Exploration-Driven Policy Optimization in RLHF: Theoretical Insights on Efficient Data Utilization

Yihan Du [1] **Anna Winnicki** [1] **Gal Dalal** [2] **Shie Mannor** [3 2] **R. Srikant** [1]

Abstract

Reinforcement Learning from Human Feedback (RLHF) has achieved impressive empirical successes while relying on a small amount of human feedback. However, there is limited theoretical justification for this phenomenon. Additionally, most recent studies focus on value-based algorithms despite the recent empirical successes of policy-based algorithms. In this work, we consider an RLHF algorithm based on policy optimization (PO-RLHF). The algorithm is based on the popular Policy Cover-Policy Gradient (PC-PG) algorithm, which assumes knowledge of the reward function. In PO-RLHF, knowledge of the reward function is not assumed, and the algorithm uses trajectory-based comparison feedback to infer the reward function. We provide performance bounds for PO-RLHF with low query complexity, which provides insight into why a small amount of human feedback may be sufficient to achieve good performance with RLHF. A key novelty is a trajectory-level elliptical potential analysis, which bounds the reward estimation error when comparison feedback (rather than numerical reward observation) is given. We provide and analyze algorithms `PG-RLHF` and `NN-PG-RLHF` for two settings: linear and neural function approximation, respectively.

1. Introduction

Reinforcement Learning (RL) (Sutton & Barto, 2018; Agarwal et al., 2021) is a classic sequential decision-making problem where an agent interacts with an unknown environment in order to maximize the expected cumulative reward. In many applications, e.g., robotics and Large Language Models (LLMs) (Ouyang et al., 2022; Achiam et al., 2023), the goal of the agent is complex and related to human evaluation. Additionally, the reward function may be hard to manually design.

To handle these challenges, a framework called Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017) has been proposed and has achieved huge empirical successes in ChatGPT (Achiam et al., 2023). In RLHF, the agent does not directly observe rewards, but has access to queries from humans on preferences based on trajectories. The agent learns the quality of trajectories (policies) from the preference feedback over time in order to optimize performance. Existing empirical works have demonstrated the practical efficiency of RLHF: human feedback can solve complex RL tasks by using fewer than 1% of the data from the agent's interactions with the environment (Christiano et al., 2017).

Recently, there have also been a number of theoretical RL papers which seek to provide analyze RLHF, e.g., (Pacchiano et al., 2021; Chen et al., 2022; Zhu et al., 2023; Wang et al., 2023). Most of these works consider value-based algorithms or a given dataset of human feedback, while in many applications, e.g., ChatGPT, policy optimization algorithms are often used. Our goal is to quantify the query and sample complexities of policy-based algorithms when used in conjunction with RLHF, and show that the query complexity is a small fraction of the overall sample complexity.

In order to address the aforementioned issues, we study Policy Optimization for RLHF (PO-RLHF) with active human feedback, through which we provide insights on the query efficiency of RLHF. The algorithm can be summarized as follows. It is an iterative process where at each iteration there is a policy, and several trajectories are drawn by following the policies obtained so far. The trajectories are compared to trajectories generated by following a baseline policy. Humans make comparisons between the trajectories generated by the two policies. Assuming a Bradley-Terry model (Bradley & Terry, 1952), the algorithm uses the results of the comparison queries to update the estimate of the underlying reward function. Then, there is an inner loop where the algorithm follows several steps of the PC-PG policy optimization algorithm (Agarwal et al., 2020) using

[1]University of Illinois Urbana-Champaign [2]NVIDIA Research [3]Technion. Correspondence to: Yihan Du <yihandu@illinois.edu>, R. Srikant <rsrikant@illinois.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

the estimated reward model. At each step of the inner loop, there is a current estimate of a parameter corresponding to the policy, and Monte Carlo simulations of the policy are used to determine the subsequent policy parameter.

Under this formulation, we consider two settings, i.e., linear function approximation and neural function approximation, where the reward function is linear and belongs to a neural function class, respectively. For both settings, we design policy gradient algorithms, `PG-RLHF` and `NN-PG-RLHF`, which can efficiently explore the unknown environment and collect human feedback adapting to the exploration. We establish sample and query complexity guarantees for these two algorithms.

While our algorithm is based on the PC-PG algorithm (Agarwal et al., 2020), unlike PC-PG where the reward function is assumed to be known, we assume the use of human feedback. In order to take into account human feedback, we first extend the PC-PG analysis techniques to incorporate sources of error in the rewards. Then we directly quantify the error from human feedback. Characterizing error from human feedback is challenging for the following reason. The standard tool to analyze policy-based methods with exploration is the elliptical potential lemma (Abbasi-Yadkori et al., 2011). However, this lemma has previously been used only in the case where the reward information is generated and observed from each individual state-action. *A key novelty in our paper is in transforming our error terms involving feature covariance matrices into a trajectory-wise form such that the elliptical potential lemma can be applied to our RLHF situation*, i.e., where rewards are not observed and two trajectories are compared based on the sum of rewards at all (state, action) pairs in each trajectory. To address this issue, we develop a novel trajectory-level elliptical potential analysis technique (for more, see Section 4.2).

Our results are consistent with the empirical observation that a small amount of human feedback is sufficient for RLHF to be successful. The reason is clear: human feedback is used to estimate the reward function which is then used in the policy-based RL algorithm. In other words, the during the policy update and policy evaluation phases of our algorithm, the reward estimate is fixed. While it may take many iterations of gradient ascent and many samples to evaluate each policy, the number of queries required to estimate the reward function is a small fraction of the overall sample complexity.

We summarize our main contributions as follows:

- Motivated by the success of RLHF, we study policy optimization for RLHF with exploration and active human feedback collection, and seek to theoretically explain the practical efficiency of RLHF.
- For linear and neural function approximation, we design provably efficient algorithms `PG-RLHF` and `NN-PG-RLHF`, which simultaneously explore the unknown environment and adaptively collect human data according to the exploration.
- We develop novel analytical techniques, including a trajectory-level elliptical potential argument and a biased MLE guarantee with neural approximation.
- We provide justification for the practical efficiency of RLHF through a rigorous comparison of sample complexity between RLHF and standard RL.

2. Related Work

In this section, we discuss works that are most closely related to ours, and defer a detailed review to Appendix B.

RLHF (Christiano et al., 2017) has shown great empirical successes, especially in LLMs (Ouyang et al., 2022; Achiam et al., 2023). Recently, a number of works have started to theoretically analyze RLHF. Xu et al. (2020); Novoseller et al. (2020); Pacchiano et al. (2021) study online RLHF for tabular MDPs. Chen et al. (2022); Wang et al. (2023) consider online RLHF with general function approximation. Wang et al. (2023) design a reduction framework for RLHF, and prove that the sample complexity for RLHF is no higher than that for standard RL. Zhu et al. (2023); Zhan et al. (2023a) study offline RLHF with function approximation. Ji et al. (2023) seek to understand the empirical success of RLHF from the perspective of intrinsic data bias.

Different from the above works which mostly consider value-based algorithms, we analyze policy gradient RLHF algorithms with exploration, and show that the amount of data needed to implement RLHF is a small fraction of the amount of data needed to train an RL algorithm.

Our work is also related to prior neural RL works, e.g., (Cai et al., 2019; Wang et al., 2019; Xu et al., 2021), which theoretically analyze neural function approximation.

3. Formulation

In this section, we formally define the PO-RLHF problem.

We consider a discounted MDP $\mathcal{M}(\mathcal{S}, \mathcal{A}, r, P, \gamma, s_{\text{init}})$. Specifically, $\mathcal{S}$ is the state space, and $\mathcal{A}$ is the action space. $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is an underlying reward function, so that $r(s, a)$ specifies the reward of taking action a in state s. In the RLHF setting, the agent cannot directly observe $r(s, a)$, and instead, can only observe comparison feedback between trajectories generated according to r (detailed shortly). $P : \mathcal{S} \times \mathcal{A} \to \triangle_{\mathcal{S}}$ is an unknown transition distribution, and $P(s'|s, a)$ gives the transition probability of transitioning to s' if action a is taken in state s. Here for any set $\mathcal{X}$, $\triangle_{\mathcal{X}}$ denotes the space of all distributions over

$\mathcal{X}$, and $\gamma \in [0,1)$ is a discount factor. We define a policy as a mapping $\pi : \mathcal{S} \to \triangle_{\mathcal{A}}$ which specifies what action to take in a state.

Let the state at step h be denoted by s_h, and the action taken at step h be denoted by a_h. The value function

$$V^{\pi}(s) := \mathbb{E}\bigg[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) | s_0 = s, \pi\bigg]$$

and the state-action value function

$$Q^{\pi}(s,a) := \mathbb{E}\bigg[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) | s_0 = s, a_0 = a, \pi\bigg]$$

denote the expected sum of discounted rewards received under policy π, starting from a given state s and state-action pair (s,a), respectively. We define the optimal policy as $\pi^* := \operatorname{argmax}_{\pi} V^{\pi}(s_{\text{init}})$.

The RLHF model is as follows. The agent starts from an initial state s_{init}. At each step h, the agent first observes the current state s_h, and then takes an action a_h according to her policy. After that, she obtains an underlying reward $r(s,a)$ (not observed), and transitions to a next state $s_{h+1} \sim P(\cdot|s_h, a_h)$. The agent can choose to terminate the current trajectory with probability $1-\gamma$ and restart from s_{init} at each step. The agent can query humans to compare trajectories $\tau^{(1)}$ and $\tau^{(2)}$, and observe preference feedback y. Following the literature (Pacchiano et al., 2021; Zhu et al., 2023), we consider the classic Bradley-Terry model (Bradley & Terry, 1952) to formulate preference generation:

$$\Pr[y=1] = \frac{1}{1+\exp(-\tilde{r}^{\tau^{(1)},\tau^{(2)}})}, \tag{1}$$

with $\Pr[y=0] = 1 - \Pr[y=1]$. Here $y=1$ represents that $\tau^{(1)}$ is preferred to $\tau^{(2)}$, and $y=0$ denotes the opposite case.

$$\tilde{r}^{\tau^{(1)},\tau^{(2)}} := \sum_{h=0}^{H(\tau^{(1)})} r(s_h^{(1)}, a_h^{(1)}) - \sum_{h=0}^{H(\tau^{(2)})} r(s_h^{(2)}, a_h^{(2)}),$$

and $H(\tau)$ denotes the length of trajectory τ.

Given a confidence parameter δ and an accuracy parameter ε, the goal of the agent is to identify an ε-optimal policy $\hat{\pi}$ which satisfies $V^{\pi^*}(s_{\text{init}}) - V^{\hat{\pi}}(s_{\text{init}}) \le \varepsilon$ with probability at least $1-\delta$. Before we describe our reward function model, we first introduce some useful notation.

Notation. For any $(s',a') \in \mathcal{S}\times\mathcal{A}$ and policy π, let $d^{\pi}_{s',a'}(s,a) := (1-\gamma)\mathbb{E}[\sum_{h=0}^{\infty}\gamma^h \Pr[s_h = s, a_h = a | s_0 = s', a_0 = a', \pi]$ denote the discounted state-action distribution of starting from (s',a') and executing π. With a slight abuse of notation, for any $s' \in \mathcal{S}$, let $d^{\pi}_{s'}(s,a) := \mathbb{E}_{a' \in \pi(\cdot|s')}[d^{\pi}_{s',a'}(s,a)]$. For any initial distribution $\rho \in \triangle_{\mathcal{S}\times\mathcal{A}}$, let $d^{\pi}_{\rho}(s,a) := \mathbb{E}_{(s',a')\sim\rho}[d^{\pi}_{s',a'}(s,a)]$. In addition, for any $(s',a') \in \mathcal{S}\times\mathcal{A}$ and policy π, let $\mathcal{O}^{\pi}_{s',a'}$ be the distribution of the trajectory generated by starting from s',a', executing π and terminating with probability $1-\gamma$ at each step, which we call a *discounted trajectory distribution*. For any $\rho \in \triangle_{\mathcal{S}\times\mathcal{A}}$, let $\mathcal{O}^{\pi}_{\rho}$ be the discounted trajectory distribution of starting from ρ and executing π.

Under this formulation, we consider linear and neural function approximation settings for the reward model.

3.1. Linear Function Approximation

In the linear setting, we consider the log-linear policy parameterization and linear reward function. Specifically, there exists a known feature mapping $\phi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$ which specifies the feature vectors of state-action pairs, and satisfies $\|\phi(s,a)\| \le 1$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. For parameter $w \in \mathbb{R}^d$, the log-linear policy is represented as

$$\pi_w(a|s) := \frac{\exp(\phi(s,a)^{\top} w)}{\sum_{a' \in \mathcal{A}} \exp(\phi(s,a')^{\top} w)}.$$

We make the following assumption on the reward function.

Assumption 3.1 (Linear Reward Function)**.** There exists some reward parameter $\mu^* \in \mathbb{R}^d$ such that

$$r(s,a) := \phi(s,a)^{\top}\mu^*.$$

3.2. Neural Function Approximation

In the neural function approximation setting, we parameterize the policy, value function and reward by neural networks.

A two-layer ReLU neural network with input feature $\phi(s,a)$, parameter w and width m is represented by (Cai et al., 2019; Xu et al., 2021)

$$f(s,a;w) = \frac{1}{\sqrt{m}} \sum_{\ell=1}^{m} b_\ell \, \mathbb{1}\{\phi(s,a)^{\top}[w]_\ell > 0\} \, \phi(s,a)^{\top}[w]_\ell,$$

where $b := [b_1, \ldots, b_m]^{\top} \in \mathbb{R}^m$, and $w := [[w]_1; \ldots; [w]_m] \in \mathbb{R}^{md}$ are the network parameters.

We initialize the parameters by $b_\ell \sim \text{Unif}([-1,1])$ and $[w^0]_\ell \sim \mathcal{D}_{\text{init}}$ for any $\ell \in [m]$. Here $\mathcal{D}_{\text{init}}$ is an initialization distribution, such that for any $w' \in \mathbb{R}^d$ in the support of $\mathcal{D}_{\text{init}}$, $\underline{c} \le \|w'\|_2 \le \bar{c}$ for some constants $\underline{c}, \bar{c} > 0$. During training, we keep b fixed and only update w.

With a temperature parameter $\alpha \in \mathbb{R}$ and a network parameter $w \in \mathbb{R}^{md}$, a policy is represented by

$$\pi_{\alpha,w} := \frac{\exp(\alpha f(s,a;w))}{\sum_{a' \in \mathcal{A}} \exp(\alpha f(s,a;w))},$$

We also use $f(s,a;\theta)$ to approximate the state-action value function Q^{π} with another parameter $\theta \in \mathbb{R}^{md}$ and the same initialization as w, i.e., $\theta^0 = w^0$.

Moreover, we approximate the reward function $r(s,a)$ by

$$h(s,a;\mu) := \frac{1}{\sqrt{m}} \sum_{\ell=1}^{m} b'_\ell \, \mathbb{1}\{\phi(s,a)^\top [\mu]_\ell > 0\} \, \phi(s,a)^\top [\mu]_\ell,$$

where $b' := [b'_1, \ldots, b'_m]^\top \in \mathbb{R}^m$ and $\mu := [[\mu]_1; \ldots; [\mu]_m] \in \mathbb{R}^{md}$ are the reward network parameters. Similarly, we initialize $b'_\ell \sim \text{Unif}([-1,1])$ and $[\mu^0]_\ell \sim \mathcal{D}_{\text{init}}$ for any $\ell \in [m]$, and only update μ during training.

For any parameter $\mu \in \mathbb{R}^{md}$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$, let $[\psi_\mu(s,a)]_\ell := \frac{b'_\ell}{\sqrt{m}} \mathbb{1}\{\phi(s,a)^\top [\mu]_\ell > 0\} \phi(s,a) \in \mathbb{R}^d$ for any $\ell \in [m]$. Let $\psi_\mu(s,a) := [[\psi_\mu(s,a)]_1; \ldots; [\psi_\mu(s,a)]_m] \in \mathbb{R}^{md}$. We can similarly define $\psi_w(s,a)$.

Define a neural function class (Rahimi & Recht, 2007):

$$\mathcal{F}^{\mu}_{R,\infty} := \Big\{ h(s,a) = h(s,a;\mu^0) + \int \mathbb{1}\{\phi(s,a)^\top \mu > 0\} \cdot \phi(s,a)^\top \nu(\mu) \, dp(\mu) : \|\nu(\mu)\|_\infty \le \frac{R}{\sqrt{d}} \Big\},$$

where $p : \mathbb{R}^d \to \mathbb{R}$ is the density function of $\mathcal{D}_{\text{init}}$, and $\nu : \mathbb{R}^d \to \mathbb{R}^d$ together with $h(s,a;\mu^0)$ parameterize the element of $\mathcal{F}^{\mu}_{R,\infty}$.

In the neural setting, we make the following assumptions.

Assumption 3.2 (Neural Realizability of r)**.** $r \in \mathcal{F}_{R,\infty}$.

This is a standard realizability assumption, and also made in prior neural RL works (Wang et al., 2019; Xu et al., 2021).

Assumption 3.3 (Regularity of State-action Distribution)**.** There exists an absolute constant $c_{\text{scale}} \in (0,1)$ such that for any $v \in \mathbb{R}^d$, $x > 0$, $(s',a') \in \mathcal{S} \times \mathcal{A}$ and policy π,

$$\mathbb{E}_{(s,a) \sim d^{\pi}_{s',a'}} \left[\mathbb{1}\left\{ \left| \phi(s,a)^\top v \right| \le x \right\} \right] \le \frac{c_{\text{scale}} x}{\|v\|_2}.$$

This is also a standard regularity assumption in the neural RL literature (Cai et al., 2019; Wang et al., 2019; Xu et al., 2021). For a random state-action pair $(s,a) \sim d^{\pi}_{s',a'}$, the probability of $|\phi(s,a)^\top v| \le x$ scales with x and $\|v\|_2^{-1}$.

3.3. Baseline Policy

We assume that we have access to a baseline policy, which will be used for comparison in our algorithms.

For any trajectory $\tau = (s_0, a_0, \ldots, s_{H(\tau)}, a_{H(\tau)})$ and feature mapping $\chi \in \{\phi, \psi_{\mu^0}\}$, let $\chi(\tau) := \sum_{h=0}^{H(\tau)} \chi(s_h, a_h)$.

Assumption 3.4 (Baseline Policy)**.** The baseline policy π^{base} satisfies that for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and policy π,

$$\mathbb{E}_{\substack{\tau^{(1)} \sim \mathcal{O}^{\pi}_{s,a} \\ \tau^{(2)} \sim \mathcal{O}^{\pi^{\text{base}}}_{s_{\text{init}}}}} [(\chi(\tau^{(1)}) - \chi(\tau^{(2)}))(\chi(\tau^{(1)}) - \chi(\tau^{(2)}))^\top] \succeq c_{\text{base}} \mathbb{E}_{\tau^{(2)} \sim \mathcal{O}^{\pi^{\text{base}}}_{s_{\text{init}}}} [\chi(\tau^{(2)}) \chi(\tau^{(2)})^\top]$$

for some absolute constant $c_{\text{base}} \in (0,1)$. Here $\chi = \phi$ in the case of linear function approximation, and $\chi = \psi$ in the case of neural function approximation.

We discuss Assumption 3.4 in more detail in Appendix D.3.4.

Algorithm 1 PG-RLHF

1: **Input:** $\varepsilon, \delta, N, K, M_{\text{HF}}, \zeta_{\text{HF}}, \zeta_{\text{cov}}, \pi^{\text{base}}, W_\mu, \pi^0$.
2: **for** $n = 0, \ldots, N-1$ **do**
3: Sample $\{s_i, a_i\}_{i=1}^K \sim d^{\pi^n}_{s_{\text{init}}}$, and $\hat{\Sigma}^n \leftarrow \frac{1}{K} \sum_{i=1}^K \phi(s_i,a_i)\phi(s_i,a_i)^\top$
4: $\hat{\Sigma}^n_{\text{cov}} \leftarrow \sum_{i=0}^n \hat{\Sigma}^i + \zeta_{\text{cov}} I$
5: Let $\rho^n_{\text{cov}} := \frac{1}{n+1} \sum_{i=0}^n d^{\pi^i}_{s_{\text{init}}}$
6: $\mathcal{O}^n_{\text{HF}} := \frac{1}{n} \sum_{i=1}^n \mathcal{O}^{\pi^i}_{\rho^{i-1}_{\text{cov}}}, \forall n \ge 1$, and $\mathcal{O}^0_{\text{HF}} := \mathcal{O}^{\pi^0}_{s_{\text{init}}}$
7: **for** $i = 1, \ldots, M_{\text{HF}}$ **do**
8: Sample trajectories $\tau_i^{(1)} \sim \mathcal{O}^n_{\text{HF}}$ and $\tau_i^{(2)} \sim \mathcal{O}^{\pi^{\text{base}}}_{s_{\text{init}}}$
9: Observe the comparison outcome y_i
10: **end for**
11: Estimate $\hat{\mu}^n$ via MLE as in Eq. (2)
12: $\pi^{n+1} \leftarrow$ NPG-Update$(\rho^n_{\text{cov}}, \hat{\Sigma}^n_{\text{cov}}, \hat{\mu}^n)$
13: **end for**
14: **return** Unif$(\pi^1, \ldots, \pi^N)$

4. PO-RLHF with Linear Function Approximation

We first study PO-RLHF with linear function approximation. We develop a policy gradient algorithm PG-RLHF which can explore the environment and adaptively collect human data.

4.1. Algorithm PG-RLHF

PG-RLHF builds upon the policy gradient algorithm PC-PG (Agarwal et al., 2020) for standard RL. Our algorithm is described in Algorithm 1. PG-RLHF runs N outer-loop phases for coverage update and reward estimation (Lines 2-13 in Algorithm 1), and T inner-loop iterations for policy optimization under given coverage and reward model (Lines 6-15 in Algorithm 2). In each phase n, PG-RLHF first estimates the feature covariance matrix $\hat{\Sigma}^n_{\text{cov}}$ and updates the state-action coverage distribution ρ^n_{cov}, which is the average of the state-action visitation distribution of all the policies $\pi^0, \ldots, \pi^n$ used so far (Line 5 in Algorithm 1). ρ^n_{cov} will be the initial state-action distribution of the policy optimization in the inner-loop, and is gradually expanded in each phase to improve the coverage.

Human Feedback Collection. Next, we collect human data for reward estimation. For any phase $n \ge 1$, let $\mathcal{O}^n_{\text{HF}}$ be the distribution of the trajectory generated by starting from state-

Algorithm 2 NPG-Update

1: **Input:** $\rho^n_{\text{cov}}, \hat{\Sigma}^n_{\text{cov}}, \hat{\mu}^n, T, \xi, \eta, W_\theta$.
2: Let $\hat{r}^n(\cdot,\cdot) := \phi(\cdot,\cdot)^\top \hat{\mu}^n$ and $\Theta := \{\theta : \|\theta\|_2 \le W_\theta\}$
3: Let $b^n(\cdot,\cdot) := \frac{1}{1-\gamma}\mathbb{1}\{\phi(\cdot,\cdot)^\top (\hat{\Sigma}^n_{\text{cov}})^{-1}\phi(\cdot,\cdot) \ge \beta\}$
4: Let $\mathcal{K}^n := \{s \in \mathcal{S} : \forall a \in \mathcal{A},\ b^n(s,a) = 0\}$
5: For $s \in \mathcal{K}^n$, initialize w^0 such that $\pi^0(\cdot|s) := \pi_{w^0}(\cdot|s) = \text{Unif}(\mathcal{A})$. For $s \notin \mathcal{K}^n$, $\pi^0(\cdot|s) := \text{Unif}(\{a \in \mathcal{A} : b^n(s,a) = \frac{1}{1-\gamma}\})$
6: **for** $t = 0, \dots, T-1$ **do**
7: Initialize $\theta^{t,0}$
8: **for** $i = 0, \dots, M_{\text{SGD}} - 1$ **do**
9: Sample $(s_i, a_i) \sim \rho^n_{\text{cov}}$ and estimate $\hat{Q}^{\pi^t}(s_i, a_i; \hat{r}^n + b^n)$ using Monte Carlo sampling
10: $\theta^{t,i+1} \leftarrow \text{Proj}_\Theta(\theta^{t,i} - 2\xi(\phi(s_i,a_i)^\top \theta^{t,i} - (\hat{Q}^{\pi^t}(s_i,a_i;\hat{r}^n + b^n) - b^n(s_i,a_i))) \cdot \phi(s_i,a_i))$
11: **end for**
12: $\theta^t \leftarrow \frac{1}{M_{\text{SGD}}} \sum_{i=0}^{M_{\text{SGD}}-1} \theta^{t,i}$
13: $w^{t+1} \leftarrow w^t + \eta\theta^t$
14: $\forall s \in \mathcal{K}^n$, $\pi^{t+1}(\cdot|s) = \pi_{w^{t+1}}(\cdot|s) \propto \exp(\phi(s,\cdot)^\top w^{t+1})$. $\forall s \notin \mathcal{K}^n$, $\pi^{t+1}(\cdot|s) = \pi^0(\cdot|s)$
15: **end for**
16: **return** $\text{Unif}(\pi^0, \dots, \pi^{T-1})$

action distribution $\rho^{\bar{n}-1}_{\text{cov}}$, executing $\pi^{\bar{n}}$ and terminating with probability $1-\gamma$ at each step, where $\bar{n} \sim \text{Unif}([n])$; For phase $n = 0$, $\mathcal{O}^n_{\text{HF}} := \mathcal{O}^{\pi^0}_{s_{\text{init}}}$ (Line 6). In addition, let $\mathcal{O}^{\pi^{\text{base}}}_{s_{\text{init}}}$ be the distribution of the trajectory generated by starting from s_{init}, executing π^{base} and stopping with probability $1-\gamma$ at each step, where π^{base} is a baseline policy (Line 8). We sample trajectories $\tau_i^{(1)}$ and $\tau_i^{(2)}$ from $\mathcal{O}^n_{\text{HF}}$ and $\mathcal{O}^{\pi^{\text{base}}}_{s_{\text{init}}}$, respectively, and observe a comparison outcome y_i. This process is independently repeated M_{HF} times, and then we obtain human data $\{\tau_i^{(1)}, \tau_i^{(2)}, y_i\}_{i=1}^{M_{\text{HF}}}$ (Lines 7-10).

With the human data, we use the maximum likelihood estimator (MLE) to estimate the reward parameter as

$$\hat{\mu}^n = \underset{\|\mu\|_2 \le W_\mu}{\text{argmin}} \Bigg(-\sum_{i=1}^{M_{\text{HF}}} \log\Bigg(\frac{\mathbb{1}\{y_i = 1\}}{1 + \exp\big(-(\tilde{\phi}^{\tau_i^{(1)},\tau_i^{(2)}})^\top \mu\big)} + \frac{\mathbb{1}\{y_i = 0\}}{1 + \exp\big((\tilde{\phi}^{\tau_i^{(1)},\tau_i^{(2)}})^\top \mu\big)} \Bigg) \Bigg), \tag{2}$$

where $\tilde{\phi}^{\tau_i^{(1)},\tau_i^{(2)}} := \sum_{h=0}^{H(\tau_i^{(1)})} \phi(s_{i,h}^{(1)}, a_{i,h}^{(1)}) - \sum_{h=0}^{H(\tau_i^{(2)})} \phi(s_{i,h}^{(2)}, a_{i,h}^{(2)})$, and $(s_{i,h}^{(\ell)}, a_{i,h}^{(\ell)})$ denotes the state-action at step h in trajectory $\tau_i^{(\ell)}$ for any $\ell \in \{1,2\}$.

Comparing to a fixed baseline policy helps to de-correlate the comparison (difference) relationship between two trajectories in the Bradley-Terry model (Eq. (1)), and provides a better control for the properties of the human data covariance matrix to cover the state-actions that we care about.

Intuition of Human Feedback Collection. The idea behind our human data collection scheme is as follows. Since we will do policy optimization with initial state-action distribution ρ^n_{cov} and obtain policy π^{n+1} in each phase n, our performance will be influenced by the reward estimation accuracy on the state-actions guided by π^{n+1} starting from ρ^n_{cov} for $n = 0, 1, \dots, N-1$. Therefore, using the human data generated by $\pi^{\bar{n}}$ and $\rho^{\bar{n}-1}_{\text{cov}}$ ($\bar{n} \sim \text{Unif}([n])$) can guarantee a small reward estimation error on the state-action space that we care about (where our performance is measured).

With the coverage distribution ρ^n_{cov}, coverage covariance matrix $\hat{\Sigma}^n_{\text{cov}}$ and estimated reward model $\hat{r}^n(\cdot,\cdot) = \phi(\cdot,\cdot)^\top \hat{\mu}^n$, we call subroutine NPG-Update to perform policy optimization. In NPG-Update (Algorithm 2), we first define the exploration bonus $b^n(s,a) := \frac{1}{1-\gamma}$ for the state-actions that are not sufficiently explored, and define $b^n(s,a) := 0$ for those that are sufficiently explored according to $\hat{\Sigma}^n_{\text{cov}}$ (Line 3). According to $b^n(s,a)$, we implicitly divide the state space into two state sets, one with well-explored state-actions (i.e., $\mathcal{K}^n$), and the other one with under-explored state-actions (Line 4).

Then, we perform natural policy gradient (NPG) (Agarwal et al., 2021) with initial state-action distribution ρ^n_{cov} and bonus-incentivized reward $\hat{r}^n + b^n$ (Lines 6-15). Formally, the optimization objective can be written as

$$\max_\pi \ \mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[Q^\pi(s,a;\hat{r}^n + b^n)\right].$$

In the t-th iteration of NPG, we use projected stochastic gradient descent (SGD) (Shalev-Shwartz & Ben-David, 2014) to fit (Lines 8-11)

$$\underset{\|\theta\|\le W_\theta}{\text{argmin}} \ \mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\Big[\Big(\phi(s,a)^\top\theta - \big(Q^{\pi^t}(s,a;\hat{r}^n + b^n) - b^n(s,a)\big)\Big)^2\Big]. \tag{3}$$

At step i of SGD, we compute the stochastic gradient by $2(\phi(s_i,a_i)^\top\theta - (\hat{Q}^{\pi^t}(s_i,a_i;\hat{r}^n + b^n) - b^n(s,a))) \cdot \phi(s,a)$, where (s_i,a_i) is sampled from ρ^n_{cov} and $\hat{Q}^{\pi^t}(s_i,a_i;\hat{r}^n + b^n)$ is estimated by Monte Carlo sampling (Line 9). After SGD, we obtain θ^t such that $\phi(s,a)^\top\theta^t + b^n(s,a)$ well fits the state-action value function $\hat{Q}^{\pi^t}(s_i,a_i;\hat{r}^n + b^n)$ (Line 12).

Then, we update the policy parameter by $w^{t+1} \leftarrow w^t + \eta\theta^t$. Furthermore, we set the policy π^{t+1} as the log-linear policy with parameter w^{t+1} for $s \in \mathcal{K}^n$, and the uniform policy over all under-explored actions for $s \notin \mathcal{K}^n$ (Line 14).

After NPG, we obtain $\pi^{n+1} = \text{Unif}(\pi^0, \dots, \pi^{T-1})$, which both optimizes the value function and has an incentive to explore the unvisited space. In the next phase, π^{n+1} is used to improve the coverage, and also expand the space where we collect human data and can guarantee accurate reward estimation.

Computational Efficiency. We remark that the computational complexity of `NPG-Update` is independent of $\mathcal{S}$. b^n, $\mathcal{K}^n$ and π^t are only implicitly maintained by computing $\hat{\Sigma}^n_{\text{cov}}$ and w^t. When we encounter some state s in Monte Carlo sampling (Line 9 in Algorithm 2), we can identify if s is in $\mathcal{K}^n$ and compute $b^n(s,a)$ by $\hat{\Sigma}^n_{\text{cov}}$ (for all a). To execute π^t in state s, if $s \in \mathcal{K}^n$, we choose an action according to π_{w^t}; If $s \notin \mathcal{K}^n$, we uniformly choose an action from the actions with $b^n(s,a) = \frac{1}{1-\gamma}$.

Sample Complexity. We note that is N is the number of times we update the coverage distribution. Between two coverage updates, (i) we observe K trajectories to update the feature covariance matrix, (ii) we perform M_{HF} human pairwise trajectory comparisons, and (iii) we run T iterations of NPG, and within each NPG iteration, we run M_{SGD} steps of SGD for policy evaluation. Further, for each step of SGD, we sample two trajectories, one to sample from the coverage distribution and one to estimate the Q-value function (Line 9 in Algorithm 2). So the overall number of trajectories used by our algorithm is $(K + 2M_{\text{HF}} + 2TM_{\text{SGD}})N$. Since the number of transitions observed for each trajectory is $\tilde{O}(\frac{1}{1-\gamma})$, the number of samples used by our algorithm is $\tilde{O}((K + M_{\text{HF}} + TM_{\text{SGD}})\frac{N}{1-\gamma})$.

4.2. Theoretical Guarantee of Algorithm `PG-RLHF`

Now we provide performance guarantees for algorithm `PG-RLHF`.

First, following (Agarwal et al., 2020), we define a bounded transfer function approximation error. Let $\theta^t_* = \text{argmin}_{\|\theta\| \le W_\theta} \mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}[(\phi(s,a)^\top\theta - (Q^{\pi^t}(s,a;r + b^n) - b^n(s,a)))^2]$, and $d^\star_{s_{\text{init}}}(s,a) := d^{\pi^*}_{s_{\text{init}}}(s) \circ \text{Unif}_{\mathcal{A}}(a)$.

Assumption 4.1 (Bounded Transfer Error)**.** For any phase $n \ge 0$ and iteration $t \ge 0$, there exists some $\varepsilon_{\text{bias}} > 0$ which satisfies

$$\mathbb{E}_{(s,a)\sim d^\star_{s_{\text{init}}}}\left[\Big(\phi(s,a)^\top\theta^t_* - \big(Q^{\pi^t}(s,a;r+b^n) - b^n(s,a)\big)\Big)^2\right] \le \varepsilon_{\text{bias}}. \quad (4)$$

$\varepsilon_{\text{bias}}$ measures the error of using the best fit θ^t_* with log-linear policies under ρ^n_{cov} to predict the state-action value function under $d^\star_{s_{\text{init}}}$. For tabular or linear MDPs (Yang & Wang, 2019; Jin et al., 2020), θ^t_* perfectly fits the value function for all (s,a) with log-linear policies, and $\varepsilon_{\text{bias}} = 0$. Then, we formally state the performance of `PG-RLHF`.

Theorem 4.2. *With probability at least $1-\delta$, the output policy of algorithm* `PG-RLHF` *satisfies*

$$V^{\pi^*}(s_{\text{init}}) - V^{\pi^{\text{out}}}(s_{\text{init}}) \le \tilde{O}\Bigg(\frac{\sqrt{|\mathcal{A}|\varepsilon_{\text{bias}}}}{1-\gamma} + \frac{W_A}{(1-\gamma)\sqrt{T}} + \frac{W_Q\sqrt{\beta N}}{(1-\gamma)(M_{\text{SGD}})^{\frac{1}{4}}} + \frac{\sqrt{\beta W_Q N d}}{(1-\gamma)\sqrt{c_{\text{MLE}}}c_{\text{base}}^{\frac{1}{4}}M_{\text{HF}}^{\frac{1}{4}}} + \frac{d}{N\beta(1-\gamma)}\Bigg).$$

Furthermore, by tuning parameters as in Eq. (25) in Appendix D.4, we can guarantee

$$V^{\pi^*}(s_{\text{init}}) - V^{\pi^{\text{out}}}(s_{\text{init}}) \le \varepsilon + \frac{2\sqrt{|\mathcal{A}|\varepsilon_{\text{bias}}}}{1-\gamma},$$

with $\tilde{O}(\text{Poly}(W_Q, W_\mu, \zeta_{\text{HF}}, d, (1-\gamma)^{-1}, \varepsilon^{-1}, c^{-1}_{\text{base}}, c^{-1}_{\text{MLE}}))$ samples. Here $W_Q := \frac{2}{(1-\gamma)^2}$, $c_{\text{MLE}} := (2 + \exp(-2W_\tau W_\mu) + \exp(2W_\tau W_\mu))^{-1}$, and $W_\tau := \tilde{O}(\frac{1}{1-\gamma})$ denotes the high probability bound of trajectory length.

See the full bounds in Eqs. (24) and (26) in Appendix D.4.

Remark. As shown in Theorem 4.2, the suboptimality can be decomposed into the following components: (i) the transfer function approximation error $\sqrt{\varepsilon_{\text{bias}}}$, (ii) the NPG regret $\frac{1}{\sqrt{T}}$, (iii) the policy evaluation error $(M_{\text{SGD}})^{-\frac{1}{4}}$, (iv) the reward estimation error $M_{\text{HF}}^{-\frac{1}{4}}$, and (v) the error due to the exploration bonus construction $\frac{1}{N}$. The statistical error (ii)-(v) will converge to zero as the number of samples increases, while the transfer function approximation error (i) can still remain even with infinite samples.

Theorem 4.2 demonstrates that algorithm `PG-RLHF` can efficiently utilize human feedback to learn a near-optimal policy up to the intrinsic function approximation error of the MDP. For tabular MDPs and linear MDPs (Yang & Wang, 2019; Jin et al., 2020), we have $\varepsilon_{\text{bias}} = 0$, and `PG-RLHF` can identify an ε-optimal policy.

Below we give a proof sketch, and introduce a novel trajectory-level elliptical potential analysis for bounding the feature vector sum of human data.

Proof Sketch. For any $r : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$, let $F^r(\theta) := \mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}[(\phi(s,a)^\top\theta - (Q^{\pi^t}(s,a;r+b^n) - b^n(s,a)))^2]$. Let θ^t_* and θ^t_{mid} be the optimal solutions to minimize $F^r(\theta)$ and $F^{\hat{r}^n}(\theta)$, respectively. Recall that θ^t is a near-optimal solution to minimize $F^{\hat{r}^n}(\theta)$ obtained by SGD in our algorithm. Applying the performance difference lemma as in (Agarwal et al., 2020), we can decompose the suboptimality into

$$\begin{aligned} &V^*(s_{\text{init}}) - V^{\pi^t}(s_{\text{init}}) \\ &\le \mathbb{E}_{(s,a)\sim d^\star_{s_{\text{init}}}}\Bigg[\frac{\mathbb{1}\{s\in\mathcal{K}^n\}}{1-\gamma}\Big(\underbrace{\bar{\phi}^t(s,a)^\top\theta^t + \bar{b}^{n,t}(s,a)}_{\Gamma_{\text{NPG}}} \\ &\quad + \underbrace{A^{\pi^t}(s,a;r+b^n) - \big(\bar{\phi}^t(s,a)^\top\theta^t_* + \bar{b}^{n,t}(s,a)\big)}_{\Gamma_{\text{bias}}} \\ &\quad + \underbrace{\bar{\phi}^t(s,a)^\top(\theta^t_* - \theta^t_{\text{mid}})}_{\Gamma_r} + \underbrace{\bar{\phi}^t(s,a)^\top(\theta^t_{\text{mid}} - \theta^t)}_{\Gamma_{\text{SGD}}}\end{aligned}$$

	PG-RLHF for RLHF	PC-PG (Agarwal et al., 2020) for Standard RL
# Samples	$\tilde{O}((NK + NTM_{\text{SGD}} + NM_{\text{HF}})\frac{1}{1-\gamma})$	$\tilde{O}((NK + NTM_{\text{SGD}})\frac{1}{1-\gamma})$
# True rewards	0	$\tilde{O}((NK + NTM_{\text{SGD}})\frac{1}{1-\gamma})$
# Queries	$O(NM_{\text{HF}})$	0

Table 1. Comparison of sample complexity, the number of true rewards and the number of queries between PG-RLHF and PC-PG (Agarwal et al., 2020) for standard RL.

$$+ \underbrace{\sum_{(s,a)\notin\mathcal{K}^n} d_{s_{\text{init}}}^{\pi^{n+1}}(s,a)\Big)\Big]}_{\Gamma_b}. \tag{5}$$

Here $A^{\pi^t}(s,a;r+b^n) := Q^{\pi^t}(s,a;r+b^n) - V^{\pi^t}(s;r+b^n)$, $\bar{\phi}^t(s,a) := \phi(s,a) - \mathbb{E}_{a'\sim\pi^t(\cdot|s)}[\phi(s,a')]$ and $\bar{b}^{n,t}(s,a) := b^n(s,a) - \mathbb{E}_{a'\sim\pi^t(\cdot|s)}[b^n(s,a')]$. Similar to (Agarwal et al., 2020), we can bound Γ_{NPG}, $\varepsilon_{\text{bias}}$, Γ_{SGD} and Γ_b due to NPG regret, transfer function approximation error, policy evaluation error and optimistic bonus construction, respectively.

Then, the remaining challenge is to bound the reward estimation error Γ_r. To tackle this, we develop a novel *trajectory-level* elliptical potential analysis to deal with human data.

