorithm 1 in the tabular setting when $i = 0$. We have*

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

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

$$
\begin{aligned}
&J_i(\pi^*) - J_i(\pi_{w_t}) \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu^*}\sum_{a\in\mathcal{A}}\pi^*(a|s)A^i_{\pi_{w_t}}(s,a) \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu^*}\sum_{a\in\mathcal{A}}\pi^*(a|s)Q^i_{\pi_{w_t}}(s,a) - \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu^*}V^i_{\pi_{w_t}}(s) \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu^*}\sum_{a\in\mathcal{A}}\pi^*(a|s)\bar{Q}_t^i(s,a) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu^*}\sum_{a\in\mathcal{A}}\pi^*(a|s)(Q^i_{\pi_{w_t}}(s,a)-\bar{Q}_t^i(s,a)) \\
&\quad - \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu^*}V^i_{\pi_{w_t}}(s) \\
&\overset{(i)}{=} \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu^*}\sum_{a\in\mathcal{A}}\pi^*(a|s)\log\frac{\pi_{w_{t+1}}(a|s)Z_t(s)}{\pi_{w_t}(a|s)} + \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu^*}\sum_{a\in\mathcal{A}}\pi^*(a|s)(Q^i_{\pi_{w_t}}(s,a)-\bar{Q}_t^i(s,a)) \\
&\quad - \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu^*}V^i_{\pi_{w_t}}(s) \\
&= \frac{1}{\alpha}\mathbb{E}_{s\sim\nu^*}(D_{\mathrm{KL}}(\pi^*||\pi_{w_t}) - D_{\mathrm{KL}}(\pi^*||\pi_{w_{t+1}})) + \frac{1}{\alpha}\mathbb{E}_{s\sim\nu^*}\left(\log Z_t(s) - \frac{\alpha}{1-\gamma}V^i_{\pi_{w_t}}(s)\right) \\
&\quad + \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu^*}\sum_{a\in\mathcal{A}}\pi^*(a|s)(Q^i_{\pi_{w_t}}(s,a)-\bar{Q}_t^i(s,a)) \\
&\leq \frac{1}{\alpha}\mathbb{E}_{s\sim\nu^*}(D_{\mathrm{KL}}(\pi^*||\pi_{w_t}) - D_{\mathrm{KL}}(\pi^*||\pi_{w_{t+1}})) \\
&\quad + \frac{1}{\alpha}\mathbb{E}_{s\sim\nu^*}\left(\log Z_t(s) - \frac{\alpha}{1-\gamma}V^i_{\pi_{w_t}}(s) + \frac{\alpha}{1-\gamma}\sum_{a\in\mathcal{A}}\pi_{w_t}(a|s)\left|\bar{Q}_t^i(s,a)-Q^i_{\pi_{w_t}}(s,a)\right|\right)
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu^*}\sum_{a\in\mathcal{A}}\pi^*(a|s)(Q^i_{\pi_{w_t}}(s,a)-\bar{Q}^i_t(s,a))\\
&\overset{(ii)}{\leq}\frac{1}{\alpha}\mathbb{E}_{s\sim\nu^*}(D_{\mathrm{KL}}(\pi^*||\pi_{w_t})-D_{\mathrm{KL}}(\pi^*||\pi_{w_{t+1}}))\\
&+\frac{1}{1-\gamma}(J_i^{\nu^*}(w_{t+1})-J_i^{\nu^*}(w_t))+\frac{1}{(1-\gamma)^2}\mathbb{E}_{s\sim\nu_{\nu^*}}\sum_{a\in\mathcal{A}}\pi_{w_{t+1}}(a|s)\left|Q^i_{\pi_{w_t}}(s,a)-\bar{Q}^i_t(s,a)\right|\\
&+\frac{1}{(1-\gamma)^2}\mathbb{E}_{s\sim\nu_{\nu^*}}\sum_{a\in\mathcal{A}}\pi_{w_t}(a|s)\left|Q^i_{\pi_{w_t}}(s,a)-\bar{Q}^i_t(s,a)\right|\\
&+\frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu^*}\sum_{a\in\mathcal{A}}\pi^*(a|s)\left|Q^i_{\pi_{w_t}}(s,a)-\bar{Q}^i_t(s,a)\right|\\
&\overset{(iii)}{\leq}\frac{1}{\alpha}\mathbb{E}_{s\sim\nu^*}(D_{\mathrm{KL}}(\pi^*||\pi_{w_t})-D_{\mathrm{KL}}(\pi^*||\pi_{w_{t+1}}))+\frac{2c_{\max}}{(1-\gamma)^2}\left\|w_{t+1}-w_t\right\|_2+\frac{3}{(1-\gamma)^2}\left\|Q^i_{\pi_{w_t}}-\bar{Q}^i_t\right\|_2\\
&=\frac{1}{\alpha}\mathbb{E}_{s\sim\nu^*}(D_{\mathrm{KL}}(\pi^*||\pi_{w_t})-D_{\mathrm{KL}}(\pi^*||\pi_{w_{t+1}}))+\frac{2\alpha c_{\max}}{(1-\gamma)^2}\left\|\bar{Q}^i_t\right\|_2+\frac{3}{(1-\gamma)^2}\left\|Q^i_{\pi_{w_t}}-\bar{Q}^i_t\right\|_2\\
&\leq\frac{1}{\alpha}\mathbb{E}_{s\sim\nu^*}(D_{\mathrm{KL}}(\pi^*||\pi_{w_t})-D_{\mathrm{KL}}(\pi^*||\pi_{w_{t+1}}))+\frac{2\alpha c_{\max}}{(1-\gamma)^2}\left\|Q^i_{\pi_{w_t}}\right\|_2+\frac{3(1+\alpha c_{\max})}{(1-\gamma)^2}\left\|Q^i_{\pi_{w_t}}-\bar{Q}^i_t\right\|_2\\
&\leq\frac{1}{\alpha}\mathbb{E}_{s\sim\nu^*}(D_{\mathrm{KL}}(\pi^*||\pi_{w_t})-D_{\mathrm{KL}}(\pi^*||\pi_{w_{t+1}}))+\frac{2\alpha c^2_{\max}|\mathcal{S}|\,|\mathcal{A}|}{(1-\gamma)^3}+\frac{3(1+\alpha c_{\max})}{(1-\gamma)^2}\left\|Q^i_{\pi_{w_t}}-\bar{Q}^i_t\right\|_2,
\end{aligned}
$$

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 $J_i^{\nu^*}(w_{t+1})-J_i^{\nu^*}(w_t)\leq\frac{2c_{\max}}{1-\gamma}\left\|w_{t+1}-w_t\right\|_2$, which is proved by Proposition 1 in (Xu et al., 2020b). □

Note that if we follow the update in line 10 of Algorithm 1, we can obtain the following result for the case $i\in\{1,\cdots,p\}$ as stated in Lemma 7:

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

Lemma 8. *Considering CRPO in Algorithm 1 in the tabular setting. Let $K_{in}=\Theta(T^{1/\sigma}\log^{2/\sigma}(|\mathcal{S}|^2|\mathcal{A}|^2T^{1+2/\sigma}/\delta))$. Define $\mathcal{N}_i$ as the set of steps that CRPO algorithm chooses to minimize the i-th constraint. With probability at least $1-\delta$, we have*

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

Proof. If $t\in\mathcal{N}_0$, by Lemma 7 we have

$$
\begin{aligned}
&\alpha(J_0(\pi^*)-J_0(\pi_{w_t}))\\
&\leq\mathbb{E}_{s\sim\nu^*}(D_{\mathrm{KL}}(\pi^*||\pi_{w_t})-D_{\mathrm{KL}}(\pi^*||\pi_{w_{t+1}}))+\frac{2\alpha^2c^2_{\max}|\mathcal{S}|\,|\mathcal{A}|}{(1-\gamma)^3}+\frac{3\alpha(1+\alpha c_{\max})}{(1-\gamma)^2}\left\|Q^0_{\pi_{w_t}}-\bar{Q}^0_t\right\|_2.
\end{aligned}
\tag{12}
$$

