Randomized Exploration in Cooperative Multi-Agent Reinforcement Learning

Hao-Lun Hsu[*‡], **Weixin Wang**[†‡], **Miroslav Pajic**[§], **Pan Xu**[¶]

Abstract

We present the first study on provably efficient randomized exploration in cooperative multi-agent reinforcement learning (MARL). We propose a unified algorithm framework for randomized exploration in parallel Markov Decision Processes (MDPs), and two Thompson Sampling (TS)-type algorithms, CoopTS-PHE and CoopTS-LMC, incorporating the perturbed-history exploration (PHE) strategy and the Langevin Monte Carlo exploration (LMC) strategy respectively, which are flexible in design and easy to implement in practice. For a special class of parallel MDPs where the transition is (approximately) linear, we theoretically prove that both CoopTS-PHE and CoopTS-LMC achieve a $\widetilde{\mathcal{O}}(d^{3/2}H^2\sqrt{MK})$ regret bound with communication complexity $\widetilde{\mathcal{O}}(dHM^2)$, where d is the feature dimension, H is the horizon length, M is the number of agents, and K is the number of episodes. This is the first theoretical result for randomized exploration in cooperative MARL. We evaluate our proposed method on multiple parallel RL environments, including a deep exploration problem (*i.e.,* N-chain), a video game, and a real-world problem in energy systems. Our experimental results support that our framework can achieve better performance, even under conditions of misspecified transition models. Additionally, we establish a connection between our unified framework and the practical application of federated learning.

1 Introduction

Multi-agent Reinforcement Learning (MARL) has emerged as a potent tool with wide-ranging applications in diverse fields including robotics (Ding et al., 2020; Liu et al., 2019), gaming (Tsay et al., 2011; Zhao et al., 2019; Ye et al., 2020), and numerous real-world systems (Bazzan, 2009; Fei & Xu, 2022; Yeh et al., 2023). This is particularly evident in cooperative scenarios, where the effectiveness of MARL is enhanced through the implementation of both direct and indirect communication channels among agents. This requires MARL algorithms to adeptly and flexibly coordinate communications to optimize the benefits of cooperation. One of the classical challenges in MARL is to balance between exploration and exploitation, i.e., ensuring that agents not only effectively utilize existing information but also acquire new knowledge. Recent literature, such as the works on cooperative exploration strategies (Hao et al., 2023) and on dynamic exploitation tactics (Rojas-Córdova et al., 2023) in MARL highlight the intricacies and importance of balancing such trade-off. As indicated by Hao et al. (2023); Chalkiadakis & Boutilier (2003); Liu et al. (2023), achieving this equilibrium is crucial for the practical deployment of MARL systems in real-world scenarios, where unpredictability and the need for rapid adaptation are prevalent.

Optimism in the Face of Uncertainty (OFU) is a popular strategy to address the exploration-exploitation problem (Abbasi-Yadkori et al., 2011). OFU strategy leads to numerous upper confi-

[*]Duke University; email: hao-lun.hsu@duke.edu
[†]Duke University; email: weixin.wang@duke.edu
[‡]Equal contribution
[§]Duke University; email: miroslav.pajic@duke.edu
[¶]Duke University; email: pan.xu@duke.edu

dence bound (UCB)-type algorithms in contextual bandits (Chu et al., 2011; Abbasi-Yadkori et al., 2011; Li et al., 2017), single-agent reinforcement learning (Jin et al., 2020; Wang et al., 2020a), and more recently multi-agent reinforcement learning (Dubey & Pentland, 2021; Min et al., 2023). These algorithms compute statistical confidence regions for the model or the value function, given the observed history, and perform the greedy policy with respect to these regions, or upper confidence bounds. Though UCB-based methods give out strong theoretical results, they often have poor performance in practice (Osband et al., 2013; Osband & Van Roy, 2017). For example, Wang et al. (2020a) demonstrates that computing the confidence bonus necessitates advanced sensitivity sampling and the expensive computation makes the practical applications inefficient. It is worth noting that UCB is mostly constructed based on a linear structure (Chu et al., 2011; Jin et al., 2020). NeuralUCB is a notable attempt at a nonlinear version while it is infeasible in terms of computational complexity (Zhou et al., 2020; Xu et al., 2021).

Inspired by Thompson Sampling (TS) (Thompson, 1933), posterior sampling for reinforcement learning (RL) (Agrawal & Jia, 2017; Zhou et al., 2019) involves maintaining a posterior distribution over the parameters of the Markov Decision Processes (MDP) model parameters. Although conceptually simple, most existing TS methods require the exact posterior or a good Laplacian approximation (Xu et al., 2022). Recently, there have been advancements in randomized exploration with approximate sampling. One important method is perturb-history exploration (PHE) strategy, which involves introducing random perturbations in the action history of the agent (Kveton et al., 2019; 2020b; Ishfaq et al., 2021). This randomized exploration approach diversifies the agent's experience, aiding in learning more robust strategies in environments with uncertainty and variability. Another effective method is Langevin Monte Carlo (LMC) method (Xu et al., 2022; Ishfaq et al., 2024; Huix et al., 2023; Karbasi et al., 2023; Mousavi-Hosseini et al., 2023). Notably, Ishfaq et al. (2024) maintains the simplicity and scalability of LMC, making it applicable in deep RL algorithms by approximating the posterior distribution of the Q function.

Despite the aforementioned advancements of randomized exploration in bandits and single-agent RL, there remains a scarcity of research on randomized exploration within cooperative MARL, which motivates us to present the first investigation into provably efficient randomized exploration in cooperative MARL, with both theoretical and empirical evidence. We specifically focus on the applicability in parallel MDPs, aiming to facilitate faster learning and to improve policy optimization with the same state and action spaces, allowing for leveraging similarities across MDPs. We theoretically and empirically demonstrate that randomized exploration strategies can be extended to the multi-agent setting and the benefit of randomized exploration instead of UCB can be significant from single-agent to multi-agent setting.