Trajectory-level Elliptical Potential Analysis. According to the definitions of θ_{mid}^t and θ_*^t, to bound term Γ_r, it suffices to bound

$$\begin{aligned}&\mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n}\left[\left|Q^{\pi^t}(s,a;\hat{r}^n+b^n) - Q^{\pi^t}(s,a;r+b^n)\right|\right]\\ \le&\mathbb{E}_{\tau\sim\mathcal{O}_{\rho_{\text{cov}}^n}^{\pi^t}}\left[\left\|\sum_{h=0}^{H(\tau)}\phi(s_h,a_h)\right\|_{(\hat{\Sigma}_{\text{HF}}^n)^{-1}}\|\hat{\mu}^n-\mu^*\|_{\hat{\Sigma}_{\text{HF}}^n}\right],\end{aligned} \tag{6}$$

Here $\hat{\Sigma}_{\text{HF}}^n := \frac{1}{M_{\text{HF}}}\sum_{i=1}^{M_{\text{HF}}}\tilde{\phi}^{\tau_i^{(1)},\tau_i^{(2)}}(\tilde{\phi}^{\tau_i^{(1)},\tau_i^{(2)}})^\top + \frac{\zeta_{\text{HF}}}{n}I$ is the feature covariance matrix of human data, and concentrates to

$$\begin{aligned}\Sigma_{\text{HF}}^n :=& \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}_{\tau^{(1)}\sim\mathcal{O}_{\rho_{\text{cov}}}^{\pi^i},\tau^{(2)}\sim\mathcal{O}_{s_{\text{init}}}^{\pi^{\text{base}}}}\left[\tilde{\phi}^{\tau^{(1)},\tau^{(2)}}(\tilde{\phi}^{\tau^{(1)},\tau^{(2)}})^\top\right]\right)\\ &+\frac{\zeta_{\text{HF}}}{n}I,\end{aligned} \tag{7}$$

for any $n \ge 1$. Let $\hat{\Sigma}_{\text{HF}}^0 = \Sigma_{\text{HF}}^0 := \zeta_{\text{HF}}I$. In addition, $\tilde{\phi}^{\tau_i^{(1)},\tau_i^{(2)}} := \sum_{h=0}^{H(\tau_i^{(1)})}\phi(s_{i,h}^{(1)},a_{i,h}^{(1)}) - \sum_{h=0}^{H(\tau_i^{(2)})}\phi(s_{i,h}^{(2)},a_{i,h}^{(2)})$.

Eq. (6) is a *key* step. Specifically, we decompose the error of state-action value function due to reward estimation into: (i) The error of reward parameter $\|\hat{\mu}^n - \mu^*\|_{\hat{\Sigma}_{\text{HF}}^n}$, which is bounded by $\tilde{O}(\frac{1}{\sqrt{M_{\text{HF}}}})$ due to the MLE guarantee; (ii) The trajectory-level feature norm $\|\sum_{h=0}^{H(\tau)}\phi(s_h,a_h)\|_{(\hat{\Sigma}_{\text{HF}}^n)^{-1}}$, *instead of* the state-action-level feature norm $\sum_{(s,a)\sim d_{\rho_{\text{cov}}^n}^{\pi^t}}\|\phi(s,a)\|_{(\hat{\Sigma}_{\text{HF}}^n)^{-1}}$.

Since π^{out} is the average of all obtained policies and $\hat{\Sigma}_{\text{HF}}^n$ concentrates to Σ_{HF}^n, with the Cauchy-Schwarz inequality, it suffices to bound the summation of the squared feature norm under Σ_{HF}^n as

$$\frac{1}{NT}\sum_{n=0}^{N-1}\sum_{t=0}^{T-1}\mathbb{E}_{\tau\sim\mathcal{O}_{\rho_{\text{cov}}^n}^{\pi^t}}\left[\left\|\sum_{h=0}^{H(\tau)}\phi(s_h,a_h)\right\|_{(\Sigma_{\text{HF}}^n)^{-1}}^2\right]. \tag{8}$$

A *nice* thing is that the covariance matrix Σ_{HF}^n (Eq. (7)) involves trajectory-level features, and here each summed term $\|\sum_{h=0}^{H(\tau)}\phi(s_h,a_h)\|$ is also a trajectory-wise feature norm. This enables us to apply the elliptical potential lemma (Abbasi-Yadkori et al., 2011) to bound this summation, which validates our decomposition scheme in Eq. (6).

Then, using $\pi^{n+1} = \text{Unif}(\{\pi^t\}_{t=0}^{T-1})$, Eq. (8) is bounded by

$$\begin{aligned}&\frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}_{\tau^{(1)}\sim\mathcal{O}_{\rho_{\text{cov}}^n}^{\pi^{n+1}}}\left[\left\|\sum_{h=0}^{H(\tau^{(1)})}\phi(s_h^{(1)},a_h^{(1)})^\top\right\|_{(\Sigma_{\text{HF}}^n)^{-1}}^2\right]\\ \overset{\text{(a)}}{\le}& \underbrace{2\sum_{n=0}^{N-1}\mathbb{E}_{\substack{\tau^{(1)}\sim\mathcal{O}_{\rho_{\text{cov}}^n}^{\pi^{n+1}}\\ \tau^{(2)}\sim\mathcal{O}_{s_{\text{init}}}^{\pi^{\text{base}}}}}\left[\left\|\tilde{\phi}^{\tau^{(1)},\tau^{(2)}}\right\|_{(n\Sigma_{\text{HF}}^n)^{-1}}^2\right]}_{\Gamma_{\text{traj}}}\\ &+2\sum_{n=0}^{N-1}\mathbb{E}_{\tau^{(2)}\sim\mathcal{O}_{s_{\text{init}}}^{\pi^{\text{base}}}}\left[\left\|\sum_{h=0}^{H(\tau^{(2)})}\phi(s_h^{(2)},a_h^{(2)})^\top\right\|_{(n\Sigma_{\text{HF}}^n)^{-1}}^2\right]\\ \overset{\text{(b)}}{=}& O\left(d\log\left(1+\frac{NW_\tau^2}{d\zeta_{\text{HF}}}\right)+\frac{d}{c_{\text{base}}}\log(N)\right).\end{aligned}$$

Here we make the convention that $(0\Sigma_{\text{HF}}^0) := \zeta_{\text{HF}}I$. Inequality (a) comes from adding and subtracting $\sum_{h=0}^{H(\tau^{(2)})}\phi(s_h^{(2)},a_h^{(2)})^\top$.

With consistency between the summed term and the covariance matrix Σ_{HF}^n (both in a trajectory and difference form), Γ_{traj} is an effective elliptical potential summation. Then, inequality (b) follows from applying the elliptical potential lemma (Abbasi-Yadkori et al., 2011) and Assumption 3.4. See Lemmas D.10, D.13 in Appendix D.3 for full proofs.□

4.3. Insight into the Practical Efficiency of RLHF

Below we compare our PG-RLHF and prior standard RL algorithm PC-PG (Agarwal et al., 2020), and provide an

insight behind the empirical success of RLHF.

Table 1 shows that PG-RLHF needs additional $\tilde{O}(\frac{NM_{\text{HF}}}{1-\gamma})$ samples due to the lack of direct reward signals. We have $O(M_{\text{HF}}) \approx O(M_{\text{SGD}})$, since their convergence rates are the same (see Theorem 4.2). Then, the additional samples needed by PG-RLHF is negligible compared to the total sample complexity. This implies that RLHF does not introduce much hardness in terms of sample complexity, which matches the finding of recent RLHF work (Wang et al., 2023).

Regarding the cost on reward observations, in standard RL, we require $\tilde{O}((NK + NTM_{\text{SGD}})\frac{1}{1-\gamma})$ observations of true rewards. However, in RLHF, we do not need any observation of true rewards, but only use $O(NM_{\text{HF}})$ human queries. The ratio of the number of queries needed to the total sample complexity is about $\frac{NM_{\text{HF}}}{NTM_{\text{SGD}}} = \frac{1}{T}$. This theoretically explains the empirical success of RLHF — RLHF only needs a small amount of comparison queries to achieve good performance as standard RL (Christiano et al., 2017).

From the perspective of improving RLHF practice, our results provide two insights. Policy optimization can consist of three phases: sampling for exploration, policy evaluation and policy improvement. One of our insights is that, inserting reward model learning before multiple iterations of policy evaluation and improvement is efficient, i.e., we get policies that are nearly as good as the case where the rewards are known, while using only a small amount of human data compared to the overall sample complexity. Another insight is that querying human feedback on the state-action space which is rarely visited or is more likely induced by the optimal policy, helps improve the exploration of RLHF algorithms in reward estimation.

5. PO-RLHF with Neural Function Approximation

In this section, we turn to the neural setting. We design an efficient algorithm NN-PG-RLHF, and derive a biased MLE guarantee with neural approximation in analysis.

5.1. Algorithm NN-PG-RLHF

A detailed description and pseudo-code are provided in Appendix C. Here we provide a brief outline of the algorithm. NN-PG-RLHF actively collects human data as exploration, learns a reward network with human data, and trains a policy network and a Q-network to optimize the policy. Similar to PG-RLHF, NN-PG-RLHF estimates the feature covariance matrix and updates the coverage with the neural feature $\psi_{w^0}(s,a)$. Then, it generates preference data by past coverage, past policies and the baseline policy. The reward and Q-function networks are trained using an MLE loss function and a least-squares loss function, respectively.

Now we provides theoretical guarantees on NN-PG-RLHF. Let $\theta_*^{\text{NN},t} = \text{argmin}_{\|\theta-\theta^0\|\le R}\ \mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n}[(\psi_{w^0}(s,a)^\top\theta - (Q^{\pi^t}(s,a;r+b^n) - b^n(s,a)))^2]$ denote the optimal solution to the approximated version of the Q-network training objective with neural feature $\psi_{w^0}(s,a)$.

Similar to Eq. (4), we assume that the error of using the best fit $\theta_*^{\text{NN},t}$ under ρ_{cov}^n to predict the state-action value function under $d^\star_{s_{\text{init}}}$ is bounded.

Assumption 5.1 (Bounded Neural Transfer Error)**.** For any phase $n \ge 0$ and iteration $t \ge 0$, there exists some $\varepsilon_{\text{bias}}^{\text{NN}} > 0$ which satisfies

$$\mathbb{E}_{(s,a)\sim d^\star_{s_{\text{init}}}}\Big[\Big(\psi_{w^0}(s,a)^\top\theta_*^{\text{NN},t} - \big(Q^{\pi^t}(s,a;r+b^n) - b^n(s,a)\big)\Big)^2\Big] \le \varepsilon_{\text{bias}}^{\text{NN}}.$$

Theorem 5.2. *With probability at least $1-\delta$, the output policy of algorithm* NN-PG-RLHF *satisfies*

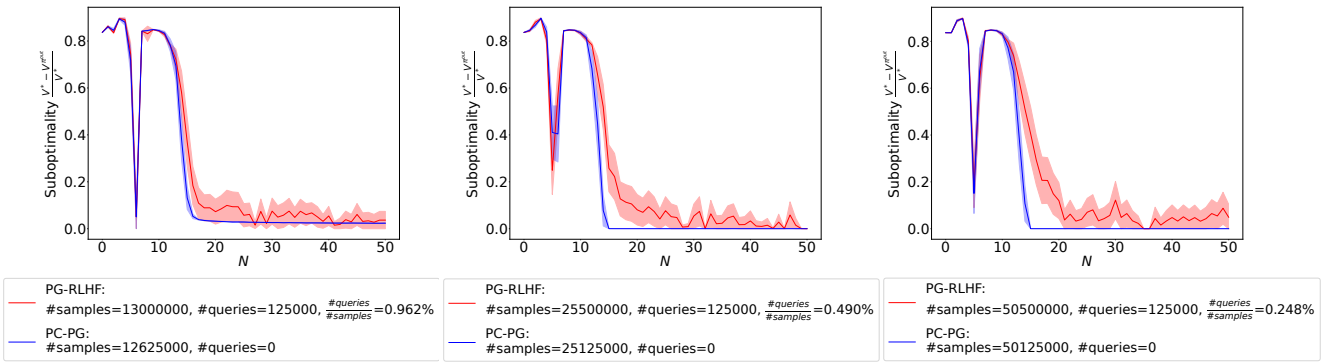
$$V^{\pi^*}(s_{\text{init}}) - V^{\pi^{\text{out}}}(s_{\text{init}}) \le \frac{2\sqrt{|\mathcal{A}|\varepsilon_{\text{bias}}^{\text{NN}}}}{1-\gamma} + \tilde{O}\Bigg(\frac{W^{\text{NN}}}{(1-\gamma)\sqrt{T}} + \frac{\sqrt{\beta NRW^{\text{NN}}}}{(1-\gamma)(M_{\text{SGD}}^\theta)^{\frac{1}{4}}} + \frac{m^{\frac{1}{4}}d^{\frac{1}{4}}\sqrt{\beta W^{\text{NN}}}}{c_{\text{base}}^{\frac{1}{4}}(1-\gamma)}\cdot\left(\frac{m^{\frac{1}{4}}d^{\frac{1}{4}}\sqrt{N}}{\sqrt{c_{\text{MLE}}^{\text{NN}}}M_{\text{HF}}^{\frac{1}{4}}} + \frac{W_\tau^{\frac{1}{4}}R^{\frac{1}{4}}\sqrt{N}}{(c_{\text{MLE}}^{\text{NN}})^{\frac{1}{4}}(M_{\text{SGD}}^\mu)^{\frac{1}{8}}}\right) + \frac{md}{(1-\gamma)N\beta} + B\left(\frac{1}{m^{\frac{1}{16}}}\right)\Bigg).$$

Here M_{SGD}^μ and M_{SGD}^θ are the numbers of iterations of the SGD for the reward network and Q-network training, respectively. $B(m^{-\frac{1}{16}})$ is a neural approximation error term scaling as $m^{-\frac{1}{16}}$. $W^{\text{NN}} := \sqrt{m}\bar{c} + R$, and $c_{\text{MLE}}^{\text{NN}} := (2+\exp(-2W_\tau W^{\text{NN}}) + \exp(2W_\tau W^{\text{NN}}))^{-1}$.

Theorem 5.2 demonstrates that the suboptimality becomes small with sufficiently large T, N, M_{HF}, M_{SGD}^θ and M_{SGD}^μ, up to the neural transfer error $O((\varepsilon_{\text{bias}}^{\text{NN}})^{\frac{1}{2}})$ and the neural approximation error $\tilde{O}(m^{-\frac{1}{16}})$. See the full bound in Eq. (44) in Appendix E.5.

Biased Neural MLE Analysis. Due to the gap between the true reward r and the functions that $h(s,a;\mu)$ can represent, our MLE reward training is biased. To tackle this difficulty, we develop a novel biased MLE analysis with neural approximation.

Specifically, let μ_r^{proj} be the network parameter of the projection of r onto neural function class $\{\psi_{\mu^0}(s,a)^\top\mu\}$. Then, we have that $\psi_{\mu^0}(s,a)^\top\mu_r^{\text{proj}}$ is close to r up to a neural approximation error scaling as $\frac{1}{m}$. Let μ_{MLE}^n be the optimal



(a) $T = 50$ (b) $T = 100$ (c) $T = 200$

Figure 1. Experimental results of algorithms PG-RLHF and PC-PG.

solution to the approximated version of the MLE objective with feature $\psi_{\mu^0}(s, a)$. Note that the human data are generated *almost* according to μ_r^{proj} (since it is close to r), and μ_{MLE}^n has a larger likelihood than μ_r^{proj}. Utilizing these two facts, we can bound $\|\mu_{\text{MLE}}^n - \mu_r^{\text{proj}}\|$ up to the standard MLE error $\tilde{O}(\frac{1}{\sqrt{M_{\text{HF}}}})$ and a neural approximation error. Furthermore, the SGD result $\hat{\mu}^n$ obtained in our algorithm is close to the MLE optimal solution μ_{MLE}^n up to the SGD error. Combining the SGD, MLE and neural approximation error, we can bound $\|\hat{\mu}^n - \mu_r^{\text{proj}}\|$. We refer interested readers to Lemma E.12 in Appendix E.4.

6. Experiments

In this section, we present experiments to demonstrate the practical efficacy of our algorithm and validate our theoretical results.

Following the experimental setup of existing algorithm PC-PG (Agarwal et al., 2020), we evaluate algorithms in an RL environment called Bidirectional Lock, which was also used in other prior works, e.g., (Zhang et al., 2021). The details of this environment are deferred to Appendix A.

In our experiments, $S = 22$, $A = 5$, $\gamma = 0.9$, $\delta = 0.005$, $\eta = 0.3$, $N = 30$, $T \in \{50, 100, 200\}$, $K = 2500$, $M_{\text{SGD}} = 2500$ and $M_{\text{HF}} = 2500$. The feature vectors $\phi(s, a)$ are one-hot vectors of state-actions, and $d = 110$. We compare our algorithm PG-RLHF with the standard RL algorithm PC-PG (Agarwal et al., 2020). Each algorithm is performed for 50 independent runs. Figure 1 plots the normalized suboptimalities of output policies $\frac{V^* - V^{\pi^{\text{out}}}}{V^*}$ with 95% confidence intervals, and reports the sample complexities and query complexities in the legend. (Since the numbers of samples and queries are computed before performing policy optimization and reward learning in algorithms PG-RLHF and PC-PG, the sample complexity and query complexity are the same for all runs.)

From Figure 1, we see that PG-RLHF effectively learns the optimal policy without observing true rewards, and achieves comparable performance to PC-PG while using a few more samples and a small amount of preference queries. When the number of iterations in policy optimization T increases, the ratio of query complexity to the overall sample complexity (scaling as $\frac{1}{T}$) decreases, which matches our theoretical results.

7. Conclusion

In this work, we study exploration-driven policy optimization for RLHF. For the linear and neural function approximation settings, we propose efficient algorithms with active human data collection. Through the comparison of results between RLHF and standard RL, we give a theoretical explanation for the query efficiency of RLHF. There is still a large space for future investigation. For example, it is interesting to explore other potential reasons behind the success of RLHF, e.g., the structural advantage of preference feedback over numerical feedback.

Acknowledgement

The work of Yihan Du, Anna Winnicki and R. Srikant is supported in part by AFOSR Grant FA9550-24-1-0002, ONR Grant N00014-19-1-2566, and NSF Grants CNS 23-12714, CNS 21-06801, CCF 19-34986, and CCF 22-07547.

Impact Statement

This work proposes efficient policy gradient RLHF algorithms with sample complexity guarantees, and provides a

theoretical insight for the empirical success of RLHF. We believe that this work may have potential societal impacts on RLHF applications, but as a theoretical work, it does not involve ethical concerns.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Agarwal, A., Henaff, M., Kakade, S., and Sun, W. PC-PG: Policy cover directed exploration for provable policy gradient learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13399–13412, 2020.

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *The Journal of Machine Learning Research*, 22(1):4431–4506, 2021.

Arora, S., Du, S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332. PMLR, 2019.

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Cai, Q., Yang, Z., Lee, J. D., and Wang, Z. Neural temporal-difference learning converges to global optima. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Chen, X., Zhong, H., Yang, Z., Wang, Z., and Wang, L. Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation. In *International Conference on Machine Learning*, pp. 3773–3793. PMLR, 2022.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Fan, J., Wang, Z., Xie, Y., and Yang, Z. A theoretical analysis of deep Q-learning. In *Learning for Dynamics and Control*, pp. 486–489. PMLR, 2020.

Hsu, D., Kakade, S., and Zhang, T. A tail inequality for quadratic forms of subgaussian random vectors. 2012.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Ji, X., Wang, H., Chen, M., Zhao, T., and Wang, M. Provable benefits of policy learning from human preferences in contextual bandit problems. *arXiv preprint arXiv:2307.12975*, 2023.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, pp. 267–274, 2002.

Kaufmann, T., Weng, P., Bengs, V., and Hüllermeier, E. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*, 2023.

Li, Z., Yang, Z., and Wang, M. Reinforcement learning with human feedback: Learning dynamic choices via pessimism. *arXiv preprint arXiv:2305.18438*, 2023.

Liu, B., Cai, Q., Yang, Z., and Wang, Z. Neural trust region/proximal policy optimization attains globally optimal policy. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Novoseller, E., Wei, Y., Sui, Y., Yue, Y., and Burdick, J. Dueling posterior sampling for preference-based reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1029–1038. PMLR, 2020.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Pacchiano, A., Saha, A., and Lee, J. Dueling RL: reinforcement learning with trajectory preferences. *arXiv preprint arXiv:2111.04850*, 2021.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Rahimi, A. and Recht, B. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in Neural Information Processing Systems*, volume 21, 2008.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Tropp, J. A. et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Wang, L., Cai, Q., Yang, Z., and Wang, Z. Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*, 2019.

Wang, Y., Liu, Q., and Jin, C. Is RLHF more difficult than standard rl? In *Advances in Neural Information Processing Systems*, 2023.

Wu, R. and Sun, W. Making RL with preference-based feedback efficient via randomization. *arXiv preprint arXiv:2310.14554*, 2023.

Xiong, W., Dong, H., Ye, C., Zhong, H., Jiang, N., and Zhang, T. Gibbs sampling from human feedback: A provable KL-constrained framework for RLHF. *arXiv preprint arXiv:2312.11456*, 2023.

Xu, T., Liang, Y., and Lan, G. CRPO: A new approach for safe reinforcement learning with convergence guarantee. In *International Conference on Machine Learning*, pp. 11480–11491. PMLR, 2021.

Xu, Y., Wang, R., Yang, L., Singh, A., and Dubrawski, A. Preference-based reinforcement learning with finite-time guarantees. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18784–18794, 2020.

Yang, L. and Wang, M. Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning*, pp. 6995–7004. PMLR, 2019.

Zanette, A., Cheng, C.-A., and Agarwal, A. Cautiously optimistic policy optimization and exploration with linear function approximation. In *Conference on Learning Theory*, pp. 4473–4525. PMLR, 2021.

Zhan, W., Uehara, M., Kallus, N., Lee, J. D., and Sun, W. Provable offline reinforcement learning with human feedback. *arXiv preprint arXiv:2305.14816*, 2023a.

Zhan, W., Uehara, M., Sun, W., and Lee, J. D. How to query human feedback efficiently in RL? *arXiv preprint arXiv:2305.18505*, 2023b.

Zhang, T., Rashidinejad, P., Jiao, J., Tian, Y., Gonzalez, J. E., and Russell, S. Made: Exploration via maximizing deviation from explored regions. *Advances in Neural Information Processing Systems*, 34:9663–9680, 2021.

Zhu, B., Jiao, J., and Jordan, M. I. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. *arXiv preprint arXiv:2301.11270*, 2023.

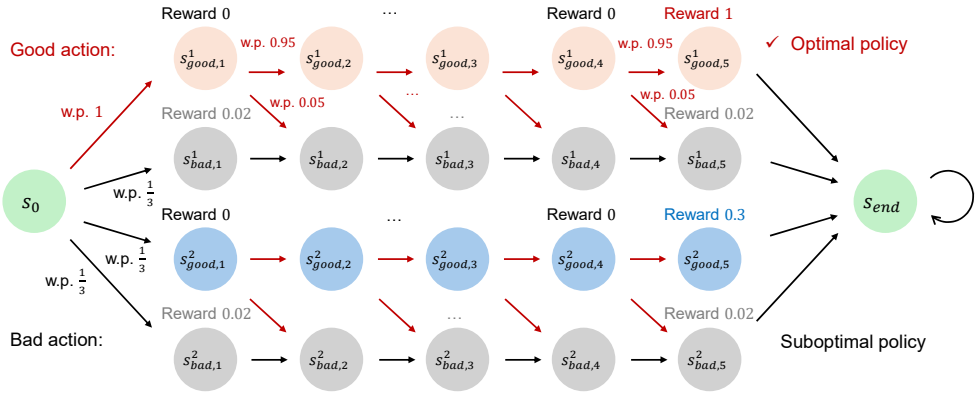


Figure 2. The Bidirectional Lock environment.

A. Details of the Experimental Environment

In this section, we describe the Bidirectional Lock environment used in our experiments.

As shown in Figure 2, there are two locks, and each of them has $2H$ states. These states are denoted by $s^{\ell}_{good,h}$ and $s^{\ell}_{bad,h}$, were $\ell \in \{1,2\}$ is the index of the lock, and $h \in [H]$ is the horizon of the lock. In addition to these $2H$ states, there are an initial state s_0 and an absorbing ending state $s_\perp$. There are 5 actions, including a good action a_{good} and 4 bad actions. The reward function depends only on states. Only the last good states of locks give high rewards, i.e., $r(s^1_{good,H},\cdot) = 1$ and $r(s^2_{good,H},\cdot) = 0.3$. Other good states of locks $\{s^{\ell}_{good,h}\}_{\ell\in\{1,2\},h\in[H-1]}$, s_0 and $s_\perp$ induce zero reward. The bad states of locks give tiny rewards, i.e., $r(s^{\ell}_{bad,h},\cdot) = \frac{0.1}{H}$ for any $\ell \in \{1,2\}$ and $h \in [H]$. We set $H = 5$ in our experiments.

The agent starts from s_0. Under a_{good}, she transitions to $s^1_{good,1}$ deterministically; Under other actions, she transitions to $s^1_{bad,1}$, $s^2_{good,1}$ and $s^2_{bad,1}$ all with probability $\frac{1}{3}$. For any $\ell \in \{1,2\}$ and $h \in [H-1]$, in state $s^{\ell}_{good,h}$, under a_{good}, the agent transitions to $s^{\ell}_{good,h+1}$ and $s^{\ell}_{bad,h+1}$ with probabilities 0.95 and 0.05, respectively; Under other actions, she transitions to $s^{\ell}_{bad,h+1}$ deterministically. For any $\ell \in \{1,2\}$ and $h \in [H-1]$, in state $s^{\ell}_{bad,h}$, the agent transitions to $s^{\ell}_{bad,h+1}$ deterministically under any action. Once the agent achieves $s^{\ell}_{good,H}$ or $s^{\ell}_{bad,H}$ for any $\ell \in \{1,2\}$, she transitions to $a_\perp$ deterministically.

This environment has sparse rewards. The optimal policy is to always take a_{good}, and obtain the high final reward of lock 1. A suboptimal (myopic) policy results in getting stuck in bad states or only obtaining the low final reward of lock 2.

B. Detailed Review of Related Works

In the following, we present a more detailed review of related works.

RLHF. RLHF (Christiano et al., 2017; Kaufmann et al., 2023) has gained a huge empirical success, especially in LLMs (Ouyang et al., 2022; Achiam et al., 2023). Recently, a number of works have emerged to theoretically analyze RLHF. Xu et al. (2020); Novoseller et al. (2020); Pacchiano et al. (2021) study online RLHF for tabular MDPs. Chen et al. (2022); Wang et al. (2023) consider online RLHF with general function approximation. Wang et al. (2023) design a reduction framework for RLHF, and prove that the sample complexity for RLHF is no higher than that for standard RL. Zhu et al. (2023); Zhan et al. (2023a); Li et al. (2023) study offline RLHF with function approximation. Xiong et al. (2023) introduce a KL-constrained framework for RLHF, and Zhan et al. (2023b); Wu & Sun (2023) consider how to optimize query complexity via experimental design and posterior sampling. Ji et al. (2023) also seek to understand the empirical success of RLHF in the offline contextual bandit setting, but different from our work, Ji et al. (2023) explain it from the perspective of intrinsic human data bias.

In contrast to the above works which most consider value-based algorithms, we analyze policy gradient RLHF algorithms with exploration, and theoretically explain why RLHF only needs a small amount of human feedback to attain good performance, from the perspective of the efficiency of RLHF (reward learning) algorithmic procedure itself.

Algorithm 3 NN-PG-RLHF

Input: $\varepsilon, \delta, N, K, M_{\text{HF}}, \zeta_{\text{HF}}, \zeta_{\text{cov}}, \pi^{\text{base}}, \pi^0$.
Initialize $\alpha^0 = 1$ and $[\mu^0]_\ell, [w^0]_\ell \sim \mathcal{D}_{\text{init}}, \forall \ell \in [m]$.
for $n = 0, \ldots, N-1$ **do**
  Sample $\{s_i, a_i\}_{i=1}^K \sim d_{s_{\text{init}}}^{\pi^n}$, and $\hat{\Sigma}^n \leftarrow \frac{1}{K}\sum_{i=1}^K \psi_{w^0}(s_i, a_i)\psi_{w^0}(s_i, a_i)^\top$
  $\hat{\Sigma}_{\text{cov}}^n \leftarrow \sum_{i=0}^n \hat{\Sigma}^i + \zeta_{\text{cov}} I$
  Let $\rho_{\text{cov}}^n := \frac{1}{n+1}\sum_{i=0}^n d_{s_{\text{init}}}^{\pi^i}$
  $\mathcal{O}_{\text{HF}}^n := \frac{1}{n}\sum_{i=1}^n \mathcal{O}_{\rho_{\text{cov}}^{i-1}}^{\pi^i}, \forall n \geq 1$, and $\mathcal{O}_{\text{HF}}^0 := \mathcal{O}_{s_{\text{init}}}^{\pi^0}$
  for $i = 1, \ldots, M_{\text{HF}}$ **do**
    Sample trajectories $\tau_i^{(1)} \sim \mathcal{O}_{\text{HF}}^n$ and $\tau_i^{(2)} \sim \mathcal{O}_{s_{\text{init}}}^{\pi^{\text{base}}}$
    Observe the preference outcome y_i^r
  end for
  Train the reward network $h(s, a; \mu^0)$ with the MLE objective Eq. (9) by projected SGD, and obtain $\hat{\mu}^n$
  $\pi^{n+1} \leftarrow$ NN-NPG-Update$(\rho_{\text{cov}}^n, \hat{\Sigma}_{\text{cov}}^n, \hat{\mu}^n)$
end for
return Unif$(\pi^1, \ldots, \pi^N)$

Algorithm 4 NN-NPG-Update

Input: $\rho_{\text{cov}}^n, \hat{\Sigma}_{\text{cov}}^n, \hat{\mu}^n, \eta, T, \beta, \alpha^0, w^0$.
Let $\hat{r}^n(\cdot, \cdot) := h(\cdot, \cdot; \hat{\mu}^n)$
$b^n(\cdot, \cdot) := \frac{1}{1-\gamma}\mathbb{1}\{\psi_{w^0}(\cdot, \cdot)^\top (\hat{\Sigma}_{\text{cov}}^n)^{-1}\psi_{w^0}(\cdot, \cdot) \geq \beta\}$
Let $\mathcal{K}^n := \{s \in \mathcal{S} : \forall a \in \mathcal{A},\ b^n(s, a) = 0\}$
For $s \in \mathcal{K}^n$, $\pi^0(\cdot|s) := \pi_{\alpha^0, w^0}$. For $s \notin \mathcal{K}^n$, $\pi^0(\cdot|s) :=$ Unif$(\{a \in \mathcal{A} : b^n(s, a) = \frac{1}{1-\gamma}\})$
for $t = 0, \ldots, T-1$ **do**
  $\theta^{t,0} \leftarrow w^0$
  Train the Q-network $f(s, a; \theta^{t,0})$ with the objective Eq. (10) by projected SGD, and obtain θ^t
  Update policy network: $\alpha^{t+1} w^{t+1} \leftarrow \alpha^t w^t + \eta\theta^t$
  $\forall s \in \mathcal{K}^n$, $\pi^{t+1}(\cdot|s) = \pi_{\alpha^{t+1}, w^{t+1}}(\cdot|s) \propto \exp(\alpha^{t+1} f(s, a; w^{t+1}))$. $\forall s \notin \mathcal{K}^n$, $\pi^{t+1}(\cdot|s) = \pi^0(\cdot|s)$
end for
return Unif$(\pi^0, \ldots, \pi^{T-1})$

RL with Neural Function Approximation. There have been several theoretical RL works, e.g., (Cai et al., 2019; Wang et al., 2019; Liu et al., 2019; Fan et al., 2020; Xu et al., 2021), use neural networks to approximate value functions and policies, and provide guarantees based on existing analysis for overparameterized neural networks (Jacot et al., 2018; Arora et al., 2019). Our work also considers neural function approximation for the RLHF environment. In addition, our work is also related to (Agarwal et al., 2020), which designs a policy gradient algorithm enabling exploration for standard RL.

C. Detailed Description of Algorithm NN-PG-RLHF

In this section, we present the pseudo-code of algorithm NN-PG-RLHF, and give a more detailed algorithm description.

Algorithm 3 illustrates the procedure of NN-PG-RLHF. Similar to PG-RLHF, in each phase n, NN-PG-RLHF first estimates the feature covariance matrix $\hat{\Sigma}_{\text{HF}}^n$ and updates the coverage distribution ρ_{cov}^n. Then, it generates M_{HF} pairs of preference data using past coverage distributions ρ_{cov}^{i-1}, past policies π^i and a baseline policy π^{base} $(i = 0, 1, \ldots, n)$. With these data, PG-RLHF trains the reward network $h(s, a; \mu^0)$ to minimize the following MLE objective by projected SGD (Line 12):

$$\underset{\|\mu - \mu^0\|_2 \leq R}{\operatorname{argmin}} \Bigg(-\sum_{i=1}^{M_{\text{HF}}} \log\Big(\frac{\mathbb{1}\{y_i = 1\}}{1 + \exp\big(-\tilde{h}(\tau_i^{(1)}, \tau_i^{(2)}; \mu)\big)} + \frac{\mathbb{1}\{y_i = 0\}}{1 + \exp\big(\tilde{h}(\tau_i^{(1)}, \tau_i^{(2)}; \mu)\big)} \Big) \Bigg), \tag{9}$$

where $\tilde{h}(\tau_i^{(1)}, \tau_i^{(2)}; \mu) := \sum_{h=0}^{H(\tau_i^{(1)})} h(s_{i,h}^{(1)}, a_{i,h}^{(1)}; \mu) - \sum_{h=0}^{H(\tau_i^{(2)})} h(s_{i,h}^{(2)}, a_{i,h}^{(2)}; \mu)$. After training, we call a subroutine NN-NPG-Update (Algorithm 4) with ρ_{cov}^n, $\hat{\Sigma}_{\text{cov}}$ and $h(\cdot, \cdot; \hat{\mu}^n)$ to perform policy optimization.

In NN-NPG-Update, we train the Q-network $f(s,a;\theta^{t,0})$ to fit the state-action value function with initial distribution ρ_{cov}^n by project SGD (Line 8):

$$\underset{\|\theta-\theta^0\|\le R}{\operatorname{argmin}} \mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n}\left[\left(f(s,a;\theta)-\left(Q^{\pi^t}(s,a;\hat{r}^n+b^n)-b^n(s,a)\right)\right)^2\right]. \tag{10}$$

With the trained Q-network $f(s,a;\theta^t)$, we update the policy network parameter $\alpha^{t+1}w^{t+1}$ using θ^t (Line 9). After the natural policy gradient, we obtain an improved policy network π^{n+1}, which is used to improve the coverage and guide the human data collection in the next phase.

D. Proofs for PO-RLHF with Linear Function Approximation

In this section, we give the proofs for algorithm PG-RLHF. In our analysis, the ideas of MDP construction and natural policy gradient (Lemmas D.1-D.6) for optimistic MDPs are originated from (Agarwal et al., 2020).

D.1. MDP construction

We consider three MDPs as follows: (i) The true MDP $\mathcal{M}$. (ii) The optimistic MDP with exploration bonuses $\mathcal{M}_{b^n}$. $\mathcal{M}_{b^n}$ replaces the reward function in $\mathcal{M}$ by $r(s,a)+b^n(s,a)$. (iii) The $(\pi^*,\mathcal{K}^n)$-modified optimistic MDP $\mathcal{M}^n$. $\mathcal{M}^n$ is the same as $\mathcal{M}_{b^n}$ except that, for any $s\notin\mathcal{K}^n$, $\mathcal{M}^n$ adds an additional action $a^\dagger$ whose reward function and transition distribution are

$$r^n(s,a^\dagger)=1,\quad p^n(s|s,a^\dagger)=1.$$

In $\mathcal{M}^n$, we consider a modified version of π^*, denoted by $\pi^{*,n}$. For any $s\in\mathcal{K}^n$, $\pi^{*,n}(\cdot|s)=\pi^*(\cdot|s)$. For any $s\notin\mathcal{K}^n$, $\pi^{*,n}(a^\dagger|s)=1$. Thus, in $\mathcal{M}^n$, under policy $\pi^{*,n}$, once the agent goes into some $s\notin\mathcal{K}^n$, she will self-loop and keep receiving the reward 1.

Lemma D.1. *For any phase $n\ge 0$, iteration $t\ge 0$, $s\in\mathcal{S}$ and $a\ne a^\dagger$,*

$$V_{\mathcal{M}^n}^{\pi^t}(s)=V_{\mathcal{M}_{b^n}}^{\pi^t}(s),\quad Q_{\mathcal{M}^n}^{\pi^t}(s,a)=Q_{\mathcal{M}_{b^n}}^{\pi^t}(s,a),\quad A_{\mathcal{M}^n}^{\pi^t}(s,a)=A_{\mathcal{M}_{b^n}}^{\pi^t}(s,a).$$

Proof. This lemma follows from the fact that $\mathcal{M}^n$ is the same as $\mathcal{M}_{b^n}$ except that $\mathcal{M}^n$ has an additional action $a^\dagger$, but π^t never picks $a^\dagger$. □

Lemma D.2 (Lemma C.1 in (Agarwal et al., 2020))**.** *For any phase $n\ge 0$, $s\in\mathcal{K}^n$ and $a\in\mathcal{A}$,*

$$d_{\mathcal{M}^n}^{\pi^{*,n}}(s,a)\le d_{\mathcal{M}}^{\pi^*}(s,a).$$

Lemma D.3 (Lemma C.2 in (Agarwal et al., 2020))**.** *For any phase $n\ge 0$ and iteration $t\ge 0$,*

$$\begin{aligned}&V_{\mathcal{M}^n}^{\pi^{*,n}}(s_{\text{init}})\ge V_{\mathcal{M}}^{\pi^*}(s_{\text{init}}),\\ &V_{\mathcal{M}^n}^{\pi^t}(s_{\text{init}})=V_{\mathcal{M}_{b^n}}^{\pi^t}(s_{\text{init}})\le V_{\mathcal{M}}^{\pi^t}(s_{\text{init}})+\frac{1}{1-\gamma}\sum_{(s,a)\notin\mathcal{K}^n}d_{s_{\text{init}}}^{\pi^t}(s,a).\end{aligned}$$

Lemma D.4 (Lemma C.3 in (Agarwal et al., 2020))**.** *For any phase $n\ge 0$,*

$$\sum_{(s,a)\notin\mathcal{K}^n}d_{s_{\text{init}}}^{\pi^{n+1}}(s,a)\le\frac{1}{\beta}\mathbb{E}_{(s,a)\sim d_{s_{\text{init}}}^{\pi^{n+1}}}\left[\phi(s,a)^\top(\hat{\Sigma}_{\text{cov}}^n)^{-1}\phi(s,a)\right].$$

Furthermore, it holds that

$$\begin{aligned}\sum_{n=0}^{N-1}\sum_{(s,a)\notin\mathcal{K}^n}d_{s_{\text{init}}}^{\pi^{n+1}}(s,a)&\le\frac{2}{\beta}\log\left(\frac{\det\left(\zeta_{\text{cov}}I+\sum_{i=1}^N\mathbb{E}_{(s,a)\sim d_{s_{\text{init}}}^{\pi^i}}\left[\phi(s,a)\phi(s,a)^\top\right]\right)}{\det\left(\zeta_{\text{cov}}I\right)}\right)\\&\le\frac{2d}{\beta}\log\left(1+\frac{N}{\zeta_{\text{cov}}d}\right).\end{aligned}$$

D.2. Performance Difference Lemma and Policy Gradient on $\mathcal{M}^n$

Lemma D.5 (Performance Difference Lemma on $\mathcal{M}^n$)**.** *For any phase $n \geq 0$ and iteration $t \geq 0$,*

$$V_{\mathcal{M}^n}^{\pi^{*,n}}(s_{\text{init}}) - V_{\mathcal{M}^n}^{\pi^t}(s_{\text{init}}) \leq \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d_{\mathcal{M}^n;s_{\text{init}}}^{\pi^{*,n}}}\left[A_{\mathcal{M}_{b^n}}^{\pi^t}(s,a)\cdot \mathbb{1}\left\{s\in\mathcal{K}^n\right\}\right].$$

Proof. For any phase $n \geq 0$ and iteration $t \geq 0$, using the standard performance difference lemma (Kakade & Langford, 2002), we have

$$\begin{aligned}
V_{\mathcal{M}^n}^{\pi^{*,n}}(s_{\text{init}}) - V_{\mathcal{M}^n}^{\pi^t}(s_{\text{init}}) &= \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d_{\mathcal{M}^n;s_{\text{init}}}^{\pi^{*,n}}}\left[A_{\mathcal{M}^n}^{\pi^t}(s,a)\right] \\
&= \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d_{\mathcal{M}^n;s_{\text{init}}}^{\pi^{*,n}}}\left[A_{\mathcal{M}^n}^{\pi^t}(s,a)\cdot \mathbb{1}\left\{s\in\mathcal{K}^n\right\}\right] \\
&\quad + \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d_{\mathcal{M}^n;s_{\text{init}}}^{\pi^{*,n}}}\left[A_{\mathcal{M}^n}^{\pi^t}(s,a)\cdot \mathbb{1}\left\{s\notin\mathcal{K}^n\right\}\right].
\end{aligned}$$

In $\mathcal{M}^n$, for any $s \notin \mathcal{K}^n$, policy $\pi^{*,n}$ chooses $a^\dagger$ deterministically. For any $s \notin \mathcal{K}^n$, we have

$$\begin{aligned}
A_{\mathcal{M}^n}^{\pi^t}(s,a^\dagger) &= Q_{\mathcal{M}^n}^{\pi^t}(s,a^\dagger) - V_{\mathcal{M}^n}^{\pi^t}(s) \\
&= 1 + \gamma V_{\mathcal{M}^n}^{\pi^t}(s) - V_{\mathcal{M}^n}^{\pi^t}(s) \\
&= 1 - (1-\gamma)V_{\mathcal{M}^n}^{\pi^t}(s) \\
&\overset{\text{(a)}}{\leq} 1 - (1-\gamma)\left(0 + \frac{1}{1-\gamma}\right) \\
&= 0,
\end{aligned}$$

where inequality (a) is due to the facts that for any $s \notin \mathcal{K}^n$, $\pi^t(\cdot|s) = \text{Unif}(\{a \in \mathcal{A} : b^n(s,a) = \frac{1}{1-\gamma}\})$, and that $V_{\mathcal{M}^n}^{\pi^t}(s)$ is no smaller than the cumulative reward 0 plus the exploration bonus $b^n(s,a) = \frac{1}{1-\gamma}$.