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

$$
\begin{aligned}
&\alpha(J_i(\pi_{w_t})-J_i(\pi^*))\\
&\leq\mathbb{E}_{s\sim\nu^*}(D_{\mathrm{KL}}(\pi^*||\pi_{w_t})-D_{\mathrm{KL}}(\pi^*||\pi_{w_{t+1}}))+\frac{2\alpha^2c^2_{\max}|\mathcal{S}|\,|\mathcal{A}|}{(1-\gamma)^3}+\frac{3\alpha(1+\alpha c_{\max})}{(1-\gamma)^2}\left\|Q^i_{\pi_{w_t}}-\bar{Q}^i_t\right\|_2.
\end{aligned}
\tag{13}
$$

Taking the summation of eq. (12) and eq. (13) from $t=0$ to $T-1$ yields

$$\begin{aligned}
&\alpha \sum_{t\in\mathcal{N}_0}\left(J_0(\pi^*)-J_0(\pi_{w_t})\right)+\alpha\sum_{i=1}^{p}\sum_{t\in\mathcal{N}_i}\left(J_i(\pi_{w_t})-J_i(\pi^*)\right)\\
&\leq \mathbb{E}_{s\sim\nu^*}D_{\mathrm{KL}}(\pi^*||\pi_{w_0})+\frac{2\alpha^2c_{\max}^2\left|\mathcal{S}\right|\left|\mathcal{A}\right|T}{(1-\gamma)^3}+\frac{3\alpha(1+\alpha c_{\max})}{(1-\gamma)^2}\sum_{i=0}^{p}\sum_{t\in\mathcal{N}_i}\left\|Q^i_{\pi_{w_t}}-\bar{Q}^i_t\right\|_2. \qquad (14)
\end{aligned}$$

Note that when $t\in\mathcal{N}_i$ ($i\neq 0$), we have $\bar{J}_i(\theta^i_t)>d_i+\eta$ (line 9 in Algorithm 1), which implies that

$$\begin{aligned}
J_i(\pi_{w_t})-J_i(\pi^*)&\geq \bar{J}_i(\theta^i_t)-J_i(\pi^*)-\left|\bar{J}_i(\theta^i_t)-J_i(\pi_{w_t})\right|\\
&\geq d_i+\eta-J_i(\pi^*)-\left|\bar{J}_i(\theta^i_t)-J_i(\pi_{w_t})\right|\\
&\geq \eta-\left\|Q^i_{\pi_w}-\bar{Q}^i_t\right\|_2. \qquad (15)
\end{aligned}$$

Substituting eq. (15) into eq. (14) yields

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

which implies

$$\begin{aligned}
&\alpha \sum_{t\in\mathcal{N}_0}\left(J_0(\pi^*)-J_0(\pi_{w_t})\right)+\alpha\eta\sum_{i=1}^{p}\left|\mathcal{N}_i\right|\\
&\leq \mathbb{E}_{s\sim\nu^*}D_{\mathrm{KL}}(\pi^*||\pi_{w_0})+\frac{2\alpha^2c_{\max}^2\left|\mathcal{S}\right|\left|\mathcal{A}\right|T}{(1-\gamma)^3}+\frac{\alpha(2+(1-\gamma)^2+3\alpha c_{\max})}{(1-\gamma)^2}\sum_{i=0}^{p}\sum_{t\in\mathcal{N}_i}\left\|Q^i_{\pi_{w_t}}-\bar{Q}^i_t\right\|_2. \qquad (16)
\end{aligned}$$

By Lemma 2, we have with probability at least $1-\delta$, the following holds

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

Thus, if we let

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

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

$$\left\|Q^i_{\pi_{w_t}}-\bar{Q}^i_t\right\|_2\leq\frac{\sqrt{(1-\gamma)\left|\mathcal{S}\right|\left|\mathcal{A}\right|}}{\sqrt{T}}. \qquad (17)$$

Applying the union bound to eq. (17) from $t=0$ to $T-1$, we have with probability at least $1-\delta$ the following holds

$$\sum_{i=0}^{p}\sum_{t\in\mathcal{N}_i}\left\|Q^i_{\pi_{w_t}}-\bar{Q}^i_t\right\|_2\leq\sqrt{(1-\gamma)\left|\mathcal{S}\right|\left|\mathcal{A}\right|T}, \qquad (18)$$

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

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

which completes the proof. □

Lemma 9. *If*

$$\frac{1}{2}\alpha\eta T \geq \mathbb{E}_{s\sim\nu^*} D_{KL}(\pi^*||\pi_{w_0}) + \frac{2\alpha^2 c_{\max}^2 |\mathcal{S}|\,|\mathcal{A}|\,T}{(1-\gamma)^3} + \frac{\alpha\sqrt{|\mathcal{S}|\,|\mathcal{A}|\,T}(2+(1-\gamma)^2+3\alpha c_{\max})}{(1-\gamma)^{1.5}}, \tag{19}$$

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 hold,*
   (a) $|\mathcal{N}_0| \geq T/2$,
   (b) $\sum_{t\in\mathcal{G}}(J_0(\pi^*) - J_0(w_t)) \leq 0$.

Proof. We prove Lemma 9 in the event that eq. (18) holds, which happens with probability at least $1-\delta$. Under such an event, the following inequality holds, which is also the result of Lemma 8.

$$\begin{aligned}
&\alpha \sum_{t\in\mathcal{N}_0} (J_0(\pi^*) - J_0(\pi_{w_t})) + \alpha\eta \sum_{i=1}^{p} |\mathcal{N}_i| \\
&\leq \mathbb{E}_{s\sim\nu^*} D_{\mathrm{KL}}(\pi^*||\pi_{w_0}) + \frac{2\alpha^2 c_{\max}^2 |\mathcal{S}|\,|\mathcal{A}|\,T}{(1-\gamma)^3} + \frac{\alpha\sqrt{|\mathcal{S}|\,|\mathcal{A}|\,T}(2+(1-\gamma)^2+2\alpha c_{\max})}{(1-\gamma)^{1.5}}.
\end{aligned} \tag{20}$$

We first verify item 1. If $\mathcal{N}_0 = \emptyset$, then $\sum_{i=1}^{p} |\mathcal{N}_i| = T$, and eq. (20) implies that

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

which contradicts eq. (19). Thus, we must have $\mathcal{N}_0 \neq \emptyset$.

We then proceed to verify item 2. If $\sum_{t\in\mathcal{N}_0}(J_0(\pi^*) - J_0(w_t)) \leq 0$, then (b) in item 2 holds. If $\sum_{t\in\mathcal{N}_0}(J_0(\pi^*) - J_0(w_t)) \leq 0$, then eq. (20) implies that

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

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

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

which contradicts eq. (19). Hence, (a) in item 2 holds. □

B.2. Proof of Theorem 1

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 $\alpha = (1-\gamma)^{1.5}/\sqrt{|\mathcal{S}|\,|\mathcal{A}|\,T}$, $\eta = \frac{2\sqrt{|\mathcal{S}||\mathcal{A}|}}{(1-\gamma)^{1.5}\sqrt{T}}(3 + \mathbb{E}_{s\sim\nu^*} D_{KL}(\pi^*||\pi_{w_0}) + 3c_{\max} + c_{\max}^2)$, 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 2 hold. Then, with probability at least $1-\delta$*, we have*

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

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

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

To prove Theorem 1 (or Theorem 3), we still consider the following event given in eq. (18) that happens with probability at least $1-\delta$:

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

which implies

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

We first consider the convergence rate of the objective function. Under the above event, the following holds

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

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

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

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

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

$$\leq \eta + \frac{2}{T}\sum_{i=0}^{p}\sum_{t\in\mathcal{N}_i}\left\|Q^i_{\pi_{w_t}} - \bar{Q}^i_t\right\|_2.$$

Under the event defined in eq. (18), we have $\sum_{i=0}^{p}\sum_{t\in\mathcal{N}_i}\left\|Q^i_{\pi_{w_t}} - \bar{Q}^i_t\right\|_2 \leq \sqrt{(1-\gamma)\,|\mathcal{S}|\,|\mathcal{A}|\,T}$. Recall the value of the tolerance $\eta = \frac{2\sqrt{|\mathcal{S}||\mathcal{A}|}}{(1-\gamma)^{1.5}\sqrt{T}}(3 + \mathbb{E}_{s\sim\nu^*} D_{\mathrm{KL}}(\pi^*||\pi_{w_0}) + 3c_{\max} + c^2_{\max})$. With probability at least $1-\delta$, we have

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

C. Proof of Lemma 1 and Theorem 2: Function Approximation Setting

For notation simplicity, we denote the state action pairs (s,a) and (s',a') as x and x', respectively. We define the weighted norm $\|f\|_{\mathcal{D}} = \sqrt{\int f(x)^2 d\mathcal{D}(x)}$ for any distribution $\mathcal{D}$ over $|\mathcal{S}|\times|\mathcal{A}|$. We will write θ^i_k as θ_k whenever there is no confusion in this subsection. We define

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

as the local linearizion of $f(x,\theta)$ at the initial point θ_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 the 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 θ^* satisfies $\bar{g}_0(\theta)^\top(\theta-\theta^*) \geq$ for any $\theta\in\boldsymbol{B}$. We define the following function spaces

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

and

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

and define $f_0(x,\theta^*_\pi)$ 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 Proof of Lemma 1

We provide the proof of supporting lemmas for Lemma 1.