Therefore,

$$\begin{aligned}
V_{\mathcal{M}^n}^{\pi^{*,n}}(s_{\text{init}}) - V_{\mathcal{M}^n}^{\pi^t}(s_{\text{init}}) &\leq \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d_{\mathcal{M}^n;s_{\text{init}}}^{\pi^{*,n}}}\left[A_{\mathcal{M}^n}^{\pi^t}(s,a)\cdot \mathbb{1}\left\{s\in\mathcal{K}^n\right\}\right] \\
&\overset{\text{(b)}}{=} \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d_{\mathcal{M}^n;s_{\text{init}}}^{\pi^{*,n}}}\left[A_{\mathcal{M}_{b^n}}^{\pi^t}(s,a)\cdot \mathbb{1}\left\{s\in\mathcal{K}^n\right\}\right],
\end{aligned}$$

where inequality (b) is due to that $A_{\mathcal{M}^n}^{\pi^t}(s,a) = A_{\mathcal{M}_{b^n}}^{\pi^t}(s,a)$ for $a \neq a^\dagger$ (Lemma D.1), and $\pi^{*,n}$ never picks $a^\dagger$ for any state $s \in \mathcal{K}^n$. □

Let $W_A := \frac{4}{(1-\gamma)^2}$ and $\eta \leq \frac{1}{W_A}$. Then, $|\hat{A}_{\mathcal{M}_{b^n}}^{\pi^t}(s,a)| \leq W_A$ for all $n \geq 0$, $t \geq 0$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$.

Lemma D.6 (Regret for Natural Policy Gradient)**.** *For any phase $n \geq 0$ and iteration $t \geq 0$,*

$$\sum_{t=0}^{T-1}\mathbb{E}_{(s,a)\sim d_{\mathcal{M}^n;s_{\text{init}}}^{\pi^{*,n}}}\left[\hat{A}_{\mathcal{M}_{b^n}}^{\pi^t}(s,a)\cdot \mathbb{1}\left\{s\in\mathcal{K}^n\right\}\right] \leq \frac{\log(|\mathcal{A}|)}{\eta} + \eta W_A^2 T.$$

Proof. For any phase $n \geq 0$, iteration $t \geq 0$, $s \in \mathcal{K}^n$ and $a \in \mathcal{A}$, we have $b^n(s,a) = 0$.

Define

$$\begin{aligned}
D_s &:= \sum_{a'\in\mathcal{A}}\left(\exp\left(\phi(s,a')^\top w^t\right)\right), \\
E_s &:= \exp\left(-\eta\mathbb{E}_{a\sim\pi^t(\cdot|s)}\left[\phi(s,a)^\top\theta^t\right]\right) \\
&= \exp\left(-\eta\mathbb{E}_{a\sim\pi^t(\cdot|s)}\left[\phi(s,a)^\top\theta^t + b^n(s,a)\right]\right)
\end{aligned}$$

$$= \exp\left(-\eta \hat{V}^{\pi^t}_{\mathcal{M}_{b^n}}(s)\right),$$

and we have

$$\begin{aligned}
\pi^{t+1}(\cdot|s) &= \frac{\exp\left(\phi(s,\cdot)^\top w^{t+1}\right)}{\sum_{a'\in\mathcal{A}}\left(\phi(s,a')^\top w^{t+1}\right)} \\
&= \frac{\exp\left(\phi(s,\cdot)^\top \left(w^t + \eta\theta^t\right)\right)}{\sum_{a'\in\mathcal{A}}\left(\phi(s,a')^\top \left(w^t + \eta\theta^t\right)\right)} \\
&= \frac{\frac{\exp\left(\phi(s,\cdot)^\top w^t\right)}{D_s}\cdot \exp\left(\eta\left(\phi(s,\cdot)^\top\theta^t + b^n(s,\cdot)\right)\right)}{\sum_{a'\in\mathcal{A}}\left(\frac{\exp(\phi(s,a')^\top w^t)}{D_s}\cdot \exp\left(\eta\left(\phi(s,a')^\top\theta^t + b^n(s,a')\right)\right)\right)} \\
&= \frac{\pi^t(\cdot|s)\cdot\exp\left(\eta\hat{Q}^{\pi^t}_{\mathcal{M}_{b^n}}(s,\cdot)\right)\cdot E_s}{\sum_{a'\in\mathcal{A}}\left(\pi^t(a'|s)\cdot\exp\left(\eta\hat{Q}^{\pi^t}_{\mathcal{M}_{b^n}}(s,a')\right)\cdot E_s\right)} \\
&= \frac{\pi^t(\cdot|s)\cdot\exp\left(\eta\hat{A}^{\pi^t}_{\mathcal{M}_{b^n}}(s,\cdot)\right)}{\sum_{a'\in\mathcal{A}}\left(\pi^t(a'|s)\cdot\exp\left(\eta\hat{A}^{\pi^t}_{\mathcal{M}_{b^n}}(s,a')\right)\right)} \\
&= \frac{\pi^t(\cdot|s)\cdot\exp\left(\eta\hat{A}^{\pi^t}_{\mathcal{M}_{b^n}}(s,\cdot)\right)}{\sum_{a'\in\mathcal{A}}\left(\pi^t(a'|s)\cdot\exp\left(\eta\hat{A}^{\pi^t}_{\mathcal{M}_{b^n}}(s,a')\right)\right)}.
\end{aligned}$$

Define $G_s := \sum_{a'\in\mathcal{A}}(\pi^t(a'|s)\cdot\exp(\eta\hat{A}^{\pi^t}_{\mathcal{M}_{b^n}}(s,a')))$, and we have

$$\begin{aligned}
\log(G_s) &= \log\left(\sum_{a'\in\mathcal{A}}\left(\pi^t(a'|s)\cdot\exp\left(\eta\hat{A}^{\pi^t}_{\mathcal{M}_{b^n}}(s,a')\right)\right)\right) \\
&\overset{\text{(a)}}{\leq} \log\left(\sum_{a'\in\mathcal{A}}\left(\pi^t(a'|s)\cdot\left(1+\eta\hat{A}^{\pi^t}_{\mathcal{M}_{b^n}}(s,a') + \left(\eta\hat{A}^{\pi^t}_{\mathcal{M}_{b^n}}(s,a')\right)^2\right)\right)\right) \\
&\leq \log\left(1+\eta^2 W_A^2\right) \\
&\leq \eta^2 W_A^2,
\end{aligned}$$

where inequality (a) is due to that $\eta\hat{A}^{\pi^t}_{\mathcal{M}_{b^n}}(s,a') \leq \eta W_A \leq 1$ and $\exp(x) \leq 1 + x + x^2$ for any $x \leq 1$.

Thus, for any $s \in \mathcal{K}^n$, we have

$$\begin{aligned}
&\mathrm{KL}(\pi^{*,n}(\cdot|s)\|\pi^{t+1}(\cdot|s)) - \mathrm{KL}(\pi^{*,n}(\cdot|s)\|\pi^t(\cdot|s)) \\
&= \mathbb{E}_{a\sim\pi^{*,n}(\cdot|s)}\left[\log\left(\frac{\pi^{*,n}(a|s)}{\pi^{t+1}(a|s)}\right)\right] - \mathbb{E}_{a\sim\pi^{*,n}(\cdot|s)}\left[\log\left(\frac{\pi^{*,n}(a|s)}{\pi^{t}(a|s)}\right)\right] \\
&= \mathbb{E}_{a\sim\pi^{*,n}(\cdot|s)}\left[\log\left(\frac{\pi^{t}(a|s)}{\pi^{t+1}(a|s)}\right)\right] \\
&= \mathbb{E}_{a\sim\pi^{*,n}(\cdot|s)}\left[\log(G_s) - \eta\hat{A}^{\pi^t}_{\mathcal{M}_{b^n}}(s,a)\right] \\
&\leq -\eta\mathbb{E}_{a\sim\pi^{*,n}(\cdot|s)}\left[\hat{A}^{\pi^t}_{\mathcal{M}_{b^n}}(s,a)\right] + \eta^2 W_A^2,
\end{aligned}$$

which is equivalent to

$$\mathbb{E}_{a\sim\pi^{*,n}(\cdot|s)}\left[\hat{A}^{\pi^t}_{\mathcal{M}_{b^n}}(s,a)\right] \leq \frac{1}{\eta}\left(\mathrm{KL}(\pi^{*,n}(\cdot|s)\|\pi^t(\cdot|s)) - \mathrm{KL}(\pi^{*,n}(\cdot|s)\|\pi^{t+1}(\cdot|s))\right) + \eta W_A^2.$$

Adding $s \sim d^{\pi^{*,n}}_{\mathcal{M}^n;s_{\text{init}}}$ on both sides and summing over $t = 0,\ldots,T-1$, we have

$$\sum_{t=0}^{T-1}\mathbb{E}_{(s,a)\sim d^{\pi^{*,n}}_{\mathcal{M}^n;s_{\text{init}}}}\left[\hat{A}^{\pi^t}_{\mathcal{M}_{b^n}}(s,a)\cdot\mathbb{1}\left\{s\in\mathcal{K}^n\right\}\right] \leq \frac{1}{\eta}\mathbb{E}_{s\sim d^{\pi^{*,n}}_{\mathcal{M}^n;s_{\text{init}}}}\left[\mathrm{KL}(\pi^{*,n}(\cdot|s)\|\pi^0(\cdot|s)) - \mathrm{KL}(\pi^{*,n}(\cdot|s)\|\pi^T(\cdot|s))\right]$$

$$
\begin{aligned}
&+ \eta W_A^2 T \\
&\leq \frac{\log(|\mathcal{A}|)}{\eta} + \eta W_A^2 T.
\end{aligned}
$$

□

D.3. Human Feedback

For any trajectory $\tau = (s_0, a_0, \ldots, s_{H(\tau)}, a_{H(\tau)})$, let $H(\tau)$ denote the length of τ, and $\phi(\tau) := \sum_{h=0}^{H(\tau)} \phi(s_h, a_h)$. For any trajectories $\tau^{(1)}, \tau^{(2)}$, let $\tilde{\phi}^{\tau^{(1)},\tau^{(2)}} := \sum_{h=0}^{H(\tau^{(1)})} \phi(s_h^{(1)}, a_h^{(1)}) - \sum_{h=0}^{H(\tau^{(2)})} \phi(s_h^{(2)}, a_h^{(2)})$.

For any $(s, a) \in \mathcal{S} \times \mathcal{A}$ and policy π, let $\mathcal{O}_{s,a}^{\pi}$ be the distribution of the trajectory which is generated by starting at (s, a), executing policy π and terminating with probability $1 - \gamma$ at each step. For any state-action distribution ρ, let $\mathcal{O}_{\rho}^{\pi} := \mathbb{E}_{\rho \sim (s,a)}[\mathcal{O}_{s,a}^{\pi}]$.

D.3.1. Trajectory Length and Covariance Matrix Concentration

To analyze the reward estimation error under human feedback, we first define the concentration events for trajectory length and the coverage and human data covariance matrices.

Define event

$$
\mathcal{E}_{\tau} := \left\{ |\tau| \leq \frac{\log\left(\frac{1}{\delta'}\right)}{1-\gamma} := W_{\tau}, \text{ for any trajectory } \tau \text{ sampled in the algorithm} \right\}. \tag{11}
$$

Lemma D.7. *It holds that* $\Pr[\mathcal{E}_{\tau}] \geq 1 - 2N(K + M_{\text{HF}} + TM_{\text{SGD}})\delta'$.

Proof. This proof is similar to Eqs. (94)-(97) in (Zanette et al., 2021).

Let H denote the length of a trajectory which is generated by terminating with probability $1 - \gamma$ at each step. Then, H is a random variable which satisfies $\Pr[H = t] = \gamma^{t-1}(1 - \gamma)$ for $t = 1, 2, \ldots$.

We have

$$
\Pr[H > h] = \sum_{t=h+1}^{\infty} \gamma^{t-1}(1-\gamma) = \gamma^h \sum_{t=1}^{\infty} \gamma^{t-1}(1-\gamma) = \gamma^h \sum_{t=1}^{\infty} \gamma^{t-1}(1-\gamma) = \gamma^h.
$$

Let $\delta' = \gamma^h$. Then,

$$
h = \frac{\ln(\delta')}{\ln(\gamma)} = \frac{-\ln(\delta')}{-\ln(\gamma)} \leq \frac{-\ln(\delta')}{-(\gamma - 1)} \leq \frac{\ln\left(\frac{1}{\delta'}\right)}{1-\gamma}.
$$

Thus, we have

$$
\Pr\left[H > \frac{\ln\left(\frac{1}{\delta'}\right)}{1-\gamma} \right] \leq \delta'.
$$

□

Let $\zeta_{\text{cov}} := 1$ and $\zeta_{\text{HF}} := 4W_{\tau}^2$. For any $n \geq 0$ and $1 \leq i \leq K$, let (s_i^n, a_i^n) denote the i-th state-action pair sampled in phase n for constructing the estimated coverage covariance matrix $\hat{\Sigma}_{\text{cov}}^n$ (Line 3 in Algorithm 1).

For any phase $n \geq 0$, let

$$
\hat{\Sigma}_{\text{cov}}^n := \sum_{i=0}^{n} \left(\frac{1}{K} \sum_{i=1}^{K} \phi(s_i^n, a_i^n) \phi(s_i^n, a_i^n)^{\top} \right) + \zeta_{\text{cov}} I,
$$

$$
\begin{aligned}
\Sigma_{\text{cov}}^n &:= \sum_{i=0}^{n} \mathbb{E}_{(s,a)\sim d_{s_{\text{init}}}^{\pi^n}} \left[\phi(s,a)\phi(s,a)^\top\right] + \zeta_{\text{cov}} I \\
&= (n+1)\mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n} \left(\phi(s,a)\phi(s,a)^\top\right) + \zeta_{\text{cov}} I. \\
\hat{\Sigma}_{\text{HF}}^n &:= \frac{1}{M_{\text{HF}}} \sum_{i=1}^{M_{\text{HF}}} \left(\phi(\tau_i^{(1)}) - \phi(\tau_i^{(2)})\right)\left(\phi(\tau_i^{(1)}) - \phi(\tau_i^{(2)})\right)^\top + \frac{\zeta_{\text{HF}}}{n} I \\
&= \frac{1}{M_{\text{HF}}} \sum_{i=1}^{M_{\text{HF}}} \tilde{\phi}^{\tau_i^{(1)},\tau_i^{(2)}} \left(\tilde{\phi}^{\tau_i^{(1)},\tau_i^{(2)}}\right)^\top + \frac{\zeta_{\text{HF}}}{n} I, \quad \forall n \geq 1 \\
\Sigma_{\text{HF}}^n &:= \frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{E}_{\substack{\tau^{(1)}\sim\mathcal{O}^{\pi^i}_{\rho_{\text{cov}}^{i-1}} \\ \tau^{(2)}\sim\mathcal{O}^{\pi^{\text{base}}}_{s_{\text{init}}}}} \left[\left(\phi(\tau^{(1)}) - \phi(\tau^{(2)})\right)\left(\phi(\tau^{(1)}) - \phi(\tau^{(2)})\right)^\top\right] \right) + \frac{\zeta_{\text{HF}}}{n} I \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{E}_{\substack{\tau^{(1)}\sim\mathcal{O}^{\pi^i}_{\rho_{\text{cov}}^{i-1}} \\ \tau^{(2)}\sim\mathcal{O}^{\pi^{\text{base}}}_{s_{\text{init}}}}} \left[\tilde{\phi}^{\tau^{(1)},\tau^{(2)}} \left(\tilde{\phi}^{\tau^{(1)},\tau^{(2)}}\right)^\top\right] \right) + \frac{\zeta_{\text{HF}}}{n} I, \quad \forall n \geq 1 \\
\hat{\Sigma}_{\text{HF}}^0 &= \Sigma_{\text{HF}}^0 := \zeta_{\text{HF}} I.
\end{aligned}
$$

Define event

$$
\begin{aligned}
\mathcal{E}_{\text{cov}} := \Bigg\{ & \frac{1}{2}\|\phi(s,a)\|_{(\Sigma_{\text{cov}}^n)^{-1}} \leq \|\phi(s,a)\|_{(\hat{\Sigma}_{\text{cov}}^n)^{-1}} \leq 2\|\phi(s,a)\|_{(\Sigma_{\text{cov}}^n)}^{-1}, \\
& \frac{1}{2}\|\phi(s,a)\|_{(\Sigma_{\text{HF}}^n)^{-1}} \leq \|\phi(s,a)\|_{(\hat{\Sigma}_{\text{HF}}^n)^{-1}} \leq 2\|\phi(s,a)\|_{(\Sigma_{\text{HF}}^n)}^{-1}, \ \forall 0 \leq n \leq N-1 \Bigg\}.
\end{aligned}
$$

Lemma D.8. *Assuming that event $\mathcal{E}_\tau$ holds, we have* $\Pr[\mathcal{E}_{\text{cov}}] \geq 1 - 2N\delta'$.

Proof. This lemma follows from Lemma F.2 and the conditions that $K \geq \frac{16(N+1)^2 \log^2\left(\frac{4dN}{\delta'}\right)}{\zeta_{\text{cov}}^2}$ and $M_{\text{HF}} \geq \frac{16W_\tau^4 \log^2\left(\frac{4d}{\delta'}\right)}{\zeta_{\text{HF}}^2}$. □

D.3.2. Reward Estimation Error in Q-value Functions

Let $W_\mu := 1$.

For any $n \geq 0$, recall that

$$
\begin{aligned}
\hat{\mu}^n := \underset{\|\mu\|_2 \leq W_\mu}{\text{argmin}} \Bigg(& -\sum_{i=1}^{M_{\text{HF}}} \log\Bigg(\frac{\mathbb{1}\{y_i = 1\}}{1 + \exp\left(\left(\sum_{h=0}^{H(\tau_i^{(2)})} \phi(s_{i,h}^{(2)}, a_{i,h}^{(2)}) - \sum_{h=0}^{H(\tau_i^{(1)})} \phi(s_{i,h}^{(1)}, a_{i,h}^{(1)})\right)^\top \mu\right)} \\
& + \frac{\mathbb{1}\{y_i = 0\}}{1 + \exp\left(\left(\sum_{h=0}^{H(\tau_i^{(1)})} \phi(s_{i,h}^{(1)}, a_{i,h}^{(1)}) - \sum_{h=0}^{H(\tau_i^{(2)})} \phi(s_{i,h}^{(2)}, a_{i,h}^{(2)})\right)^\top \mu\right)} \Bigg)\Bigg).
\end{aligned}
$$

Lemma D.9 (MLE, Lemma 5.1 in (Zhu et al., 2023))**.** *For any phase $n \geq 0$, with probability at least $1 - \delta'$, we have*

$$
\|\hat{\mu}^n - \mu^*\|_{\hat{\Sigma}_{\text{HF}}^n} \leq 8\sqrt{\frac{d + \log\left(\frac{1}{\delta'}\right)}{c_{\text{MLE}}^2 M_{\text{HF}}} + \frac{\zeta_{\text{HF}} W_\mu^2}{n}} := \varepsilon_{\text{HF}}^n.
$$

where $c_{\text{MLE}} := \frac{1}{2+\exp(-2W_\tau W_\mu)+\exp(2W_\tau W_\mu)}$.

In other words, defining event

$$
\mathcal{E}_{\text{MLE}} = \left\{ \|\hat{\mu}^n - \mu^*\|_{\hat{\Sigma}_{\text{HF}}^n} \leq \varepsilon_{\text{HF}}^n, \ \forall 0 \leq n \leq N-1 \right\},
$$

we have $\Pr[\mathcal{E}_{\text{MLE}}] \geq 1 - N\delta'$.

Lemma D.10. *Assume that event $\mathcal{E}_\tau \cap \mathcal{E}_{\text{cov}} \cap \mathcal{E}_{\text{MLE}}$ holds. Then, for any phase $n \geq 0$, iteration $t \geq 0$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$\left|Q^{\pi^t}(s,a;\hat{r}^n + b^n) - Q^{\pi^t}(s,a;r+b^n)\right| \leq 2\varepsilon_{\text{HF}}^n \mathbb{E}_{\tau \sim \mathcal{O}_{s,a}^{\pi^t}} \left[\left\|\sum_{h=0}^{H(\tau)} \phi(s_h,a_h)\right\|_{\left(\Sigma_{\text{HF}}^n\right)^{-1}}\right] := \varsigma_{s,a}^{\pi^t}.$$

Proof. Since $Q^{\pi^t}(s,a;\hat{r}^n + b^n) = \mathbb{E}_{\tau \sim \mathcal{O}_{s,a}^{\pi^t}}[\sum_{h=0}^{H(\tau)}(\hat{r}^n(s_h,a_h) + b^n(s_h,a_h))]$ and $Q^{\pi^t}(s,a;r+b^n) = \mathbb{E}_{\tau \sim \mathcal{O}_{s,a}^{\pi^t}}[\sum_{h=0}^{H(\tau)}(r(s_h,a_h) + b^n(s_h,a_h))]$, we have

$$\begin{aligned}
\left|Q^{\pi^t}(s,a;\hat{r}^n + b^n) - Q^{\pi^t}(s,a;r+b^n)\right| &= \left|\mathbb{E}_{\tau \sim \mathcal{O}_{s,a}^{\pi^t}}\left[\sum_{h=0}^{H(\tau)}(\hat{r}^n(s_h,a_h) - r(s_h,a_h))\right]\right| \\
&\leq \mathbb{E}_{\tau \sim \mathcal{O}_{s,a}^{\pi^t}}\left[\left|\sum_{h=0}^{H(\tau)}(\hat{r}^n(s_h,a_h) - r(s_h,a_h))\right|\right] \\
&= \mathbb{E}_{\tau \sim \mathcal{O}_{s,a}^{\pi^t}}\left[\left|\sum_{h=0}^{H(\tau)}\phi(s_h,a_h)^\top(\hat{\mu}^n - \mu^*)\right|\right] \\
&\leq \mathbb{E}_{\tau \sim \mathcal{O}_{s,a}^{\pi^t}}\left[\left\|\sum_{h=0}^{H(\tau)}\phi(s_h,a_h)\right\|_{\left(\hat{\Sigma}_{\text{HF}}^n\right)^{-1}} \|\hat{\mu}^n - \mu^*\|_{\hat{\Sigma}_{\text{HF}}^n}\right] \\
&\overset{\text{(a)}}{\leq} 2\varepsilon_{\text{HF}}^n \mathbb{E}_{\tau \sim \mathcal{O}_{s,a}^{\pi^t}}\left[\left\|\sum_{h=0}^{H(\tau)}\phi(s_h,a_h)\right\|_{\left(\Sigma_{\text{HF}}^n\right)^{-1}}\right],
\end{aligned}$$

where inequality (a) is due to the definition of event $\mathcal{E}_{\text{cov}}$. □

Let $\varsigma_{\rho_{\text{cov}}^n}^{\pi^t} := \mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n}[\varsigma_{s,a}^{\pi^t}] = 2\varepsilon_{\text{HF}}^n \mathbb{E}_{\tau \sim \mathcal{O}_{\rho_{\text{cov}}^n}^{\pi^t}}[\|\sum_{h=0}^{H(\tau)}\phi(s_h,a_h)\|_{\left(\Sigma_{\text{HF}}^n\right)^{-1}}]$, $W_\theta := \frac{2}{(1-\gamma)^2} - \frac{1}{1-\gamma}$ and $W_Q := \frac{2}{(1-\gamma)^2}$.

Lemma D.11. *Assume that event $\mathcal{E}_\tau \cap \mathcal{E}_{\text{cov}} \cap \mathcal{E}_{\text{MLE}}$ holds. Then, for any phase $n \geq 0$, iteration $t \geq 0$, $s \in \mathcal{K}^n$ and $a \in \mathcal{A}$,*

$$\left|\phi(s,a)^\top\left(\theta_*^t - \theta_{\text{mid}}^t\right)\right| \leq \sqrt{32\beta W_Q \varepsilon_{\text{HF}}^n (n+1)\mathbb{E}_{\tau \sim \mathcal{O}_{\rho_{\text{cov}}^n}^{\pi^t}}\left[\left\|\sum_{h=0}^{H(\tau)}\phi(s_h,a_h)\right\|_{\left(\Sigma_{\text{HF}}^n\right)^{-1}}\right]} + W_\theta\sqrt{8\beta\zeta_{\text{HF}}}.$$

Proof. For any phase $n \geq 0$ and iteration $t \geq 0$, for any fixed θ and (s,a), using Lemma F.3, we have

$$\begin{aligned}
&\left(Q^{\pi^t}(s,a;\hat{r}^n + b^n) - b^n(s,a) - \phi(s,a)^\top\theta\right)^2 - \left(Q^{\pi^t}(s,a;r+b^n) - b^n(s,a) - \phi(s,a)^\top\theta\right)^2 \\
&\leq 4W_Q\left|Q^{\pi^t}(s,a;\hat{r}^n + b^n) - Q^{\pi^t}(s,a;r+b^n)\right| \\
&\leq 4W_Q\varsigma_{s,a}^{\pi^t},
\end{aligned} \tag{12}$$

where W_Q satisfies that $\max\{|Q^{\pi^t}(s,a;\hat{r}^n + b^n)|, |Q^{\pi^t}(s,a;r+b^n)|, |\phi(s,a)^\top\theta + b^n(s,a)|\} \leq W_Q$ for all $n \geq 0$, $t \geq 0$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Taking $\mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n}[\cdot]$ on both sides, we have

$$\mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n}\left[\left(Q^{\pi^t}(s,a;\hat{r}^n + b^n) - b^n(s,a) - \phi(s,a)^\top\theta\right)^2\right] \tag{13}$$

$$
\begin{aligned}
&\quad - \mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left(Q^{\pi^t}(s,a;r+b^n) - b^n(s,a) - \phi(s,a)^\top\theta\right)^2\right] \\
&\le 4W_Q\mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\varsigma^{\pi^t}_{s,a}\right] \\
&= 4W_Q\varsigma^{\pi^t}_{\rho^n_{\text{cov}}}. \qquad (14)
\end{aligned}
$$

Plugging θ^t_* into θ, we have that for any fixed (s,a),

$$
\begin{aligned}
&\mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left(Q^{\pi^t}(s,a;r+b^n) - b^n(s,a) - \phi(s,a)^\top\theta^t_*\right)^2\right] \\
&\ge \mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left(Q^{\pi^t}(s,a;\hat{r}^n+b^n) - b^n(s,a) - \phi(s,a)^\top\theta^t_*\right)^2\right] - 4W_Q\varsigma^{\pi^t}_{\rho^n_{\text{cov}}} \\
&\overset{\text{(a)}}{\ge} \mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left(Q^{\pi^t}(s,a;\hat{r}^n+b^n) - b^n(s,a) - \phi(s,a)^\top\theta^t_{\text{mid}}\right)^2\right] - 4W_Q\varsigma^{\pi^t}_{\rho^n_{\text{cov}}} \qquad (15)
\end{aligned}
$$

where inequality (a) is due to the definition of θ^t_{mid}.

Furthermore, we have

$$
\begin{aligned}
&\mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left(Q^{\pi^t}(s,a;r+b^n) - b^n(s,a) - \phi(s,a)^\top\theta^t_{\text{mid}}\right)^2\right] \\
&\quad - \mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left(Q^{\pi^t}(s,a;r+b^n) - b^n(s,a) - \phi(s,a)^\top\theta^t_*\right)^2\right] \\
&= \mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left(Q^{\pi^t}(s,a;\hat{r}^n+b^n) - b^n(s,a) - \phi(s,a)^\top\theta^t_{\text{mid}}\right)^2\right] \\
&\quad - \mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left(Q^{\pi^t}(s,a;r+b^n) - b^n(s,a) - \phi(s,a)^\top\theta^t_*\right)^2\right] \\
&\quad + \mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left(Q^{\pi^t}(s,a;r+b^n) - b^n(s,a) - \phi(s,a)^\top\theta^t_{\text{mid}}\right)^2\right] \\
&\quad - \mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left(Q^{\pi^t}(s,a;\hat{r}^n+b^n) - b^n(s,a) - \phi(s,a)^\top\theta^t_{\text{mid}}\right)^2\right] \\
&\overset{\text{(a)}}{\le} 4W_Q\varsigma^{\pi^t}_{\rho^n_{\text{cov}}} + 4W_Q\mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left|Q^{\pi^t}(s,a;\hat{r}^n+b^n) - Q^{\pi^t}(s,a;r+b^n)\right|\right] \\
&\le 8W_Q\varsigma^{\pi^t}_{\rho^n_{\text{cov}}}, \qquad (16)
\end{aligned}
$$

where inequality (a) uses Lemma F.3.

On the other hand, it holds that

$$
\begin{aligned}
&\mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left(Q^{\pi^t}(s,a;r+b^n) - b^n(s,a) - \phi(s,a)^\top\theta^t_{\text{mid}}\right)^2\right] \\
&\quad - \mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left(Q^{\pi^t}(s,a;r+b^n) - b^n(s,a) - \phi(s,a)^\top\theta^t_*\right)^2\right] \\
&= \mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left(\phi(s,a)^\top\left(\theta^t_* - \theta^t_{\text{mid}}\right)\right)^2\right] \\
&\quad + 2\underbrace{\mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left(Q^{\pi^t}(s,a;r+b^n) - b^n(s,a) - \phi(s,a)^\top\theta^t_*\right)\phi(s,a)^\top\left(\theta^t_* - \theta^t_{\text{mid}}\right)\right]}_{\text{Term }\Gamma\,\ge\,0}, \qquad (17)
\end{aligned}
$$

where Term Γ is non-negative due to the the first-order optimality of θ^t_*.

Thus, we have

$$
\mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left(\phi(s,a)^\top\left(\theta^t_* - \theta^t_{\text{mid}}\right)\right)^2\right]
$$

$$
\begin{aligned}
&\leq \mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left(Q^{\pi^t}(s,a;r+b^n)-b^n(s,a)-\phi(s,a)^\top\theta^t_{\text{mid}}\right)^2\right]\\
&\quad-\mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left(Q^{\pi^t}(s,a;r+b^n)-b^n(s,a)-\phi(s,a)^\top\theta^t_{*}\right)^2\right]\\
&\leq 8W_Q\varsigma^{\pi^t}_{\rho^n_{\text{cov}}}.
\end{aligned}
$$

Since $\mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}[(\phi(s,a)^\top(\theta^t_*-\theta^t_{\text{mid}}))^2] = \mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}[(\theta^t_*-\theta^t_{\text{mid}})^\top\phi(s,a)\phi(s,a)^\top(\theta^t_*-\theta^t_{\text{mid}})] = (\theta^t_*-\theta^t_{\text{mid}})^\top\mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}[\phi(s,a)\phi(s,a)^\top](\theta^t_*-\theta^t_{\text{mid}})$, we have

$$
(\theta^t_*-\theta^t_{\text{mid}})^\top\mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}[\phi(s,a)\phi(s,a)^\top](\theta^t_*-\theta^t_{\text{mid}})\leq 8W_Q\varsigma^{\pi^t}_{\rho^n_{\text{cov}}}.
$$

Moreover,

$$
\begin{aligned}
&\left\|\theta^t_*-\theta^t_{\text{mid}}\right\|^2_{\Sigma^n_{\text{cov}}}\\
&=\left(\theta^t_*-\theta^t_{\text{mid}}\right)^\top\left(\sum_{i=0}^n\mathbb{E}_{(s,a)\sim d^{\pi^i}_{s_{\text{init}}}}\left[\phi(s,a)\phi(s,a)^\top\right]+\zeta_{\text{cov}}I\right)\left(\theta^t_*-\theta^t_{\text{mid}}\right)\\
&=\left(\theta^t_*-\theta^t_{\text{mid}}\right)^\top\left(\sum_{i=0}^n\sum_{(s,a)}d^{\pi^i}_{s_{\text{init}}}(s,a)\cdot\phi(s,a)\phi(s,a)^\top+\zeta_{\text{cov}}I\right)\left(\theta^t_*-\theta^t_{\text{mid}}\right)\\
&=(n+1)\left(\theta^t_*-\theta^t_{\text{mid}}\right)^\top\left(\sum_{(s,a)}\frac{1}{n+1}\sum_{i=0}^n d^{\pi^i}_{s_{\text{init}}}(s,a)\cdot\phi(s,a)\phi(s,a)^\top+\frac{\zeta_{\text{cov}}}{n+1}I\right)\left(\theta^t_*-\theta^t_{\text{mid}}\right)\\
&=(n+1)\left(\theta^t_*-\theta^t_{\text{mid}}\right)^\top\left(\sum_{(s,a)}\rho^n_{\text{cov}}(s,a)\cdot\phi(s,a)\phi(s,a)^\top+\frac{\zeta_{\text{cov}}}{n+1}I\right)\left(\theta^t_*-\theta^t_{\text{mid}}\right)\\
&\leq 8(n+1)W_Q\varsigma^{\pi^t}_{\rho^n_{\text{cov}}}+4\zeta_{\text{cov}}W^2_\theta.
\end{aligned}
$$

For any $s\in\mathcal{K}^n$, using the definitions of $\mathcal{K}^n$ and event $\mathcal{E}_{\text{cov}}$, we have

$$
\frac{1}{\sqrt{2}}\left\|\phi(s,a)\right\|_{(\Sigma^n_{\text{cov}})^{-1}}\leq\left\|\phi(s,a)\right\|_{(\hat{\Sigma}^n_{\text{cov}})^{-1}}\leq\sqrt{\beta}.
$$

Therefore, we obtain

$$
\begin{aligned}
\left|\phi(s,a)^\top\left(\theta^t_*-\theta^t_{\text{mid}}\right)\right|&\leq\left\|\phi(s,a)\right\|_{(\Sigma^n_{\text{cov}})^{-1}}\left\|\theta^t_*-\theta^t_{\text{mid}}\right\|_{\Sigma^n_{\text{cov}}}\\
&\leq\sqrt{2\beta\left(8(n+1)W_Q\varsigma^{\pi^t}_{\rho^n_{\text{cov}}}+4\zeta_{\text{cov}}W^2_\theta\right)}\\
&\leq\sqrt{32\beta W_Q\varepsilon^n_{\text{HF}}(n+1)\mathbb{E}_{\tau\sim\mathcal{O}^{\pi^t}_{\rho^n_{\text{cov}}}}\left[\left\|\sum_{h=0}^{H(\tau)}\phi(s_h,a_h)\right\|_{\left(\Sigma^n_{\text{HF}}\right)^{-1}}\right]}+W_\theta\sqrt{8\beta\zeta_{\text{cov}}}.
\end{aligned}
$$

□

D.3.3. Elliptical Potential Analysis for Human Data

Lemma D.12 (Elliptical Potential for the Baseline Policy)**.** *For any phase $n\geq 0$,*

$$
\sum_{n=0}^{N-1}\mathbb{E}_{\tau\sim\mathcal{O}^{\pi^{\text{base}}}_{s_{\text{init}}}}\left[\left\|\phi(\tau)\right\|^2_{\left(n\Sigma^n_{\text{HF}}\right)^{-1}}\right]\leq\frac{2d}{c_{\text{base}}}\log(N).
$$

Proof. We have

$$
\begin{aligned}
&\sum_{n=0}^{N-1} \mathbb{E}_{\tau\sim\mathcal{O}^{\pi^{\text{base}}}_{s_{\text{init}}}} \left[\|\phi(\tau)\|^2_{\left(n\Sigma^n_{\text{HF}}\right)^{-1}} \right] \\
&= \sum_{n=0}^{N-1} \mathbb{E}_{\tau\sim\mathcal{O}^{\pi^{\text{base}}}_{s_{\text{init}}}} \left[\phi(\tau)^\top \left(\sum_{i=1}^{n} \mathbb{E}_{\substack{\tau^{(1)}\sim\mathcal{O}^{\pi^i}_{\rho^{i-1}_{\text{cov}}} \\ \tau^{(2)}\sim\mathcal{O}^{\pi^{\text{base}}}_{s_{\text{init}}}}} \left[\left(\phi(\tau^{(1)}) - \phi(\tau^{(2)})\right) \left(\phi(\tau^{(1)}) - \phi(\tau^{(2)})\right)^\top \right] + \zeta_{\text{HF}} I \right)^{-1} \phi(\tau) \right] \\
&\overset{(a)}{\leq} \frac{1}{c_{\text{base}}} \sum_{n=0}^{N-1} \mathbb{E}_{\tau\sim\mathcal{O}^{\pi^{\text{base}}}_{s_{\text{init}}}} \left[\phi(\tau)^\top \left(\sum_{i=1}^{n} \mathbb{E}_{\tau^{(2)}\sim\mathcal{O}^{\pi^{\text{base}}}_{s_{\text{init}}}} \left[\phi(\tau^{(2)})\phi(\tau^{(2)})^\top \right] + \zeta_{\text{HF}} I \right)^{-1} \phi(\tau) \right] \\
&= \frac{1}{c_{\text{base}}} \sum_{n=1}^{N-1} \frac{1}{n} \text{tr} \left(\left(\mathbb{E}_{\tau^{(2)}\sim\mathcal{O}^{\pi^{\text{base}}}_{s_{\text{init}}}} \left[\phi(\tau^{(2)})\phi(\tau^{(2)})^\top \right] + \frac{\zeta_{\text{HF}}}{n} I \right)^{-1} \mathbb{E}_{\tau\sim\mathcal{O}^{\pi^{\text{base}}}_{s_{\text{init}}}} \left[\phi(\tau)\phi(\tau)^\top \right] \right) + \frac{W_\tau^2}{c_{\text{base}}\zeta_{\text{HF}}} \\
&= \frac{d}{c_{\text{base}}} \sum_{n=1}^{N-1} \frac{1}{n} + \frac{W_\tau^2}{c_{\text{base}}\zeta_{\text{HF}}} \\
&\leq \frac{d}{c_{\text{base}}} \left(\log(N) + 1\right) + \frac{1}{c_{\text{base}}} \\
&\overset{(b)}{\leq} \frac{2d}{c_{\text{base}}} \log(N).
\end{aligned}
$$

where inequality (a) uses Assumption 3.4, and inequality (b) holds if $\log(N) \geq 2$ which can be easily guaranteed in our problem. □

Lemma D.13 (Elliptical Potential for Preference-based Data)**.** *It holds that*

$$
\frac{1}{N} \sum_{n=0}^{N-1} \left(\frac{1}{T} \sum_{t=0}^{T-1} \sqrt{ \mathbb{E}_{\tau\sim\mathcal{O}^{\pi^t}_{\rho^n_{\text{cov}}}} \left[\left\| \sum_{h=0}^{H(\tau)} \phi(s_h, a_h) \right\|_{\left(\Sigma^n_{\text{HF}}\right)^{-1}} \right] } \right) \leq 2d^{\frac{1}{4}} \log^{\frac{1}{4}} \left(1 + \frac{4NW_\tau^2}{\zeta_{\text{HF}} d} \right) + \frac{2d^{\frac{1}{4}} \log^{\frac{1}{4}}(N)}{c_{\text{base}}^{\frac{1}{4}}}.
$$

Proof. For any phase $n \geq 0$, we have

$$
\begin{aligned}
&\frac{1}{T} \sum_{t=0}^{T-1} \sqrt{ \mathbb{E}_{\tau\sim\mathcal{O}^{\pi^t}_{\rho^n_{\text{cov}}}} \left[\left\| \sum_{h=0}^{H(\tau)} \phi(s_h, a_h) \right\|_{\left(\Sigma^n_{\text{HF}}\right)^{-1}} \right] } \\
&\leq \frac{1}{T} \sqrt{ T \cdot \sum_{t=0}^{T-1} \mathbb{E}_{\tau\sim\mathcal{O}^{\pi^t}_{\rho^n_{\text{cov}}}} \left[\left\| \sum_{h=0}^{H(\tau)} \phi(s_h, a_h) \right\|_{\left(\Sigma^n_{\text{HF}}\right)^{-1}} \right] } \\
&= \frac{1}{T} \sqrt{ T^2 \cdot \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}_{\tau\sim\mathcal{O}^{\pi^t}_{\rho^n_{\text{cov}}}} \left[\left\| \sum_{h=0}^{H(\tau)} \phi(s_h, a_h) \right\|_{\left(\Sigma^n_{\text{HF}}\right)^{-1}} \right] } \\
&= \sqrt{ \mathbb{E}_{\tau\sim\mathcal{O}^{\pi^{n+1}}_{\rho^n_{\text{cov}}}} \left[\left\| \sum_{h=0}^{H(\tau)} \phi(s_h, a_h) \right\|_{\left(\Sigma^n_{\text{HF}}\right)^{-1}} \right] }.
\end{aligned}
$$

We make the convention that $n\Sigma_{\text{HF}}^n := \zeta_{\text{HF}} I$ for $n = 0$. Then, we obtain

$$
\begin{aligned}
&\frac{1}{N}\sum_{n=0}^{N-1}\left(\frac{1}{T}\sum_{t=0}^{T-1}\sqrt{\mathbb{E}_{\tau\sim\mathcal{O}^{\pi^t}_{\rho^n_{\text{cov}}}}\left[\left\|\sum_{h=0}^{H(\tau)}\phi(s_h,a_h)\right\|^2_{\left(\Sigma^n_{\text{HF}}\right)^{-1}}\right]}\right)\\
&\le \frac{1}{N}\sum_{n=0}^{N-1}\sqrt{\mathbb{E}_{\tau\sim\mathcal{O}^{\pi^{n+1}}_{\rho^n_{\text{cov}}}}\left[\sqrt{\left(\sum_{h=0}^{H(\tau)}\phi(s_h,a_h)\right)^\top\left(\Sigma^n_{\text{HF}}\right)^{-1}\left(\sum_{h=0}^{H(\tau)}\phi(s_h,a_h)\right)}\right]}\\
&\overset{\text{(a)}}{\le} \frac{1}{N}\sum_{n=0}^{N-1}\sqrt{\sqrt{\mathbb{E}_{\tau\sim\mathcal{O}^{\pi^{n+1}}_{\rho^n_{\text{cov}}}}\left[\left(\sum_{h=0}^{H(\tau)}\phi(s_h,a_h)\right)^\top\left(\Sigma^n_{\text{HF}}\right)^{-1}\left(\sum_{h=0}^{H(\tau)}\phi(s_h,a_h)\right)\right]}}\\
&\le \frac{1}{N}\sqrt{N\cdot\sum_{n=0}^{N-1}\sqrt{\mathbb{E}_{\tau\sim\mathcal{O}^{\pi^{n+1}}_{\rho^n_{\text{cov}}}}\left[\left(\sum_{h=0}^{H(\tau)}\phi(s_h,a_h)\right)^\top\left(\Sigma^n_{\text{HF}}\right)^{-1}\left(\sum_{h=0}^{H(\tau)}\phi(s_h,a_h)\right)\right]}}\\
&\le \frac{1}{\sqrt{N}}\sqrt{\sqrt{N\cdot\sum_{n=0}^{N-1}\mathbb{E}_{\tau\sim\mathcal{O}^{\pi^{n+1}}_{\rho^n_{\text{cov}}}}\left[\left(\sum_{h=0}^{H(\tau)}\phi(s_h,a_h)\right)^\top\left(\Sigma^n_{\text{HF}}\right)^{-1}\left(\sum_{h=0}^{H(\tau)}\phi(s_h,a_h)\right)\right]}}\\
&= N^{-\frac{1}{4}}\left(\sum_{n=1}^{N-1}\mathbb{E}_{\tau\sim\mathcal{O}^{\pi^{n+1}}_{\rho^n_{\text{cov}}}}\left[\left(\sum_{h=0}^{H(\tau)}\phi(s_h,a_h)\right)^\top\left(\Sigma^n_{\text{HF}}\right)^{-1}\left(\sum_{h=0}^{H(\tau)}\phi(s_h,a_h)\right)\right]\right)^{\frac{1}{4}}\\
&\le \left(\sum_{n=0}^{N-1}\mathbb{E}_{\tau\sim\mathcal{O}^{\pi^{n+1}}_{\rho^n_{\text{cov}}}}\left[\left(\sum_{h=0}^{H(\tau)}\phi(s_h,a_h)\right)^\top\left(n\Sigma^n_{\text{HF}}\right)^{-1}\left(\sum_{h=0}^{H(\tau)}\phi(s_h,a_h)\right)\right]\right)^{\frac{1}{4}}\\
&= \left(\sum_{n=0}^{N-1}\mathbb{E}_{\tau\sim\mathcal{O}^{\pi^{n+1}}_{\rho^n_{\text{cov}}}}\left[\left\|\sum_{h=0}^{H(\tau)}\phi(s_h,a_h)\right\|^2_{\left(n\Sigma^n_{\text{HF}}\right)^{-1}}\right]\right)^{\frac{1}{4}}.
\end{aligned}
$$

where inequality (a) uses the Jensen inequality.

It holds that

$$
\begin{aligned}
&\sum_{n=0}^{N-1}\mathbb{E}_{\tau\sim\mathcal{O}^{\pi^{n+1}}_{\rho^n_{\text{cov}}}}\left[\left\|\sum_{h=0}^{H(\tau)}\phi(s_h,a_h)\right\|^2_{\left(n\Sigma^n_{\text{HF}}\right)^{-1}}\right]\\
&= \sum_{n=0}^{N-1}\mathbb{E}_{\substack{\tau^{(1)}\sim\mathcal{O}^{\pi^{n+1}}_{\rho^n_{\text{cov}}}\\ \tau^{(2)}\sim\mathcal{O}^{\pi^{\text{base}}}_{s_{\text{init}}}}}\left[\left\|\phi(\tau^{(1)})-\phi(\tau^{(2)})+\phi(\tau^{(2)})\right\|^2_{\left(n\Sigma^n_{\text{HF}}\right)^{-1}}\right]\\
&\le \sum_{n=0}^{N-1}\left(2\mathbb{E}_{\substack{\tau^{(1)}\sim\mathcal{O}^{\pi^{n+1}}_{\rho^n_{\text{cov}}}\\ \tau^{(2)}\sim\mathcal{O}^{\pi^{\text{base}}}_{s_{\text{init}}}}}\left[\left\|\phi(\tau^{(1)})-\phi(\tau^{(2)})\right\|^2_{\left(n\Sigma^n_{\text{HF}}\right)^{-1}}\right]+2\mathbb{E}_{\tau^{(2)}\sim\mathcal{O}^{\pi^{\text{base}}}_{s_{\text{init}}}}\left[\left\|\phi(\tau^{(2)})\right\|^2_{\left(n\Sigma^n_{\text{HF}}\right)^{-1}}\right]\right)
\end{aligned}
$$

$$\overset{\text{(a)}}{\leq} 2\sum_{n=1}^{N}\mathbb{E}_{\substack{\tau^{(1)}\sim\mathcal{O}^{\pi^n}_{\rho_{\text{cov}}^{n-1}}\\ \tau^{(2)}\sim\mathcal{O}^{\pi_{\text{base}}}_{s_{\text{init}}}}}\left[\left(\phi(\tau^{(1)})-\phi(\tau^{(2)})\right)^\top\left(\sum_{i=1}^{n-1}\mathbb{E}_{\substack{\tau^{(1)}\sim\mathcal{O}^{\pi^i}_{\rho_{\text{cov}}^{i-1}}\\ \tau^{(2)}\sim\mathcal{O}^{\pi_{\text{base}}}_{s_{\text{init}}}}}\left[\left(\phi(\tau^{(1)})-\phi(\tau^{(2)})\right)\left(\phi(\tau^{(1)})-\phi(\tau^{(2)})\right)^\top\right]+\zeta_{\text{HF}}I\right)^{-1}\cdot\right.$$
$$\left.\left(\phi(\tau^{(1)})-\phi(\tau^{(2)})\right)\right]+\frac{4d}{c_{\text{base}}}\log(N)$$
$$\overset{\text{(b)}}{\leq} 4d\log\left(\frac{\det\left(\sum_{i=1}^{N}\mathbb{E}_{\tau^{(1)}\sim\mathcal{O}^{\pi^i}_{\rho_{\text{cov}}^{i-1}},\tau^{(2)}\sim\mathcal{O}^{\pi_{\text{base}}}_{s_{\text{init}}}}\left[\left(\phi(\tau^{(1)})-\phi(\tau^{(2)})\right)\left(\phi(\tau^{(1)})-\phi(\tau^{(2)})\right)^\top\right]+\zeta_{\text{HF}}I\right)}{\det\left(\zeta_{\text{HF}}I\right)}\right)+\frac{4d}{c_{\text{base}}}\log(N)$$
$$\leq 4d\log\left(1+\frac{4NW_\tau^2}{\zeta_{\text{HF}}d}\right)+\frac{4d}{c_{\text{base}}}\log(N).$$

Here inequality (a) uses Lemma D.12 and Assumption 3.4. Inequality (b) follows from the elliptical potential lemma (Lemma F.5) and the fact that $\zeta_{\text{HF}} := 4W_\tau^2$.

Therefore, we have

$$\frac{1}{N}\sum_{n=0}^{N-1}\left(\frac{1}{T}\sum_{t=0}^{T-1}\sqrt{\mathbb{E}_{\tau\sim\mathcal{O}^{\pi^t}_{\rho_{\text{cov}}^n}}\left[\left\|\sum_{h=0}^{H(\tau)}\phi(s_h,a_h)\right\|_{\left(\Sigma^n_{\text{HF}}\right)^{-1}}\right]}\right)$$
$$\leq\left(4d\log\left(1+\frac{4NW_\tau^2}{\zeta_{\text{HF}}d}\right)+\frac{4d}{c_{\text{base}}}\log(N)\right)^{\frac{1}{4}}$$
$$\leq 2d^{\frac{1}{4}}\log^{\frac{1}{4}}\left(1+\frac{4NW_\tau^2}{\zeta_{\text{HF}}d}\right)+\frac{2d^{\frac{1}{4}}\log^{\frac{1}{4}}(N)}{c_{\text{base}}^{\frac{1}{4}}}.$$

□

D.3.4. Discussion on Assumption 3.4

Assumption 3.4 can be abstracted from the RLHF framework, and serve as a technical condition for an independent mathematical problem.

We provide Lemma D.14 to demonstrate that under Assumption 3.4, one can systematically utilize the elliptical potential lemma (Abbasi-Yadkori et al., 2011) to obtain a mathematical conclusion that is independent of the RLHF framework.

Our RLHF analysis (Lemmas D.12 and D.13) is an application of this systematical analytical procedure.

Lemma D.14. *Let $\Phi := \{\phi\in\mathbb{R}^d : \|\phi\|_2\leq W_\phi\}$. There are random distributions $\mathcal{D}_1,\ldots,\mathcal{D}_N$ and $\mathcal{D}_{\text{base}}$ over Φ, and a regularization parameter $\zeta\geq W_\phi^2$.*

Assume that $\mathcal{D}_{\text{base}}$ satisfies that for any $n\in[N]$,

$$\mathbb{E}_{\phi\sim\mathcal{D}_n,\phi'\sim\mathcal{D}_{\text{base}}}\left[(\phi-\phi')(\phi-\phi')^\top\right]\succeq c_{\text{base}}\mathbb{E}_{\phi'\sim\mathcal{D}_{\text{base}}}\left[\phi'\phi'^\top\right] \tag{18}$$

for some constant $c_{\text{base}}\in(0,1)$.

Then, we can use the elliptical potential lemma (Lemma F.5) (Abbasi-Yadkori et al., 2011) to bound

$$\sum_{n=1}^{N}\mathbb{E}_{\phi_n\sim\mathcal{D}_n,\phi'_n\sim\mathcal{D}_{\text{base}}}\left[\|\phi_n\|^2_{\left(\sum_{i=1}^{n-1}\mathbb{E}_{\phi_i\sim\mathcal{D}_i,\phi'_i\sim\mathcal{D}_{\text{base}}}\left[(\phi_i-\phi'_i)(\phi_i-\phi'_i)^\top\right]+\zeta I\right)^{-1}}\right]$$
$$\leq 2\sum_{n=1}^{N}\left(\mathbb{E}_{\phi_n\sim\mathcal{D}_n,\phi'_n\sim\mathcal{D}_{\text{base}}}\left[\|\phi_n-\phi'_n\|^2_{\left(\sum_{i=1}^{n-1}\mathbb{E}_{\phi_i\sim\mathcal{D}_i,\phi'_i\sim\mathcal{D}_{\text{base}}}\left[(\phi_i-\phi'_i)(\phi_i-\phi'_i)^\top\right]+\zeta I\right)^{-1}}\right]\right.$$
$$\left.+\frac{2}{c_{\text{base}}}\mathbb{E}_{\phi'_n\sim\mathcal{D}_{\text{base}}}\left[\|\phi'_n\|^2_{\left(\sum_{i=1}^{n-1}\mathbb{E}_{\phi'_i\sim\mathcal{D}_{\text{base}}}\left[\phi'_i(\phi'_i)^\top\right]+\zeta I\right)^{-1}}\right]\right)$$

$$\leq 4d \log\left(1 + \frac{NW_\phi^2}{\zeta d}\right) + \frac{2d}{c_{\text{base}}}\left(\log(N) + 2\right).$$

Proof. According to the assumption Eq. (18), we have that for any $i \in [N]$,

$$\mathbb{E}_{\phi\sim\mathcal{D}_i,\phi'\sim\mathcal{D}_{\text{base}}}\left[(\phi-\phi')(\phi-\phi')^\top\right] + \zeta I \succeq c_{\text{base}}\mathbb{E}_{\phi'\sim\mathcal{D}_{\text{base}}}\left[\phi'\phi'^\top\right] + \zeta I \succeq c_{\text{base}}\mathbb{E}_{\phi'\sim\mathcal{D}_{\text{base}}}\left[\phi'\phi'^\top\right] + c_{\text{base}}\zeta I,$$

which implies that

$$\left(\mathbb{E}_{\phi\sim\mathcal{D}_i,\phi'\sim\mathcal{D}_{\text{base}}}\left[(\phi-\phi')(\phi-\phi')^\top\right] + \zeta I\right)^{-1} \preceq \frac{1}{c_{\text{base}}}\left(\mathbb{E}_{\phi'\sim\mathcal{D}_{\text{base}}}\left[\phi'\phi'^\top\right] + \zeta I\right)^{-1}.$$

Hence, we have that for any $v \in \mathbb{R}^d$,

$$v^\top\left(\mathbb{E}_{\phi\sim\mathcal{D}_i,\phi'\sim\mathcal{D}_{\text{base}}}\left[(\phi-\phi')(\phi-\phi')^\top\right] + \zeta I\right)^{-1} v \leq \frac{1}{c_{\text{base}}} v^\top\left(\mathbb{E}_{\phi'\sim\mathcal{D}_{\text{base}}}\left[\phi'\phi'^\top\right] + \zeta I\right)^{-1} v.$$

Furthermore, we have

$$\begin{aligned}
&\sum_{n=1}^N \mathbb{E}_{\phi_n\sim\mathcal{D}_n,\phi'_n\sim\mathcal{D}_{\text{base}}}\left[\|\phi_n\|^2_{\left(\sum_{i=1}^{n-1}\mathbb{E}_{\phi_i\sim\mathcal{D}_i,\phi'_i\sim\mathcal{D}_{\text{base}}}\left[(\phi_i-\phi'_i)(\phi_i-\phi'_i)^\top\right]+\zeta I\right)^{-1}}\right] \\
&= \sum_{n=1}^N \mathbb{E}_{\phi_n\sim\mathcal{D}_n,\phi'_n\sim\mathcal{D}_{\text{base}}}\left[\|\phi_n-\phi'_n+\phi'_n\|^2_{\left(\sum_{i=1}^{n-1}\mathbb{E}_{\phi_i\sim\mathcal{D}_i,\phi'_i\sim\mathcal{D}_{\text{base}}}\left[(\phi_i-\phi'_i)(\phi_i-\phi'_i)^\top\right]+\zeta I\right)^{-1}}\right] \\
&\leq \sum_{n=1}^N \left(2\mathbb{E}_{\phi_n\sim\mathcal{D}_n,\phi'_n\sim\mathcal{D}_{\text{base}}}\left[\|\phi_n-\phi'_n\|^2_{\left(\sum_{i=1}^{n-1}\mathbb{E}_{\phi_i\sim\mathcal{D}_i,\phi'_i\sim\mathcal{D}_{\text{base}}}\left[(\phi_i-\phi'_i)(\phi_i-\phi'_i)^\top\right]+\zeta I\right)^{-1}}\right]\right. \\
&\quad \left. + 2\mathbb{E}_{\phi'_n\sim\mathcal{D}_{\text{base}}}\left[\|\phi'_n\|^2_{\left(\sum_{i=1}^{n-1}\mathbb{E}_{\phi_i\sim\mathcal{D}_i,\phi'_i\sim\mathcal{D}_{\text{base}}}\left[(\phi_i-\phi'_i)(\phi_i-\phi'_i)^\top\right]+\zeta I\right)^{-1}}\right]\right) \\
&\overset{\text{(a)}}{\leq} 2\sum_{n=1}^N \left(\mathbb{E}_{\phi_n\sim\mathcal{D}_n,\phi'_n\sim\mathcal{D}_{\text{base}}}\left[\|\phi_n-\phi'_n\|^2_{\left(\sum_{i=1}^{n-1}\mathbb{E}_{\phi_i\sim\mathcal{D}_i,\phi'_i\sim\mathcal{D}_{\text{base}}}\left[(\phi_i-\phi'_i)(\phi_i-\phi'_i)^\top\right]+\zeta I\right)^{-1}}\right]\right. \\
&\quad \left. + \frac{2}{c_{\text{base}}}\mathbb{E}_{\phi'_n\sim\mathcal{D}_{\text{base}}}\left[\|\phi'_n\|^2_{\left(\sum_{i=1}^{n-1}\mathbb{E}_{\phi'_i\sim\mathcal{D}_{\text{base}}}\left[\phi'_i(\phi'_i)^\top\right]+\zeta I\right)^{-1}}\right]\right) \\
&= 2\sum_{n=1}^N \text{tr}\left(\left(\sum_{i=1}^{n-1}\mathbb{E}_{\phi_i\sim\mathcal{D}_i,\phi'_n\sim\mathcal{D}_{\text{base}}}\left[(\phi_i-\phi'_i)(\phi_i-\phi'_i)^\top\right]+\zeta I\right)^{-1}\mathbb{E}_{\phi_n\sim\mathcal{D}_n,\phi'_n\sim\mathcal{D}_{\text{base}}}\left[(\phi_n-\phi'_n)(\phi_n-\phi'_n)^\top\right]\right) \\
&\quad + \frac{2}{c_{\text{base}}}\sum_{n=2}^N \frac{1}{n-1}\cdot\text{tr}\left(\left(\mathbb{E}_{\phi'\sim\mathcal{D}_{\text{base}}}\left[\phi'(\phi')^\top\right]+\frac{\zeta}{n-1}I\right)^{-1}\mathbb{E}_{\phi'_n\sim\mathcal{D}_{\text{base}}}\left[\phi'_n(\phi'_n)^\top\right]\right) + \frac{2W_\phi^2}{c_{\text{base}}\zeta} \\
&\overset{\text{(b)}}{\leq} 4d\log\left(1+\frac{NW_\phi^2}{\zeta d}\right) + \frac{2d}{c_{\text{base}}}\sum_{n=2}^N\frac{1}{n-1} + \frac{2W_\phi^2}{c_{\text{base}}\zeta} \\
&\leq 4d\log\left(1+\frac{NW_\phi^2}{\zeta d}\right) + \frac{2d}{c_{\text{base}}}\left(\log(N)+2\right),
\end{aligned}$$

where inequality (a) uses Assumption 3.4, and inequality (b) applies the elliptical potential lemma (Lemma F.5) (Abbasi-Yadkori et al., 2011).

□

D.4. Proof of Theorem 4.2

For any phase $n = 0, \ldots, N-1$ and iteration $t = 0, \ldots, T-1$, define

$$\theta_*^t := \operatorname*{argmin}_{\|\theta\|_2 \le W_\theta} \mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n}\left[\left(\phi(s,a)^\top\theta - \left(Q^{\pi^t}(s,a;r+b^n) - b^n(s,a)\right)\right)^2\right],$$

$$\theta_{\text{mid}}^t := \operatorname*{argmin}_{\|\theta\|_2 \le W_\theta} \mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n}\left[\left(\phi(s,a)^\top\theta - \left(Q^{\pi^t}(s,a;\hat{r}^n+b^n) - b^n(s,a)\right)\right)^2\right],$$

$$\theta^t \overset{\text{SGD}}{\approx} \operatorname*{argmin}_{\|\theta\|_2 \le W_\theta} \mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n}\left[\left(\phi(s,a)^\top\theta - \left(Q^{\pi^t}(s,a;\hat{r}^n+b^n) - b^n(s,a)\right)\right)^2\right].$$

For any $n \ge 0$, $t \ge 0$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$, let $\bar{b}^{n,t}(s,a) := b^n(s,a) - \mathbb{E}_{a'\sim\pi^t(\cdot|s)}[b^n(s,a')]$ and $\bar{\phi}^t(s,a) := \phi(s,a) - \mathbb{E}_{a'\sim\pi^t(\cdot|s)}[\phi(s,a')]$.

Proof of Theorem 4.2. Using Lemma D.5, we have that for any phase $n = 0, \ldots, N-1$ and iteration $t = 0, \ldots, T-1$,

$$\begin{aligned}
&V_{\mathcal{M}^n}^{\pi^*,n}(s_{\text{init}}) - V_{\mathcal{M}^n}^{\pi^t}(s_{\text{init}}) \\
&\le \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d_{\mathcal{M}^n;s_{\text{init}}}^{\pi^*,n}}\left[A_{\mathcal{M}_{b^n}}^{\pi^t}(s,a)\cdot\mathbb{1}\{s\in\mathcal{K}^n\}\right] \\
&= \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d_{\mathcal{M}^n;s_{\text{init}}}^{\pi^*,n}}\Bigg[\hat{A}_{\mathcal{M}_{b^n}}^{\pi^t}(s,a)\cdot\mathbb{1}\{s\in\mathcal{K}^n\} \\
&\quad + \underbrace{\left(A_{\mathcal{M}_{b^n}}^{\pi^t}(s,a) - \left(\bar{\phi}^t(s,a)^\top\theta_*^t + \bar{b}^{n,t}(s,a)\right)\right)\cdot\mathbb{1}\{s\in\mathcal{K}^n\}}_{\text{Term 1}} \\
&\quad + \underbrace{\bar{\phi}^t(s,a)^\top\left(\theta_*^t - \theta_{\text{mid}}^t\right)\cdot\mathbb{1}\{s\in\mathcal{K}^n\}}_{\text{Term 2}} + \underbrace{\bar{\phi}^t(s,a)^\top\left(\theta_{\text{mid}}^t - \theta^t\right)\cdot\mathbb{1}\{s\in\mathcal{K}^n\}}_{\text{Term 3}}\Bigg].
\end{aligned} \tag{19}$$

Following the proof of Lemma D.1 in (Agarwal et al., 2020), we can bound Terms 1 and 3 as follows.

$$\begin{aligned}
\text{Term 1} &= \mathbb{E}_{(s,a)\sim d_{\mathcal{M}^n;s_{\text{init}}}^{\pi^*,n}}\left[\left(A_{\mathcal{M}_{b^n}}^{\pi^t}(s,a) - \left(\bar{\phi}^t(s,a)^\top\theta_*^t + \bar{b}^{n,t}(s,a)\right)\right)\cdot\mathbb{1}\{s\in\mathcal{K}^n\}\right] \\
&= \mathbb{E}_{(s,a)\sim d_{\mathcal{M}^n;s_{\text{init}}}^{\pi^*,n}}\left[\left(Q_{\mathcal{M}_{b^n}}^{\pi^t}(s,a) - \left(\phi(s,a)^\top\theta_*^t + b(s,a)\right)\right)\cdot\mathbb{1}\{s\in\mathcal{K}^n\}\right] \\
&\quad + \mathbb{E}_{s\sim d_{\mathcal{M}^n;s_{\text{init}}}^{\pi^*,n},a'\sim\pi^t(\cdot|s)}\left[\left(Q_{\mathcal{M}_{b^n}}^{\pi^t}(s,a') - \left(\phi(s,a)^\top\theta_*^t + b(s,a')\right)\right)\cdot\mathbb{1}\{s\in\mathcal{K}^n\}\right] \\
&\overset{\text{(a)}}{\le} \sqrt{\mathbb{E}_{(s,a)\sim d_{s_{\text{init}}}^{\pi^*}}\left[\left(Q_{\mathcal{M}_{b^n}}^{\pi^t}(s,a) - (\phi(s,a)^\top\theta_*^t + b^n(s,a))\right)^2\right]} \\
&\quad + \sqrt{\mathbb{E}_{s\sim d_{s_{\text{init}}}^{\pi^*},a'\sim\pi^t(\cdot|s)}\left[\left(Q_{\mathcal{M}_{b^n}}^{\pi^t}(s,a') - (\phi(s,a')^\top\theta_*^t + b^n(s,a'))\right)^2\right]} \\
&\le 2\sqrt{|\mathcal{A}|\mathbb{E}_{(s,a)\sim d_{s_{\text{init}}}^{\star}}\left[\left(Q_{\mathcal{M}_{b^n}}^{\pi^t}(s,a) - (\phi(s,a)^\top\theta_*^t + b^n(s,a))\right)^2\right]} \\
&\le 2\sqrt{|\mathcal{A}|\varepsilon_{\text{bias}}},
\end{aligned} \tag{20}$$

where inequality (a) uses Lemma D.2.

Define the Q-value function fitting error as

$$\varepsilon_Q := 8W_Q^2\sqrt{\frac{\log\left(\frac{1}{\delta'}\right)}{M_{\text{SGD}}}}.$$

With probability at least $1 - 2NT\delta'$,

$$\text{Term 3} = \mathbb{E}_{(s,a)\sim d_{s_{\text{init}}}^{\pi^*}}\left[\bar{\phi}^t(s,a)^\top\left(\theta_{\text{mid}}^t - \theta^t\right)\cdot\mathbb{1}\{s\in\mathcal{K}^n\}\right]$$

$$\leq 2\sqrt{\beta\zeta_{\text{cov}}W_\theta^2 + \beta(n+1)\varepsilon_Q}$$
$$= 2\sqrt{\beta\zeta_{\text{cov}}W_\theta^2 + 8\beta W_Q^2(n+1)\sqrt{\frac{\log\left(\frac{1}{\delta'}\right)}{M_{\text{SGD}}}}}$$
$$\leq 2W_\theta\sqrt{\beta\zeta_{\text{cov}}} + 4W_Q\sqrt{\beta(n+1)}\left(\frac{\log\left(\frac{1}{\delta'}\right)}{M_{\text{SGD}}}\right)^{\frac{1}{4}}. \tag{21}$$

Define event

$$\mathcal{E}_\theta := \left\{\text{Term 3} \leq 2\sqrt{\beta\zeta_{\text{cov}}W_\theta^2 + \beta(n+1)\varepsilon_Q}\right\}.$$

Then, $\Pr[\mathcal{E}_\theta] \geq 1 - 2NT\delta'$.

Now we have $\Pr[\mathcal{E}_\theta \cap \mathcal{E}_\tau \cap \mathcal{E}_{\text{MLE}} \cap \mathcal{E}_{\text{cov}}] \geq 1 - 4 \cdot 2N(K + M_{\text{HF}} + TM_{\text{SGD}}) \cdot 2\delta' \geq 1 - \delta$. In the following, we assume that event $\mathcal{E}_\theta \cap \mathcal{E}_\tau \cap \mathcal{E}_{\text{MLE}} \cap \mathcal{E}_{\text{cov}}$ holds, and derive the suboptimality guarantee.

Applying Lemma D.11, Term 2 can be bounded as follows.

$$\begin{aligned}
\text{Term 2} &= \mathbb{E}_{(s,a)\sim d^{\pi^*,n}_{\mathcal{M}^n;s_{\text{init}}}}\left[\bar{\phi}^t(s,a)^\top\left(\theta_*^t - \theta_{\text{mid}}^t\right)\cdot\mathbb{1}\left\{s\in\mathcal{K}^n\right\}\right] \\
&\leq \mathbb{E}_{(s,a)\sim d^{\pi^*,n}_{\mathcal{M}^n;s_{\text{init}}}}\left[\left|\phi(s,a)^\top\left(\theta_*^t - \theta_{\text{mid}}^t\right)\right|\cdot\mathbb{1}\left\{s\in\mathcal{K}^n\right\}\right] \\
&\quad + \mathbb{E}_{s\sim d^{\pi^*,n}_{\mathcal{M}^n;s_{\text{init}}},a'\sim\pi^t(\cdot|s)}\left[\left|\phi(s,a')^\top\left(\theta_*^t - \theta_{\text{mid}}^t\right)\right|\cdot\mathbb{1}\left\{s\in\mathcal{K}^n\right\}\right] \\
&\leq 16\sqrt{\beta W_Q\varepsilon_{\text{HF}}^n(n+1)\mathbb{E}_{\tau\sim\mathcal{O}^{\pi^t}_{\rho^n_{\text{cov}}}}\left[\left\|\sum_{h=0}^{H(\tau)}\phi(s_h,a_h)\right\|_{\left(\Sigma^n_{\text{HF}}\right)^{-1}}\right]} + 8W_\theta\sqrt{\beta\zeta_{\text{cov}}}.
\end{aligned} \tag{22}$$

Plugging the connection result between $\mathcal{M}^n$ and $\mathcal{M}$ (Lemma D.3) into the suboptimality decomposition (Eq. (19)), we have

$$V^{\pi^*}(s_{\text{init}}) - V^{\pi^t}(s_{\text{init}}) \leq \text{RHS in Eq. (19)} + \frac{1}{1-\gamma}\sum_{(s,a)\notin\mathcal{K}^n} d^{\pi^t}_{s_{\text{init}}}(s,a).$$

Summing over $t = 0, \ldots, T-1$ and dividing T, we have

$$\begin{aligned}
&V^{\pi^*}(s_{\text{init}}) - V^{\pi^{n+1}}(s_{\text{init}}) \\
&= \frac{1}{T}\sum_{t=0}^{T-1}\left(V^{\pi^*}(s_{\text{init}}) - V^{\pi^t}(s_{\text{init}})\right) \\
&\leq \frac{1}{T}\sum_{t=0}^{T-1}\text{RHS in Eq. (19)} + \frac{1}{1-\gamma}\sum_{(s,a)\notin\mathcal{K}^n} d^{\pi^{n+1}}_{s_{\text{init}}}(s,a) \\
&\overset{\text{(a)}}{\leq} \frac{\log(|\mathcal{A}|)}{(1-\gamma)\eta T} + \frac{\eta W_A^2}{1-\gamma} + \frac{2\sqrt{|\mathcal{A}|\varepsilon_{\text{bias}}}}{1-\gamma} + \frac{4W_Q\sqrt{\beta(n+1)}}{1-\gamma}\left(\frac{\log\left(\frac{1}{\delta'}\right)}{M_{\text{SGD}}}\right)^{\frac{1}{4}} + \frac{10W_\theta\sqrt{\beta\zeta_{\text{cov}}}}{16-\gamma} \\
&\quad + \frac{16\sqrt{\beta W_Q}}{1-\gamma}\cdot 2\left(\frac{(n+1)^2\left(d+\log\left(\frac{1}{\delta'}\right)\right)}{c_{\text{MLE}}^2 M_{\text{HF}}} + 2(n+1)\zeta_{\text{HF}}W_\mu^2\right)^{\frac{1}{4}}\cdot \\
&\quad \frac{1}{T}\sum_{t=0}^{T-1}\sqrt{\mathbb{E}_{\tau\sim\mathcal{O}^{\pi^t}_{\rho^n_{\text{cov}}}}\left[\left\|\sum_{h=0}^{H(\tau)}\phi(s_h,a_h)\right\|_{\left(\Sigma^n_{\text{HF}}\right)^{-1}}\right]} + \frac{1}{1-\gamma}\sum_{(s,a)\notin\mathcal{K}^n} d^{\pi^{n+1}}_{s_{\text{init}}}(s,a),
\end{aligned} \tag{23}$$

where inequality (a) combines the natural policy gradient regret (Lemma D.6) and Terms 1-3 (Eqs. (20)-(22)).

Summing over $n = 0, \ldots, N-1$ and dividing N, we have

$$
\begin{aligned}
& V^{\pi^*}(s_{\text{init}}) - V^{\pi^{\text{out}}}(s_{\text{init}}) \\
&= \frac{1}{N}\sum_{n=0}^{N-1}\left(V^{\pi^*}(s_{\text{init}}) - V^{\pi^{n+1}}(s_{\text{init}})\right) \\
&\le \frac{\log(|\mathcal{A}|)}{(1-\gamma)\eta T} + \frac{\eta W_A^2}{1-\gamma} + \frac{2\sqrt{|\mathcal{A}|\varepsilon_{\text{bias}}}}{1-\gamma} + \frac{8W_Q\sqrt{\beta N}}{1-\gamma}\left(\frac{\log\left(\frac{1}{\delta'}\right)}{M_{\text{SGD}}}\right)^{\frac{1}{4}} + \frac{10W_\theta\sqrt{\beta\zeta_{\text{cov}}}}{1-\gamma} \\
&\quad + \frac{32\sqrt{\beta W_Q}}{1-\gamma}\cdot 2\left(\frac{4N^2\left(d+\log\left(\frac{1}{\delta'}\right)\right)}{c_{\text{MLE}}^2 M_{\text{HF}}} + 4N\zeta_{\text{HF}}W_\mu^2\right)^{\frac{1}{4}}\cdot \\
&\quad \frac{1}{N}\sum_{n=0}^{N-1}\frac{1}{T}\sum_{t=0}^{T-1}\sqrt{\mathbb{E}_{\tau\sim\mathcal{O}^{\pi^t}\rho_{\text{cov}}^n}\left[\left\|\sum_{h=0}^{H(\tau)}\phi(s_h,a_h)\right\|_{(\Sigma_{\text{HF}}^n)^{-1}}\right]} \\
&\quad + \frac{1}{(1-\gamma)N}\sum_{n=0}^{N-1}\sum_{(s,a)\notin\mathcal{K}^n} d_{s_{\text{init}}}^{\pi^{n+1}}(s,a) \\
&\overset{\text{(a)}}{\le} \frac{\log(|\mathcal{A}|)}{(1-\gamma)\eta T} + \frac{\eta W_A^2}{1-\gamma} + \frac{2\sqrt{|\mathcal{A}|\varepsilon_{\text{bias}}}}{1-\gamma} + \frac{8W_Q\sqrt{\beta N}}{1-\gamma}\left(\frac{\log\left(\frac{1}{\delta'}\right)}{M_{\text{SGD}}}\right)^{\frac{1}{4}} + \frac{10W_\theta\sqrt{\beta\zeta_{\text{cov}}}}{1-\gamma} \\
&\quad + \frac{256\sqrt{\beta W_Q}}{1-\gamma}\cdot\left(\frac{N^2\left(d+\log\left(\frac{1}{\delta'}\right)\right)}{c_{\text{MLE}}^2 M_{\text{HF}}} + N\zeta_{\text{HF}}W_\mu^2\right)^{\frac{1}{4}}\cdot\left(d^{\frac{1}{4}}\log^{\frac{1}{4}}\left(1+\frac{4NW_\tau^2}{\zeta_{\text{HF}}d}\right) + \frac{d^{\frac{1}{4}}\log^{\frac{1}{4}}(N)}{c_{\text{base}}^{\frac{1}{4}}}\right) \\
&\quad + \frac{1}{(1-\gamma)N}\cdot\frac{2d}{\beta}\log\left(1+\frac{N}{\zeta_{\text{cov}}d}\right) \\
&\le \frac{2\sqrt{|\mathcal{A}|\varepsilon_{\text{bias}}}}{1-\gamma} + \frac{W_A\sqrt{\log(|\mathcal{A}|)}}{(1-\gamma)\sqrt{T}} + \frac{8W_Q\sqrt{\beta N}}{1-\gamma}\cdot\frac{\log^{\frac{1}{4}}\left(\frac{1}{\delta'}\right)}{(M_{\text{SGD}})^{\frac{1}{4}}} + \frac{10W_\theta\sqrt{\beta\zeta_{\text{cov}}}}{1-\gamma} \\
&\quad + \frac{2\cdot 256\sqrt{\beta W_Q}}{1-\gamma}\cdot\frac{d^{\frac{1}{4}}\log^{\frac{1}{4}}(5N)}{c_{\text{base}}^{\frac{1}{4}}}\cdot\left(\frac{2N^2 d\log\left(\frac{1}{\delta'}\right)}{c_{\text{MLE}}^2 M_{\text{HF}}} + N\zeta_{\text{HF}}W_\mu^2\right)^{\frac{1}{4}} + \frac{2d}{(1-\gamma)N\beta}\log(2N), \qquad (24)
\end{aligned}
$$

where inequality (a) uses Lemmas D.13, D.4, and the fact $\eta := \frac{\sqrt{\log(|\mathcal{A}|)}}{W_A\sqrt{T}}$.

In addition, due to the condition of concentration event $\mathcal{E}_{\text{cov}}$, we should guarantee $K \ge \frac{16(N+1)^2\log^2\left(\frac{4dN}{\delta'}\right)}{\zeta_{\text{cov}}^2}$ and $M_{\text{HF}} \ge \frac{16W_\tau^4\log^2\left(\frac{4d}{\delta'}\right)}{\zeta_{\text{HF}}^2}$.

Recall that $W_\tau := \frac{\log(\frac{1}{\delta'})}{1-\gamma}$, $W_\mu := 1$, $W_A := \frac{4}{(1-\gamma)^2}$, $W_\theta := \frac{2}{(1-\gamma)^2} - \frac{1}{1-\gamma}$, $W_Q := \frac{2}{(1-\gamma)^2}$, $c_{\text{MLE}} := \frac{1}{2+\exp(-2W_\tau W_\mu)+\exp(2W_\tau W_\mu)}$, $\xi := \frac{W_\theta}{(W_Q+W_\theta)\sqrt{T}}$, $\zeta_{\text{cov}} := 1$ and $\zeta_{\text{HF}} := 4W_\tau^2 = \frac{4\log^2(\frac{1}{\delta'})}{(1-\gamma)^2}$.

We set

$$
\begin{aligned}
T &:= \frac{6^2 W_A^2\log(|\mathcal{A}|)}{(1-\gamma)^2\varepsilon^2}, \\
\eta &:= \frac{\sqrt{\log(|\mathcal{A}|)}}{W_A\sqrt{T}} = \frac{(1-\gamma)\varepsilon}{6W_A^2}, \\
\beta &:= \frac{(1-\gamma)^5\varepsilon^5 c_{\text{base}}}{5000\cdot 6^5\cdot 2^4\cdot 256^4 W_Q^2 W_\mu^2\zeta_{\text{HF}}d^2}\log^{-2}\left(\frac{800\cdot 256^2 d^3 W_Q W_\mu\sqrt{10\zeta_{\text{HF}}}}{(1-\gamma)^{4.5}\sqrt{c_{\text{base}}}}\right) = \tilde{O}\left(\frac{(1-\gamma)^5\varepsilon^5 c_{\text{base}}}{W_Q^2 W_\mu^2 d^2\zeta_{\text{HF}}}\right),
\end{aligned}
$$

$$
\begin{aligned}
N &:= \frac{6\cdot 10d}{(1-\gamma)\varepsilon\beta}\log\left(\frac{6\cdot 4d}{(1-\gamma)\varepsilon\beta}\right) = \tilde{O}\left(\frac{d^3W_Q^2W_\mu^2\zeta_{\text{HF}}}{(1-\gamma)^6\varepsilon^6 c_{\text{base}}}\right),\\
M_{\text{SGD}} &:= 1200\cdot \underbrace{\frac{6^4\cdot 8^4\cdot W_Q^4\beta^2N^2}{(1-\gamma)^4\varepsilon^4}}_{:=L_1}\log^2\left(\frac{L_1L_2^2L_3}{\delta}\right) = \tilde{O}\left(\frac{W_Q^4d^2}{(1-\gamma)^6\varepsilon^6}\right),\\
M_{\text{HF}} &:= 1200\cdot \underbrace{\frac{6^4\cdot 2^5\cdot 256^4\beta^2W_Q^2N^2d^2\log(5N)}{(1-\gamma)^4\varepsilon^4c_{\text{MLE}}^2c_{\text{base}}}}_{:=L_3}\log^2\left(\frac{L_1L_2^2L_3}{\delta}\right) = \tilde{O}\left(\frac{W_Q^2d^4}{(1-\gamma)^6\varepsilon^6c_{\text{MLE}}^2c_{\text{base}}}\right),\\
K &:= \frac{64N^2}{\zeta_{\text{cov}}^2}\cdot\log^2\underbrace{\left(\frac{4dN\cdot 12N(K+1+T)M_{\text{HF}}M_{\text{SGD}}}{\delta}\right)}_{:=L_2} = \tilde{O}\left(\frac{d^6W_Q^4W_\mu^4\zeta_{\text{HF}}^2}{(1-\gamma)^{12}\varepsilon^{12}c_{\text{base}}^2}\right),\\
\delta' &:= \frac{\delta}{12N(K+1+T)M_{\text{HF}}M_{\text{SGD}}}.
\end{aligned}
\tag{25}
$$

Then, we have

$$
V^{\pi^*}(s_{\text{init}}) - V^{\pi^{\text{out}}}(s_{\text{init}}) \le \varepsilon + \frac{2\sqrt{|\mathcal{A}|\varepsilon_{\text{bias}}}}{1-\gamma}.
$$

Finally, the number of samples is bounded by

$$
\begin{aligned}
&\tilde{O}\left(N\left(K + M_{\text{HF}} + TM_{\text{SGD}}\right)\cdot\frac{1}{1-\gamma}\right)\\
&= \tilde{O}\left(\frac{W_Q^2W_\mu^2\zeta_{\text{HF}}d^3}{(1-\gamma)^6\varepsilon^6c_{\text{base}}}\cdot\left(\frac{W_Q^4W_\mu^4\zeta_{\text{HF}}^2d^6}{(1-\gamma)^{12}\varepsilon^{12}c_{\text{base}}^2} + \frac{W_Q^2d^4}{(1-\gamma)^6\varepsilon^6c_{\text{MLE}}^2c_{\text{base}}} + \frac{W_A^2}{(1-\gamma)^2\varepsilon^2}\cdot\frac{W_Q^4d^2}{(1-\gamma)^6\varepsilon^6}\right)\cdot\frac{1}{1-\gamma}\right)\\
&= \tilde{O}\left(\frac{W_Q^6W_\mu^6\zeta_{\text{HF}}^3d^9}{(1-\gamma)^{19}\varepsilon^{18}c_{\text{base}}^3}\right).
\end{aligned}
\tag{26}
$$

□

E. Proofs for PO-RLHF with Neural Function Approximation

In this section, we provide the proofs for algorithm `NN-PG-RLHF`.