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*

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

where $\mathcal{D}$ is any distribution over $\mathcal{S}\times\mathcal{A}$.

For the following Lemma 11 and Lemma 12, we provide slightly different proofs from those in (Cai et al., 2019), which are included here for completeness.

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

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

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

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

which further implies

$$\left|\mathbb{1}\left(\theta_{k,r}^\top \psi(x) > 0\right) - \mathbb{1}\left(\theta_{0,r}^\top \psi(x) > 0\right)\right| \leq \mathbb{1}(\left|\theta_{0,r}^\top \psi(x)\right| \leq \left\|\theta_{k,r} - \theta_{0,r}\right\|_2). \tag{21}$$

Then, we can derive the following upper bound

$$\begin{aligned}
&\mathbb{E}_{\mu_\pi}\left[\frac{1}{m}\sum_{r=1}^m \left|\mathbb{1}\left(\theta_{k,r}^\top \psi(x) > 0\right) - \mathbb{1}\left(\theta_{0,r}^\top \psi(x) > 0\right)\right|\right] \\
&\leq \mathbb{E}_{\mu_\pi}\left[\frac{1}{m}\sum_{r=1}^m \mathbb{1}(\left|\theta_{0,r}^\top \psi(x)\right| \leq \left\|\theta_{k,r} - \theta_{0,r}\right\|_2)\right] \qquad (22) \\
&= \frac{1}{m}\sum_{r=1}^m \boldsymbol{P}_{\mu_\pi}(\left|\theta_{0,r}^\top \psi(x)\right| \leq \left\|\theta_{k,r} - \theta_{0,r}\right\|_2) \\
&\overset{(i)}{\leq} \frac{C_0}{m}\sum_{r=1}^m \frac{\left\|\theta_{k,r} - \theta_{0,r}\right\|_2}{\left\|\theta_{0,r}\right\|_2} \\
&\leq \frac{C_0}{m}\left(\sum_{r=1}^m \left\|\theta_{k,r} - \theta_{0,r}\right\|_2^2\right)^{1/2}\left(\sum_{r=1}^m \frac{1}{\left\|\theta_{0,r}\right\|_2^2}\right)^{1/2} \\
&\overset{(ii)}{\leq} \frac{C_0 R}{d_1\sqrt{m}}. \qquad (23)
\end{aligned}$$

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

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

$$\mathbb{E}_{\mu_\pi}\left[\left|f((s,a);\theta_k) - f_0((s,a);\theta_k)\right|^2\right] \leq \frac{4C_0R^3}{d_1\sqrt{m}}.$$

Proof. By definition, we have

$$\begin{aligned}
&\left|f((s,a);\theta_t) - f_0((s,a);\theta_t)\right| \\
&= \frac{1}{\sqrt{m}}\left|\sum_{r=1}^m \left(\mathbb{1}(\theta_{k,r}^\top \psi(x) > 0) - \mathbb{1}(\theta_{0,r}^\top \psi(x) > 0)\right) b_r \theta_{k,r}^\top \psi(x)\right| \\
&\leq \frac{1}{\sqrt{m}}\sum_{r=1}^m \left|\left(\mathbb{1}(\theta_{k,r}^\top \psi(x) > 0) - \mathbb{1}(\theta_{0,r}^\top \psi(x) > 0)\right)\right| \left|b_r\right| \left\|\theta_{k,r}^\top \psi(x)\right\|_2 \\
&\overset{(i)}{\leq} \frac{1}{\sqrt{m}}\sum_{r=1}^m \mathbb{1}(\left|\theta_{0,r}^\top \psi(x)\right| \leq \left\|\theta_{k,r} - \theta_{0,r}\right\|_2) \left\|\theta_{k,r}^\top \psi(x)\right\|_2 \\
&\leq \frac{1}{\sqrt{m}}\sum_{r=1}^m \mathbb{1}(\left|\theta_{0,r}^\top \psi(x)\right| \leq \left\|\theta_{k,r} - \theta_{0,r}\right\|_2)\left(\left\|\theta_{0,r} - \theta_{k,r}\right\|_2 + \left\|\theta_{0,r}^\top \psi(x)\right\|_2\right) \\
&\leq \frac{2}{\sqrt{m}}\sum_{r=1}^m \mathbb{1}(\left|\theta_{0,r}^\top \psi(x)\right| \leq \left\|\theta_{k,r} - \theta_{0,r}\right\|_2) \left\|\theta_{0,r} - \theta_{k,r}\right\|_2. \qquad (24)
\end{aligned}$$

where (i) follows from eq. (21). We can then obtain the following upper bound.

$$\mathbb{E}_{\mu_\pi}\left[\left|f((s,a);\theta_t) - f_0((s,a);\theta_t)\right|^2\right]$$

$$
\begin{aligned}
&\leq \frac{4}{m}\mathbb{E}_{\mu_\pi}\left[\left(\sum_{r=1}^{m}\mathbb{1}(\left|\theta_{0,r}^\top\psi(x)\right| \leq \|\theta_{k,r}-\theta_{0,r}\|_2)\,\|\theta_{0,r}-\theta_{k,r}\|_2\right)^2\right]\\
&\overset{(i)}{\leq} \frac{4}{m}\mathbb{E}_{\mu_\pi}\left[\sum_{r=1}^{m}\mathbb{1}(\left|\theta_{0,r}^\top\psi(x)\right| \leq \|\theta_{k,r}-\theta_{0,r}\|_2)\sum_{r=1}^{m}\|\theta_{0,r}-\theta_{k,r}\|_2^2\right]\\
&= \frac{4R^2}{m}\mathbb{E}_{\mu_\pi}\left[\sum_{r=1}^{m}\mathbb{1}(\left|\theta_{0,r}^\top\psi(x)\right| \leq \|\theta_{k,r}-\theta_{0,r}\|_2)\right]\\
&\overset{(ii)}{\leq} \frac{4C_0R^3}{d_1\sqrt{m}},
\end{aligned}
\tag{25}
$$

where (i) follows from Holder's inequality, and (ii) follows from the derivation in Lemma 11 after eq. (22). □