Definitions for Neural Function Approximation. We first introduce or recall some definitions.

Let $\mathcal{S}_R := \{w \in \mathbb{R}^{md} : \left\|w - w^0\right\|_2 \le R\}$ and $\mathcal{U}_R := \{\mu \in \mathbb{R}^{md} : \left\|\mu - \mu^0\right\|_2 \le R\}$.

For any $w \in \mathbb{R}^{md}$, recall that

$$
\begin{aligned}
[\psi_w]_\ell(s,a) &:= \frac{b_\ell}{\sqrt{m}}\cdot\mathbb{1}\left\{\phi(s,a)^\top[w]_\ell > 0\right\}\phi(s,a) \ \in \mathbb{R}^d, \quad \forall \ell \in [m],\\
\psi_w(s,a) &:= [[\psi_w]_1(s,a); \ldots; [\psi_w]_m(s,a)] \ \in \mathbb{R}^{md}.
\end{aligned}
$$

Here $\underline{c} \le \|[w^0]_\ell\|_2 \le \bar{c}$ for all $\ell \in [m]$ for some constants $\underline{c}, \bar{c} > 0$.

Recall the Q-network, policy network and reward network as follow:

$$
f(s,a;\theta) := \frac{1}{\sqrt{m}}\sum_{\ell=1}^m b_\ell\cdot\mathbb{1}\left\{\phi(s,a)^\top[\theta]_\ell > 0\right\}\phi(s,a)^\top[\theta]_\ell = \psi_\theta(s,a)^\top\theta,
$$

$$\pi_{\alpha,w}(a|s) := \frac{\exp\left(\alpha f(s,a;w)\right)}{\sum_{a'\in\mathcal{A}}\exp\left(\alpha f(s,a';w)\right)} = \frac{\exp(\alpha\psi_w(s,a)^\top w)}{\sum_{a'\in\mathcal{A}}\exp\left(\alpha\psi_w(s,a')^\top w\right)},$$
$$h(s,a;\mu) := \frac{1}{\sqrt{m}}\sum_{\ell=1}^{m} b'_\ell \cdot \mathbb{1}\left\{\phi(s,a)^\top[\mu]_\ell > 0\right\}\phi(s,a)^\top[\mu]_\ell = \psi_\mu(s,a)^\top\mu.$$

For any $t \geq 0$, we use π^t and π_{α^t,w^t} interchangeably.

Let

$$f_0(s,a;w) := \frac{1}{\sqrt{m}}\sum_{\ell=1}^{m} b_\ell \cdot \mathbb{1}\left\{\phi(s,a)^\top[w^0]_\ell > 0\right\}\phi(s,a)^\top[w]_\ell = \psi_{w^0}(s,a)^\top w,$$
$$h_0(s,a;\mu) := \frac{1}{\sqrt{m}}\sum_{\ell=1}^{m} b'_\ell \cdot \mathbb{1}\left\{\phi(s,a)^\top[\mu^0]_\ell > 0\right\}\phi(s,a)^\top[\mu]_\ell = \psi_{\mu^0}(s,a)^\top w.$$

Define the neural kernel spaces as

$$\mathcal{F}^w_{R,\infty} := \left\{f(s,a) = f(s,a;w^0) + \int \mathbb{1}\left\{\phi(s,a)^\top w > 0\right\}\phi(s,a)^\top\nu^w(w)dp^w(w) :\ \|\nu^w(w)\|_\infty \leq \frac{R}{\sqrt{d}}\right\},$$
$$\mathcal{F}^\mu_{R,\infty} := \left\{f(s,a) = h(s,a;\mu^0) + \int \mathbb{1}\left\{\phi(s,a)^\top\mu > 0\right\}\phi(s,a)^\top\nu^\mu(\mu)dp^\mu(\mu) :\ \|\nu^\mu(\mu)\|_\infty \leq \frac{R}{\sqrt{d}}\right\},$$

Here $\nu^w : \mathbb{R}^d \to \mathbb{R}^d$ and $f(s,a;w^0)$ parameterize the element of $\mathcal{F}^w_{R,\infty}$, and $p^w : \mathbb{R}^d \to \mathbb{R}$ is the density function of the initialization distribution of w^0. Similarly, $\nu^\mu : \mathbb{R}^d \to \mathbb{R}^d$ and $h(s,a;\mu^0)$ parameterize the element of $\mathcal{F}^\mu_{R,\infty}$, and $p^\mu : \mathbb{R}^d \to \mathbb{R}$ is the density function of the initialization distribution of μ^0. In this work, for simplicity, we set the initialization distribution for w^0 and μ^0 as $\mathcal{D}_{\text{init}}$.

Define

$$\mathcal{F}^\mu_{R,m} := \left\{\frac{1}{\sqrt{m}}\sum_{\ell=1}^{m} b'_\ell \cdot \mathbb{1}\left\{\phi(s,a)^\top[\mu^0]_\ell > 0\right\}\phi(s,a)^\top[\mu]_\ell :\ \left\|\mu-\mu^0\right\|_2 \leq R\right\},$$
$$\bar{\mathcal{F}}^\mu_{R,m} := \left\{\frac{1}{\sqrt{m}}\sum_{\ell=1}^{m} b'_\ell \cdot \mathbb{1}\left\{\phi(s,a)^\top[\mu^0]_\ell > 0\right\}\phi(s,a)^\top[\mu]_\ell :\ \left\|[\mu]_\ell-[\mu^0]_\ell\right\|_\infty \leq \frac{R}{\sqrt{md}}\right\}.$$

$\bar{\mathcal{F}}^\mu_{R,m}$ is the subset of $\mathcal{F}^\mu_{R,m}$.

Let $\mu_r^{\text{proj}} \in \mathcal{U}_R$ be the parameter such that

$$\text{Proj}_{\mathcal{F}_{R,m}} r(s,a) = \psi_0(s,a)^\top\mu_r^{\text{proj}}.$$

Covariance Matrix Concentration. Next, we define the concentration event for the coverage and human data covariance matrices.

For any trajectory $\tau = (s_0, a_0, \ldots, s_{H(\tau)}, a_{H(\tau)})$ and $\mu \in \mathcal{U}_R$, let $\psi_\mu(\tau) := \sum_{h=0}^{H(\tau)}\psi_\mu(s_h,a_h)$. For any trajectories $\tau^{(1)},\tau^{(2)}$ and $\mu \in \mathcal{U}_R$, let $\tilde{\psi}_\mu^{\tau^{(1)},\tau^{(2)}} := \sum_{h=0}^{H(\tau^{(1)})}\psi_\mu(s_h^{(1)},a_h^{(1)}) - \sum_{h=0}^{H(\tau^{(2)})}\psi_\mu(s_h^{(2)},a_h^{(2)})$.

For any $n \geq 0$ and $t \geq 0$, let (s_i^n, a_i^n) denote the i-th state-action pair sampled in phase n for constructing the estimated coverage covariance matrix $\hat{\Sigma}^{\text{NN},n}_{\text{cov}}$ (Line 4 in Algorithm 3).

For any phase $n \geq 0$, define

$$\hat{\Sigma}^{\text{NN},n}_{\text{cov}} := \sum_{i=0}^{n}\left(\frac{1}{K}\sum_{i=1}^{K}\psi_0(s_i^n,a_i^n)\psi_0(s_i^n,a_i^n)^\top\right) + \zeta_{\text{cov}}I,$$

$$
\begin{aligned}
\Sigma_{\text{cov}}^{\text{NN},n} &:= \sum_{i=0}^{n} \mathbb{E}_{(s,a)\sim d_{s_{\text{init}}}^{\pi^i}} \left[\psi_0(s,a)\psi_0(s,a)^\top\right] + \zeta_{\text{cov}} I \\
&= (n+1)\mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n} \left(\psi_0(s,a)\psi_0(s,a)^\top\right) + \zeta_{\text{cov}} I. \\
\hat{\Sigma}_{\text{HF}}^{\text{NN},n} &:= \frac{1}{M_{\text{HF}}} \sum_{i=1}^{M_{\text{HF}}} \left(\psi_0(\tau_i^{(1)}) - \psi_0(\tau_i^{(2)})\right) \left(\psi_0(\tau_i^{(1)}) - \psi_0(\tau_i^{(2)})\right)^\top + \frac{\zeta_{\text{HF}}}{n} I \\
&= \frac{1}{M_{\text{HF}}} \sum_{i=1}^{M_{\text{HF}}} \tilde{\psi}_0^{\tau_i^{(1)},\tau_i^{(2)}} \left(\tilde{\psi}_0^{\tau_i^{(1)},\tau_i^{(2)}}\right)^\top + \frac{\zeta_{\text{HF}}}{n} I, \quad \forall n \geq 1 \\
\Sigma_{\text{HF}}^{\text{NN},n} &:= \frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{E}_{\substack{\tau^{(1)}\sim\mathcal{O}^{\pi^i}_{\rho_{\text{cov}}^{i-1}} \\ \tau^{(2)}\sim\mathcal{O}^{\pi^{\text{base}}}_{s_{\text{init}}}}} \left[\left(\psi_0(\tau^{(1)}) - \psi_0(\tau^{(2)})\right) \left(\psi_0(\tau^{(1)}) - \psi_0(\tau^{(2)})\right)^\top \right] \right) + \frac{\zeta_{\text{HF}}}{n} I \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{E}_{\substack{\tau^{(1)}\sim\mathcal{O}^{\pi^i}_{\rho_{\text{cov}}^{i-1}} \\ \tau^{(2)}\sim\mathcal{O}^{\pi^{\text{base}}}_{s_{\text{init}}}}} \left[\tilde{\psi}_0^{\tau^{(1)},\tau^{(2)}} \left(\tilde{\psi}_0^{\tau^{(1)},\tau^{(2)}}\right)^\top \right] \right) + \frac{\zeta_{\text{HF}}}{n} I, \quad \forall n \geq 1 \\
\hat{\Sigma}_{\text{HF}}^{\text{NN},n} &= \Sigma_{\text{HF}}^{\text{NN},n} := \zeta_{\text{HF}} I.
\end{aligned}
$$

Recall $W_\tau := \frac{\log\left(\frac{1}{\delta'}\right)}{1-\gamma}$ and the definition of event $\mathcal{E}_\tau$ (Eq. (11)).

Define event

$$
\begin{aligned}
\mathcal{E}_{\text{cov}}^{\text{NN}} := \Bigg\{ & \frac{1}{2} \left\|\psi_0(s,a)\right\|_{(\Sigma_{\text{cov}}^{\text{NN},n})^{-1}} \leq \left\|\psi_0(s,a)\right\|_{(\hat{\Sigma}_{\text{cov}}^{\text{NN},n})^{-1}} \leq 2 \left\|\psi_0(s,a)\right\|^{-1}_{(\Sigma_{\text{cov}}^{\text{NN},n})}, \\
& \frac{1}{2} \left\|\psi_0(s,a)\right\|_{(\Sigma_{\text{HF}}^{\text{NN},n})^{-1}} \leq \left\|\psi_0(s,a)\right\|_{(\hat{\Sigma}_{\text{HF}}^{\text{NN},n})^{-1}} \leq 2 \left\|\psi_0(s,a)\right\|^{-1}_{(\Sigma_{\text{HF}}^{\text{NN},n})}, \ \forall 0 \leq n \leq N-1 \Bigg\}.
\end{aligned}
$$

Lemma E.1. *Assuming that event $\mathcal{E}_\tau$ holds, then we have $\Pr[\mathcal{E}_{\text{cov}}^{\text{NN}}] \geq 1 - 2N\delta'$.*

Proof. This lemma follows from Lemma F.2 and the condition that $K \geq \frac{16(N+1)^2 \log^2\left(\frac{4dN}{\delta'}\right)}{\zeta_{\text{cov}}^2}$ and $M_{\text{HF}} \geq \frac{16 W_\tau^4 \log^2\left(\frac{4dN}{\delta'}\right)}{\zeta_{\text{HF}}^2}$. □

E.1. Neural Function Approximation

In the following, we present useful technical lemmas for neural function approximation. Lemmas E.2-E.5 borrow the ideas from prior neural network theory works (Rahimi & Recht, 2008; Cai et al., 2019; Wang et al., 2019; Xu et al., 2021).

For brevity of presentation, Lemmas E.2 and E.3 are written with parameter w and function f, but it works for parameters w, θ, μ and their corresponding functions f, h.

For ease of notation, we simplify the notations ψ_{θ^0} and ψ_{μ^0} as ψ_0, which can be easily recovered from the context.

Lemma E.2. *For any $w, w' \in \mathbb{R}^{md}$ such that $\|w - w^0\|_2 \leq R$ and $\|w' - w^0\|_2 \leq R$,*

$$
\begin{aligned}
\mathbb{E}_\rho \left[\left| \psi_0(s,a)^\top w' - \psi_w(s,a)^\top w' \right|^2 \right] &\leq \frac{4 c_{\text{scale}} R^3}{\underline{c}\sqrt{m}}, \\
\mathbb{E}_\rho \left[\left\| \psi_0(s,a) - \psi_w(s,a) \right\|_2^2 \right] &\leq \frac{c_{\text{scale}} R}{\underline{c}\sqrt{m}}.
\end{aligned}
$$

Proof. We prove the first statement as follows.

$$
\left| \psi_0(s,a)^\top w' - \psi_w(s,a)^\top w' \right|
$$

$$= \frac{1}{\sqrt{m}} \sum_{\ell=1}^{m} b_\ell \cdot \left(\mathbb{1}\left\{ \phi(s,a)^\top [w^0]_\ell > 0 \right\} - \mathbb{1}\left\{ \phi(s,a)^\top [w]_\ell > 0 \right\} \right) \phi(s,a)^\top [w']_\ell$$
$$\leq \frac{1}{\sqrt{m}} \sum_{\ell=1}^{m} \left| \mathbb{1}\left\{ \phi(s,a)^\top [w^0]_\ell > 0 \right\} - \mathbb{1}\left\{ \phi(s,a)^\top [w]_\ell > 0 \right\} \right| \left| \phi(s,a)^\top [w']_\ell \right|.$$

Since $\left| \mathbb{1}\left\{ \phi(s,a)^\top [w^0]_\ell > 0 \right\} - \mathbb{1}\left\{ \phi(s,a)^\top [w]_\ell > 0 \right\} \right|$ implies

$$\left| \phi(s,a)^\top [w^0]_\ell \right| \leq \left| \phi(s,a)^\top [w]_\ell - \phi(s,a)^\top [w^0]_\ell \right| \leq \|\phi(s,a)\|_2 \left\| [w^0]_\ell - [w]_\ell \right\|_2,$$

we have

$$\left| \mathbb{1}\left\{ \phi(s,a)^\top [w^0]_\ell > 0 \right\} - \mathbb{1}\left\{ \phi(s,a)^\top [w]_\ell > 0 \right\} \right| \leq \mathbb{1}\left\{ \left| \phi(s,a)^\top [w^0]_\ell \right| \leq \|\phi(s,a)\|_2 \left\| [w^0]_\ell - [w]_\ell \right\|_2 \right\}. \tag{27}$$

Hence, we have

$$\begin{aligned}
&\left| \psi_0(s,a)^\top w' - \psi_w(s,a)^\top w' \right| \\
&\leq \frac{1}{\sqrt{m}} \sum_{\ell=1}^{m} \mathbb{1}\left\{ \left| \phi(s,a)^\top [w^0]_\ell \right| \leq \|\phi(s,a)\|_2 \left\| [w^0]_\ell - [w]_\ell \right\|_2 \right\} \left| \phi(s,a)^\top [w']_\ell \right| \\
&\leq \frac{1}{\sqrt{m}} \sum_{\ell=1}^{m} \mathbb{1}\left\{ \left| \phi(s,a)^\top [w^0]_\ell \right| \leq \|\phi(s,a)\|_2 \left\| [w^0]_\ell - [w]_\ell \right\|_2 \right\} \left(\left| \phi(s,a)^\top [w^0]_\ell \right| + \left| \phi(s,a)^\top \left([w']_\ell - [w^0]_\ell \right) \right| \right) \\
&\overset{\text{(a)}}{\leq} \frac{1}{\sqrt{m}} \sum_{\ell=1}^{m} \mathbb{1}\left\{ \left| \phi(s,a)^\top [w^0]_\ell \right| \leq \|\phi(s,a)\|_2 \left\| [w^0]_\ell - [w]_\ell \right\|_2 \right\} \cdot \\
&\quad \left(\|\phi(s,a)\|_2 \left\| [w^0]_\ell - [w]_\ell \right\|_2 + \|\phi(s,a)\|_2 \left\| [w']_\ell - [w^0]_\ell \right\|_2 \right),
\end{aligned}$$

where inequality (a) is due to $\mathbb{1}\{|x| \leq y\} |x| \leq \mathbb{1}\{|x| \leq y\} y$.

Using the Cauchy-Schwartz inequality, we have

$$\begin{aligned}
&\left| \psi_0(s,a)^\top w' - \psi_w(s,a)^\top w' \right|^2 \\
&\leq \frac{1}{m} \sum_{\ell=1}^{m} \mathbb{1}\left\{ \left| \phi(s,a)^\top [w^0]_\ell \right| \leq \|\phi(s,a)\|_2 \left\| [w^0]_\ell - [w]_\ell \right\|_2 \right\} \cdot \\
&\quad \sum_{\ell=1}^{m} \left(2 \|\phi(s,a)\|_2^2 \left\| [w^0]_\ell - [w]_\ell \right\|_2^2 + 2 \|\phi(s,a)\|_2^2 \left\| [w']_\ell - [w^0]_\ell \right\|_2^2 \right) \\
&\leq \frac{4R^2}{m} \sum_{\ell=1}^{m} \mathbb{1}\left\{ \left| \phi(s,a)^\top [w^0]_\ell \right| \leq \|\phi(s,a)\|_2 \left\| [w^0]_\ell - [w]_\ell \right\|_2 \right\} \\
&\leq \frac{4R^2}{m} \sum_{\ell=1}^{m} \mathbb{1}\left\{ \left| \phi(s,a)^\top [w^0]_\ell \right| \leq \left\| [w^0]_\ell - [w]_\ell \right\|_2 \right\}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}_\rho \left[\left| \psi_0(s,a)^\top w' - \psi_w(s,a)^\top w' \right|^2 \right] &\leq \frac{4R^2}{m} \sum_{\ell=1}^{m} \mathbb{E}_\rho \left[\mathbb{1}\left\{ \left| \phi(s,a)^\top [w^0]_\ell \right| \leq \left\| [w^0]_\ell - [w]_\ell \right\|_2 \right\} \right] \\
&\overset{\text{(a)}}{\leq} \frac{4c_{\text{scale}} R^2}{m} \sum_{\ell=1}^{m} \frac{\left\| [w^0]_\ell - [w]_\ell \right\|_2}{\left\| [w^0]_\ell \right\|_2} \\
&\leq \frac{4c_{\text{scale}} R^2}{m} \sqrt{\sum_{\ell=1}^{m} \left\| [w^0]_\ell - [w]_\ell \right\|_2^2} \sqrt{\sum_{\ell=1}^{m} \frac{1}{\left\| [w^0]_\ell \right\|_2^2}}
\end{aligned}$$

$$\leq \frac{4c_{\text{scale}}R^3}{\underline{c}\sqrt{m}},$$

where inequality (a) uses Assumption 3.3.

Next, we prove the second statement using the similar argument.

$$\begin{aligned}
&\|\psi_0(s,a) - \psi_w(s,a)\|_2^2 \\
&= \sum_{\ell=1}^{m} \frac{b_\ell^2}{m} \cdot \left(\mathbb{1}\left\{\phi(s,a)^\top [w^0]_\ell > 0\right\} - \mathbb{1}\left\{\phi(s,a)^\top [w]_\ell > 0\right\}\right)^2 \|\phi(s,a)\|_2^2 \\
&\overset{\text{(a)}}{\leq} \frac{1}{m}\sum_{\ell=1}^{m} \mathbb{1}\left\{\left|\phi(s,a)^\top [w^0]_\ell\right| \leq \left\|[w^0]_\ell - [w]_\ell\right\|_2\right\},
\end{aligned}$$

where inequality (a) uses Eq. (27).

Taking $\mathbb{E}_\rho[\cdot]$, we have

$$\begin{aligned}
\mathbb{E}_\rho\left[\|\psi_0(s,a) - \psi_w(s,a)\|_2^2\right] &\leq \frac{1}{m}\sum_{\ell=1}^{m} \mathbb{E}_\rho\left[\mathbb{1}\left\{\left|\phi(s,a)^\top [w^0]_\ell\right| \leq \left\|[w^0]_\ell - [w]_\ell\right\|_2\right\}\right] \\
&\leq \frac{c_{\text{scale}}}{m}\sum_{\ell=1}^{m} \frac{\left\|[w^0]_\ell - [w]_\ell\right\|_2}{\left\|[w^0]_\ell\right\|_2} \\
&\leq \frac{c_{\text{scale}}}{m}\sqrt{\sum_{\ell=1}^{m}\left\|[w^0]_\ell - [w]_\ell\right\|_2^2}\sqrt{\sum_{\ell=1}^{m}\frac{1}{\left\|[w^0]_\ell\right\|_2^2}} \\
&\leq \frac{c_{\text{scale}}R}{m}\sqrt{\sum_{\ell=1}^{m}\frac{1}{\left\|[w^0]_\ell\right\|_2^2}} \\
&\leq \frac{c_{\text{scale}}R}{\underline{c}\sqrt{m}}.
\end{aligned}$$

□

Lemma E.3. *For any $w \in \mathcal{S}_R$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$\begin{aligned}
\|\psi_w(s,a)\|_2 &\leq 1, \\
\|w\|_2 &\leq \sqrt{m}\bar{c} + R, \\
|f(s,a;w)| &\leq \sqrt{m}\bar{c} + R.
\end{aligned}$$

Proof. We have

$$\|\psi_w(s,a)\|_2 = \sqrt{\sum_{\ell=1}^{m}\|[\psi_w(s,a)]_\ell\|_2^2} = \sqrt{\sum_{\ell=1}^{m}\frac{b_\ell^2}{m}\cdot\mathbb{1}\left\{\phi(s,a)^\top[w]_\ell > 0\right\}\|\phi(s,a)\|_2^2} \leq 1.$$

In addition,

$$\left\|w^0\right\|_2 = \sqrt{\sum_{\ell=1}^{m}\|[w^0]_\ell\|_2^2} \leq \sqrt{m}\bar{c}.$$

Then,

$$\|w\|_2 \leq \left\|w^0\right\|_2 + \left\|w - w^0\right\|_2$$

$$\leq \sqrt{m}\bar{c} + R.$$

Furthermore,

$$|f(s,a;w)| = |\psi_w(s,a)^\top w| \leq \|\psi_w(s,a)\|_2 \|w\|_2 \leq \sqrt{m}\bar{c} + R.$$

□

Lemma E.4 (Projection Error for $\bar{\mathcal{F}}^{\mu}_{R,m}$ (Rahimi & Recht, 2008))**.** *Let $h \in \mathcal{F}^{\mu}_{R,\infty}$. For any $\delta' > 0$, with probability at least $1-\delta'$,*

$$\left\|\text{Proj}_{\bar{\mathcal{F}}_{R,m}} h - h\right\|_\rho \leq \frac{R\left(1+\sqrt{2\log\left(\frac{1}{\delta'}\right)}\right)}{\sqrt{m}},$$

where ρ is a distribution over $\mathcal{S}\times\mathcal{A}$.

Lemma E.5 (Distance between $r(s,a)$ and $\psi_0(s,a)^\top \mu_r^{\text{proj}}$)**.** *Assume that event $\mathcal{E}_{\text{init}}$ holds. Then,*

$$\left\|\psi_0^\top \mu_r^{\text{proj}} - r\right\|_\rho \leq 4R\sqrt{\frac{\log\left(\frac{1}{\delta'}\right)}{m}}.$$

Proof. Recall that $r \in \mathcal{F}^{\mu}_{R,\infty}$ and $\text{Proj}_{\mathcal{F}^{\mu}_{R,m}} r(s,a) = \psi_0(s,a)^\top \mu_r^{\text{proj}}$. Since $\bar{\mathcal{F}}^{\mu}_{R,m}$ is a subset of $\mathcal{F}^{\mu}_{R,m}$, we have

$$\left\|\psi_0^\top \mu_r^{\text{proj}} - r\right\|_\rho = \left\|\text{Proj}_{\mathcal{F}_{R,m}} r - r\right\|_\rho \leq \left\|\text{Proj}_{\bar{\mathcal{F}}_{R,m}} r - r\right\|_\rho \leq 4R\sqrt{\frac{\log\left(\frac{1}{\delta'}\right)}{m}}.$$

□

Define event

$$\mathcal{E}_{\text{init}} : \left\{ \left\|\psi_0^\top \mu_r^{\text{proj}} - r\right\|_{d^{\pi^t}_{\rho^n_{\text{cov}}}} \leq 4R\sqrt{\frac{\log\left(\frac{1}{\delta'}\right)}{m}},\ \forall t \in [T],\ \forall n \in [N] \right\}.$$

Lemma E.6. *It holds that $\Pr[\mathcal{E}_{\text{init}}] \geq NT\delta'$.*

Proof. This lemma follows from Lemma E.5 and a union bound. □

E.2. Neural Neural Policy Gradient

Let $W_\theta^{\text{NN}} := \sqrt{m}\bar{c} + R$. According to Remark 28 in (Agarwal et al., 2021), since $\|\psi_0(s,a)\|_2 \leq 1$, $\log(\pi_{\alpha,w})$ is a smooth function with smoothness parameter $W_S = 1$.

Lemma E.7 (Neural Neural Policy Gradient)**.** *For any phase $n \geq 0$ and iteration $t \geq 0$,*

$$\sum_{t=0}^{T-1} \mathbb{E}_{(s,a)\sim d^{\pi^{*,n}}_{\mathcal{M}^n;s_{\text{init}}}} \left[\left(\bar{\psi}^t_{w^t}(s,a)^\top \theta^t + \bar{b}^{n,t}(s,a)\right)\cdot \mathbb{1}\left\{s \in \mathcal{K}^n\right\}\right] \leq \frac{\log(|\mathcal{A}|)}{\eta} + \eta W_S (W_\theta^{\text{NN}})^2 T.$$

Proof. Following the analysis in (Agarwal et al., 2021), according to the W_S-smoothness of $\log(\pi_{\alpha^t,w^t})$, we have

$$\begin{aligned}
&\log\left(\pi_{\alpha^{t+1},w^{t+1}}(a|s)\right) - \log\left(\pi_{\alpha^t,w^t}(a|s)\right) \\
&\geq \nabla_w \log\left(\pi_{\alpha^t,w^t}(a|s)\right)^\top \left(\alpha^{t+1}w^{t+1} - \alpha^t w^t\right) - W_S\left\|\alpha^{t+1}w^{t+1} - \alpha^t w^t\right\|_2^2.
\end{aligned}$$

For any $s \in \mathcal{K}^n$, we have

$$\text{KL}(\pi^{*,n}(\cdot|s)\|\pi^t(\cdot|s)) - \text{KL}(\pi^{*,n}(\cdot|s)\|\pi^{t+1}(\cdot|s))$$

Algorithm 5 Q-network Training via Projected SGD (with the objective Eq. (10))

Input: $f(s,a;w^0)$, ξ_θ.
for $i=0,\dots,M_{\text{SGD}}^{\theta}-1$ **do**
  $g^{t,i} \leftarrow 2\left(f(s_i,a_i;\theta^{t,i}) - \left(\hat{Q}^{\pi^t}(s_i,a_i;\hat{r}^n+b^n) - b^n(s_i,a_i)\right)\right)\nabla_\theta f(s_i,a_i;\theta^{t,i})$, where $(s_i,a_i) \sim \rho_{\text{cov}}^n$ and $\hat{Q}^{\pi^t}(s_i,a_i;\hat{r}^n+b^n)$ is estimated by Monte Carlo sampling
  $\tilde{\theta}^{t,i+1} := \theta^{t,i} - \xi_\theta g^{t,i}$
  $\theta^{t,i+1} \leftarrow \text{Proj}_{\mathcal{U}_R}(\tilde{\theta}^{t,i+1})$
end for
return $\theta^t = \sum_{i=0}^{M_{\text{SGD}}^{\theta}-1}\theta^{t,i}$

$$
\begin{aligned}
&= \mathbb{E}_{a\sim\pi^{*,n}(\cdot|s)}\left[\log\left(\frac{\pi^{*,n}(a|s)}{\pi^t(a|s)}\right)\right] - \mathbb{E}_{a\sim\pi^{*,n}(\cdot|s)}\left[\log\left(\frac{\pi^{*,n}(a|s)}{\pi^{t+1}(a|s)}\right)\right]\\
&= \mathbb{E}_{a\sim\pi^{*,n}(\cdot|s)}\left[\log\left(\frac{\pi^{t+1}(a|s)}{\pi^t(a|s)}\right)\right]\\
&\geq \mathbb{E}_{a\sim\pi^{*,n}(\cdot|s)}\left[\nabla_w \log\left(\pi_{\alpha^t,w^t}(a|s)\right)^\top\left(\alpha^{t+1}w^{t+1}-\alpha^t w^t\right) - W_S\left\|\alpha^{t+1}w^{t+1}-\alpha^t w^t\right\|_2^2\right]\\
&= \eta\mathbb{E}_{a\sim\pi^{*,n}(\cdot|s)}\left[\bar{\psi}_{w^t}^t(s,a)^\top\theta^t\right] - \eta^2 W_S\left\|\theta^t\right\|_2^2,
\end{aligned}
$$

which is equivalent to

$$
\mathbb{E}_{a\sim\pi^{*,n}(\cdot|s)}\left[\bar{\psi}_{w^t}^t(s,a)^\top\theta^t\right] \leq \frac{1}{\eta}\left(\text{KL}(\pi^{*,n}(\cdot|s)\|\pi^t(\cdot|s)) - \text{KL}(\pi^{*,n}(\cdot|s)\|\pi^{t+1}(\cdot|s))\right) + \eta W_S\left\|\theta^t\right\|_2^2.
$$

For any phase $n\geq 0$, $s\in\mathcal{K}^n$ and $a\in\mathcal{A}$, we have $b^n(s,a)=0$, and then $\bar{b}^{n,t}(s,a) := b^n(s,a) - \mathbb{E}_{a'\sim\pi^t(\cdot|s)}\left[b^n(s,a')\right] = 0$. Adding $s\sim d_{\mathcal{M}^n;s_{\text{init}}}^{\pi^{*,n}}$ on both sides and summing over $t=0,\dots,T-1$, we have

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}_{(s,a)\sim d_{\mathcal{M}^n;s_{\text{init}}}^{\pi^{*,n}}}\left[\left(\bar{\psi}_{w^t}^t(s,a)^\top\theta^t + \bar{b}^{n,t}(s,a)\right)\cdot\mathbb{1}\left\{s\in\mathcal{K}^n\right\}\right]\\
=&\sum_{t=0}^{T-1}\mathbb{E}_{(s,a)\sim d_{\mathcal{M}^n;s_{\text{init}}}^{\pi^{*,n}}}\left[\bar{\psi}_{w^t}^t(s,a)^\top\theta^t\cdot\mathbb{1}\left\{s\in\mathcal{K}^n\right\}\right]\\
\leq&\frac{1}{\eta}\mathbb{E}_{s\sim d_{\mathcal{M}^n;s_{\text{init}}}^{\pi^{*,n}}}\left[\text{KL}(\pi^{*,n}(\cdot|s)\|\pi^0(\cdot|s)) - \text{KL}(\pi^{*,n}(\cdot|s)\|\pi^T(\cdot|s))\right] + \eta W_S(W_\theta^{\text{NN}})^2 T\\
\leq&\frac{\log(|\mathcal{A}|)}{\eta} + \eta W_S(W_\theta^{\text{NN}})^2 T.
\end{aligned}
$$

□

E.3. Q-value Function Fitting

For any fixed phase $n=0,\dots,N-1$ and fixed iteration $t=0,\dots,T-1$, define

$$
\begin{aligned}
F^{\hat{r}^n}(\theta) &:= \mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n}\left[\left(f_0(s,a;\theta) - \left(Q^{\pi^t}(s,a;\hat{r}^n+b^n) - b^n(s,a)\right)\right)^2\right],\\
\theta_{\text{mid}}^{t,\hat{r}^n} &:= \operatorname*{argmin}_{\theta\in\mathcal{S}_R} F^{\hat{r}^n}(\theta).
\end{aligned}
$$

Then,

$$
\nabla_\theta F^{\hat{r}^n}(\theta) := \mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n}\left[2\left(f_0(s,a;\theta) - \left(Q^{\pi^t}(s,a;\hat{r}^n+b^n) - b^n(s,a)\right)\right)\nabla_\theta f_0(s,a;\theta)\right].
$$

Furthermore, for any $i = 0, \ldots, M_{\text{SGD}}^{\theta} - 1$, define

$$\begin{aligned} g^{t,i} &:= 2\left(f(s_i, a_i; \theta^{t,i}) - \left(\hat{Q}^{\pi^t}(s_i, a_i; \hat{r}^n + b^n) - b^n(s_i, a_i)\right)\right) \nabla_\theta f(s_i, a_i; \theta^{t,i}), \\ \tilde{\theta}^{t,i+1} &:= \theta^{t,i} - \xi_\theta g^{t,i}, \\ \bar{g}^{t,i} &:= \mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n}\left[2\left(f(s, a; \theta^{t,i}) - \left(Q^{\pi^t}(s, a; \hat{r}^n + b^n) - b^n(s, a)\right)\right) \nabla_\theta f(s, a; \theta^{t,i})\right], \end{aligned}$$

and it holds that

$$\theta^{i+1} = \text{Proj}_{\mathcal{S}_R}(\tilde{\theta}^{i+1}).$$

Let $W_{\nabla F}^{\text{NN}} := \frac{4}{(1-\gamma)^2} + \frac{4(\sqrt{m}\bar{c}+R)}{1-\gamma}$.

Define event

$$\mathcal{E}_\theta^{\text{NN}} := \Bigg\{ \left| \sum_{i=0}^{M_{\text{SGD}}^{\theta}-1} (g^{t,i})^\top \left(\theta^{t,i} - \theta_{\text{mid}}^{t,\hat{r}^n}\right) - \sum_{i=0}^{M_{\text{SGD}}^{\theta}-1} (\bar{g}^{t,i})^\top \left(\theta^{t,i} - \theta_{\text{mid}}^{t,\hat{r}^n}\right) \right| \le 2W_{\nabla F}^{\text{NN}} R \sqrt{M_{\text{SGD}}^{\theta} \log\left(\frac{1}{\delta'}\right)}, \\ \forall 0 \le n \le N-1, \forall 0 \le t \le T-1, \forall 0 \le i \le M_{\text{SGD}}^{\theta} - 1 \Bigg\}.$$

Lemma E.8. *It holds that* $\Pr[\mathcal{E}_\theta^{\text{NN}}] \ge 1 - 2NT\delta'$.

Proof. This lemma can be obtained by using the Azuma-Hoeffding inequality and the union bound. □

Let $W_f^{\text{NN}} := \sqrt{m}\bar{c} + R$, $W_Q^{\text{NN}} := \frac{\sqrt{m}\bar{c}+R}{1-\gamma} + \frac{2}{(1-\gamma)^2}$ and $\xi_\theta := \frac{R}{W_{\nabla F}^{\text{NN}}\sqrt{M_{\text{SGD}}^{\theta}}}$.

Below we give the guarantee for the projected SGD of Q-network training, which is described in algorithm 5.

Lemma E.9 (SGD for Q-value Function Fitting)**.** *Assume that event $\mathcal{E}_\theta^{\text{NN}}$ holds. Then, for any phase $n \ge 0$ and iteration $t \ge 0$,*

$$F^{\hat{r}^n}(\theta^t) - F^{\hat{r}^n}(\theta_{\text{mid}}^{t,\hat{r}^n}) \le 4W_{\nabla F}^{\text{NN}} R \sqrt{\frac{\log\left(\frac{1}{\delta'}\right)}{M_{\text{SGD}}^{\theta}}} + \frac{12R^2(W_f^{\text{NN}} + W_Q^{\text{NN}})\sqrt{c_{\text{scale}}R}}{\sqrt{\underline{c}}m^{\frac{1}{4}}} := \varepsilon_Q^{\text{NN}}.$$

Proof. Fix phase n and iteration t. For any $i = 0, \ldots, M_{\text{SGD}}^{\theta} - 1$, since $F^{\hat{r}^n}(\theta)$ is convex with respect to θ, we have

$$\begin{aligned} F^{\hat{r}^n}(\theta^{t,i}) - F^{\hat{r}^n}(\theta_{\text{mid}}^{t,\hat{r}^n}) &\le \nabla_\theta F^{\hat{r}^n}(\theta^{t,i})^\top \left(\theta^{t,i} - \theta_{\text{mid}}^{t,\hat{r}^n}\right) \\ &= (g^{t,i})^\top \left(\theta^{t,i} - \theta_{\text{mid}}^{t,\hat{r}^n}\right) + \left(\nabla_\theta F^{\hat{r}^n}(\theta^{t,i}) - g^{t,i}\right)^\top \left(\theta^{t,i} - \theta_{\text{mid}}^{t,\hat{r}^n}\right) \\ &= \frac{1}{\xi_\theta}(\theta^{t,i} - \tilde{\theta}^{t+1,i})^\top \left(\theta^{t,i} - \theta_{\text{mid}}^{t,\hat{r}^n}\right) + \left(\nabla_\theta F^{\hat{r}^n}(\theta^{t,i}) - g^{t,i}\right)^\top \left(\theta^{t,i} - \theta_{\text{mid}}^{t,\hat{r}^n}\right) \\ &= \frac{1}{2\xi_\theta}\left(\left\|\theta^{t,i} - \tilde{\theta}^{t,i+1}\right\|_2^2 + \left\|\theta^{t,i} - \theta_{\text{mid}}^{t,\hat{r}^n}\right\|_2^2 - \left\|\tilde{\theta}^{t,i+1} - \theta_{\text{mid}}^{t,\hat{r}^n}\right\|_2^2\right) \\ &\quad + \left(\nabla_\theta F^{\hat{r}^n}(\theta^{t,i}) - g^{t,i}\right)^\top \left(\theta^{t,i} - \theta_{\text{mid}}^{t,\hat{r}^n}\right) \\ &\le \frac{\xi_\theta}{2}\left\|g^{t,i}\right\|_2^2 + \frac{1}{2\xi_\theta}\left(\left\|\theta^{t,i} - \theta_{\text{mid}}^{t,\hat{r}^n}\right\|_2^2 - \left\|\theta^{t,i+1} - \theta_{\text{mid}}^{t,\hat{r}^n}\right\|_2^2\right) \\ &\quad + \left(\nabla_\theta F^{\hat{r}^n}(\theta^{t,i}) - g^{t,i}\right)^\top \left(\theta^{t,i} - \theta_{\text{mid}}^{t,\hat{r}^n}\right) \end{aligned}$$

Summing $i = 0, \ldots, M_{\text{SGD}}^{\theta} - 1$ and dividing M_{SGD}^{θ}, we have

$$F^{\hat{r}^n}(\theta^t) - F^{\hat{r}^n}(\theta_{\text{mid}}^{t,\hat{r}^n})$$

$$
\begin{aligned}
&= F\left(\frac{1}{M_{\text{SGD}}^{\theta}}\sum_{i=0}^{M_{\text{SGD}}^{\theta}-1}\theta^{t,i}\right) - F^{\hat{r}^n}(\theta_{\text{mid}}^{t,\hat{r}^n}) \\
&\overset{\text{(a)}}{\leq} \frac{1}{M_{\text{SGD}}^{\theta}}\sum_{i=0}^{M_{\text{SGD}}^{\theta}-1} F^{\hat{r}^n}(\theta^{t,i}) - F^{\hat{r}^n}(\theta_{\text{mid}}^{t,\hat{r}^n}) \\
&\leq \frac{\xi_\theta}{2M_{\text{SGD}}^{\theta}}\sum_{i=0}^{M_{\text{SGD}}^{\theta}-1}\left\|g^{t,i}\right\|_2^2 + \frac{1}{2\xi_\theta M_{\text{SGD}}^{\theta}}\left(\left\|\theta^{t,0}-\theta_{\text{mid}}^{t,\hat{r}^n}\right\|_2^2 - \left\|\theta^{t,M_{\text{SGD}}^{\theta}}-\theta_{\text{mid}}^{t,\hat{r}^n}\right\|_2^2\right) \\
&\quad + \frac{1}{M_{\text{SGD}}^{\theta}}\sum_{i=0}^{M_{\text{SGD}}^{\theta}-1}\left(\nabla_\theta F^{\hat{r}^n}(\theta^{t,i}) - \bar{g}^{t,i} + \bar{g}^{t,i} - g^{t,i}\right)^\top\left(\theta^{t,i}-\theta_{\text{mid}}^{t,\hat{r}^n}\right) \\
&\leq \frac{\xi_\theta}{2M_{\text{SGD}}^{\theta}}\sum_{i=0}^{M_{\text{SGD}}^{\theta}-1}\left\|g^{t,i}\right\|_2^2 + \frac{R^2}{2\xi_\theta M_{\text{SGD}}^{\theta}} + \frac{1}{M_{\text{SGD}}^{\theta}}\sum_{i=0}^{M_{\text{SGD}}^{\theta}-1}\left(\bar{g}^{t,i}-g^{t,i}\right)^\top\left(\theta^{t,i}-\theta_{\text{mid}}^{t,\hat{r}^n}\right) \\
&\quad + \frac{2R}{M_{\text{SGD}}^{\theta}}\sum_{i=0}^{M_{\text{SGD}}^{\theta}-1}\left\|\nabla_\theta F^{\hat{r}^n}(\theta^{t,i}) - \bar{g}^{t,i}\right\|_2,
\end{aligned} \tag{28}
$$

where inequality (a) uses the Jensen inequality.

For any $i \geq 0$, let $\mathcal{H}_i$ be all histories of steps $0, \ldots, i$, and we make the convention that $\mathcal{H}_{i-1} = \emptyset$ for $i = 0$. Let $\mathbb{E}_i[\cdot|\mathcal{H}_{i-1}]$ denote the expectation with respect to the randomness at step i conditioning on all histories of steps $0, \ldots, i-1$. Then, for any $i \geq 0$, we have $\mathbb{E}_i[\nabla_\theta \hat{F}^i(\theta^{t,i})^\top\left(\theta^{t,i}-\theta_{\text{mid}}^{t,\hat{r}^n}\right)|\mathcal{H}_{i-1}] = \nabla_\theta F^{\hat{r}^n}(\theta^{t,i})^\top\left(\theta^{t,i}-\theta_{\text{mid}}^{t,\hat{r}^n}\right)$.

According to the definition of event $\mathcal{E}_\theta^{\text{NN}}$, we have

$$
\begin{aligned}
&\left|\sum_{i=0}^{M_{\text{SGD}}^{\theta}-1}(g^{t,i})^\top\left(\theta^{t,i}-\theta_{\text{mid}}^{t,\hat{r}^n}\right) - \sum_{i=0}^{M_{\text{SGD}}^{\theta}-1}(\bar{g}^{t,i})^\top\left(\theta^{t,i}-\theta_{\text{mid}}^{t,\hat{r}^n}\right)\right| \\
&\leq 2W_{\nabla F}^{\text{NN}} R\sqrt{M_{\text{SGD}}^{\theta}\log\left(\frac{1}{\delta'}\right)}.
\end{aligned} \tag{29}
$$

Then, we have

$$
\begin{aligned}
&\left\|\nabla_\theta F^{\hat{r}^n}(\theta^{t,i}) - \bar{g}^{t,i}\right\|_2 \\
&= \Big\|\mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n}\Big[2\left(f_0(s,a;\theta^{t,i}) - \left(Q^{\pi^t}(s,a;\hat{r}^n+b^n) - b^n(s,a)\right)\right)\nabla_\theta f_0(s,a;\theta^{t,i}) \\
&\quad - 2\left(f(s,a;\theta^{t,i}) - \left(Q^{\pi^t}(s,a;\hat{r}^n+b^n) - b^n(s,a)\right)\right)\nabla_\theta f(s,a;\theta^{t,i})\Big]\Big\|_2 \\
&\leq 2\mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n}\Big[\left\|f_0(s,a;\theta^{t,i})\nabla_\theta f_0(s,a;\theta^{t,i}) - f(s,a;\theta^{t,i})\nabla_\theta f(s,a;\theta^{t,i})\right\|_2 \\
&\quad + W_Q^{\text{NN}}\left\|\nabla_\theta f_0(s,a;\theta^{t,i}) - \nabla_\theta f(s,a;\theta^{t,i})\right\|_2\Big] \\
&\leq 2\mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n}\Big[\left\|f_0(s,a;\theta^{t,i})\nabla_\theta f_0(s,a;\theta^{t,i}) - f_0(s,a;\theta^{t,i})\nabla_\theta f(s,a;\theta^{t,i})\right\|_2 \\
&\quad + \left\|f_0(s,a;\theta^{t,i})\nabla_\theta f(s,a;\theta^{t,i}) - f(s,a;\theta^{t,i})\nabla_\theta f(s,a;\theta^{t,i})\right\|_2 \\
&\quad + W_Q^{\text{NN}}\left\|\nabla_\theta f_0(s,a;\theta^{t,i}) - \nabla_\theta f(s,a;\theta^{t,i})\right\|_2\Big] \\
&\leq 2\mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n}\Big[\left|f_0(s,a;\theta^{t,i}) - f(s,a;\theta^{t,i})\right| \\
&\quad + (W_f^{\text{NN}} + W_Q^{\text{NN}})\left\|\psi_0(s,a) - \psi_{\theta^{t,i}}(s,a)\right\|_2\Big].
\end{aligned} \tag{30}
$$

Plugging Eqs. (29) and (30) into Eq. (28), we have

$$\begin{aligned}
F^{\hat{r}^n}(\theta^t) - F^{\hat{r}^n}(\theta^{t,\hat{r}^n}_{\text{mid}}) &\le 4W^{\text{NN}}_{\nabla F} R\sqrt{\frac{\log\left(\frac{1}{\delta'}\right)}{M^\theta_{\text{SGD}}}} + \frac{4R}{M^\theta_{\text{SGD}}} \sum_{i=0}^{M^\theta_{\text{SGD}}-1} \mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\Big[\,\big|f_0(s,a;\theta^{t,i}) - f(s,a;\theta^{t,i})\big| \\
&\quad + (W^{\text{NN}}_f + W^{\text{NN}}_Q)\,\|\psi_0(s,a) - \psi_{\theta^{t,i}}(s,a)\|_2\,\Big] \\
&\overset{\text{(a)}}{\le} 4W^{\text{NN}}_{\nabla F} R\sqrt{\frac{\log\left(\frac{1}{\delta'}\right)}{M^\theta_{\text{SGD}}}} + 4R\left(\frac{2\sqrt{c_{\text{scale}}R^3}}{\sqrt{\underline{c}}m^{\frac{1}{4}}} + \frac{(W^{\text{NN}}_f + W^{\text{NN}}_Q)\sqrt{c_{\text{scale}}R}}{\sqrt{\underline{c}}m^{\frac{1}{4}}}\right) \\
&\le 4W^{\text{NN}}_{\nabla F} R\sqrt{\frac{\log\left(\frac{1}{\delta'}\right)}{M^\theta_{\text{SGD}}}} + \frac{12R^2(W^{\text{NN}}_f + W^{\text{NN}}_Q)\sqrt{c_{\text{scale}}R}}{\sqrt{\underline{c}}m^{\frac{1}{4}}},
\end{aligned}$$

where inequality (a) uses Assumption 3.3.

□

E.4. Human Feedback

Recall that for any trajectories $\tau^{(1)}, \tau^{(2)}$ and $\mu \in \mathcal{U}_R$, let $\tilde{\psi}^{\tau^{(1)},\tau^{(2)}}_\mu := \sum_{h=0}^{H(\tau^{(1)})} \psi_\mu(s^{(1)}_h, a^{(1)}_h) - \sum_{h=0}^{H(\tau^{(2)})} \psi_\mu(s^{(2)}_h, a^{(2)}_h)$, $\tilde{h}(\tau^{(1)}, \tau^{(2)}; \mu) := \sum_{h=0}^{H(\tau^{(1)})} h(s^{(1)}_h, a^{(1)}_h; \mu) - \sum_{h=0}^{H(\tau^{(2)})} h(s^{(2)}_h, a^{(2)}_h; \mu)$ and $\tilde{r}(\tau^{(1)}, \tau^{(2)}) := \sum_{h=0}^{H(\tau^{(1)})} r(s^{(1)}_h, a^{(1)}_h) - \sum_{h=0}^{H(\tau^{(2)})} r(s^{(2)}_h, a^{(2)}_h)$.

For any fixed phase $n = 0, \ldots, N-1$, define the approximated MLE objective function and its optimal solution as follows:

$$\begin{aligned}
L(\mu) &:= \frac{1}{M_{\text{HF}}} \sum_{i=1}^{M_{\text{HF}}} \Bigg(-\log\Bigg(\frac{\mathbb{1}\{y_i = 1\}}{1 + \exp\left(\sum_{h=0}^{H(\tau^{(2)}_i)} f_0(s^{(2)}_{i,h}, a^{(2)}_{i,h}; \mu) - \sum_{h=0}^{H(\tau^{(1)}_i)} f_0(s^{(1)}_{i,h}, a^{(1)}_{i,h}; \mu)\right)} \\
&\qquad + \frac{\mathbb{1}\{y_i = 0\}}{1 + \exp\left(\sum_{h=0}^{H(\tau^{(1)}_i)} f_0(s^{(1)}_{i,h}, a^{(1)}_{i,h}; \mu) - \sum_{h=0}^{H(\tau^{(2)}_i)} f_0(s^{(2)}_{i,h}, a^{(2)}_{i,h}; \mu)\right)} \Bigg)\Bigg) \\
&= \frac{1}{M_{\text{HF}}} \sum_{i=1}^{M_{\text{HF}}} \left(-\log\left(\frac{\mathbb{1}\{y_i = 1\}}{1 + \exp\left(-(\tilde{\psi}_0^{\tau^{(1)}_i,\tau^{(2)}_i})^\top \mu\right)} + \frac{\mathbb{1}\{y_i = 0\}}{1 + \exp\left((\tilde{\psi}_0^{\tau^{(1)}_i,\tau^{(2)}_i})^\top \mu\right)} \right)\right), \\
\mu^*_{\text{MLE}} &:= \operatorname*{argmin}_{\mu \in \mathcal{U}_R} L(\mu).
\end{aligned}$$

Then, it holds that

$$\begin{aligned}
\nabla_\mu L(\mu) &= \frac{1}{M_{\text{HF}}} \sum_{i=1}^{M_{\text{HF}}} \Bigg(\underbrace{\left(-\frac{\mathbb{1}\{y_i = 1\} \exp\left(-(\tilde{\psi}_0^{\tau^{(1)}_i,\tau^{(2)}_i})^\top \mu\right)}{1 + \exp\left(-(\tilde{\psi}_0^{\tau^{(1)}_i,\tau^{(2)}_i})^\top \mu\right)} + \frac{\mathbb{1}\{y_i = 0\}}{1 + \exp\left((\tilde{\psi}_0^{\tau^{(1)}_i,\tau^{(2)}_i})^\top \mu\right)} \right)}_{:= q^i_0(\mu)} \tilde{\psi}_0^{\tau^{(1)}_i,\tau^{(2)}_i} \Bigg), \\
\nabla^2_\mu L(\mu) &= \frac{1}{M_{\text{HF}}} \sum_{i=1}^{M_{\text{HF}}} \left(\left(\frac{\mathbb{1}\{y_i = 1\} \exp\left(-(\tilde{\psi}_0^{\tau^{(1)}_i,\tau^{(2)}_i})^\top \mu\right)}{\left(1 + \exp\left(-(\tilde{\psi}_0^{\tau^{(1)}_i,\tau^{(2)}_i})^\top \mu\right)\right)^2} + \frac{\mathbb{1}\{y_i = 0\}\, (\tilde{\psi}_0^{\tau^{(1)}_i,\tau^{(2)}_i})^\top \mu}{\left(1 + \exp\left((\tilde{\psi}_0^{\tau^{(1)}_i,\tau^{(2)}_i})^\top \mu\right)\right)^2} \right) \tilde{\psi}_0^{\tau^{(1)}_i,\tau^{(2)}_i} (\tilde{\psi}_0^{\tau^{(1)}_i,\tau^{(2)}_i})^\top \right).
\end{aligned}$$

Algorithm 6 Reward Network Training via Projected SGD (with the objective Eq. (9))

Input: $h(s,a;\mu^0)$, ξ_μ.
for $j=0,\dots,M_{\text{SGD}}^{\mu}-1$ **do**
  $z^j \leftarrow \left(-\frac{\mathbb{1}\{y_j=1\}\exp\left(-\tilde{h}(\tau_j^{(1)},\tau_j^{(2)};\mu^j)\right)}{1+\exp\left(-\tilde{h}(\tau_j^{(1)},\tau_j^{(2)};\mu^j)\right)}+\frac{\mathbb{1}\{y_j=0\}}{1+\exp\left(-\tilde{h}(\tau_j^{(1)},\tau_j^{(2)};\mu^j)\right)}\right)\nabla_\mu\tilde{h}(\tau_j^{(1)},\tau_j^{(2)};\mu^j)$
  $\tilde{\mu}^{j+1}\leftarrow\mu^j-\xi_\mu z^j$
  $\mu^{i+1}\leftarrow\text{Proj}_{\mathcal{U}_R}(\tilde{\mu}^{j+1})$
end for
return $\hat{\mu}^n=\sum_{j=0}^{M_{\text{SGD}}^{\mu}-1}\mu^j$

For any $j=0,\dots,M_{\text{SGD}}^{\mu}-1$, define

$$z^j:=\left(-\frac{\mathbb{1}\{y_j=1\}\exp\left(-\tilde{h}(\tau_j^{(1)},\tau_j^{(2)};\mu^j)\right)}{1+\exp\left(-\tilde{h}(\tau_j^{(1)},\tau_j^{(2)};\mu^j)\right)}+\frac{\mathbb{1}\{y_j=0\}}{1+\exp\left(-\tilde{h}(\tau_j^{(1)},\tau_j^{(2)};\mu^j)\right)}\right)\nabla_\mu\tilde{h}(\tau_j^{(1)},\tau_j^{(2)};\mu^j),$$

$$\bar{z}^j:=\frac{1}{M_{\text{HF}}}\sum_{i=1}^{M_{\text{HF}}}\Bigg(\underbrace{\left(-\frac{\mathbb{1}\{y_i=1\}\exp\left(-\tilde{h}(\tau_i^{(1)},\tau_i^{(2)};\mu^j)\right)}{1+\exp\left(-\tilde{h}(\tau_i^{(1)},\tau_i^{(2)};\mu^j)\right)}+\frac{\mathbb{1}\{y_i=0\}}{1+\exp\left(-\tilde{h}(\tau_i^{(1)},\tau_i^{(2)};\mu^j)\right)}\right)}_{:=q^i(\mu^j)}\nabla_\mu\tilde{h}(\tau_i^{(1)},\tau_i^{(2)};\mu^j)\Bigg),$$

$$\tilde{\mu}^{j+1}:=\mu^j-\xi_\mu z^j,$$

where $(\tau_j^{(1)},\tau_j^{(2)},y_j)$ is uniformly drawn from $\{(\tau_i^{(1)},\tau_i^{(2)},y_i)\}_{i=1}^{M_{\text{HF}}}$.

Then, we have

$$\mu^{j+1}=\text{Proj}_{\mathcal{U}_R}(\tilde{\mu}^{j+1}).$$

Define event

$$\mathcal{E}_\mu^{\text{NN}}:=\Bigg\{\Bigg|\sum_{j=0}^{M_{\text{SGD}}^{\mu}-1}\nabla_\mu\hat{L}^j(\mu^j)^\top\left(\mu^j-\mu_{\text{MLE}}^*\right)-\sum_{j=0}^{M_{\text{SGD}}^{\mu}-1}\nabla_\mu L(\mu^j)^\top\left(\mu^j-\mu_{\text{MLE}}^*\right)\Bigg|$$
$$\leq 8W_\tau R\sqrt{M_{\text{SGD}}^{\mu}\log\left(\frac{1}{\delta'}\right)},\ \forall 0\leq n\leq N-1,\forall 0\leq j\leq M_{\text{SGD}}^{\mu}-1\Bigg\}.$$

Lemma E.10. *It holds that* $\Pr[\mathcal{E}_\mu^{\text{NN}}]\geq 1-2N\delta'$.

Proof. This lemma can be obtained by applying the Azuma-Hoeffding inequality and the union bound. □

Let $\xi_\mu:=\frac{R}{W_\tau\sqrt{M_{\text{SGD}}^{\mu}}}$.

Below we provide the guarantee for the projected SGD of reward training, which is illustrated in algorithm 6.

Lemma E.11 (SGD for the Reward Model)**.** *Assume that event* $\mathcal{E}_{\text{init}}\cap\mathcal{E}_\tau\cap\mathcal{E}_\mu^{\text{NN}}$ *holds. Then, for any phase* n*,*

$$L(\mu^n)-L(\mu_{\text{MLE}}^*)$$
$$\leq 17W_\tau R\sqrt{\frac{\log\left(\frac{1}{\delta'}\right)}{M_{\text{SGD}}^{\mu}}}+\frac{2R}{M_{\text{SGD}}^{\mu}}\sum_{j=0}^{M_{\text{SGD}}^{\mu}-1}\Bigg(\frac{1}{M_{\text{HF}}}\sum_{i=1}^{M_{\text{HF}}}\Bigg(2\left\|\tilde{\psi}_0^{\tau_i^{(1)},\tau_i^{(2)}}-\tilde{\psi}_{\mu^j}^{\tau_i^{(1)},\tau_i^{(2)}}\right\|_2$$
$$+4W_\tau\left|\tilde{h}_0(\tau_i^{(1)},\tau_i^{(2)};\mu^j)-\tilde{h}(\tau_i^{(1)},\tau_i^{(2)};\mu^j)\right|\Bigg)\Bigg):=\varepsilon_{\text{SGD}}^{\text{NN},n}.$$

Furthermore,

$$\mathbb{E}_{\substack{\{\tau_i^{(1)}\}_{i=1}^{M_{\text{HF}}}\sim\mathcal{O}_{\text{HF}}^n \\ \{\tau_i^{(2)}\}_{i=1}^{M_{\text{HF}}}\sim\mathcal{O}_{s_{\text{init}}}^{\pi^{\text{base}}}}}\left[\varepsilon_{\text{SGD}}^{\text{NN},n}\right] \le 17W_\tau R\sqrt{\frac{\log\left(\frac{1}{\delta'}\right)}{M_{\text{SGD}}^\mu}} + \frac{40R^2W_\tau\sqrt{c_{\text{scale}}\overline{R}}}{(1-\gamma)\sqrt{\underline{c}}m^{\frac{1}{4}}}.$$

Proof. For any $j = 0, \ldots, M_{\text{SGD}}^\mu - 1$,

$$\begin{aligned}
L(\mu^j) - L(\mu_{\text{MLE}}^*) &\le \nabla_\mu L(\mu^j)^\top \left(\mu^j - \mu_{\text{MLE}}^*\right) \\
&= (z^j)^\top \left(\mu^j - \mu_{\text{MLE}}^*\right) + \left(\nabla_\mu L(\mu^j) - z^j\right)^\top \left(\mu^j - \mu_{\text{MLE}}^*\right) \\
&= \frac{1}{\xi_\mu}\left(\mu^j - \tilde{\mu}^{j+1}\right)^\top \left(\mu^j - \mu_{\text{MLE}}^*\right) + \left(\nabla_\mu L(\mu^j) - z^j\right)^\top \left(\mu^j - \mu_{\text{MLE}}^*\right) \\
&= \frac{1}{2\xi_\mu}\left(\left\|\mu^j - \tilde{\mu}^{j+1}\right\|_2^2 + \left\|\mu^j - \mu_{\text{MLE}}^*\right\|_2^2 - \left\|\tilde{\mu}^{j+1} - \mu_{\text{MLE}}^*\right\|_2^2\right) \\
&\quad + \left(\nabla_\mu L(\mu^j) - z^j\right)^\top \left(\mu^j - \mu_{\text{MLE}}^*\right) \\
&\le \frac{\xi_\mu}{2}\left\|z^j\right\|_2^2 + \frac{1}{2\xi_\mu}\left(\left\|\mu^j - \mu_{\text{MLE}}^*\right\|_2^2 - \left\|\mu^{j+1} - \mu_{\text{MLE}}^*\right\|_2^2\right) \\
&\quad + \left(\nabla_\mu L(\mu^j) - z^j\right)^\top \left(\mu^j - \mu_{\text{MLE}}^*\right).
\end{aligned}$$

Summing $j = 0, \ldots, M_{\text{SGD}}^\mu - 1$ and dividing M_{SGD}^μ, we have

$$\begin{aligned}
&L(\mu^n) - L(\mu_{\text{MLE}}^*) \\
&= L\left(\frac{1}{M_{\text{SGD}}^\mu}\sum_{j=0}^{M_{\text{SGD}}^\mu-1}\mu^j\right) - L(\mu_{\text{MLE}}^*) \\
&\overset{\text{(a)}}{\le} \frac{1}{M_{\text{SGD}}^\mu}\sum_{j=0}^{M_{\text{SGD}}^\mu-1}\left(L(\mu^j) - L(\mu_{\text{MLE}}^*)\right) \\
&\le \frac{\xi_\mu}{2M_{\text{SGD}}^\mu}\sum_{j=0}^{M_{\text{SGD}}^\mu-1}\left\|z^j\right\|_2^2 + \frac{1}{2\xi_\mu M_{\text{SGD}}^\mu}\left(\left\|\mu^0 - \mu_{\text{MLE}}^*\right\|_2^2 - \left\|\mu^{M_{\text{SGD}}^\mu} - \mu_{\text{MLE}}^*\right\|_2^2\right) \\
&\quad + \frac{1}{M_{\text{SGD}}^\mu}\sum_{j=0}^{M_{\text{SGD}}^\mu-1}\left(\nabla_\mu L(\mu^j) - \bar{z}^j + \bar{z}^j - z^j\right)^\top \left(\mu^j - \mu_{\text{MLE}}^*\right) \\
&\le \frac{\xi_\mu}{2M_{\text{SGD}}^\mu}\sum_{j=0}^{M_{\text{SGD}}^\mu-1}\left\|z^j\right\|_2^2 + \frac{R^2}{2\xi_\mu M_{\text{SGD}}^\mu} + \frac{1}{M_{\text{SGD}}^\mu}\sum_{j=0}^{M_{\text{SGD}}^\mu-1}\left(\bar{z}^j - z^j\right)^\top \left(\mu^j - \mu_{\text{MLE}}^*\right) \\
&\quad + \frac{2R}{M_{\text{SGD}}^\mu}\sum_{j=0}^{M_{\text{SGD}}^\mu-1}\left\|\nabla_\mu L(\mu^j) - \bar{z}^j\right\|_2,
\end{aligned} \tag{31}$$

where inequality (a) uses the Jensen inequality.

We have

$$\begin{aligned}
\left\|z^j\right\|_2 &\le 2\left\|\tilde{\psi}_{\mu^j}^{\tau_j^{(1)},\tau_j^{(2)}}\right\|_2 \\
&= 2\left\|\sum_{h=0}^{H(\tau_j^{(1)})}\psi_{\mu^j}(s_{j,h}^{(1)}, a_{j,h}^{(1)}) - \sum_{h=0}^{H(\tau_j^{(2)})}\psi_{\mu^j}(s_{j,h}^{(2)}, a_{j,h}^{(2)})\right\|_2 \\
&\le 4W_\tau.
\end{aligned} \tag{32}$$

For any $j \geq 0$, let $\mathcal{H}_j$ be all histories of steps $0, \ldots, j$, and we make the convention that $\mathcal{H}_{j-1} = \emptyset$ for $j = 0$. Let $\mathbb{E}_j[\cdot|\mathcal{H}_{j-1}]$ denote the expectation with respect to the randomness at step j (i.e., $(\tau_j^{(1)}, \tau_j^{(2)}, y_j) \sim \text{Unif}(\{(\tau_i^{(1)}, \tau_i^{(2)}, y_i)\}_{i=1}^{M_{\text{HF}}})$ conditioning on all histories of steps $0, \ldots, j-1$. Then, for any $j \geq 0$, we have $\mathbb{E}_j[\nabla_\mu \hat{L}^j(\mu^j)^\top \left(\mu^j - \mu^*_{\text{MLE}}\right)|\mathcal{H}_{j-1}] = \nabla_\mu L(\mu^j)^\top \left(\mu^j - \mu^*_{\text{MLE}}\right)$.

According to the definition of $\mathcal{E}_\mu^{\text{NN}}$, we have

$$\begin{aligned}
&\left| \sum_{j=0}^{M^\mu_{\text{SGD}}-1} (z^j)^\top \left(\mu^j - \mu^*_{\text{MLE}}\right) - \sum_{j=0}^{M^\mu_{\text{SGD}}-1} (\bar{z}^j)^\top \left(\mu^j - \mu^*_{\text{MLE}}\right) \right| \\
&\leq 8W_\tau R \sqrt{M^\mu_{\text{SGD}} \log\left(\frac{1}{\delta'}\right)}. \qquad (33)
\end{aligned}$$

We have

$$\begin{aligned}
\left|q_0^i(\mu^j) - q^i(\mu^j)\right| \leq &\left| - \frac{\mathbb{1}\{y_i = 1\} \exp\left(-(\tilde{\psi}_0^{\tau_i^{(1)}, \tau_i^{(2)}})^\top \mu^j\right)}{1 + \exp\left(-(\tilde{\psi}_0^{\tau_i^{(1)}, \tau_i^{(2)}})^\top \mu^j\right)} + \frac{\mathbb{1}\{y_i = 0\}}{1 + \exp\left(-(\tilde{\psi}_0^{\tau_i^{(1)}, \tau_i^{(2)}})^\top \mu^j\right)} \right. \\
&\left. + \frac{\mathbb{1}\{y_i = 1\} \exp\left(-\tilde{h}(\tau_i^{(1)}, \tau_i^{(2)}; \mu^j)\right)}{1 + \exp\left(-\tilde{h}(\tau_i^{(1)}, \tau_i^{(2)}; \mu^j)\right)} - \frac{\mathbb{1}\{y_i = 0\}}{1 + \exp\left(-\tilde{h}(\tau_i^{(1)}, \tau_i^{(2)}; \mu^j)\right)} \right| \\
\leq &\left| \frac{\exp\left(-\tilde{h}(\tau_i^{(1)}, \tau_i^{(2)}; \mu^j)\right)}{1 + \exp\left(-\tilde{h}(\tau_i^{(1)}, \tau_i^{(2)}; \mu^j)\right)} - \frac{\exp\left(-(\tilde{\psi}_0^{\tau_i^{(1)}, \tau_i^{(2)}})^\top \mu^j\right)}{1 + \exp\left(-(\tilde{\psi}_0^{\tau_i^{(1)}, \tau_i^{(2)}})^\top \mu^j\right)} \right| \\
&+ \left| \frac{1}{1 + \exp\left(-(\tilde{\psi}_0^{\tau_i^{(1)}, \tau_i^{(2)}})^\top \mu^j\right)} - - \frac{1}{1 + \exp\left(-\tilde{h}(\tau_i^{(1)}, \tau_i^{(2)}; \mu^j)\right)} \right| \\
\overset{\text{(a)}}{\leq} &2 \left|\tilde{h}_0(\tau_i^{(1)}, \tau_i^{(2)}; \mu^j) - \tilde{h}(\tau_i^{(1)}, \tau_i^{(2)}; \mu^j)\right|,
\end{aligned}$$

where inequality (a) uses the fact that the derivative of functions $\frac{\exp(x)}{1+\exp(x)}$ and $\frac{1}{1+\exp(x)}$ lies in $(0, 1)$.

Then, it holds that

$$\begin{aligned}
&\left\|\nabla_\mu L(\mu^j) - \bar{z}^j\right\|_2 \qquad (34) \\
&= \left\| \frac{1}{M_{\text{HF}}} \sum_{i=1}^{M_{\text{HF}}} \left(q_0^i(\mu^j) \tilde{\psi}_0^{\tau_i^{(1)}, \tau_i^{(2)}} - q^i(\mu^j) \tilde{\psi}_{\mu^j}^{\tau_i^{(1)}, \tau_i^{(2)}} \right) \right\|_2 \\
&\leq \frac{1}{M_{\text{HF}}} \sum_{i=1}^{M_{\text{HF}}} \left\| q_0^i(\mu^j) \tilde{\psi}_0^{\tau_i^{(1)}, \tau_i^{(2)}} - q_0^i(\mu^j) \tilde{\psi}_{\mu^j}^{\tau_i^{(1)}, \tau_i^{(2)}} + q_0^i(\mu^j) \tilde{\psi}_{\mu^j}^{\tau_i^{(1)}, \tau_i^{(2)}} - q^i(\mu^j) \tilde{\psi}_{\mu^j}^{\tau_i^{(1)}, \tau_i^{(2)}} \right\|_2 \\
&\leq \frac{1}{M_{\text{HF}}} \sum_{i=1}^{M_{\text{HF}}} \left(2 \left\| \tilde{\psi}_0^{\tau_i^{(1)}, \tau_i^{(2)}} - \tilde{\psi}_{\mu^j}^{\tau_i^{(1)}, \tau_i^{(2)}} \right\|_2 + 2W_\tau \left| q_0^i(\mu^j) - q^i(\mu^j) \right| \right) \\
&\leq \frac{1}{M_{\text{HF}}} \sum_{i=1}^{M_{\text{HF}}} \left(2 \left\| \tilde{\psi}_0^{\tau_i^{(1)}, \tau_i^{(2)}} - \tilde{\psi}_{\mu^j}^{\tau_i^{(1)}, \tau_i^{(2)}} \right\|_2 + 4W_\tau \left| \tilde{h}_0(\tau_i^{(1)}, \tau_i^{(2)}; \mu^j) - \tilde{h}(\tau_i^{(1)}, \tau_i^{(2)}; \mu^j) \right| \right). \qquad (35)
\end{aligned}$$

Plugging Eqs. (32)-(35) into Eq. (31), we have

$$L(\mu^n) - L(\mu^*_{\text{MLE}})$$

$$
\begin{aligned}
&\leq 8\xi_\mu W_\tau^2 + \frac{R^2}{2\xi_\mu M_{\text{SGD}}^\mu} + 8W_\tau R\sqrt{\frac{\log\left(\frac{1}{\delta'}\right)}{M_{\text{SGD}}^\mu}} \\
&\quad + \frac{2R}{M_{\text{SGD}}^\mu}\sum_{j=0}^{M_{\text{SGD}}^\mu-1}\left(\frac{1}{M_{\text{HF}}}\sum_{i=1}^{M_{\text{HF}}}\left(2\left\|\tilde{\psi}_0^{\tau_i^{(1)},\tau_i^{(2)}} - \tilde{\psi}_{\mu^j}^{\tau_i^{(1)},\tau_i^{(2)}}\right\|_2\right.\right. \\
&\quad \left.\left. + 4W_\tau\left|\tilde{h}_0(\tau_i^{(1)},\tau_i^{(2)};\mu^j) - \tilde{h}(\tau_i^{(1)},\tau_i^{(2)};\mu^j)\right|\right)\right) \\
&\leq 17W_\tau R\sqrt{\frac{\log\left(\frac{1}{\delta'}\right)}{M_{\text{SGD}}^\mu}} + \frac{2R}{M_{\text{SGD}}^\mu}\sum_{j=0}^{M_{\text{SGD}}^\mu-1}\left(\frac{1}{M_{\text{HF}}}\sum_{i=1}^{M_{\text{HF}}}\left(2\left\|\tilde{\psi}_0^{\tau_i^{(1)},\tau_i^{(2)}} - \tilde{\psi}_{\mu^j}^{\tau_i^{(1)},\tau_i^{(2)}}\right\|_2\right.\right. \\
&\quad \left.\left. + 4W_\tau\left|\tilde{h}_0(\tau_i^{(1)},\tau_i^{(2)};\mu^j) - \tilde{h}(\tau_i^{(1)},\tau_i^{(2)};\mu^j)\right|\right)\right) := \varepsilon_{\text{SGD}}^{\text{NN},n}.
\end{aligned}
$$

In addition, we have

$$
\begin{aligned}
&\mathbb{E}_{\{\tau_i^{(1)}\}_{i=1}^{M_{\text{HF}}}\sim\mathcal{O}_{\text{HF}}^n,\ \{\tau_i^{(2)}\}_{i=1}^{M_{\text{HF}}}\sim\mathcal{O}^{\pi^{\text{base}}}_{s_{\text{init}}}}\left[\varepsilon_{\text{SGD}}^{\text{NN},n}\right] \\
&\leq 17W_\tau R\sqrt{\frac{\log\left(\frac{1}{\delta'}\right)}{M_{\text{SGD}}^\mu}} + 2R\left(\mathbb{E}_{\substack{\tau^{(1)}\sim\mathcal{O}_{\text{HF}}^n \\ \tau^{(2)}\sim\mathcal{O}^{\pi^{\text{base}}}_{s_{\text{init}}}}}\left[2\left(\sum_{h=0}^{H(\tau^{(1)})}\left\|\psi_0(s_h^{(1)},a_h^{(1)}) - \psi_{\mu^j}(s_h^{(1)},a_h^{(1)})\right\|_2\right.\right.\right. \\
&\quad + \sum_{h=0}^{H(\tau^{(2)})}\left\|\psi_0(s_h^{(2)},a_h^{(2)}) - \psi_{\mu^j}(s_h^{(2)},a_h^{(2)})\right\|_2\Bigg) + 4W_\tau\Bigg(\sum_{h=0}^{H(\tau^{(1)})}\left|h_0(s_h^{(1)},a_h^{(1)};\mu^j) - h(s_h^{(1)},a_h^{(1)};\mu^j)\right| \\
&\quad \left.\left.\left. + \sum_{h=0}^{H(\tau^{(2)})}\left|h_0(s_h^{(2)},a_h^{(2)};\mu^j) - h(s_h^{(2)},a_h^{(2)};\mu^j)\right|\Bigg)\right]\right) \\
&\leq 17W_\tau R\sqrt{\frac{\log\left(\frac{1}{\delta'}\right)}{M_{\text{SGD}}^\mu}} + \frac{2R}{1-\gamma}\left(\mathbb{E}_{\substack{s^{(1)}\sim d_{\text{HF}}^n \\ a^{(2)}\sim d_{\text{base}}}}\left[2\Big(\left\|\psi_0(s^{(1)},a^{(1)}) - \psi_{\mu^j}(s^{(1)},a^{(1)})\right\|_2\right.\right. \\
&\quad + \left\|\psi_0(s^{(2)},a^{(2)}) - \psi_{\mu^j}(s^{(2)},a^{(2)})\right\|_2\Big) + 4W_\tau\Big(\left|h_0(s^{(1)},a^{(1)};\mu^j) - h(s^{(1)},a^{(1)};\mu^j)\right| \\
&\quad \left.\left. + \left|h_0(s^{(2)},a^{(2)};\mu^j) - h(s^{(2)},a^{(2)};\mu^j)\right|\Big)\right]\right) \\
&\overset{\text{(a)}}{\leq} 17W_\tau R\sqrt{\frac{\log\left(\frac{1}{\delta'}\right)}{M_{\text{SGD}}^\mu}} + \frac{2R}{1-\gamma}\left(\frac{4\sqrt{c_{\text{scale}}R}}{\sqrt{\underline{c}}m^{\frac{1}{4}}} + \frac{16W_\tau\sqrt{c_{\text{scale}}R^3}}{\sqrt{\underline{c}}m^{\frac{1}{4}}}\right) \\
&\leq 17W_\tau R\sqrt{\frac{\log\left(\frac{1}{\delta'}\right)}{M_{\text{SGD}}^\mu}} + \frac{40R^2W_\tau\sqrt{c_{\text{scale}}R}}{(1-\gamma)\sqrt{\underline{c}}m^{\frac{1}{4}}},
\end{aligned}
$$

where inequality (a) uses Assumption 3.3. □

Let $c_{\text{MLE}}^{\text{NN}} := (2+\exp(-2W_\tau(\sqrt{m\bar{c}}+R)) + \exp(2W_\tau(\sqrt{m\bar{c}}+R)))^{-1}$.

Lemma E.12 (MLE). *Assume that event $\mathcal{E}_{\text{init}}\cap\mathcal{E}_\tau\cap\mathcal{E}_\mu^{\text{NN}}$ holds. Then, for any phase $n\geq 0$, we have that with probability at least $1-2\delta'$,*

$$
\left\|\mu^n - \mu_r^{\text{proj}}\right\|_{\hat{\Sigma}_{\text{HF}}^{\text{NN},n}} \leq \frac{1}{2c_{\text{MLE}}^{\text{NN}}}\sqrt{\frac{5md\log\left(\frac{1}{\delta'}\right)}{M_{\text{HF}}}} + \frac{3}{2c_{\text{MLE}}^{\text{NN}}}\sqrt{\sum_{i=1}^{M_{\text{HF}}}(\mathbb{E}_{y_i}[V_i])^2}\left(\frac{\log\left(\frac{1}{\delta'}\right)}{M_{\text{HF}}}\right)^{\frac{1}{4}}
$$

$$+\sqrt{\frac{\varepsilon_{\text{SGD}}^{\text{NN},n}}{c_{\text{MLE}}^{\text{NN}}}}+2R\sqrt{\frac{\zeta_{\text{HF}}}{n}}:=\varepsilon_{\text{MLE}}^{\text{NN},n}.$$

In other words, defining event

$$\mathcal{E}_{\text{MLE}}^{\text{NN}}:=\left\{\left\|\mu^n-\mu_r^{\text{proj}}\right\|_{\hat{\Sigma}_{\text{HF}}^{\text{NN},n}}\le\varepsilon_{\text{MLE}}^{\text{NN},n},\ \forall 0\le n\le N-1\right\},$$

we have $\Pr[\mathcal{E}_{\text{MLE}}^{\text{NN}}]\ge 1-2N\delta'$.

Furthermore, we have

$$\begin{aligned}\mathbb{E}_{\substack{\{\tau_i^{(1)}\}_{i=1}^{M_{\text{HF}}}\sim\mathcal{O}_{\text{HF}}^n\\ \{\tau_i^{(2)}\}_{i=1}^{M_{\text{HF}}}\sim\mathcal{O}^{\pi_{s_{\text{init}}}^{\text{base}}}}}\left[\varepsilon_{\text{MLE}}^{\text{NN},n}\right]\le&\frac{1}{2c_{\text{MLE}}^{\text{NN}}}\sqrt{\frac{5md\log\left(\frac{1}{\delta'}\right)}{M_{\text{HF}}}}+\frac{1}{\sqrt{c_{\text{MLE}}^{\text{NN}}}}\left(17W_\tau R\sqrt{\frac{\log\left(\frac{1}{\delta'}\right)}{M_{\text{SGD}}^{\mu}}}\right)^{\frac{1}{2}}\\&+2R\sqrt{\frac{\zeta_{\text{HF}}}{n}}+\frac{19c_{\text{scale}}^{\frac{1}{4}}R^{\frac{5}{4}}M_{\text{HF}}^{\frac{1}{4}}\sqrt{W_\tau\exp(4W_\tau)}}{\underline{c}^{\frac{1}{4}}c_{\text{MLE}}^{\text{NN}}\sqrt{1-\gamma}m^{\frac{1}{8}}}\log\left(\frac{1}{\delta'}\right).\end{aligned}$$

Here we make the convention that $\frac{\zeta_{\text{HF}}}{n}:=\zeta_{\text{HF}}$.