Lemma 13. *Suppose Assumption 1 holds. For any policy π 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

$$
\begin{aligned}
&\|\bar{g}_k(\theta_k) - \bar{g}_0(\theta_k)\|_2\\
&= \|\mathbb{E}_{\mu_\pi}[\delta_k(x,x'.\theta_k)\nabla_\theta f(x,\theta_k)] - \mathbb{E}_{\mu_\pi}[\delta_0(x,x'.\theta_k)\nabla_\theta f_0(x,\theta_k)]\|_2\\
&= \|\mathbb{E}_{\mu_\pi}[(\delta_k(x,x'.\theta_k)-\delta_0(x,x'.\theta_k))\,\nabla_\theta f(x,\theta_k) + \delta_0(x,x'.\theta_k)\,(\nabla_\theta f(x,\theta_k)-\nabla_\theta f_0(x,\theta_k))]\|_2\\
&\leq \mathbb{E}_{\mu_\pi}[|\delta_k(x,x'.\theta_k)-\delta_0(x,x'.\theta_k)|\,\|\nabla_\theta f(x,\theta_k)\|_2 + |\delta_0(x,x'.\theta_k)|\,\|\nabla_\theta f(x,\theta_k)-\nabla_\theta f_0(x,\theta_k)\|_2]\\
&\overset{(i)}{\leq} \mathbb{E}_{\mu_\pi}[|\delta_k(x,x'.\theta_k)-\delta_0(x,x'.\theta_k)|] + \mathbb{E}_{\mu_\pi}[|\delta_0(x,x'.\theta_k)|\,\|\nabla_\theta f(x,\theta_k)-\nabla_\theta f_0(x,\theta_k)\|_2],
\end{aligned}
\tag{26}
$$

where (i) follows from the fact that $\|\nabla_\theta f(x,\theta_k)\|_2 \leq 1$. Then, eq. (26) implies that

$$
\begin{aligned}
&\|\bar{g}_k(\theta_k) - \bar{g}_0(\theta_k)\|_2^2\\
&\leq 2\mathbb{E}_{\mu_\pi}[|\delta_k(x,x'.\theta_k)-\delta_0(x,x'.\theta_k)|^2] + 2\left(\mathbb{E}_{\mu_\pi}[|\delta_0(x,x'.\theta_k)|\,\|\nabla_\theta f(x,\theta_k)-\nabla_\theta f_0(x,\theta_k)\|_2]\right)^2\\
&\leq 2\mathbb{E}_{\mu_\pi}[|\delta_k(x,x'.\theta_k)-\delta_0(x,x'.\theta_k)|^2] + 2\mathbb{E}_{\mu_\pi}[|\delta_0(x,x'.\theta_k)|^2]\mathbb{E}_{\mu_\pi}[\|\nabla_\theta f(x,\theta_k)-\nabla_\theta f_0(x,\theta_k)\|_2^2].
\end{aligned}
\tag{27}
$$

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

$$
\begin{aligned}
&\mathbb{E}_{\mu_\pi}[|\delta_k(x,x'.\theta_k)-\delta_0(x,x'.\theta_k)|^2]\\
&\leq 2\mathbb{E}_{\mu_\pi}[|f(x,\theta_k)-f_0(x,\theta_k)|^2] + 2\mathbb{E}_{\mu_\pi}[|f(x',\theta_k)-f_0(x',\theta_k)|^2]\\
&= 4\mathbb{E}_{\mu_\pi}[|f(x,\theta_k)-f_0(x,\theta_k)|^2]\\
&\overset{(i)}{\leq} \frac{16C_0R^2}{d_1\sqrt{m}},
\end{aligned}
\tag{28}
$$

where (i) follows from Lemma 12. 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

$$
\|\nabla_\theta f(x,\theta_k)-\nabla_\theta f_0(x,\theta_k)\|_2
$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{m}} \left\| \sum_{r=1}^{m} \left[\mathbb{1}\left(\theta_{k,r}^{\top}\psi(x) > 0\right) - \mathbb{1}\left(\theta_{0,r}^{\top}\psi(x) > 0\right) \right] b_r \theta_{0,r}^{\top}\psi(x) \right\|_2 \\
&\overset{(i)}{\leq} \frac{1}{\sqrt{m}} \sum_{r=1}^{m} \left| \mathbb{1}\left(\theta_{k,r}^{\top}\psi(x) > 0\right) - \mathbb{1}\left(\theta_{0,r}^{\top}\psi(x) > 0\right) \right| \left\|\theta_{0,r}\right\|_2 \\
&\overset{(ii)}{\leq} \frac{1}{\sqrt{m}} \sum_{r=1}^{m} \mathbb{1}(\left|\theta_{0,r}^{\top}\psi(x)\right| \leq \left\|\theta_{k,r} - \theta_{0,r}\right\|_2) \left\|\theta_{0,r}\right\|_2, \qquad (29)
\end{aligned}
$$

where (i) follows because $|b_r| \leq 1$ and $\|\psi(s)\|_2 \leq 1$, and (ii) follows from eq. (21). Further, eq. (29) implies that

$$
\begin{aligned}
&\mathbb{E}_{\mu_\pi}[\|\nabla_\theta f(x,\theta_k) - \nabla_\theta f_0(x,\theta_k)\|_2^2] \\
&\leq \frac{1}{m}\mathbb{E}_{\mu_\pi}\left[\left(\sum_{r=1}^{m} \mathbb{1}(\left|\theta_{0,r}^{\top}\psi(x)\right| \leq \left\|\theta_{k,r} - \theta_{0,r}\right\|_2) \right) \left(\sum_{r=1}^{m} \left\|\theta_{0,r}\right\|_2^2 \right) \right] \\
&\leq \frac{R^2}{m} \sum_{r=1}^{m} \mathbb{1}(\left|\theta_{0,r}^{\top}\psi(x)\right| \leq \left\|\theta_{k,r} - \theta_{0,r}\right\|_2) \\
&\overset{(i)}{\leq} \frac{C_0 R^3}{d_1\sqrt{m}}, \qquad (30)
\end{aligned}
$$

where (i) follows from the derivation in Lemma 11 after eq. (22).

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

$$
\begin{aligned}
&\mathbb{E}_{\mu_\pi}[|\delta_0(x,x'.\theta_k)|^2] \\
&\leq \mathbb{E}_{\mu_\pi}[|f_0(x,\theta_k) - r(x,x') - \gamma f_0(x',\theta_k)|^2] \\
&\leq 3\mathbb{E}_{\mu_\pi}[|f_0(x,\theta_k)|^2] + 3\mathbb{E}_{\mu_\pi}[r^2(x,x')] + 3\gamma^2\mathbb{E}_{\mu_\pi}[\|f_0(x',\theta_k)\|_2^2] \\
&\leq 6\mathbb{E}_{\mu_\pi}[|f_0(x,\theta_k)|^2] + 3c_{\max}^2 \\
&= 6\mathbb{E}_{\mu_\pi}[|f_0(x,\theta_k) - f_0(x,\theta_\pi^*) + f_0(x,\theta_\pi^*) - Q_\pi(x) + Q_\pi(x)|^2] + 3c_{\max}^2 \\
&= 18\mathbb{E}_{\mu_\pi}[|f_0(x,\theta_k) - f_0(x,\theta_\pi^*)|^2] + 18\mathbb{E}_{\mu_\pi}[|f_0(x,\theta_\pi^*) - Q_\pi(x)|^2] + 18\mathbb{E}_{\mu_\pi}[|Q_\pi(x)|^2] + 3c_{\max}^2 \\
&\overset{(i)}{\leq} 18R^2 + \frac{21c_{\max}^2}{(1-\gamma)^2} + 18\mathbb{E}_{\mu_\pi}[|f_0(x,\theta_\pi^*) - Q_\pi(x)|^2], \qquad (31)
\end{aligned}
$$

where (i) follows from the fact that $Q_\pi(x) \leq \frac{c_{\max}}{1-\gamma}$, $\|\theta_k\|_2 \leq R$ and $\|\theta_\pi^*\|_2 \leq R$.