Proof. Since

$$\begin{aligned}\nabla_\mu^2L(\mu)=&\frac{1}{M_{\text{HF}}}\sum_{i=1}^{M_{\text{HF}}}\left(\left(\frac{\mathbb{1}\left\{y_i=1\right\}\exp\left(-(\tilde{\psi}_0^{\tau_i^{(1)},\tau_i^{(2)}})^\top\mu\right)}{\left(1+\exp\left(-(\tilde{\psi}_0^{\tau_i^{(1)},\tau_i^{(2)}})^\top\mu\right)\right)^2}+\frac{\mathbb{1}\left\{y_i=0\right\}(\tilde{\psi}_0^{\tau_i^{(1)},\tau_i^{(2)}})^\top\mu}{\left(1+\exp\left((\tilde{\psi}_0^{\tau_i^{(1)},\tau_i^{(2)}})^\top\mu\right)\right)^2}\right)\tilde{\psi}_0^{\tau_i^{(1)},\tau_i^{(2)}}(\tilde{\psi}_0^{\tau_i^{(1)},\tau_i^{(2)}})^\top\right)\\\succeq&\frac{c_{\text{MLE}}^{\text{NN}}}{M_{\text{HF}}}\sum_{i=1}^{M_{\text{HF}}}\tilde{\psi}_0^{\tau_i^{(1)},\tau_i^{(2)}}(\tilde{\psi}_0^{\tau_i^{(1)},\tau_i^{(2)}})^\top,\end{aligned}$$

we have that for any $\Delta\in\mathbb{R}^d$,

$$\begin{aligned}&L(\mu_r^{\text{proj}}+\Delta)-L(\mu_r^{\text{proj}})-(\mu_r^{\text{proj}})^\top\Delta\\\ge&\Delta^\top\nabla_\mu^2L(\mu_r^{\text{proj}})\Delta\\\ge&c_{\text{MLE}}^{\text{NN}}\Delta^\top\left(\frac{1}{M_{\text{HF}}}\sum_{i=1}^{M_{\text{HF}}}\left(\psi_0(\tau_i^{(1)})-\psi_0(\tau_i^{(2)})\right)\left(\psi_0(\tau_i^{(1)})-\psi_0(\tau_i^{(2)})\right)^\top\right)\Delta.\end{aligned}$$

Using Lemma E.11 and the definition of μ_{MLE}^*, we have

$$L(\mu^n)\le L(\mu_{\text{MLE}}^*)+\varepsilon_{\text{SGD}}^{\text{NN},n}\le L(\mu_r^{\text{proj}})+\varepsilon_{\text{SGD}}^{\text{NN},n}.$$

Then,

$$\begin{aligned}&c_{\text{MLE}}^{\text{NN}}\left(\mu^n-\mu_r^{\text{proj}}\right)^\top\left(\frac{1}{M_{\text{HF}}}\sum_{i=1}^{M_{\text{HF}}}\left(\psi_0(\tau_i^{(1)})-\psi_0(\tau_i^{(2)})\right)\left(\psi_0(\tau_i^{(1)})-\psi_0(\tau_i^{(2)})\right)^\top\right)\left(\mu^n-\mu_r^{\text{proj}}\right)\\\le&L(\mu^n)-L(\mu_r^{\text{proj}})-\nabla_\mu L(\mu_r^{\text{proj}})^\top\left(\mu^n-\mu_r^{\text{proj}}\right)\\\le&-\nabla_\mu L(\mu_r^{\text{proj}})^\top\left(\mu^n-\mu_r^{\text{proj}}\right)+\varepsilon_{\text{SGD}}^{\text{NN},n},\end{aligned}$$

which implies

$$c_{\text{MLE}}^{\text{NN}}\left\|\mu^n-\mu_r^{\text{proj}}\right\|_{\hat{\Sigma}_{\text{HF}}^{\text{NN},n}}^2\le\left\|\nabla_\mu L(\mu_r^{\text{proj}})\right\|_{\left(\hat{\Sigma}_{\text{HF}}^{\text{NN},n}\right)^{-1}}\left\|\mu^n-\mu_r^{\text{proj}}\right\|_{\hat{\Sigma}_{\text{HF}}^{\text{NN},n}}+\varepsilon_{\text{SGD}}^{\text{NN},n}+\frac{4c_{\text{MLE}}^{\text{NN}}\zeta_{\text{HF}}R^2}{n}.$$

By analysis for quadratic functions, we have

$$\left\|\mu^n-\mu_r^{\text{proj}}\right\|_{\hat{\Sigma}_{\text{HF}}^{\text{NN},n}} \le \frac{1}{2c_{\text{MLE}}^{\text{NN}}}\left\|\nabla_\mu L(\mu_r^{\text{proj}})\right\|_{(\hat{\Sigma}_{\text{HF}}^{\text{NN},n})^{-1}} + \sqrt{\frac{\varepsilon_{\text{SGD}}^{\text{NN},n}}{c_{\text{MLE}}^{\text{NN}}}} + 2R\sqrt{\frac{\zeta_{\text{HF}}}{n}}. \tag{36}$$

Let

$$\begin{aligned}
V_i &= -\frac{\mathbb{1}\{y_i=1\}\exp\left(-(\tilde{\psi}_0^{\tau_i^{(1)},\tau_i^{(2)}})^\top\mu_r^{\text{proj}}\right)}{1+\exp\left(-(\tilde{\psi}_0^{\tau_i^{(1)},\tau_i^{(2)}})^\top\mu_r^{\text{proj}}\right)} + \frac{\mathbb{1}\{y_i=0\}}{1+\exp\left(-(\tilde{\psi}_0^{\tau_i^{(1)},\tau_i^{(2)}})^\top\mu_r^{\text{proj}}\right)},\ \forall i\in[M_{\text{HF}}],\\
V &= [V_1,\dots,V_{M_{\text{HF}}}]^\top \in \mathbb{R}^{M_{\text{HF}}},\\
X &= [(\psi_0^{\tau_1^{(1)},\tau_1^{(2)}})^\top;\dots;(\psi_0^{\tau_{M_{\text{HF}}}^{(1)},\tau_{M_{\text{HF}}}^{(2)}})^\top] \in \mathbb{R}^{M_{\text{HF}}\times d},\\
X^\top &= [\psi_0^{\tau_1^{(1)},\tau_1^{(2)}},\dots,\psi_0^{\tau_{M_{\text{HF}}}^{(1)},\tau_{M_{\text{HF}}}^{(2)}}] \in \mathbb{R}^{d\times M_{\text{HF}}},
\end{aligned}$$

and then

$$\begin{aligned}
\nabla_\mu L(\mu_r^{\text{proj}}) &= \frac{1}{M_{\text{HF}}}\sum_{i=1}^{M_{\text{HF}}} V_i\tilde{\psi}_0^{\tau_i^{(1)},\tau_i^{(2)}} = \frac{1}{M_{\text{HF}}}X^\top V,\\
\hat{\Sigma}_{\text{HF}}^{\text{NN},n} &= \frac{1}{M_{\text{HF}}}X^\top X + \frac{\zeta_{\text{HF}}}{n}I.
\end{aligned}$$

For any $i\in[M_{\text{HF}}]$, we have $|V_i|\le 1$ and

$$\begin{aligned}
\mathbb{E}_{y_i}[V_i] &= \mathbb{E}_{y_i}\left[-\frac{\mathbb{1}\{y_i=1\}\exp\left(-(\tilde{\psi}_0^{\tau_i^{(1)},\tau_i^{(2)}})^\top\mu_r^{\text{proj}}\right)}{1+\exp\left(-(\tilde{\psi}_0^{\tau_i^{(1)},\tau_i^{(2)}})^\top\mu_r^{\text{proj}}\right)} + \frac{\mathbb{1}\{y_i=0\}}{1+\exp\left(-(\tilde{\psi}_0^{\tau_i^{(1)},\tau_i^{(2)}})^\top\mu_r^{\text{proj}}\right)}\right]\\
&= -\frac{1}{1+\exp\left(-\tilde{r}(\tau_i^{(1)},\tau_i^{(2)})\right)}\cdot\frac{\exp\left(-(\tilde{\psi}_0^{\tau_i^{(1)},\tau_i^{(2)}})^\top\mu_r^{\text{proj}}\right)}{1+\exp\left(-(\tilde{\psi}_0^{\tau_i^{(1)},\tau_i^{(2)}})^\top\mu_r^{\text{proj}}\right)}\\
&\quad + \frac{\exp\left(-\tilde{r}(\tau_i^{(1)},\tau_i^{(2)})\right)}{1+\exp\left(-\tilde{r}(\tau_i^{(1)},\tau_i^{(2)})\right)}\cdot\frac{1}{1+\exp\left(-(\tilde{\psi}_0^{\tau_i^{(1)},\tau_i^{(2)}})^\top\mu_r^{\text{proj}}\right)}\\
&= \frac{\exp\left(-\tilde{r}(\tau_i^{(1)},\tau_i^{(2)})\right)-\exp\left(-(\tilde{\psi}_0^{\tau_i^{(1)},\tau_i^{(2)}})^\top\mu_r^{\text{proj}}\right)}{\left(1+\exp\left(-\tilde{r}(\tau_i^{(1)},\tau_i^{(2)})\right)\right)\left(1+\exp\left(-(\tilde{\psi}_0^{\tau_i^{(1)},\tau_i^{(2)}})^\top\mu_r^{\text{proj}}\right)\right)}.
\end{aligned}$$

Then,

$$\begin{aligned}
|\mathbb{E}_{y_i}[V_i]| &\le \exp(2W_\tau)\left|(\tilde{\psi}_0^{\tau_i^{(1)},\tau_i^{(2)}})^\top\mu_r^{\text{proj}} - \tilde{r}(\tau_i^{(1)},\tau_i^{(2)})\right|\\
&= \exp(2W_\tau)\left|\left(\sum_{h=0}^{H(\tau_i^{(1)})} f_0(s_{i,h}^{(1)},a_{i,h}^{(1)};\mu_r^{\text{proj}}) - \sum_{h=0}^{H(\tau_i^{(2)})} f_0(s_{i,h}^{(2)},a_{i,h}^{(2)};\mu_r^{\text{proj}})\right)\right.\\
&\quad \left. - \left(\sum_{h=0}^{H(\tau_i^{(1)})} r(s_{i,h}^{(1)},a_{i,h}^{(1)}) - \sum_{h=0}^{H(\tau_i^{(2)})} r(s_{i,h}^{(2)},a_{i,h}^{(2)})\right)\right|
\end{aligned}$$

$$\leq \exp(2W_\tau)\left(\sum_{h=0}^{H(\tau_i^{(1)})}\left|f_0(s_{i,h}^{(1)},a_{i,h}^{(1)};\mu_r^{\text{proj}})-r(s_{i,h}^{(1)},a_{i,h}^{(1)})\right|\right.$$
$$\left.+\sum_{h=0}^{H(\tau_i^{(2)})}\left|f_0(s_{i,h}^{(2)},a_{i,h}^{(2)};\mu_r^{\text{proj}})-r(s_{i,h}^{(2)},a_{i,h}^{(2)})\right|\right).$$

Let $D := \frac{1}{M_{\text{HF}}^2}X(\hat{\Sigma}_{\text{HF}}^{\text{NN},n})^{-1}X^\top = \frac{1}{M_{\text{HF}}^2}X(\frac{1}{M_{\text{HF}}}X^\top X+\frac{\zeta_{\text{HF}}}{n}I)^{-1}X^\top \in \mathbb{R}^{M_{\text{HF}}\times M_{\text{HF}}}$.

Then,

$$\begin{aligned}\left\|\nabla_\mu L(\mu_r^{\text{proj}})\right\|^2_{\left(\hat{\Sigma}_{\text{HF}}^{\text{NN},n}\right)^{-1}} &= \nabla_\mu L(\mu_r^{\text{proj}})^\top\left(\hat{\Sigma}_{\text{HF}}^{\text{NN},n}\right)^{-1}\nabla_\mu L(\mu_r^{\text{proj}})\\ &= \frac{1}{M_{\text{HF}}^2}V^\top X\left(\hat{\Sigma}_{\text{HF}}^{\text{NN},n}\right)^{-1}X^\top V\\ &= V^\top DV.\end{aligned}$$

Since D is positive semi-definite, let $\lambda_1\geq\cdots\geq\lambda_{M_{\text{HF}}}\geq 0$ denote the eigenvalues of D.

We bound $\text{tr}(D)$, $\|D\|$, $\text{tr}\left(D\mathbb{E}[V]\mathbb{E}[V]^\top\right)$ and $\frac{\|D\|^2}{\text{tr}(D^2)}$ as follows.

$$\text{tr}(D)=\text{tr}\left(\frac{1}{M_{\text{HF}}^2}X\left(\frac{1}{M_{\text{HF}}}X^\top X+\frac{\zeta_{\text{HF}}}{n}I\right)^{-1}X^\top\right)=\frac{1}{M_{\text{HF}}}\text{tr}\left(\left(X^\top X+\frac{M_{\text{HF}}\zeta_{\text{HF}}}{n}I\right)^{-1}X^\top X\right)\leq\frac{d}{M_{\text{HF}}},$$

$$\|D\|=\left\|\frac{1}{M_{\text{HF}}^2}X\left(\frac{1}{M_{\text{HF}}}X^\top X+\frac{\zeta_{\text{HF}}}{n}I\right)^{-1}X^\top\right\|=\left\|\frac{1}{M_{\text{HF}}}X\left(X^\top X+\frac{M_{\text{HF}}\zeta_{\text{HF}}}{n}I\right)^{-1}X^\top\right\|\leq\frac{1}{M_{\text{HF}}},$$

$$\text{tr}\left(D\mathbb{E}[V]\mathbb{E}[V]^\top\right)=\text{tr}\left(\mathbb{E}[V]^\top D\mathbb{E}[V]\right)=\mathbb{E}[V]^\top D\mathbb{E}[V]\leq\|\mathbb{E}[V]\|_2^2\|D\|\leq\frac{\sum_{i=1}^{M_{\text{HF}}}(\mathbb{E}_{y_i}[V_i])^2}{M_{\text{HF}}},$$

and

$$\frac{\|D\|^2}{\text{tr}\left(D^2\right)}\overset{\text{(a)}}{\leq}\frac{M_{\text{HF}}\|D\|^2}{\left(\text{tr}\left(D\right)\right)^2}=\frac{M_{\text{HF}}\lambda_1^2(D)}{\left(\sum_{i=1}^{M_{\text{HF}}}\lambda_i(D)\right)^2}\leq\frac{M_{\text{HF}}\lambda_1^2(D)}{\sum_{i=1}^{M_{\text{HF}}}\lambda_i^2(D)}\leq M_{\text{HF}},$$

where inequality (a) is due to $\text{tr}\left(D^2\right)\geq\frac{(\text{tr}(D))^2}{M_{\text{HF}}}$.

Let $\delta'\leq\frac{1}{e}$. According to Lemma F.4, we have that with probability at least $1-2\delta'$,

$$\begin{aligned}\left\|\nabla_\mu L(\mu_r^{\text{proj}})\right\|^2_{\left(\hat{\Sigma}_{\text{HF}}^{\text{NN},n}\right)^{-1}} &\leq \text{tr}\left(D\right)+2\sqrt{\left(\text{tr}\left(D\right)\right)^2\log\left(\frac{1}{\delta'}\right)}+2\|D\|\log\left(\frac{1}{\delta'}\right)\\ &\quad+\text{tr}\left(D\mathbb{E}[V]\mathbb{E}[V]^\top\right)\left(1+2\sqrt{M_{\text{HF}}\log\left(\frac{1}{\delta'}\right)}\right)\\ &\leq\frac{md}{M_{\text{HF}}}+\frac{2md}{M_{\text{HF}}}\sqrt{\log\left(\frac{1}{\delta'}\right)}+\frac{2}{M_{\text{HF}}}\log\left(\frac{1}{\delta'}\right)+\frac{\sum_{i=1}^{M_{\text{HF}}}(\mathbb{E}_{y_i}[V_i])^2}{M_{\text{HF}}}\left(1+2\sqrt{M_{\text{HF}}\log\left(\frac{1}{\delta'}\right)}\right)\\ &\leq\frac{5md\log\left(\frac{1}{\delta'}\right)}{M_{\text{HF}}}+3\left(\sum_{i=1}^{M_{\text{HF}}}(\mathbb{E}_{y_i}[V_i])^2\right)\sqrt{\frac{\log\left(\frac{1}{\delta'}\right)}{M_{\text{HF}}}}.\end{aligned}\tag{37}$$

Plugging Eq. (37) into Eq. (36), we have

$$\begin{aligned}
&\left\|\mu^n - \mu_r^{\text{proj}}\right\|_{\hat{\Sigma}_{\text{HF}}^{\text{NN},n}} \\
&\leq \frac{1}{2c_{\text{MLE}}^{\text{NN}}}\sqrt{\frac{5md\log\left(\frac{1}{\delta'}\right)}{M_{\text{HF}}} + 3\left(\sum_{i=1}^{M_{\text{HF}}}(\mathbb{E}_{y_i}[V_i])^2\right)\sqrt{\frac{\log\left(\frac{1}{\delta'}\right)}{M_{\text{HF}}}}} + \sqrt{\frac{\varepsilon_{\text{SGD}}^{\text{NN},n}}{c_{\text{MLE}}^{\text{NN}}}} + 2R\sqrt{\frac{\zeta_{\text{HF}}}{n}} \\
&\leq \frac{1}{2c_{\text{MLE}}^{\text{NN}}}\sqrt{\frac{5md\log\left(\frac{1}{\delta'}\right)}{M_{\text{HF}}}} + \frac{3}{2c_{\text{MLE}}^{\text{NN}}}\sqrt{\sum_{i=1}^{M_{\text{HF}}}(\mathbb{E}_{y_i}[V_i])^2}\left(\frac{\log\left(\frac{1}{\delta'}\right)}{M_{\text{HF}}}\right)^{\frac{1}{4}} + \sqrt{\frac{\varepsilon_{\text{SGD}}^{\text{NN},n}}{c_{\text{MLE}}^{\text{NN}}}} + 2R\sqrt{\frac{\zeta_{\text{HF}}}{n}} := \varepsilon_{\text{MLE}}^{\text{NN},n}.
\end{aligned}$$

Next, we handle the term $\mathbb{E}_{y_i}[V_i]$. For any $i \in [M_{\text{HF}}]$,

$$\begin{aligned}
\mathbb{E}_{\substack{\tau_i^{(1)}\sim\mathcal{O}_{\text{HF}}^n \\ \tau_i^{(2)}\sim\mathcal{O}_{s_{\text{init}}}^{\pi^{\text{base}}}}}\left[(\mathbb{E}_{y_i}[V_i])^2\right] &\leq \exp(4W_\tau)W_\tau\mathbb{E}_{\substack{\mathbb{E}_{\tau_i^{(1)}\sim\mathcal{O}_{\text{HF}}^n} \\ \tau_i^{(2)}\sim\mathcal{O}_{s_{\text{init}}}^{\pi^{\text{base}}}}}\left[\sum_{h=0}^{H(\tau_i^{(1)})}\left(f_0(s_{i,h}^{(1)},a_{i,h}^{(1)};\mu_r^{\text{proj}}) - r(s_{i,h}^{(1)},a_{i,h}^{(1)})\right)^2\right. \\
&\quad \left. + \sum_{h=0}^{H(\tau_i^{(2)})}\left(f_0(s_{i,h}^{(2)},a_{i,h}^{(2)};\mu_r^{\text{proj}}) - r(s_{i,h}^{(2)},a_{i,h}^{(2)})\right)^2\right] \\
&= \frac{\exp(4W_\tau)W_\tau}{1-\gamma}\mathbb{E}_{\substack{(s^{(1)},a^{(1)})\sim d_{\text{HF}}^n \\ (s^{(2)},a^{(2)})\sim d_{\text{base}}}}\left[\left(f_0(s_h^{(1)},a_h^{(1)};\mu_r^{\text{proj}}) - r(s_h^{(1)},a_h^{(1)})\right)^2\right. \\
&\quad \left. + \left(f_0(s_h^{(2)},a_h^{(2)};\mu_r^{\text{proj}}) - r(s_h^{(2)},a_h^{(2)})\right)^2\right] \\
&\leq \frac{32R^2W_\tau\exp(4W_\tau)}{(1-\gamma)m}\log\left(\frac{1}{\delta'}\right).
\end{aligned}$$

Then, we have

$$\begin{aligned}
\mathbb{E}_{\substack{\{\tau_i^{(1)}\}_{i=1}^{M_{\text{HF}}}\sim\mathcal{O}_{\text{HF}}^n \\ \{\tau_i^{(2)}\}_{i=1}^{M_{\text{HF}}}\sim\mathcal{O}_{s_{\text{init}}}^{\pi^{\text{base}}}}}\left[\varepsilon_{\text{MLE}}^{\text{NN},n}\right] &\leq \frac{1}{2c_{\text{MLE}}^{\text{NN}}}\sqrt{\frac{5md\log\left(\frac{1}{\delta'}\right)}{M_{\text{HF}}}} + \frac{12RM_{\text{HF}}^{\frac{1}{4}}}{c_{\text{MLE}}^{\text{NN}}}\sqrt{\frac{W_\tau\exp(4W_\tau)}{(1-\gamma)m}\log\left(\frac{1}{\delta'}\right)} \\
&\quad + \sqrt{\frac{\mathbb{E}_{\{\tau_i^{(1)}\}_{i=1}^{M_{\text{HF}}}\sim\mathcal{O}_{\text{HF}}^n,\ \{\tau_i^{(2)}\}_{i=1}^{M_{\text{HF}}}\sim\mathcal{O}_{s_{\text{init}}}^{\pi^{\text{base}}}}\left[\varepsilon_{\text{SGD}}^{\text{NN},n}\right]}{c_{\text{MLE}}^{\text{NN}}}} + 2R\sqrt{\frac{\zeta_{\text{HF}}}{n}} \\
&\leq \frac{1}{2c_{\text{MLE}}^{\text{NN}}}\sqrt{\frac{5md\log\left(\frac{1}{\delta'}\right)}{M_{\text{HF}}}} + \frac{12RM_{\text{HF}}^{\frac{1}{4}}}{c_{\text{MLE}}^{\text{NN}}}\sqrt{\frac{W_\tau\exp(4W_\tau)}{(1-\gamma)m}\log\left(\frac{1}{\delta'}\right)} \\
&\quad + \frac{1}{\sqrt{c_{\text{MLE}}^{\text{NN}}}}\left(17W_\tau R\sqrt{\frac{\log\left(\frac{1}{\delta'}\right)}{M_{\text{SGD}}^{\mu}}}\right)^{\frac{1}{2}} + \frac{1}{\sqrt{c_{\text{MLE}}^{\text{NN}}}}\left(\frac{40R^2W_\tau\sqrt{c_{\text{scale}}R}}{(1-\gamma)\sqrt{\underline{c}}m^{\frac{1}{4}}}\right)^{\frac{1}{2}} + 2R\sqrt{\frac{\zeta_{\text{HF}}}{n}} \\
&\leq \frac{1}{2c_{\text{MLE}}^{\text{NN}}}\sqrt{\frac{5md\log\left(\frac{1}{\delta'}\right)}{M_{\text{HF}}}} + \frac{1}{\sqrt{c_{\text{MLE}}^{\text{NN}}}}\left(17W_\tau R\sqrt{\frac{\log\left(\frac{1}{\delta'}\right)}{M_{\text{SGD}}^{\mu}}}\right)^{\frac{1}{2}} \\
&\quad + 2R\sqrt{\frac{\zeta_{\text{HF}}}{n}} + \frac{19c_{\text{scale}}^{\frac{1}{4}}R^{\frac{5}{4}}M_{\text{HF}}^{\frac{1}{4}}\sqrt{W_\tau\exp(4W_\tau)}}{\underline{c}^{\frac{1}{4}}c_{\text{MLE}}^{\text{NN}}\sqrt{1-\gamma}m^{\frac{1}{8}}}\log\left(\frac{1}{\delta'}\right).
\end{aligned}$$

□

Lemma E.13. *Assume that event $\mathcal{E}_{\text{init}} \cap \mathcal{E}_\tau \cap \mathcal{E}_\mu^{\text{NN}} \cap \mathcal{E}_{\text{MLE}}^{\text{NN}} \cap \mathcal{E}_{\text{cov}}^{\text{NN}}$ holds. Then, for any phase $n \geq 0$ and iteration $t \geq 0$,*

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n,\ \{\tau_i^{(1)}\}_{i=1}^{M_{\text{HF}}}\sim\mathcal{O}_{\text{HF}}^n,\ \{\tau_i^{(2)}\}_{i=1}^{M_{\text{HF}}}}\left[\left|Q^{\pi^t}(s,a;\hat{r}^n+b^n)-Q^{\pi^t}(s,a;r+b^n)\right|\right]\\
&\leq 2\mathbb{E}_{\substack{\{\tau_i^{(1)}\}_{i=1}^{M_{\text{HF}}}\sim\mathcal{O}_{\text{HF}}^n\\ \{\tau_i^{(2)}\}_{i=1}^{M_{\text{HF}}}\sim\mathcal{O}_{s_{\text{init}}}^{\pi^{\text{base}}}}}\left[\varepsilon_{\text{MLE}}^{\text{NN},n}\right]\cdot\mathbb{E}_{\tau\sim\mathcal{O}_{\rho_{\text{cov}}^n}^{\pi^t}}\left[\left\|\sum_{h=0}^{H(\tau)}\psi_0(s_h,a_h)\right\|_{(\Sigma_{\text{HF}}^{\text{NN},n})^{-1}}\right]+\frac{6\sqrt{c_{\text{scale}}R^3\log\left(\frac{1}{\delta'}\right)}}{(1-\gamma)\sqrt{\underline{c}}m^{\frac{1}{4}}}:=\varsigma_{\rho_{\text{cov}}^n}^{\text{NN},\pi^t}.
\end{aligned}$$

Proof. We have

$$\begin{aligned}
&\left|Q^{\pi^t}(s,a;\hat{r}^n+b^n)-Q^{\pi^t}(s,a;r+b^n)\right|\\
&=\left|\mathbb{E}_{\tau\sim\mathcal{O}_{s,a}^{\pi^t}}\left[\sum_{h=0}^{H(\tau)}\left(h(s_h,a_h;\mu^n)-r(s_h,a_h)\right)\right]\right|\\
&=\left|\mathbb{E}_{\tau\sim\mathcal{O}_{s,a}^{\pi^t}}\left[\sum_{h=0}^{H(\tau)}\Big(h(s_h,a_h;\mu^n)-h_0(s_h,a_h;\mu^n)+h_0(s_h,a_h;\mu^n)-h_0(s_h,a_h;\mu_r^{\text{proj}})\right.\right.\\
&\quad\left.\left.+h_0(s_h,a_h;\mu_r^{\text{proj}})-r(s_h,a_h)\Big)\right]\right|\\
&\leq\mathbb{E}_{\tau\sim\mathcal{O}_{s,a}^{\pi^t}}\left[\left|\sum_{h=0}^{H(\tau)}\left(h(s_h,a_h;\mu^n)-h_0(s_h,a_h;\mu^n)\right)\right|\right]+\mathbb{E}_{\tau\sim\mathcal{O}_{s,a}^{\pi^t}}\left[\left|\sum_{h=0}^{H(\tau)}\psi_0(s_h,a_h)^\top\left(\mu^n-\mu_r^{\text{proj}}\right)\right|\right]\\
&\quad+\mathbb{E}_{\tau\sim\mathcal{O}_{s,a}^{\pi^t}}\left[\left|\sum_{h=0}^{H(\tau)}\left(h_0(s_h,a_h;\mu_r^{\text{proj}})-r(s_h,a_h)\right)\right|\right]\\
&\leq\mathbb{E}_{\tau\sim\mathcal{O}_{s,a}^{\pi^t}}\left[\left\|\sum_{h=0}^{H(\tau)}\psi_0(s_h,a_h)\right\|_{(\hat{\Sigma}_{\text{HF}}^{\text{NN},n})^{-1}}\left\|\mu^n-\mu_r^{\text{proj}}\right\|_{\hat{\Sigma}_{\text{HF}}^{\text{NN},n}}\right]+\frac{1}{1-\gamma}\mathbb{E}_{(s',a')\sim d_{s,a}^{\pi^t}}\left[\left|h(s',a';\mu^n)-h_0('s,a';\mu^n)\right|\right]\\
&\quad+\frac{1}{1-\gamma}\mathbb{E}_{(s',a')\sim d_{s,a}^{\pi^t}}\left[\left|h_0(s',a';\mu_r^{\text{proj}})-r(s',a')\right|\right]\\
&\overset{\text{(a)}}{\leq}2\varepsilon_{\text{MLE}}^{\text{NN},n}\mathbb{E}_{\tau\sim\mathcal{O}_{s,a}^{\pi^t}}\left[\left\|\sum_{h=0}^{H(\tau)}\psi_0(s_h,a_h)\right\|_{(\Sigma_{\text{HF}}^{\text{NN},n})^{-1}}\right]+\frac{1}{1-\gamma}\mathbb{E}_{(s',a')\sim d_{s,a}^{\pi^t}}\left[\left|h(s',a';\mu^n)-h_0('s,a';\mu^n)\right|\right]\\
&\quad+\frac{1}{1-\gamma}\mathbb{E}_{(s',a')\sim d_{s,a}^{\pi^t}}\left[\left|h_0(s',a';\mu_r^{\text{proj}})-r(s',a')\right|\right],
\end{aligned}$$

where inequality (a) uses the definition of $\mathcal{E}_{\text{cov}}^{\text{NN}}$ and Lemma E.12.

Then, taking $\mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n}[\cdot]$ on both sides, we have

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n}\left[\left|Q^{\pi^t}(s,a;\hat{r}^n+b^n)-Q^{\pi^t}(s,a;r+b^n)\right|\right]\\
&\leq 2\varepsilon_{\text{MLE}}^{\text{NN},n}\mathbb{E}_{\tau\sim\mathcal{O}_{\rho_{\text{cov}}^n}^{\pi^t}}\left[\left\|\sum_{h=0}^{H(\tau)}\psi_0(s_h,a_h)\right\|_{(\Sigma_{\text{HF}}^{\text{NN},n})^{-1}}\right]+\frac{1}{1-\gamma}\mathbb{E}_{(s',a')\sim d_{\rho_{\text{cov}}^n}^{\pi^t}}\left[\left|h(s',a';\mu^n)-h_0('s,a';\mu^n)\right|\right]\\
&\quad+\frac{1}{1-\gamma}\mathbb{E}_{(s',a')\sim d_{\rho_{\text{cov}}^n}^{\pi^t}}\left[\left|h_0(s',a';\mu_r^{\text{proj}})-r(s',a')\right|\right]\\
&\overset{\text{(a)}}{\leq}2\varepsilon_{\text{MLE}}^{\text{NN},n}\mathbb{E}_{\tau\sim\mathcal{O}_{\rho_{\text{cov}}^n}^{\pi^t}}\left[\left\|\sum_{h=0}^{H(\tau)}\psi_0(s_h,a_h)\right\|_{(\Sigma_{\text{HF}}^{\text{NN},n})^{-1}}\right]+\frac{2\sqrt{c_{\text{scale}}R^3}}{(1-\gamma)\sqrt{\underline{c}}m^{\frac{1}{4}}}+\frac{4R}{1-\gamma}\sqrt{\frac{\log\left(\frac{1}{\delta'}\right)}{m}}
\end{aligned}$$

$$\leq 2\varepsilon_{\text{MLE}}^{\text{NN},n}\mathbb{E}_{\tau\sim\mathcal{O}^{\pi^t}_{\rho^n_{\text{cov}}}}\left[\left\|\sum_{h=0}^{H(\tau)}\psi_0(s_h,a_h)\right\|_{(\Sigma_{\text{HF}}^{\text{NN},n})^{-1}}\right]+\frac{6\sqrt{c_{\text{scale}}R^3\log\left(\frac{1}{\delta'}\right)}}{(1-\gamma)\sqrt{\underline{c}}m^{\frac{1}{4}}},$$

where inequality (a) uses Lemma E.2 and the definition of event $\mathcal{E}_{\text{init}}$.

Furthermore, taking $\mathbb{E}_{\{\tau_i^{(1)}\}_{i=1}^{M_{\text{HF}}}\sim\mathcal{O}^n_{\text{HF}},\ \{\tau_i^{(2)}\}_{i=1}^{M_{\text{HF}}}\sim\mathcal{O}^{\pi^{\text{base}}}_{s_{\text{init}}}}[\cdot]$ on both sides, we have

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left|Q^{\pi^t}(s,a;\hat{r}^n+b^n)-Q^{\pi^t}(s,a;r+b^n)\right|\right]\\
&\leq 2\mathbb{E}_{\substack{\{\tau_i^{(1)}\}_{i=1}^{M_{\text{HF}}}\sim\mathcal{O}^n_{\text{HF}}\\ \{\tau_i^{(2)}\}_{i=1}^{M_{\text{HF}}}\sim\mathcal{O}^{\pi^{\text{base}}}_{s_{\text{init}}}}}\left[\varepsilon_{\text{MLE}}^{\text{NN},n}\right]\mathbb{E}_{\tau\sim\mathcal{O}^{\pi^t}_{\rho^n_{\text{cov}}}}\left[\left\|\sum_{h=0}^{H(\tau)}\psi_0(s_h,a_h)\right\|_{(\Sigma_{\text{HF}}^{\text{NN},n})^{-1}}\right]+\frac{6\sqrt{c_{\text{scale}}R^3\log\left(\frac{1}{\delta'}\right)}}{(1-\gamma)\sqrt{\underline{c}}m^{\frac{1}{4}}}:=\varsigma^{\text{NN},\pi^t}_{\rho^n_{\text{cov}}}.
\end{aligned}$$

□

In the following, for ease of notation, we use $\mathbb{E}_{\{\tau_i^{(1)}\}_{i=1}^{M_{\text{HF}}}\sim\mathcal{O}^n_{\text{HF}},\ \{\tau_i^{(2)}\}_{i=1}^{M_{\text{HF}}}\sim\mathcal{O}^{\pi^{\text{base}}}_{s_{\text{init}}}}[\cdot]$ and $\mathbb{E}_{\hat{r}^n}[\cdot]$ interchangeably.

Lemma E.14. *Assume that event $\mathcal{E}_{\text{init}}\cap\mathcal{E}_\tau\cap\mathcal{E}_\mu^{\text{NN}}\cap\mathcal{E}_{\text{MLE}}^{\text{NN}}\cap\mathcal{E}_{\text{cov}}^{\text{NN}}$ holds. Then, for any phase $n\geq 0$, iteration $t\geq 0$, $s\in\mathcal{K}^n$ and $a\in\mathcal{A}$,*

$$\left|\psi_0(s,a)^\top\left(\theta_*^t-\theta_{\text{mid}}^t\right)\right|\leq\sqrt{2\beta\left(8(n+1)W_Q^{\text{NN}}\varsigma^{\text{NN},\pi^t}_{\rho^n_{\text{cov}}}+4\zeta_{\text{cov}}R^2\right)}.$$

Proof. For any phase $n\geq 0$,

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}},\hat{r}^n}\left[\left(Q^{\pi^t}(s,a;\hat{r}^n+b^n)-b^n(s,a)-\psi_0(s,a)^\top\theta\right)^2-\left(Q^{\pi^t}(s,a;r+b^n)-b^n(s,a)-\psi_0(s,a)^\top\theta\right)^2\right]\\
&\leq 4W_Q^{\text{NN}}\mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}},\hat{r}^n}\left[\left|Q^{\pi^t}(s,a;\hat{r}^n+b^n)-Q^{\pi^t}(s,a;r+b^n)\right|\right]\\
&\leq 4W_Q^{\text{NN}}\varsigma^{\text{NN},\pi^t}_{\rho^n_{\text{cov}}},
\end{aligned}\tag{38}$$

Here W_Q^{NN} satisfies $\max\{|Q^{\pi^t}(s,a;r+b^n)-b^n(s,a)|,|\psi_0(s,a)^\top\theta_{\text{mid}}^t|,|\psi_0(s,a)^\top\theta_*^t|\}\leq W_Q^{\text{NN}}$.

Plugging θ_*^t into θ, we have that for any fixed (s,a),

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left(Q^{\pi^t}(s,a;r+b^n)-b^n(s,a)-\psi_0(s,a)^\top\theta_*^t\right)^2\right]\\
&\geq\mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}},\hat{r}^n}\left[\left(Q^{\pi^t}(s,a;\hat{r}^n+b^n)-b^n(s,a)-\psi_0(s,a)^\top\theta_*^t\right)^2\right]-4W_Q^{\text{NN}}\varsigma^{\text{NN},\pi^t}_{\rho^n_{\text{cov}}}\\
&\overset{\text{(a)}}{\geq}\mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}},\hat{r}^n}\left[\left(Q^{\pi^t}(s,a;\hat{r}^n+b^n)-b^n(s,a)-\psi_0(s,a)^\top\theta_{\text{mid}}^t\right)^2\right]-4W_Q^{\text{NN}}\varsigma^{\text{NN},\pi^t}_{\rho^n_{\text{cov}}},
\end{aligned}\tag{39}$$

where inequality (a) is due to the definition of θ_{mid}^t.

Furthermore, we have

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left(Q^{\pi^t}(s,a;r+b^n)-b^n(s,a)-\psi_0(s,a)^\top\theta_{\text{mid}}^t\right)^2\right]\\
&\quad-\mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left(Q^{\pi^t}(s,a;r+b^n)-b^n(s,a)-\psi_0(s,a)^\top\theta_*^t\right)^2\right]\\
&=\mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}},\hat{r}^n}\left[\left(Q^{\pi^t}(s,a;\hat{r}^n+b^n)-b^n(s,a)-\psi_0(s,a)^\top\theta_{\text{mid}}^t\right)^2\right]\\
&\quad-\mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left(Q^{\pi^t}(s,a;r+b^n)-b^n(s,a)-\psi_0(s,a)^\top\theta_*^t\right)^2\right]
\end{aligned}$$

$$
\begin{aligned}
&+ \mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left(Q^{\pi^t}(s,a;r+b^n) - b^n(s,a) - \psi_0(s,a)^\top\theta^t_{\text{mid}}\right)^2\right] \\
&- \mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}},\hat{r}^n}\left[\left(Q^{\pi^t}(s,a;\hat{r}^n+b^n) - b^n(s,a) - \psi_0(s,a)^\top\theta^t_{\text{mid}}\right)^2\right] \\
&\overset{\text{(a)}}{\leq} 4W_Q^{\text{NN}}\varsigma^{\text{NN},\pi^t}_{\rho^n_{\text{cov}}} + 4W_Q^{\text{NN}}\mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}},\hat{r}^n}\left[\left|Q^{\pi^t}(s,a;\hat{r}^n+b^n) - Q^{\pi^t}(s,a;r+b^n)\right|\right] \\
&\leq 8W_Q^{\text{NN}}\varsigma^{\text{NN},\pi^t}_{\rho^n_{\text{cov}}},
\end{aligned} \tag{40}
$$

where inequality (a) uses Lemma F.3.

On the other hand, it holds that

$$
\begin{aligned}
&\mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left(Q^{\pi^t}(s,a;r+b^n) - b^n(s,a) - \psi_0(s,a)^\top\theta^t_{\text{mid}}\right)^2\right] \\
&- \mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left(Q^{\pi^t}(s,a;r+b^n) - b^n(s,a) - \psi_0(s,a)^\top\theta^t_{*}\right)^2\right] \\
&= \mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left(\psi_0(s,a)^\top\left(\theta^t_* - \theta^t_{\text{mid}}\right)\right)^2\right] \\
&\quad + 2\underbrace{\mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left(Q^{\pi^t}(s,a;r+b^n) - b^n(s,a) - \psi_0(s,a)^\top\theta^t_*\right)\psi_0(s,a)^\top\left(\theta^t_* - \theta^t_{\text{mid}}\right)\right]}_{\text{Term } \Gamma^{\text{NN}} \geq 0},
\end{aligned} \tag{41}
$$

where Term Γ^{NN} is non-negative due to the the first-order optimality of θ^t_*.

Then,

$$
\begin{aligned}
&(\theta^t_* - \theta^t_{\text{mid}})^\top\mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}[\psi_0(s,a)\psi_0(s,a)^\top](\theta^t_* - \theta^t_{\text{mid}}) \\
&= \mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left(\psi_0(s,a)^\top\left(\theta^t_* - \theta^t_{\text{mid}}\right)\right)^2\right] \\
&\overset{\text{(a)}}{\leq} \mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left(Q^{\pi^t}(s,a;r+b^n) - b^n(s,a) - \psi_0(s,a)^\top\theta^t_{\text{mid}}\right)^2\right] \\
&\quad - \mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}\left[\left(Q^{\pi^t}(s,a;r+b^n) - b^n(s,a) - \psi_0(s,a)^\top\theta^t_{*}\right)^2\right] \\
&\leq 8W_Q^{\text{NN}}W_Q^{\text{NN}}\varsigma^{\text{NN},\pi^t}_{\rho^n_{\text{cov}}},
\end{aligned}
$$

where inequality (a) uses the same argument as Eq. (17) (i.e., the first optimality of θ^t_*).

The above equation implies

$$
\begin{aligned}
\left\|\theta^t_* - \theta^t_{\text{mid}}\right\|^2_{\Sigma^{\text{NN},n}_{\text{cov}}} &= (\theta^t_* - \theta^t_{\text{mid}})^\top\left((n+1)\mathbb{E}_{(s,a)\sim\rho^n_{\text{cov}}}[\psi_0(s,a)\psi_0(s,a)^\top] + \zeta_{\text{cov}}I\right)(\theta^t_* - \theta^t_{\text{mid}}) \\
&\leq 8(n+1)W_Q^{\text{NN}}\varsigma^{\text{NN},\pi^t}_{\rho^n_{\text{cov}}} + 4\zeta_{\text{cov}}R^2.
\end{aligned}
$$

For any $s \in \mathcal{K}^n$, using the definition of $\mathcal{K}^n$ and event $\mathcal{E}^{\text{NN}}_{\text{cov}}$, we have

$$
\frac{1}{\sqrt{2}}\left\|\psi_0(s,a)\right\|_{(\Sigma^{\text{NN},n}_{\text{cov}})^{-1}} \leq \left\|\psi_0(s,a)\right\|_{(\hat{\Sigma}^{\text{NN},n}_{\text{cov}})^{-1}} \leq \sqrt{\beta}. \tag{42}
$$

Thus, for any $s \in \mathcal{K}^n$,

$$
\begin{aligned}
\left|\psi_0(s,a)^\top\left(\theta^t_* - \theta^t_{\text{mid}}\right)\right| &\leq \left\|\psi_0(s,a)\right\|_{(\Sigma^{\text{NN},n}_{\text{cov}})^{-1}}\left\|\theta^t_* - \theta^t_{\text{mid}}\right\|_{\Sigma^{\text{NN},n}_{\text{cov}}} \\
&\leq \sqrt{2\beta\left(8(n+1)W_Q^{\text{NN}}\varsigma^{\text{NN},\pi^t}_{\rho^n_{\text{cov}}} + 4\zeta_{\text{cov}}R^2\right)}.
\end{aligned}
$$

□

E.5. Proof of Theorem 5.2

For any phase $n = 0, \ldots, N-1$ and iteration $t = 0, \ldots, T-1$, let

$$\theta_*^t := \operatorname*{argmin}_{\theta \in \mathcal{S}_R} \mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n} \left[\left(f_0(s,a;\theta) - \left(Q^{\pi^t}(s,a; r+b^n) - b^n(s,a) \right) \right)^2 \right],$$
$$\theta_{\text{mid}}^t := \operatorname*{argmin}_{\theta \in \mathcal{S}_R} \mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n, \hat{r}^n} \left[\left(f_0(s,a;\theta) - \left(Q^{\pi^t}(s,a; \hat{r}^n+b^n) - b^n(s,a) \right) \right)^2 \right],$$
$$\theta^t \overset{\text{SGD}}{\approx} \operatorname*{argmin}_{\theta \in \mathcal{S}_R} \mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n} \left[\left(f_0(s,a;\theta) - \left(Q^{\pi^t}(s,a; \hat{r}^n+b^n) - b^n(s,a) \right) \right)^2 \right].$$

Let $\delta' := \frac{\delta}{24N(K+M_{\text{HF}}+M_{\text{SGD}}^{\mu}+TM_{\text{SGD}}^{\theta})}$. For any $n \geq 0$, $t \geq 0$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$, let $\bar{b}^{n,t}(s,a) := b^n(s,a) - \mathbb{E}_{a'\sim\pi^t(\cdot|s)}[b^n(s,a')]$, and for any $w \in \mathbb{R}^{md}$, let $\bar{\psi}_w^t(s,a) := \psi_w(s,a) - \mathbb{E}_{a'\sim\pi^t(\cdot|s)}[\psi_w(s,a')]$.