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

$$
\mathbb{E}_{\mu_\pi}[|f_0(x,\theta_\pi^*) - Q_\pi(x)|^2] \leq \frac{4R^2\log\left(\frac{1}{\delta}\right)}{m} \leq 4R^2\log\left(\frac{1}{\delta}\right). \qquad (32)
$$

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

$$
\mathbb{E}_{\mu_\pi}[|\delta_0(x,x'.\theta_k)|^2] \leq 18R^2 + \frac{21c_{\max}^2}{(1-\gamma)^2} + 72R^2\log\left(\frac{1}{\delta}\right). \qquad (33)
$$

Combining eq. (28), eq. (30) and eq. (33), 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 completes the proof. □

C.2. Proof of Lemma 1

We consider the convergence of θ_k^i for a given i under a fixed policy π. For the iteration of θ_k, we proceed as follows.

$$
\begin{aligned}
&\|\theta_{k+1} - \theta^*\|_2^2 \\
&= \|\Pi_{\boldsymbol{B}}(\theta_k - \beta g_k(\theta_k)) - \Pi_{\boldsymbol{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 \\
&\overset{(i)}{\leq} \|\theta_k - \theta^*\|_2^2 - 2(1-\gamma)\beta \mathbb{E}_{\mu_\pi}\left[(f_0((s,a);\theta_k) - f_0((s,a);\theta^*))^2\right] \\
&\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 \\
&\overset{(ii)}{\leq} \|\theta_k - \theta^*\|_2^2 - [2\beta(1-\gamma) - 12\beta^2]\mathbb{E}_{\mu_\pi}\left[(f_0((s,a);\theta_k) - f_0((s,a);\theta^*))^2\right] \\
&\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} \tag{34}
$$

where (i) follows from the fact that

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

and (ii) follows from the fact that

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

Rearranging eq. (34) yields

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

Taking summation of eq. (35) over $t = 0$ to $K-1$ yields

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

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.

$$
\begin{aligned}
&\boldsymbol{P}_{\mu_\pi}\left(\sum_{t=0}^{K-1}\|\zeta_k(\theta_k)\|_2^2 \geq (1+\Lambda)C_\zeta^2 K\right)\\
&= \boldsymbol{P}_{\mu_\pi}\left(\frac{\sum_{t=0}^{K-1}\|\zeta_k(\theta_k)\|_2^2}{C_\zeta^2 K} \geq 1+\Lambda\right)\\
&= \boldsymbol{P}_{\mu_\pi}\left(\exp\left(\frac{\sum_{t=0}^{K-1}\|\zeta_k(\theta_k)\|_2^2}{C_\zeta^2 K}\right) \geq \exp(1+\Lambda)\right)\\
&\leq \boldsymbol{P}_{\mu_\pi}\left(\frac{1}{K}\sum_{t=0}^{K-1}\exp\left(\frac{\|\zeta_k(\theta_k)\|_2^2}{C_\zeta^2}\right) \geq \exp(1+\Lambda)\right)\\
&\overset{(i)}{\leq} \frac{1}{K}\sum_{t=0}^{K-1}\mathbb{E}_{\mu_\pi}\left[\exp\left(\frac{\|\zeta_k(\theta_k)\|_2^2}{C_\zeta^2}\right)\right]/\exp(1+\Lambda)\\
&\overset{(ii)}{\leq} \exp(-\Lambda),
\end{aligned}
\tag{36}
$$

where (i) follows from Markov's inequality, (ii) follows from Assumption 3. Then, eq. (36) implies that with probability at least $1-\delta_1$, we have

$$
\sum_{t=0}^{K-1}\|\zeta_k(\theta_k)\|_2^2 \leq \left(1+\log\left(\frac{1}{\delta_1}\right)\right)C_\zeta^2 K \leq 2\log\left(\frac{1}{\delta_1}\right)C_\zeta^2 K. \tag{37}
$$

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

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

which implies

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

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

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

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

$$
\left|\sum_{t=0}^{K-1}\zeta_k(\theta_k)^\top(\theta_k-\theta^*)\right| \leq \sqrt{2}\left(1+\sqrt{3\log\left(\frac{1}{\delta_2}\right)}\right)C_\zeta\sqrt{K} \leq 5C_\zeta\sqrt{\log\left(\frac{1}{\delta_2}\right)}\sqrt{K}. \tag{38}
$$

We then consider the terms $\sum_{t=0}^{K-1}\|\xi_k(\theta_k)\|_2$ 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 then union bound we can obtain that with probability at least $1-\delta_3$, we have

$$\sum_{t=0}^{K-1} \|\xi_k(\theta_k)\|_2 \leq \Theta\left(\frac{K\sqrt{\log(\frac{K}{\delta_3})}}{(1-\gamma)m^{1/4}}\right). \tag{39}$$

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

$$\sum_{t=0}^{K-1} \|\xi_k(\theta_k)\|_2^2 \leq \Theta\left(\frac{K\log(\frac{K}{\delta_3})}{(1-\gamma)^2m^{1/2}}\right). \tag{40}$$

Combining eq. (37), eq. (38), eq. (39) and eq. (40) and applying the union bound, we can obtain that with probability at least $1-(\delta_1+\delta_2+\delta_3+\delta_4)$, we have

$$\begin{aligned}
&[2\beta(1-\gamma)-12\beta^2]\sum_{t=0}^{K-1}\mathbb{E}_{\mu_\pi}\left[(f_0((s,a);\theta_k)-f_0((s,a);\theta^*))^2\right]\\
&\leq R^2+10\beta C_\zeta\sqrt{\log\left(\frac{1}{\delta_2}\right)}\sqrt{K}+6\beta^2\log\left(\frac{1}{\delta_1}\right)C_\zeta^2K+\beta K\Theta\left(\frac{\sqrt{\log(\frac{K}{\delta_3})}}{(1-\gamma)m^{1/4}}\right)\\
&\quad+\beta^2K\Theta\left(\frac{\log(\frac{K}{\delta_3})}{(1-\gamma)^2m^{1/2}}\right).
\end{aligned} \tag{41}$$

Divide both sides of eq. (41) 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

$$\begin{aligned}
&\left\|f_0((s,a);\bar{\theta}_K)-f_0((s,a);\theta^*)\right\|_{\mu_\pi}^2\\
&\leq\frac{1}{K}\sum_{t=0}^{K-1}\mathbb{E}_{\mu_\pi}\left[(f_0((s,a);\theta_k)-f_0((s,a);\theta^*))^2\right]\\
&\leq\frac{R^2}{[2\beta(1-\gamma)-12\beta^2]K}+\frac{10\beta C_\zeta\sqrt{\log\left(\frac{1}{\delta_2}\right)}}{[2\beta(1-\gamma)-12\beta^2]\sqrt{K}}+\frac{6\beta\log\left(\frac{1}{\delta_1}\right)C_\zeta^2}{[2\beta(1-\gamma)-12\beta^2]\sqrt{K}}\\
&\quad+\Theta\left(\frac{\sqrt{\log(\frac{K}{\delta_3})}}{(1-\gamma)m^{1/4}}\right)\frac{1}{[2\beta(1-\gamma)-12\beta^2]\sqrt{K}}\\
&\quad+\Theta\left(\frac{\log(\frac{K}{\delta_3})}{(1-\gamma)^2m^{1/2}}\right)\frac{1}{[2\beta(1-\gamma)-12\beta^2]\sqrt{K}}\\
&\leq\Theta\left(\frac{1}{(1-\gamma)^2\sqrt{K}}\right)+\Theta\left(\frac{1}{(1-\gamma)^2\sqrt{K}}\sqrt{\log\left(\frac{1}{\delta_1}\right)}\right)+\Theta\left(\frac{1}{(1-\gamma)^2\sqrt{K}}\sqrt{\log\left(\frac{1}{\delta_2}\right)}\right)\\
&\quad+\Theta\left(\frac{\sqrt{\log(\frac{K}{\delta_3})}}{(1-\gamma)^3m^{1/4}}\right)+\Theta\left(\frac{\sqrt{\log(\frac{K}{\delta_4})}}{(1-\gamma)^3m^{1/4}}\right)\\
&=\Theta\left(\frac{1}{(1-\gamma)^2\sqrt{K}}\left(\sqrt{\log\left(\frac{1}{\delta_1}\right)}+\sqrt{\log\left(\frac{1}{\delta_1}\right)}\right)\right)\\
&\quad+\Theta\left(\frac{1}{(1-\gamma)^3m^{1/4}}\left(\sqrt{\log\left(\frac{K}{\delta_3}\right)}+\sqrt{\log\left(\frac{K}{\delta_4}\right)}\right)\right).
\end{aligned} \tag{42}$$

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

$$\begin{aligned}
&\left\|f((s,a);\bar{\theta}_K) - Q_\pi(s,a)\right\|_{\mu_\pi}^2 \\
&\leq 3\left\|f((s,a);\bar{\theta}_K) - f_0((s,a);\bar{\theta}_K)\right\|_{\mu_\pi}^2 + 3\left\|f_0((s,a);\bar{\theta}_K) - f_0((s,a);\theta^*)\right\|_{\mu_\pi}^2 \\
&\quad + 3\left\|f_0((s,a);\theta^*) - Q_\pi(s,a)\right\|_{\mu_\pi}^2 \\
&\overset{(i)}{\leq} \Theta\left(\frac{1}{\sqrt{m}}\right) + 3\left\|f_0((s,a);\bar{\theta}_K) - f_0((s,a);\theta^*)\right\|_{\mu_\pi}^2 + \frac{3}{1-\gamma}\left\|f_0((s,a);\theta_\pi^*) - Q_\pi(s,a)\right\|_{\mu_\pi}^2,
\end{aligned} \tag{43}$$

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

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

which is given in (Cai et al., 2019). Then, eq. (32) implies that, with probability at least δ_5, we have

$$\left\|f_0((s,a);\theta_\pi^*) - Q_\pi(s,a)\right\|_{\mu_\pi}^2 \leq \frac{4R^2\log\left(\frac{1}{\delta_5}\right)}{m}. \tag{44}$$

Substituting eq. (42) and eq. (44) into eq. (43), we have with probability at least $1-(\delta_1+\delta_2+\delta_3+\delta_4+\delta_5)$, the following holds:

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

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:

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

which completes the proof.

C.3. Supporting Lemmas for Proof of 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:

$$\pi_W(a|s) := \frac{\exp(f((s,a);W))}{\sum_{a'\mathcal{A}}\exp(f((s,a');W))}, \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A}, \tag{45}$$

and $J_i(W)$ as the accumulated cost with policy π_W. We denote $\phi_W^i(s,a) = \nabla_W f_i((s,a),W)$. We define the diameter of $\boldsymbol{B}_W$ as R_W. When performing each NPG update, we will need to solve the linear regression problem specified in eq. (10). 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. (10). Thus, instead of solving the problem eq. (10) directly, here we simply use $\bar{\theta}_t$ as the approximated NPG update at each iteration:

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

Without loss of generality, we assume that for the visitation distribution of the global optimal policy ν^*, there exists a constants C_{RN} such that for all π_W, the following holds

$$\int_x \left(\frac{d\nu^*(x)}{d\mu_{\pi_W}(x)}\right)^2 d\mu_{\pi_W}(x) \leq C_{RN}^2. \tag{46}$$

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

$$\left\|\phi_\theta(s,a)^\top \theta' - \phi_{\theta_0}(s,a)^\top \theta'\right\|_{\mu_\pi}^2 \leq \frac{4C_0R^3}{d_1\sqrt{m}}.$$

Proof. By definition, we have

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

where (i) follows from eq. (21). Following from Holder's inequality, we obtain from eq. (47) that

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

which implies

$$\left\|\phi_\theta(s,a)^\top \theta' - \phi_{\theta_0}(s,a)^\top \theta'\right\|_{\mu_\pi}^2 = \mathbb{E}_{\mu_\pi}[\left|\phi_\theta(s,a)^\top \theta' - \phi_{\theta_0}(s,a)^\top \theta'\right|^2] \leq \frac{4C_0R^3}{d_1\sqrt{m}}, \tag{48}$$

where (i) follows from the derivation in Lemma 11 after eq. (22). □

Lemma 15 (Upper bound on optimality gap for neural NPG)**.** *Consider the approximated NPG updates in the neural network approximation setting. We have*

$$\alpha(1-\gamma)(J_0(\pi^*) - J_0(\pi_{\tau_t W_t}))$$

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

Proof. It has been verified that the feature mapping $\phi^r_W(s,a)$ is bounded (Wang et al., 2019; Cai et al., 2019). By following the argument similar to that 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^*}\left[D_{\mathrm{KL}}(\pi^*||\pi_{\tau_t W_t})\right] - \mathbb{E}_{\nu^*}\left[D_{\mathrm{KL}}(\pi^*||\pi_{\tau_{t+1} W_{t+1}})\right]\\
&= \mathbb{E}_{\nu^*}\left[\log(\pi_{\tau_{t+1} W_{t+1}}(a|s)) - \log(\pi_{\tau_t W_t}(a|s))\right]\\
&\overset{(i)}{\geq} \mathbb{E}_{\nu^*}\left[\nabla_W \log(\pi_{\tau_t W_t}(a|s))\right]^\top (\tau_{t+1}W_{t+1} - \tau_t W_t) - \frac{L_f}{2}\left\|\tau_{t+1}W_{t+1} - \tau_t W_t\right\|_2^2\\
&= \alpha\mathbb{E}_{\nu^*}\left[\nabla_W \log(\pi_{\tau_t W_t}(a|s))\right]^\top \bar{\theta}_t - \frac{\alpha^2 L_f}{2}\left\|\bar{\theta}_t\right\|_2^2\\
&= \alpha\mathbb{E}_{\nu^*}\left[\phi_{W_t}(s,a) - \mathbb{E}_{\pi_{\tau_t W_t}}\left[\phi_{W_t}(s,a')\right]\right]^\top \bar{\theta}_t - \frac{\alpha^2 L_f}{2}\left\|\bar{\theta}_t\right\|_2^2\\
&= \alpha\mathbb{E}_{\nu^*}\left[Q_{\pi_{\tau_t W_t}}(s,a) - \mathbb{E}_{\pi_{\tau_t W_t}}\left[Q_{\pi_{\tau_t W_t}}(s,a')\right]\right] + \alpha\mathbb{E}_{\nu^*}\left[\phi_{W_t}(s,a)^\top\bar{\theta}_t - Q_{\pi_{\tau_t W_t}}(s,a)\right]\\
&\quad + \alpha\mathbb{E}_{\nu^*}\mathbb{E}_{\pi_{\tau_t W_t}}\left[Q_{\pi_{\tau_t W_t}}(s,a') - \phi_{W_t}(s,a')^\top\bar{\theta}_t\right] - \frac{\alpha^2 L_f}{2}\left\|\bar{\theta}_t\right\|_2^2\\
&= \alpha(1-\gamma)(J_0(\pi^*) - J_0(\pi_{\tau_t W_t})) + \alpha\mathbb{E}_{\nu^*}\left[\phi_{W_t}(s,a)^\top\bar{\theta}_t - f((s,a),\bar{\theta}_t)\right]\\
&\quad + \alpha\mathbb{E}_{\nu^*}\left[f((s,a),\bar{\theta}_t) - Q_{\pi_{\tau_t W_t}}(s,a)\right] + \alpha\mathbb{E}_{\nu^*}\mathbb{E}_{\pi_{\tau_t W_t}}\left[Q_{\pi_{\tau_t W_t}}(s,a') - f((s,a'),\bar{\theta}_t)\right]\\
&\quad + \alpha\mathbb{E}_{\nu^*}\mathbb{E}_{\pi_{\tau_t W_t}}\left[f((s,a'),\bar{\theta}_t) - \phi_{W_t}(s,a')^\top\bar{\theta}_t\right] - \frac{\alpha^2 L_f}{2}\left\|\bar{\theta}_t\right\|_2^2\\
&= \alpha(1-\gamma)(J_0(\pi^*) - J_0(\pi_{\tau_t W_t})) + \alpha\mathbb{E}_{\nu^*}\left[\phi_{W_t}(s,a)^\top\bar{\theta}_t - f((s,a),\bar{\theta}_t)\right]\\
&\quad + \alpha\mathbb{E}_{\nu^*}\left[f((s,a),\bar{\theta}_t) - Q_{\pi_{\tau_t W_t}}(s,a)\right] + \alpha\mathbb{E}_{\nu^*}\mathbb{E}_{\pi_{\tau_t W_t}}\left[Q_{\pi_{\tau_t W_t}}(s,a') - f((s,a'),\bar{\theta}_t)\right]\\
&\quad + \alpha\mathbb{E}_{\nu^*}\mathbb{E}_{\pi_{\tau_t W_t}}\left[f((s,a'),\bar{\theta}_t) - \phi_{W_t}(s,a')^\top\bar{\theta}_t\right] - \frac{\alpha^2 L_f}{2}\left\|\bar{\theta}_t\right\|_2^2\\
&= \alpha(1-\gamma)(J_0(\pi^*) - J_0(\pi_{\tau_t W_t})) + \alpha\mathbb{E}_{\nu^*}\mathbb{E}_{\bar{\theta}_t}\left[\phi_{W_t}(s,a)^\top\bar{\theta}_t - f((s,a),\bar{\theta}_t)\right]\\
&\quad + \alpha\mathbb{E}_{\nu^*}\left[f((s,a),\bar{\theta}_t) - Q_{\pi_{\tau_t W_t}}(s,a)\right] + \alpha\mathbb{E}_{\nu^*}\mathbb{E}_{\pi_{\tau_t W_t}}\left[Q_{\pi_{\tau_t W_t}}(s,a') - f((s,a'),\bar{\theta}_t)\right]\\
&\quad + \alpha\mathbb{E}_{\nu^*}\mathbb{E}_{\pi_{\tau_t W_t}}\left[f((s,a'),\bar{\theta}_t) - \phi_{W_t}(s,a')^\top\bar{\theta}_t\right] - \frac{\alpha^2 L_f}{2}\left\|\bar{\theta}_t\right\|_2^2\\
&\geq \alpha(1-\gamma)(J_0(\pi^*) - J_0(\pi_{\tau_t W_t})) - \alpha\sqrt{\mathbb{E}_{\nu^*}\left[\left(\phi_{W_t}(s,a)^\top\bar{\theta}_t - f((s,a),\bar{\theta}_t)\right)^2\right]}\\
&\quad - \alpha\sqrt{\mathbb{E}_{\nu^*}\left[(f((s,a),\bar{\theta}_t) - Q_{\pi_{\tau_t W_t}}(s,a))^2\right]} - \alpha\sqrt{\mathbb{E}_{\nu^*}\mathbb{E}_{\pi_{\tau_t W_t}}\left[(Q_{\pi_{\tau_t W_t}}(s,a') - f((s,a'),\bar{\theta}_t))^2\right]}\\
&\quad - \alpha\sqrt{\mathbb{E}_{\nu^*}\mathbb{E}_{\pi_{\tau_t W_t}}\left[(f((s,a'),\bar{\theta}_t) - \phi_{W_t}(s,a')^\top\bar{\theta}_t)^2\right]} - \frac{\alpha^2 L_f}{2}\left\|\bar{\theta}_t\right\|_2^2, \qquad (49)
\end{aligned}$$