Proof of Theorem 5.2. First, we have $\Pr[\mathcal{E}_{\text{init}} \cap \mathcal{E}_\theta^{\text{NN}} \cap \mathcal{E}_\tau \cap \mathcal{E}_\mu^{\text{NN}} \cap \mathcal{E}_{\text{MLE}}^{\text{NN}} \cap \mathcal{E}_{\text{cov}}^{\text{NN}}] \geq 1 - 6 \cdot 2N(K+M_{\text{HF}}+M_{\text{SGD}}^{\mu}+TM_{\text{SGD}}^{\theta}) \cdot 2\delta' = 1-\delta$. In the following, we assume that event $\mathcal{E}_{\text{init}} \cap \mathcal{E}_\theta^{\text{NN}} \cap \mathcal{E}_\tau \cap \mathcal{E}_\mu^{\text{NN}} \cap \mathcal{E}_{\text{MLE}}^{\text{NN}} \cap \mathcal{E}_{\text{cov}}^{\text{NN}}$ holds.

For any phase $n = 0, \ldots, N-1$ and iteration $t = 0, \ldots, T-1$, we have

$$\begin{aligned}
& V_{\mathcal{M}^n}^{\pi^*,n}(s_{\text{init}}) - V_{\mathcal{M}^n}^{\pi^t}(s_{\text{init}}) \\
\leq\ & \frac{1}{1-\gamma} \mathbb{E}_{(s,a)\sim d_{\mathcal{M}^n;s_{\text{init}}}^{\pi^*,n}} \left[A_{\mathcal{M}_{b^n}}^{\pi^t}(s,a) \cdot \mathbb{1}\{s \in \mathcal{K}^n\} \right] \\
=\ & \frac{1}{1-\gamma} \mathbb{E}_{(s,a)\sim d_{\mathcal{M}^n;s_{\text{init}}}^{\pi^*,n}} \Bigg[\left(\bar{\psi}_{w^t}^t(s,a)^\top \theta^t + \bar{b}^{n,t}(s,a) \right) \cdot \mathbb{1}\{s \in \mathcal{K}^n\} \\
& + \underbrace{\left(A_{\mathcal{M}_{b^n}}^{\pi^t}(s,a) - \left(\bar{\psi}_0^t(s,a)^\top \theta_*^t + \bar{b}^{n,t}(s,a) \right) \right) \cdot \mathbb{1}\{s \in \mathcal{K}^n\}}_{\text{Term 1}} \\
& + \underbrace{\bar{\psi}_0^t(s,a)^\top \left(\theta_*^t - \theta_{\text{mid}}^t \right) \cdot \mathbb{1}\{s \in \mathcal{K}^n\}}_{\text{Term 2}} + \underbrace{\bar{\psi}_0^t(s,a)^\top \left(\theta_{\text{mid}}^t - \theta^t \right) \cdot \mathbb{1}\{s \in \mathcal{K}^n\}}_{\text{Term 3}} \\
& + \underbrace{\left(\bar{\psi}_0^t(s,a) - \bar{\psi}_{w^t}^t(s,a) \right)^\top \theta^t}_{\text{Term 4}} \Bigg].
\end{aligned} \tag{43}$$

Below we bound Terms 1-4.

Term 1. We first bound Term 1.

$$\begin{aligned}
\text{Term 1} =\ & \mathbb{E}_{(s,a)\sim d_{\mathcal{M}^n;s_{\text{init}}}^{\pi^*,n}} \left[\left(Q_{\mathcal{M}_{b^n}}^{\pi^t}(s,a) - \left(\psi_0(s,a)^\top \theta_*^t + b^n(s,a) \right) \right) \cdot \mathbb{1}\{s \in \mathcal{K}^n\} \right] \\
& - \mathbb{E}_{s\sim d_{\mathcal{M}^n;s_{\text{init}}}^{\pi^*,n}, a'\sim\pi^t(\cdot|s)} \left[\left(Q_{\mathcal{M}_{b^n}}^{\pi^t}(s,a') - \left(\psi_0(s,a')^\top \theta_*^t + b^n(s,a') \right) \right) \cdot \mathbb{1}\{s \in \mathcal{K}^n\} \right] \\
\overset{\text{(a)}}{\leq}\ & \sqrt{\mathbb{E}_{(s,a)\sim d_{s_{\text{init}}}^{\pi^*}} \left[\left(Q_{\mathcal{M}_{b^n}}^{\pi^t}(s,a) - (\psi_0(s,a)^\top \theta_*^t + b^n(s,a)) \right)^2 \right]} \\
& + \sqrt{\mathbb{E}_{s\sim d_{s_{\text{init}}}^{\pi^*}, a'\sim\pi^t(\cdot|s)} \left[\left(Q_{\mathcal{M}_{b^n}}^{\pi^t}(s,a') - (\psi_0(s,a')^\top \theta_*^t + b^n(s,a')) \right)^2 \right]} \\
\leq\ & 2\sqrt{|\mathcal{A}| \mathbb{E}_{(s,a)\sim d_{s_{\text{init}}}^{\star}} \left[\left(Q_{\mathcal{M}_{b^n}}^{\pi^t}(s,a) - (\psi_0(s,a)^\top \theta_*^t + b^n(s,a)) \right)^2 \right]} \\
=\ & 2\sqrt{|\mathcal{A}| \varepsilon_{\text{bias}}^{\text{NN}}},
\end{aligned}$$

where inequality (a) uses Lemma D.2.

Term 2. Then, we bound Term 2.

Using Lemma E.14, we have that for any $s \in \mathcal{K}^n$ and $a \in \mathcal{A}$,

$$\left|\psi_0(s,a)^\top \left(\theta_*^t - \theta_{\text{mid}}^t\right)\right| \le \sqrt{2\beta\left(8(n+1)W_Q^{\text{NN}}\varsigma_{\rho_{\text{cov}}^n}^{\text{NN},\pi^t} + 4\zeta_{\text{cov}}R^2\right)}.$$

Thus,

$$\begin{aligned}
\text{Term 2} &= \mathbb{E}_{(s,a)\sim d^{\pi_*^n}_{\mathcal{M}^n;s_{\text{init}}}}\left[\bar{\psi}_0^t(s,a)^\top\left(\theta_*^t - \theta_{\text{mid}}^t\right)\cdot \mathbb{1}\left\{s\in\mathcal{K}^n\right\}\right] \\
&\le \mathbb{E}_{(s,a)\sim d^{\pi_*^n}_{\mathcal{M}^n;s_{\text{init}}}}\left[\left|\psi_0(s,a)^\top\left(\theta_*^t - \theta_{\text{mid}}^t\right)\right|\cdot \mathbb{1}\left\{s\in\mathcal{K}^n\right\}\right] \\
&\quad + \mathbb{E}_{s\sim d^{\pi_*^n}_{\mathcal{M}^n;s_{\text{init}}},a'\sim\pi^t(\cdot|s)}\left[\left|\psi_0(s,a')^\top\left(\theta_*^t - \theta_{\text{mid}}^t\right)\right|\cdot \mathbb{1}\left\{s\in\mathcal{K}^n\right\}\right] \\
&\le 2\sqrt{2\beta\left(8(n+1)W_Q^{\text{NN}}\varsigma_{\rho_{\text{cov}}^n}^{\text{NN},\pi^t} + 4\zeta_{\text{cov}}R^2\right)}.
\end{aligned}$$

Term 3. Next, we bound Term 3.

Using the same argument as Eq. (41) (i.e., the first optimality of θ_{mid}^t),

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n,\hat{r}^n}\left[\left(\psi_0(s,a)^\top\theta_{\text{mid}}^t - \psi_0(s,a)^\top\theta^t\right)^2\right] \\
&\le \mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n,\hat{r}^n}\left[\left(\psi_0(s,a)^\top\theta^t - \left(Q^{\pi^t}(s,a;\hat{r}^n+b^n) - b^n(s,a)\right)\right)^2\right] \\
&\quad - \mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n,\hat{r}^n}\left[\left(\psi_0(s,a)^\top\theta_{\text{mid}}^t - \left(Q^{\pi^t}(s,a;\hat{r}^n+b^n) - b^n(s,a)\right)\right)^2\right] \\
&= \mathbb{E}_{\hat{r}^n}\left[\mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n}\left[\left(\psi_0(s,a)^\top\theta^t - \left(Q^{\pi^t}(s,a;\hat{r}^n+b^n) - b^n(s,a)\right)\right)^2\right.\right. \\
&\quad \left.\left. - \left(\psi_0(s,a)^\top\theta_{\text{mid}}^t - \left(Q^{\pi^t}(s,a;\hat{r}^n+b^n) - b^n(s,a)\right)\right)^2\right]\Big|\hat{r}^n\right] \\
&= \mathbb{E}_{\hat{r}^n}\left[F^{\hat{r}^n}(\theta^t) - F^{\hat{r}^n}(\theta_{\text{mid}}^t)|\hat{r}^n\right] \\
&\le \mathbb{E}_{\hat{r}^n}\left[F^{\hat{r}^n}(\theta^t) - F^{\hat{r}^n}(\theta_{\text{mid}}^{t,\hat{r}^n})|\hat{r}^n\right] \\
&\overset{\text{(a)}}{\le} \varepsilon_Q^{\text{NN}}.
\end{aligned}$$

where inequality (a) is due to Lemma E.9.

Then, we have

$$\begin{aligned}
\left\|\theta_{\text{mid}}^t - \theta^t\right\|^2_{\Sigma_{\text{cov}}^{\text{NN},n}} &\le \left(\theta_{\text{mid}}^t - \theta^t\right)^\top\left((n+1)\mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n}\left[\psi_0(s,a)\psi_0(s,a)^\top\right] + \zeta_{\text{cov}}I\right)\left(\theta_{\text{mid}}^t - \theta^t\right) \\
&= (n+1)\mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n}\left[\left(\psi_0(s,a)^\top\theta_{\text{mid}}^t - \psi_0(s,a)^\top\theta^t\right)^2\right] + 4R^2\zeta_{\text{cov}} \\
&\le (n+1)\varepsilon_Q^{\text{NN}} + 4R^2\zeta_{\text{cov}}.
\end{aligned}$$

For any $s \in \mathcal{K}^n$ and $a \in \mathcal{A}$,

$$\begin{aligned}
\left|\psi_0(s,a)^\top\left(\theta_{\text{mid}}^t - \theta^t\right)\right| &\le \|\psi_0(s,a)\|_{\left(\Sigma_{\text{cov}}^{\text{NN},n}\right)^{-1}}\left\|\theta_{\text{mid}}^t - \theta^t\right\|_{\Sigma_{\text{cov}}^{\text{NN},n}} \\
&\overset{\text{(a)}}{\le} \sqrt{2\beta\left((n+1)\varepsilon_Q^{\text{NN}} + 4R^2\zeta_{\text{cov}}\right)},
\end{aligned}$$

where inequality (a) uses Eq. (42).

Hence, we have

$$\begin{aligned}\text{Term 3} &= \mathbb{E}_{(s,a)\sim d^{\pi^*,n}_{\mathcal{M}^n;s_{\text{init}}}}\left[\bar{\psi}_0^t(s,a)^\top\left(\theta^t_{\text{mid}}-\theta^t\right)\cdot\mathbb{1}\left\{s\in\mathcal{K}^n\right\}\right]\\ &\le 2\sqrt{\beta(n+1)\varepsilon_Q^{\text{NN}}}+4R\sqrt{\beta\zeta_{\text{cov}}}.\end{aligned}$$

Term 4. Finally, we bound Term 4 as follows.

$$\begin{aligned}\text{Term 4} &= \mathbb{E}_{(s,a)\sim d^{\pi^*,n}_{\mathcal{M}^n;s_{\text{init}}}}\left[\left(\bar{\psi}_0^t(s,a)-\bar{\psi}^t_{w^t}(s,a)\right)^\top\theta^t\right]\\ &\le \mathbb{E}_{(s,a)\sim d^{\pi^*}_{\mathcal{M};s_{\text{init}}}}\left[\left|\left(\psi_0(s,a)-\psi_{w^t}(s,a)\right)^\top\theta^t\right|\right]\\ &\quad+|\mathcal{A}|\mathbb{E}_{(s,a)\sim d^{\pi^*}_{\mathcal{M};s_{\text{init}}}}\left[\left|\left(\psi_0(s,a')-\psi_{w^t}(s,a')\right)^\top\theta^t\right|\right]\\ &\le \frac{4|\mathcal{A}|\sqrt{c_{\text{scale}}R^3}}{\sqrt{\underline{c}}m^{\frac{1}{4}}}.\end{aligned}$$

The Total Suboptimality. Combining Lemma D.3 and Eq. (43), we have

$$V^{\pi^*}(s_{\text{init}})-V^{\pi^t}(s_{\text{init}})\le\text{RHS in Eq. (43)}+\frac{1}{1-\gamma}\sum_{(s,a)\notin\mathcal{K}^n}d^{\pi^t}_{s_{\text{init}}}(s,a).$$

Summing over $t=0,\dots,T-1$, dividing T and applying the regret for natural policy gradient (Lemma E.7), we have

$$\begin{aligned}&V^{\pi^*}(s_{\text{init}})-V^{\pi^{n+1}}(s_{\text{init}})\\ &=\frac{1}{T}\sum_{t=0}^{T-1}\left(V^{\pi^*}(s_{\text{init}})-V^{\pi^t}(s_{\text{init}})\right)\\ &\le\frac{\log(|\mathcal{A}|)}{(1-\gamma)\eta T}+\frac{\eta W_S(W_\theta^{\text{NN}})^2}{(1-\gamma)}+\frac{2\sqrt{|\mathcal{A}|\varepsilon_{\text{bias}}^{\text{NN}}}}{1-\gamma}+\frac{1}{1-\gamma}8\sqrt{\beta(n+1)W_Q^{\text{NN}}}\cdot\frac{1}{T}\sum_{t=0}^{T-1}\sqrt{\varsigma^{\text{NN},\pi^t}_{\rho^n_{\text{cov}}}}\\ &\quad+\frac{12R\sqrt{\beta\zeta_{\text{cov}}}}{1-\gamma}+\frac{2}{1-\gamma}\sqrt{\beta(n+1)\varepsilon_Q^{\text{NN}}}+\frac{4|\mathcal{A}|\sqrt{c_{\text{scale}}R^3}}{\sqrt{\underline{c}}m^{\frac{1}{4}}}+\frac{1}{1-\gamma}\sum_{(s,a)\notin\mathcal{K}^n}d^{\pi^{n+1}}_{s_{\text{init}}}(s,a).\end{aligned}$$

Next, we handle the term $\frac{1}{T}\sum_{t=0}^{T-1}\sqrt{\varsigma^{\text{NN},\pi^t}_{\rho^n_{\text{cov}}}}$.

$$\begin{aligned}\frac{1}{N}\sum_{n=0}^{N-1}\sqrt{n+1}\cdot\frac{1}{T}\sum_{t=0}^{T-1}\sqrt{\varsigma^{\text{NN},\pi^t}_{\rho^n_{\text{cov}}}}&\le\frac{1}{NT}\sum_{n=0}^{N-1}\sum_{t=0}^{T-1}\sqrt{(n+1)\varsigma^{\text{NN},\pi^t}_{\rho^n_{\text{cov}}}}\\ &\le\frac{1}{NT}\sum_{n=0}^{N-1}\sum_{t=0}^{T-1}\left(\left(\frac{N}{c_{\text{MLE}}^{\text{NN}}}\sqrt{\frac{5md\log\left(\frac{N}{\delta'}\right)}{M_{\text{HF}}}}+\frac{2N}{\sqrt{c_{\text{MLE}}^{\text{NN}}}}\left(17W_\tau R\sqrt{\frac{\log\left(\frac{1}{\delta'}\right)}{M^\mu_{\text{SGD}}}}\right)^{\frac{1}{2}}\right.\right.\\ &\quad\left.+4(n+1)R\sqrt{\frac{\zeta_{\text{HF}}}{n}}+\frac{38Nc_{\text{scale}}^{\frac{1}{4}}R^{\frac{5}{4}}M_{\text{HF}}^{\frac{1}{4}}\sqrt{W_\tau\exp(4W_\tau)}}{\underline{c}^{\frac{1}{4}}c_{\text{MLE}}^{\text{NN}}\sqrt{1-\gamma}m^{\frac{1}{8}}}\log\left(\frac{1}{\delta'}\right)\right)\cdot\\ &\quad\left.\mathbb{E}_{\tau\sim\mathcal{O}^{\pi^t}_{\rho^n_{\text{cov}}}}\left[\left\|\sum_{h=0}^{H(\tau)}\psi_0(s_h,a_h)\right\|_{(\Sigma_{\text{HF}}^{\text{NN},n})^{-1}}\right]+\frac{6N\sqrt{c_{\text{scale}}R^3\log\left(\frac{1}{\delta'}\right)}}{(1-\gamma)\sqrt{\underline{c}}m^{\frac{1}{4}}}\right)^{\frac{1}{2}}\\ &\le\left(\sqrt{\frac{N}{c_{\text{MLE}}^{\text{NN}}}}\left(\frac{5md\log\left(\frac{N}{\delta'}\right)}{M_{\text{HF}}}\right)^{\frac{1}{4}}+\frac{\sqrt{2N}}{\left(c_{\text{MLE}}^{\text{NN}}\right)^{\frac{1}{4}}}\left(17W_\tau R\sqrt{\frac{\log\left(\frac{1}{\delta'}\right)}{M^\mu_{\text{SGD}}}}\right)^{\frac{1}{4}}\right.\end{aligned}$$

$$
\begin{aligned}
&+ 4\sqrt{R}N^{\frac{1}{4}}\zeta_{\text{HF}}^{\frac{1}{4}} + \frac{7\sqrt{N}c_{\text{scale}}^{\frac{1}{8}}R^{\frac{5}{8}}M_{\text{HF}}^{\frac{1}{8}}\left(W_\tau\exp(4W_\tau)\right)^{\frac{1}{4}}}{\underline{c}^{\frac{1}{8}}\sqrt{c_{\text{MLE}}^{\text{NN}}}\left(1-\gamma\right)^{\frac{1}{4}}m^{\frac{1}{16}}}\sqrt{\log\left(\frac{1}{\delta'}\right)}\Bigg)\cdot \\
&\frac{1}{NT}\sum_{n=0}^{N-1}\sum_{t=0}^{T-1}\sqrt{\mathbb{E}_{\tau\sim\mathcal{O}^{\pi^t}_{\rho_{\text{cov}}^n}}\left[\left\|\sum_{h=0}^{H(\tau)}\psi_0(s_h,a_h)\right\|_{(\Sigma_{\text{HF}}^{\text{NN},n})^{-1}}\right]} + \frac{3\sqrt{N}\left(c_{\text{scale}}R^3\log\left(\frac{1}{\delta'}\right)\right)^{\frac{1}{4}}}{\sqrt{1-\gamma}\underline{c}^{\frac{1}{4}}m^{\frac{1}{8}}} \\
&\overset{\text{(a)}}{\leq}\left(\sqrt{\frac{N}{c_{\text{MLE}}^{\text{NN}}}}\left(\frac{5md\log\left(\frac{N}{\delta'}\right)}{M_{\text{HF}}}\right)^{\frac{1}{4}} + \frac{\sqrt{2N}}{\left(c_{\text{MLE}}^{\text{NN}}\right)^{\frac{1}{4}}}\left(17W_\tau R\sqrt{\frac{\log\left(\frac{1}{\delta'}\right)}{M_{\text{SGD}}^{\mu}}}\right)^{\frac{1}{4}}\right. \\
&\quad + 4\sqrt{R}N^{\frac{1}{4}}\zeta_{\text{HF}}^{\frac{1}{4}} + \frac{7\sqrt{N}c_{\text{scale}}^{\frac{1}{8}}R^{\frac{5}{8}}M_{\text{HF}}^{\frac{1}{8}}\left(W_\tau\exp(4W_\tau)\right)^{\frac{1}{4}}}{\underline{c}^{\frac{1}{8}}\sqrt{c_{\text{MLE}}^{\text{NN}}}\left(1-\gamma\right)^{\frac{1}{4}}m^{\frac{1}{16}}}\sqrt{\log\left(\frac{1}{\delta'}\right)}\Bigg)\cdot \\
&\underbrace{\left(2m^{\frac{1}{4}}d^{\frac{1}{4}}\log^{\frac{1}{4}}\left(1+\frac{4NW_\tau^2}{\zeta_{\text{HF}}md}\right)+\frac{2m^{\frac{1}{4}}d^{\frac{1}{4}}\log^{\frac{1}{4}}(N)}{c_{\text{base}}^{\frac{1}{4}}}\right)}_{\tilde{d}_{\text{HF}}} + \frac{3\sqrt{N}\left(c_{\text{scale}}R^3\log\left(\frac{1}{\delta'}\right)\right)^{\frac{1}{4}}}{\sqrt{1-\gamma}\underline{c}^{\frac{1}{4}}m^{\frac{1}{8}}},
\end{aligned}
$$

where inequality (a) uses Lemma D.13 with feature dimension md.

Recall that $\delta' := \frac{\delta}{24N(K+M_{\text{HF}}+M_{\text{SGD}}^{\mu}+TM_{\text{SGD}}^{\theta})}$, $\zeta_{\text{cov}} := 1$, $\zeta_{\text{HF}} := 4W_\tau^2$, $W_S := 1$, $W_\theta^{\text{NN}} := \sqrt{m}\bar{c}+R$, $W_{\nabla F}^{\text{NN}} := \frac{4}{(1-\gamma)^2} + \frac{4(\sqrt{m}\bar{c}+R)}{1-\gamma}$, $W_f^{\text{NN}} := \sqrt{m}\bar{c}+R$, $W_Q^{\text{NN}} := \frac{\sqrt{m}\bar{c}+R}{1-\gamma} + \frac{2}{(1-\gamma)^2}$, $\xi_\theta := \frac{R}{W_{\nabla F}^{\text{NN}}\sqrt{M_{\text{SGD}}^{\theta}}}$, $\xi_\mu := \frac{R}{W_\tau\sqrt{M_{\text{SGD}}^{\mu}}}$, $\eta := \frac{\log(|\mathcal{A}|)}{W_\theta^{\text{NN}}\sqrt{W_S T}}$ and $c_{\text{MLE}}^{\text{NN}} := (2+\exp(-2W_\tau(\sqrt{m}\bar{c}+R))+\exp(2W_\tau(\sqrt{m}\bar{c}+R)))^{-1}$. K and M_{HF} should satisfy that $K \geq \frac{16(N+1)^2\log^2\left(\frac{4dN}{\delta'}\right)}{\zeta_{\text{cov}}^2}$ and $M_{\text{HF}} \geq \frac{16W_\tau^4\log^2\left(\frac{4dN}{\delta'}\right)}{\zeta_{\text{HF}}^2}$, respectively.

Therefore, summing over $n = 0, \ldots, N-1$, dividing N, and applying Lemma D.4, we have

$$
\begin{aligned}
&V^{\pi^*}(s_{\text{init}}) - V^{\pi^{\text{out}}}(s_{\text{init}}) \\
&= \frac{1}{N}\sum_{n=0}^{N-1}\left(V^{\pi^*}(s_{\text{init}}) - V^{\pi^{n+1}}(s_{\text{init}})\right) \\
&\leq \frac{2\sqrt{|\mathcal{A}|\varepsilon_{\text{bias}}^{\text{NN}}}}{1-\gamma} + \underbrace{\frac{\log(|\mathcal{A}|)}{(1-\gamma)\eta T} + \frac{\eta W_S(W_\theta^{\text{NN}})^2}{(1-\gamma)}}_{=\frac{W_\theta^{\text{NN}}\sqrt{W_S\log(|\mathcal{A}|)}}{(1-\gamma)\sqrt{T}}} + \frac{12R\sqrt{\beta\zeta_{\text{cov}}}}{1-\gamma} + \frac{2md}{(1-\gamma)N\beta}\log\left(1+\frac{N}{\zeta_{\text{cov}}md}\right) \\
&\quad + \frac{8\sqrt{\beta NW_{\nabla F}^{\text{NN}}R}}{1-\gamma}\left(\frac{\log\left(\frac{1}{\delta'}\right)}{M_{\text{SGD}}^{\theta}}\right)^{\frac{1}{4}} + \frac{8\tilde{d}_{\text{HF}}\sqrt{\beta W_Q^{\text{NN}}}}{1-\gamma}\left(\frac{4\sqrt{N}m^{\frac{1}{4}}d^{\frac{1}{4}}\log^{\frac{1}{4}}\left(\frac{N}{\delta'}\right)}{\sqrt{c_{\text{MLE}}^{\text{NN}}}M_{\text{HF}}^{\frac{1}{4}}} + \frac{5\sqrt{N}W_\tau^{\frac{1}{4}}R^{\frac{1}{4}}\log^{\frac{1}{8}}\left(\frac{1}{\delta'}\right)}{(c_{\text{MLE}}^{\text{NN}})^{\frac{1}{4}}(M_{\text{SGD}}^{\mu})^{\frac{1}{8}}}\right. \\
&\quad + 4\sqrt{R}N^{\frac{1}{4}}\zeta_{\text{HF}}^{\frac{1}{4}} + \frac{9\sqrt{N}c_{\text{scale}}^{\frac{1}{8}}R^{\frac{5}{8}}M_{\text{HF}}^{\frac{1}{8}}\left(W_\tau\exp(4W_\tau)\right)^{\frac{1}{4}}}{\underline{c}^{\frac{1}{8}}\sqrt{c_{\text{MLE}}^{\text{NN}}}\left(1-\gamma\right)^{\frac{1}{4}}m^{\frac{1}{16}}}\sqrt{\log\left(\frac{1}{\delta'}\right)}\Bigg) \\
&\quad + \frac{8\sqrt{\beta W_Q^{\text{NN}}}}{1-\gamma}\cdot\frac{3\sqrt{N}\left(c_{\text{scale}}R^3\log\left(\frac{1}{\delta'}\right)\right)^{\frac{1}{4}}}{\sqrt{1-\gamma}\underline{c}^{\frac{1}{4}}m^{\frac{1}{8}}} + \frac{16R\sqrt{\beta N(W_f^{\text{NN}}+W_Q^{\text{NN}})}c_{\text{scale}}^{\frac{1}{4}}R^{\frac{1}{4}}}{(1-\gamma)\underline{c}^{\frac{1}{4}}m^{\frac{1}{8}}} + \frac{4|\mathcal{A}|\sqrt{c_{\text{scale}}R^3}}{\sqrt{\underline{c}}m^{\frac{1}{4}}}.
\end{aligned} \tag{44}
$$

□

F. Technical Tools

Lemma F.1. *Let $\mathcal{D}$ be a distribution of random vector $\phi \in \mathbb{R}^d$ such that $\|\phi\|_2 \leq W$ and $\Sigma = \mathbb{E}_{\phi\sim\mathcal{D}}[\phi\phi^\top]$. Given K i.i.d. samples $\phi_1, \ldots, \phi_K \sim \mathcal{D}$, then with probability at least $1-\delta'$,*

$$\Pr\left[\left\|\frac{1}{K}\sum_{i=1}^{K}\phi_i\phi_i^\top - \Sigma\right\| \leq \frac{2W^2\log\left(\frac{4d}{\delta'}\right)}{\sqrt{K}}\right].$$

Proof. This analysis is originated from Lemma H.3 in (Agarwal et al., 2020).

Let $X_i = \phi_i\phi_i^\top - \Sigma$, and it holds that $\mathbb{E}[X_i] = 0$ and $\|X_i\| \leq W^2$.

Then, we have

$$\begin{aligned}
\mathbb{E}\left[X_i^2\right] &= \mathbb{E}\left[\left(\phi_i\phi_i^\top - \Sigma\right)^2\right] \\
&= \mathbb{E}\left[\left(\phi_i\phi_i^\top\right)^2 + \Sigma^2 - 2\Sigma\phi_i\phi_i^\top\right] \\
&= \mathbb{E}\left[\left(\phi_i\phi_i^\top\right)^2\right] + \Sigma^2 - 2\Sigma\mathbb{E}\left[\phi_i\phi_i^\top\right] \\
&= \mathbb{E}\left[\left(\phi_i\phi_i^\top\right)^2\right] + \Sigma^2 - 2\Sigma^2 \\
&= \mathbb{E}\left[\left(\phi_i\phi_i^\top\right)^2\right] - \Sigma^2.
\end{aligned}$$

For any $x \in \mathbb{R}^d$,

$$x^\top\left(\mathbb{E}\left[\left(\phi_i\phi_i^\top\right)^2\right] - \mathbb{E}\left[X_i^2\right]\right)x = x^\top\Sigma^2 x = \left(\Sigma x\right)^\top\Sigma x \geq 0,$$

which implies

$$\mathbb{E}\left[\left(\phi_i\phi_i^\top\right)^2\right] \succeq \mathbb{E}\left[X_i^2\right].$$

For any $x \in \mathbb{R}^d$,

$$\begin{aligned}
x^\top\left(W^2\mathbb{E}\left[\phi_i\phi_i^\top\right] - \mathbb{E}\left[\left(\phi_i\phi_i^\top\right)^2\right]\right)x &= W^2\cdot x^\top\mathbb{E}\left[\phi_i\phi_i^\top\right]x - x^\top\mathbb{E}\left[\left(\phi_i\phi_i^\top\right)\left(\phi_i\phi_i^\top\right)\right]x \\
&\geq W^2\cdot x^\top\mathbb{E}\left[\phi_i\phi_i^\top\right]x - W^2\cdot x^\top\mathbb{E}\left[\phi_i\phi_i^\top\right]x \\
&= 0,
\end{aligned}$$

which implies

$$W^2\mathbb{E}\left[\phi_i\phi_i^\top\right] \succeq \mathbb{E}\left[\left(\phi_i\phi_i^\top\right)^2\right].$$

Then, we have

$$\mathbb{E}\left[X_i^2\right] \preceq \mathbb{E}\left[\left(\phi_i\phi_i^\top\right)^2\right] \preceq W^2\mathbb{E}\left[\phi_i\phi_i^\top\right] = W^2\Sigma,$$

and thus,

$$\begin{aligned}
\sum_{i=1}^{K}\mathbb{E}\left[X_i^2\right] &\preceq KW^2\Sigma, \\
\left\|\sum_{i=1}^{K}\mathbb{E}\left[X_i^2\right]\right\| &\leq KW^4.
\end{aligned}$$

Using the Matrix Bernstein inequality (Theorem 7.7.1 in (Tropp et al., 2015)), we have that for any $t \geq W^2\sqrt{K} + \frac{1}{3}W^2$,

$$\Pr\left[\left\|\sum_{i=1}^{K} X_i\right\| \geq t\right] \leq 4d \exp\left(\frac{-\frac{1}{2}t^2}{W^4 K + \frac{1}{3}W^2 t}\right),$$

which is equivalent to that for any $z \geq \frac{W^2}{\sqrt{K}} + \frac{1}{3}\frac{W^2}{K}$,

$$\Pr\left[\left\|\frac{1}{K}\sum_{i=1}^{K} X_i\right\| \geq z\right] \leq 4d \exp\left(\frac{-\frac{1}{2}K^2 z^2}{W^4 K + \frac{1}{3}W^2 K z}\right).$$

Let

$$z = \frac{2W^2 \log\left(\frac{4d}{\delta'}\right)}{\sqrt{K}}.$$

Then,

$$\begin{aligned}
&\Pr\left[\left\|\frac{1}{K}\sum_{i=1}^{K} X_i\right\| \geq \frac{2W^2 \log\left(\frac{4d}{\delta'}\right)}{\sqrt{K}}\right] \\
&\leq 4d \exp\left(\frac{-\frac{1}{2}K^2 \cdot \frac{4W^4 \log^2\left(\frac{4d}{\delta'}\right)}{K}}{W^4 K + \frac{1}{3}WK \cdot \frac{2W^2 \log\left(\frac{4d}{\delta'}\right)}{\sqrt{K}}}\right) \\
&= 4d \exp\left(-\frac{2W^4 K \log^2\left(\frac{4d}{\delta'}\right)}{W^4 K + \frac{2}{3}W^3\sqrt{K}\log\left(\frac{4d}{\delta'}\right)}\right) \\
&\leq 4d \exp\left(-\log\left(\frac{4d}{\delta'}\right)\right) \\
&= \delta'.
\end{aligned}$$

Thus, with probability at least $1 - \delta'$,

$$\Pr\left[\left\|\frac{1}{K}\sum_{i=1}^{K} \phi_i \phi_i^\top - \Sigma\right\| \leq \frac{2W^2 \log\left(\frac{4d}{\delta'}\right)}{\sqrt{K}}\right].$$

□

Lemma F.2. *For any $n \in [N]$, let $\mathcal{D}^n$ be a distribution of random vector $\phi \in \mathbb{R}^d$ such that $\|\phi\|_2 \leq W$, and define $\Sigma^n = \mathbb{E}_{\phi\sim\mathcal{D}^n}[\phi\phi^\top]$ and $\Sigma = \sum_{n=1}^{N}\Sigma^n$. For any $n \in [N]$, given K i.i.d. samples $\phi_1^n, \ldots, \phi_K^n \sim \mathcal{D}^n$, let $\hat{\Sigma}^n = \frac{1}{K}\sum_{j=1}^{K}\phi_j^n(\phi_j^n)^\top$ and $\hat{\Sigma} = \sum_{n=1}^{N}\hat{\Sigma}^n$. Letting $K \geq \frac{16N^2W^4\log^2\left(\frac{4dN}{\delta'}\right)}{\zeta^2}$, then with probability at least $1-\delta'$, we have that for any $x \in \mathbb{R}^d$,*

$$\frac{1}{2}x^\top(\Sigma + \zeta I)^{-1}x \leq x^\top\left(\hat{\Sigma} + \zeta I\right)^{-1}x \leq 2x^\top(\Sigma + \zeta I)^{-1}x.$$

Proof. This proof is originated from Lemma H.4 in (Agarwal et al., 2020).

According to Lemma F.1, we have that for any $n \in [N]$, with probability at least $1 - \frac{\delta'}{N}$,

$$\Pr\left[\left\|\frac{1}{K}\sum_{j=1}^{K}\phi_j^n(\phi_j^n)^\top - \Sigma^n\right\| \leq \frac{2W^2\log\left(\frac{4dN}{\delta'}\right)}{\sqrt{K}}\right].$$

Thus,

$$\Sigma^n - \frac{2W^2 \log\left(\frac{4dN}{\delta'}\right)}{\sqrt{K}} I \preceq \frac{1}{K}\sum_{j=1}^{K} \phi_j^n (\phi_j^n)^\top \preceq \Sigma^n + \frac{2W^2 \log\left(\frac{4dN}{\delta'}\right)}{\sqrt{K}} I,$$

and then summing over $n \in [N]$, we have

$$\Sigma - \frac{2NW^2 \log\left(\frac{4dN}{\delta'}\right)}{\sqrt{K}} I + \zeta I \preceq \hat{\Sigma} + \zeta I \preceq \Sigma + \frac{2NW^2 \log\left(\frac{4dN}{\delta'}\right)}{\sqrt{K}} I + \zeta I.$$

This implies that for $\zeta \geq \frac{2NW^2 \log\left(\frac{4dN}{\delta'}\right)}{\sqrt{K}}$,

$$\left(\Sigma + \frac{2NW^2 \log\left(\frac{4dN}{\delta'}\right)}{\sqrt{K}} I + \zeta I\right)^{-1} \preceq \left(\hat{\Sigma} + \zeta I\right)^{-1} \preceq \left(\Sigma - \frac{2NW^2 \log\left(\frac{4dN}{\delta'}\right)}{\sqrt{K}} I + \zeta I\right)^{-1}.$$

Let $U\Lambda U^\top$ be the eigendecomposition of Σ, where $\Lambda = diag([\lambda_1, \ldots, \lambda_d])$ and $U = [u_1, \ldots, u_d]$. Then, we have

$$\begin{aligned}
&x^\top \left(\hat{\Sigma} + \zeta I\right)^{-1} x - x^\top (\Sigma + \zeta I)^{-1} x \\
&\leq x^\top \left(\Sigma - \frac{2NW^2 \log\left(\frac{4dN}{\delta'}\right)}{\sqrt{K}} I + \zeta I\right)^{-1} x - x^\top (\Sigma + \zeta I)^{-1} x \\
&= \sum_{i \in [d]} \left(\left(\sigma_i + \zeta - \frac{2NW^2 \log\left(\frac{4dN}{\delta'}\right)}{\sqrt{K}}\right)^{-1} - (\sigma_i + \zeta)^{-1} \right) (u_i x)^2 .
\end{aligned}$$

Since $\zeta \geq \frac{4NW^2 \log\left(\frac{4dN}{\delta'}\right)}{\sqrt{K}}$, we have

$$\begin{aligned}
2\left(\sigma_i + \zeta - \frac{2NW^2 \log\left(\frac{4dN}{\delta'}\right)}{\sqrt{K}}\right) &= \sigma_i + \zeta - \frac{4NW^2 \log\left(\frac{4dN}{\delta'}\right)}{\sqrt{K}} + \sigma_i + \zeta \\
&\geq \sigma_i + \zeta,
\end{aligned}$$

and thus

$$\left(\sigma_i + \zeta - \frac{2NW^2 \log\left(\frac{4dN}{\delta'}\right)}{\sqrt{K}}\right)^{-1} \leq 2(\sigma_i + \zeta)^{-1}.$$

Hence,

$$\begin{aligned}
x^\top \left(\hat{\Sigma} + \zeta I\right)^{-1} x - x^\top (\Sigma + \zeta I)^{-1} x &\leq \sum_{i \in [d]} (\sigma_i + \zeta)^{-1} (u_i x)^2 \\
&= x^\top (\Sigma + \zeta I)^{-1} x.
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
&x^\top (\Sigma + \zeta I)^{-1} x - x^\top \left(\hat{\Sigma} + \zeta I\right)^{-1} x \\
&\leq x^\top (\Sigma + \zeta I)^{-1} x - x^\top \left(\Sigma + \frac{2NW^2 \log\left(\frac{4dN}{\delta'}\right)}{\sqrt{K}} I + \zeta I\right)^{-1} x
\end{aligned}$$

$$= \sum_{i\in[d]} \left((\sigma_i + \zeta)^{-1} - \left(\sigma_i + \zeta + \frac{2NW^2 \log\left(\frac{4dN}{\delta'}\right)}{\sqrt{K}} \right)^{-1} \right) (u_i x)^2 .$$

Since $\zeta \geq \frac{2NW^2 \log\left(\frac{4dN}{\delta'}\right)}{\sqrt{K}}$, we have

$$\begin{aligned} 2\left(\sigma_i + \zeta\right) &= \sigma_i + \zeta + \sigma_i + \zeta \\ &\geq \sigma_i + \zeta + \frac{2NW^2 \log\left(\frac{4dN}{\delta'}\right)}{\sqrt{K}}, \end{aligned}$$

and thus

$$(\sigma_i + \zeta)^{-1} \leq 2 \left(\sigma_i + \zeta + \frac{2NW^2 \log\left(\frac{4dN}{\delta'}\right)}{\sqrt{K}} \right)^{-1} .$$

Hence,

$$\begin{aligned} & x^\top \left(\Sigma + \zeta I\right)^{-1} x - x^\top \left(\hat{\Sigma} + \zeta I\right)^{-1} x \\ \leq & \sum_{i\in[d]} \left(\sigma_i + \zeta + \frac{2NW^2 \log\left(\frac{4dN}{\delta'}\right)}{\sqrt{K}} \right)^{-1} (u_i x)^2 \\ = & \, x^\top \left(\hat{\Sigma} + \zeta I\right)^{-1} x. \end{aligned}$$

□

Lemma F.3. *For any $a, b, c \in \mathbb{R}$, we have*

$$(b-a)^2 - (c-a)^2 \leq 4 \max\{|a|, |b|, |c|\} |b-c|.$$

Proof. It holds that

$$\begin{aligned} (b-a)^2 - (c-a)^2 &= (a^2 + b^2 - 2ab) - (a^2 + c^2 - 2ac) \\ &= b^2 - c^2 - 2a(b-c) \\ &= (b+c)(b-c) - 2a(b-c) \\ &= (b + c - 2a)(b-c) \\ &\leq 4 \max\{|a|, |b|, |c|\} |b-c|. \end{aligned}$$

□

Lemma F.4 (Theorem 2.1 in (Hsu et al., 2012))**.** *Let $A \in \mathbb{R}^{n\times n}$ be a matrix, and let $\Sigma := A^\top A$. Suppose that $x = (x_1, \ldots, x_n)$ is a random vector such that, for some $\nu \in \mathbb{R}^n$ and $\sigma \geq 0$,*

$$\mathbb{E}\left[\exp\left(\alpha^\top (x - \nu)\right)\right] \leq \exp\left(\frac{\|\alpha\|^2 \sigma^2}{2}\right)$$

for all $\alpha \in \mathbb{R}^n$. For all $t > 0$,

$$\Pr\left[\|Ax\|^2 > \sigma^2 \left(\mathrm{tr}(\Sigma) + 2\sqrt{\mathrm{tr}(\Sigma^2) t} + 2\|\Sigma\| t \right) + \mathrm{tr}(\Sigma \nu \nu^\top) \left(1 + 2 \left(\frac{\|\Sigma\|^2}{\mathrm{tr}\left(\Sigma^2\right)} t \right)^{\frac{1}{2}} \right) \right] \leq \exp(-t).$$

Lemma F.5 (Lemma 11 in (Abbasi-Yadkori et al., 2011)). *Let $X_1, \ldots, X_N$ be a sequence of $d \times d$-dimensional positive semi-definite matrices, and $\|X_n\| \leq W_x$ for all $n \in [N]$. Let $A_0 = \zeta I_d$ with $\zeta \geq \max\{1, W_x\}$. For any $n \in [N]$, let $A_n = A_0 + \sum_{i=1}^{n} X_i$. Then, we have*

$$\sum_{n=1}^{N} \text{tr}\left(A_{n-1}^{-1} X_n\right) \leq 2 \log\left(\frac{\det(A_N)}{\det(A_0)}\right).$$