where (i) follows from the L_f-Lipschitz property of $\log(\pi_w(a|s))$. Note that for any $x \sim \nu_{\pi_W}$, and any function $h(x)$, we have

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

where (i) follows from Holder's inequality, and (ii) follows from eq. (46). Similarly, we can obtain

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

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

$$\begin{aligned}
&\mathbb{E}_{\nu^*}\left[D_{\text{KL}}(\pi^*||\pi_{\tau_t W_t})\right] - \mathbb{E}_{\nu^*}\left[D_{\text{KL}}(\pi^*||\pi_{\tau_{t+1} W_{t+1}})\right] \\
&\geq \alpha(1-\gamma)(J_0(\pi^*) - J_0(\pi_{\tau_t W_t})) - \alpha C_{RN}\sqrt{\mathbb{E}_{\nu_{\pi_{\tau_t W_t}}}\left[\left(\phi_{W_t}(s,a)^\top \bar{\theta}_t - f((s,a),\bar{\theta}_t)\right)^2\right]} \\
&\quad - \alpha C_{RN}\sqrt{\mathbb{E}_{\mu_{\pi_{\tau_t W_t}}}\left[(f((s,a),\bar{\theta}_t) - Q_{\pi_{\tau_t W_t}}(s,a))^2\right]} \\
&\quad - \alpha C_{RN}\sqrt{\mathbb{E}_{\mu_{\pi_{\tau_t W_t}}}\left[(Q_{\pi_{\tau_t W_t}}(s,a') - f((s,a'),\bar{\theta}_t))^2\right]} \\
&\quad - \alpha C_{RN}\sqrt{\mathbb{E}_{\mu_{\pi_{\tau_t W_t}}}\left[(f((s,a'),\bar{\theta}_t) - \phi_{W_t}(s,a')^\top\bar{\theta}_t)^2\right]} - \alpha^2 L_f(R^2 + md_2^2) \\
&= \alpha(1-\gamma)(J_0(\pi^*) - J_0(\pi_{\tau_t W_t})) - 2\alpha C_{RN}\sqrt{\mathbb{E}_{\mu_{\pi_{\tau_t W_t}}}\left[\left(\phi_{W_t}(s,a)^\top \bar{\theta}_t - f((s,a),\bar{\theta}_t)\right)^2\right]} \\
&\quad - 2\alpha C_{RN}\sqrt{\mathbb{E}_{\mu_{\pi_{\tau_t W_t}}}\left[(f((s,a),\bar{\theta}_t) - Q_{\pi_{\tau_t W_t}}(s,a))^2\right]} - \alpha^2 L_f(R^2 + md_2^2) \\
&= \alpha(1-\gamma)(J_0(\pi^*) - J_0(\pi_{\tau_t W_t})) - 2\alpha C_{RN}\left\|\phi_{W_t}(s,a)^\top \bar{\theta}_t - f((s,a),\bar{\theta}_t)\right\|_{\mu_{\pi_{\tau_t W_t}}} \\
&\quad - 2\alpha C_{RN}\left\|f((s,a),\bar{\theta}_t) - Q_{\pi_{\tau_t W_t}}(s,a)\right\|_{\mu_{\pi_{\tau_t W_t}}} - \alpha^2 L_f(R^2 + md_2^2).
\end{aligned} \tag{52}$$

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

$$\begin{aligned}
&\left\|\phi_{W_t}(s,a)^\top \bar{\theta}_t - f((s,a),\bar{\theta}_t)\right\|^2_{\mu_{\pi_{\tau_t W_t}}} \\
&= \left\|\phi_{W_t}(s,a)^\top \bar{\theta}_t - \phi_{W_0}(s,a)^\top \bar{\theta}_t + \phi_{W_0}(s,a)^\top \bar{\theta}_t - f((s,a),\bar{\theta}_t)\right\|^2_{\mu_{\pi_{\tau_t W_t}}} \\
&\leq 2\left\|\phi_{W_t}(s,a)^\top \bar{\theta}_t - \phi_{W_0}(s,a)^\top \bar{\theta}_t\right\|^2_{\mu_{\pi_{\tau_t W_t}}} + 2\left\|\phi_{W_0}(s,a)^\top \bar{\theta}_t - f((s,a),\bar{\theta}_t)\right\|^2_{\mu_{\pi_{\tau_t W_t}}} \\
&\overset{(i)}{\leq} \frac{16C_0R^3}{d_1\sqrt{m}},
\end{aligned} \tag{53}$$

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

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

Rearranging the above inequality yields the desired result. □

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

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

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*

$$\begin{aligned}
&\alpha(1-\gamma)\sum_{t\in\mathcal{N}_0}(J_0(\pi^*) - J_0(\pi_{w_t})) + \alpha(1-\gamma)\eta\sum_{i=1}^{p}|\mathcal{N}_i| \\
&\leq \mathbb{E}_{s\sim\nu^*} D_{KL}(\pi^*||\pi_{w_0}) + C_3\left(\frac{\alpha T}{m^{1/4}}\right) + C_4(\alpha^2 mT) \\
&\quad + C_5\left(\frac{\alpha T}{(1-\gamma)^{1.5}m^{1/8}}\log^{\frac{1}{4}}\left(\frac{T^3}{\delta}\right)\right) + C_6\left(\alpha(1-\gamma)\sqrt{T}\right).
\end{aligned}$$

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

Proof. We define $\mathcal{N}_i$ 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

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

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

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

Taking summation of eq. (12) and eq. (13) from $t=0$ to $T-1$ yields

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

Note that when $t\in\mathcal{N}_i$ ($i\neq 0$), we have $\bar{J}_i(\theta_t^i) > d_i + \eta$ (line 9 in Algorithm 1), which implies that

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

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

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

$$\overset{(i)}{\leq} \left|\bar{J}_i(\theta_t^i) - \mathbb{E}_{\nu_{\pi_{\tau_t W_t}}}[f_i((s,a),\bar{\theta}_t)]\right| + C_{RN}\left\|f_i((s,a),\bar{\theta}_t) - Q^i_{\pi_{\tau_t W_t}}(s,a)\right\|_{\mu_{\pi_{\tau_t W_t}}}, \tag{58}$$

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

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

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

$$\begin{aligned}
&\alpha(1-\gamma)\sum_{t\in\mathcal{N}_0}(J_0(\pi^*) - J_0(\pi_{w_t})) + \alpha(1-\gamma)\eta\sum_{i=1}^{p}|\mathcal{N}_i| \\
&\leq \mathbb{E}_{s\sim\nu^*} D_{\mathrm{KL}}(\pi^*||\pi_{w_0}) + \frac{8\alpha C_{RN}\sqrt{C_0}R^{1.5}T}{\sqrt{d_1}m^{1/4}} + \alpha^2 L_f(R^2 + md_2^2)T \\
&\quad + 3\alpha C_{RN}\sum_{i=0}^{p}\sum_{t\in\mathcal{N}_i}\left\|f_i((s,a),\bar{\theta}_t) - Q^i_{\pi_{\tau_t W_t}}(s,a)\right\|_{\mu_{\pi_{\tau_t W_t}}} \\
&\quad + \alpha(1-\gamma)\sum_{i=1}^{p}\sum_{t\in\mathcal{N}_i}\left|\bar{J}_i(\theta_t^i) - \mathbb{E}_{\nu_{\pi_{\tau_t W_t}}}[f_i((s,a),\bar{\theta}_t)]\right| \\
&\leq \mathbb{E}_{s\sim\nu^*} D_{\mathrm{KL}}(\pi^*||\pi_{w_0}) + \frac{8\alpha C_{RN}\sqrt{C_0}R^{1.5}T}{\sqrt{d_1}m^{1/4}} + \alpha^2 L_f(R^2 + md_2^2)T \\
&\quad + 3\alpha C_{RN}\sum_{t=0}^{T-1}\left\|f_i((s,a),\bar{\theta}_t) - Q^i_{\pi_{\tau_t W_t}}(s,a)\right\|_{\mu_{\pi_{\tau_t W_t}}} \\
&\quad + \alpha(1-\gamma)\sum_{t=0}^{T-1}\left|\bar{J}_i(\theta_t^i) - \mathbb{E}_{\nu_{\pi_{\tau_t W_t}}}[f_i((s,a),\bar{\theta}_t)]\right|.
\end{aligned} \tag{60}$$

We then upper bound the term $\sum_{t=0}^{T-1}\left\|f_i((s,a),\bar{\theta}_t) - Q^i_{\pi_{\tau_t W_t}}(s,a)\right\|_{\mu_{\pi_{\tau_t W_t}}}$. Lemma 1 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 the union bound, we have with probability at least $1-\delta_1$,

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

We then 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 $J_i'(\bar{\theta}_t) = \mathbb{E}_{\xi\cdot\mu_{\pi_{\tau_t W_t}}}[f_i((s,a),\bar{\theta}_t)]$. 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 $t\geq 0$, we bound the error $\bar{J}_i(\theta_t^i) - J_i'(\bar{\theta}_t)$ as follows:

$$\begin{aligned}
&\boldsymbol{P}\left(\left(\frac{1}{N}\sum_{j=1}^{N} f_i((s_j,a_j),\bar{\theta}_t) - J_i'(\bar{\theta}_t)\right)^2 \geq \frac{(1+\Lambda)C_f^2}{N}\right) \\
&\leq \boldsymbol{P}\left(\frac{1}{N}\sum_{j=1}^{N}\frac{\left[f_i((s_j,a_j),\bar{\theta}_t) - J_i'(\bar{\theta}_t)\right]^2}{C_f^2} \geq 1+\Lambda\right)
\end{aligned}$$

$$
\begin{aligned}
&= \boldsymbol{P}\left(\exp\left(\frac{1}{N}\sum_{j=1}^{N}\frac{\left[f_i((s_j,a_j),\bar{\theta}_t)-J_i'(\bar{\theta}_t)\right]^2}{C_f^2}\right)\geq 1+\Lambda\right)\\
&\leq \boldsymbol{P}\left(\frac{1}{N}\sum_{j=1}^{N}\exp\left(\frac{\left[f_i((s_j,a_j),\bar{\theta}_t)-J_i'(\bar{\theta}_t)\right]^2}{C_f^2}\right)\geq 1+\Lambda\right)\\
&\overset{(i)}{\leq} \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left[\exp\left(\frac{\left[f_i((s_j,a_j),\bar{\theta}_t)-J_i'(\bar{\theta}_t)\right]^2}{C_f^2}\right)\right]/\exp(1+\Lambda)\\
&\leq \exp(-\Lambda),
\end{aligned}
\tag{62}
$$

where (i) follows from Markov's inequality. Then, 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 the union bound, we have with probability at least $1-\delta_2$,

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

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

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

where $C_3=\frac{8C_{RN}\sqrt{C_0}R^{1.5}}{\sqrt{d_1}}$, $C_4=L_f(R^2+d_2^2)$, $C_5=3\alpha C_2C_{RN}$, and $C_6=2C_f$ are positive constants. □

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

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

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 hold,*
   - *(a) $|\mathcal{N}_0|\geq T/2$,*
   - *(b) $\sum_{t\in\mathcal{G}}(J_0(\pi^*)-J_0(w_t))\leq 0$.*

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

$$
\alpha(1-\gamma)\sum_{t\in\mathcal{N}_0}(J_0(\pi^*)-J_0(\pi_{w_t}))+\alpha(1-\gamma)\eta\sum_{i=1}^{p}|\mathcal{N}_i|
$$

$$\leq \mathbb{E}_{s\sim\nu^*} D_{\mathrm{KL}}(\pi^*||\pi_{w_0}) + C_3\left(\frac{\alpha T}{m^{1/4}}\right) + C_4(\alpha^2 mT) + C_5\left(\frac{\alpha T}{(1-\gamma)^{1.5}m^{1/8}}\log^{\frac{1}{4}}\left(\frac{(1-\gamma)^2T\sqrt{m}}{\delta}\right)\right) + C_6\left(\alpha(1-\gamma)\sqrt{T}\right). \tag{65}$$

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

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

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

We then proceed to verify the item 2. If $\sum_{t\in\mathcal{G}}(J_0(\pi^*) - J_0(w_t)) \leq 0$, then (b) in item 2 holds. If $\sum_{t\in\mathcal{G}}(J_0(\pi^*) - J_0(w_t)) \leq 0$, then eq. (65) implies that

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

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

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

which contradicts eq. (64). Hence, (a) in item 2 holds. □

C.4. Proof of Theorem 2

We restate Theorem 2 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 Assumptions 1-4 hold. Let* $\alpha = \frac{1}{2C_4\sqrt{T}}$ *and*

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

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*

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

where

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

and

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

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

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

To proceed the proof of Theorem 2/Theorem 4, we consider the event given in Lemma 16, which happens with probability at least $1-\delta$:

$$\begin{aligned}
&\alpha(1-\gamma)\sum_{t\in\mathcal{N}_0}(J_0(\pi^*) - J_0(\pi_{w_t})) + \alpha(1-\gamma)\eta\sum_{i=1}^{p}|\mathcal{N}_i| \\
&\leq \mathbb{E}_{s\sim\nu^*} D_{\mathrm{KL}}(\pi^*||\pi_{w_0}) + C_3\left(\frac{\alpha T}{m^{1/4}}\right) + C_4(\alpha^2 mT) \\
&\quad + C_5\left(\frac{\alpha T}{(1-\gamma)^{1.5}m^{1/8}}\log^{\frac{1}{4}}\left(\frac{(1-\gamma)^2 T\sqrt{m}}{\delta}\right)\right) + C_6\left(\alpha(1-\gamma)\sqrt{T}\right). \qquad (66)
\end{aligned}$$

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

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

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

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

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

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

and

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

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

$$\mathbb{E}[J_i(\pi_{w_{\text{out}}})] - d_i = \frac{1}{|\mathcal{N}_0|}\sum_{t\in\mathcal{N}_0} J_i(\pi_{w_t}) - d_i$$

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

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

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

and

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

Let the value of the tolerance η be

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

We have

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

which satisfies the requirement specified in Lemma 17. Combining eq. (67), eq. (68) and eq. (69), and using 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

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