In summary, **our contributions** are as follows:

- We propose a unified algorithm framework for learning parallel MDPs, and apply two TS-related strategies PHE and LMC for exploration, which leads to the CoopTS-PHE and CoopTS-LMC algorithms. Different from conventional TS where the computation of posterior is expensive and the Laplace approximation causes sampling errors (Riquelme et al., 2018; Kveton et al., 2020a), our proposed algorithms only require adding standard Gaussian noises to the dataset (CoopTS-PHE) or to the gradient (CoopTS-LMC) when performing Least-Square Value Iteration, which is efficient in computation and avoids sampling bias due to the Laplace approximation. Notably, both algorithms can be easily implemented with deep neural networks which are more practical than UCB-based algorithms in deep MARL.
- When reduced to linear parallel MDPs, we theoretically prove that both CoopTS-PHE and CoopTS-LMC with linear function approximation can achieve a regret bound $\widetilde{\mathcal{O}}\big(d^{3/2}H^2\sqrt{M}\big(\sqrt{dM\gamma}+\sqrt{K}\big)\big)$ with communication complexity $\widetilde{\mathcal{O}}\big((d+K/\gamma)MH\big)$, where d is the feature dimension, H is the horizon length, M is the number of agents, K is the number of episodes for each agent, and γ is a parameter controlling the communication frequency. When $\gamma=\mathcal{O}(K/dM)$, our algorithms attain $\widetilde{\mathcal{O}}\big(d^{3/2}H^2\sqrt{MK}\big)$ regret with $\widetilde{\mathcal{O}}(dHM^2)$ communication complexity. This result matches the best communication complexity in cooperative MARL (Min et al., 2023), and the best regret bounds for randomized RL in the single-agent setting ($M=1$)

Table 1: Comparison on episodic, non-stationary, linear MDPs. We define the average regret as the cumulative regret divided by the total number of samples used by the algorithm. Here d is the feature dimension, H is the episode length, K is the number of episodes, and M is the number of agents in a multi-agent setting.

Setting	Algorithm	Regret	Average Regret	Randomized Exploration	Generalizable to Deep RL	Communication Complexity
single-agent	OPT-RLSVI (Zanette et al., 2020)	$\widetilde{\mathcal{O}}(d^2H^{\frac{5}{2}}\sqrt{K})$	$\widetilde{\mathcal{O}}(d^{\frac{3}{2}}H^{\frac{3}{2}}\sqrt{1/K})$	✓	✗	–
	LSVI-UCB (Jin et al., 2020)	$\widetilde{\mathcal{O}}(d^{\frac{3}{2}}H^2\sqrt{K})$	$\widetilde{\mathcal{O}}(d^{\frac{3}{2}}H\sqrt{1/K})$	✗	✗	–
	LSVI-PHE (Ishfaq et al., 2021)	$\widetilde{\mathcal{O}}(d^{\frac{3}{2}}H^2\sqrt{K})$	$\widetilde{\mathcal{O}}(d^{\frac{3}{2}}H\sqrt{1/K})$	✓	✓	–
	LMC-LSVI (Ishfaq et al., 2024)	$\widetilde{\mathcal{O}}(d^{\frac{3}{2}}H^2\sqrt{K})$	$\widetilde{\mathcal{O}}(d^{\frac{3}{2}}H\sqrt{1/K})$	✓	✓	–
multi-agent	Coop-LSVI (Dubey & Pentland, 2021)	$\widetilde{\mathcal{O}}(d^{\frac{3}{2}}H^2\sqrt{MK})$	$\widetilde{\mathcal{O}}(d^{\frac{3}{2}}H\sqrt{1/MK})$	✗	✗	dHM^3
	Asyn-LSVI (Min et al., 2023)	$\widetilde{\mathcal{O}}(d^{\frac{3}{2}}H^2\sqrt{K})$	$\widetilde{\mathcal{O}}(d^{\frac{3}{2}}H\sqrt{1/K})$	✗	✗	dHM^2
	CoopTS-PHE (Ours)	$\widetilde{\mathcal{O}}(d^{\frac{3}{2}}H^2\sqrt{MK})$	$\widetilde{\mathcal{O}}(d^{\frac{3}{2}}H\sqrt{1/MK})$	✓	✓	dHM^2
	CoopTS-LMC (Ours)	$\widetilde{\mathcal{O}}(d^{\frac{3}{2}}H^2\sqrt{MK})$	$\widetilde{\mathcal{O}}(d^{\frac{3}{2}}H\sqrt{1/MK})$	✓	✓	dHM^2

(Ishfaq et al., 2021; 2024). A comprehensive comparison with baseline algorithms on episodic, non-sationary, linear MDPs is presented in Table 1.

- We further extend our theoretical analysis to the misspecified setting where both the transition and reward are approximately linear up to an error ζ and the MDPs could be heterogeneous across agents, which is a generalized notion of misspecification (Jin et al., 2020). We theoretically prove when $\zeta = \mathcal{O}\big(\sqrt{d/MK}\big)$, the cumulative regret for CoopTS-PHE matches the result in the linear homogeneous MDP setting. Simultaneously, when $\zeta = \mathcal{O}\big(\sqrt{1/MK}\big)$, the cumulative regret for CoopTS-LMC matches the result in the linear homogeneous MDP setting. This result indicates that CoopTS-PHE has a slightly higher tolerance on the model misspecification than CoopTS-LMC.
- We conduct extensive experiments on various benchmarks with comprehensive ablation studies, including N-chain that requires deep exploration, Super Mario Bros task in a misspecified setting, and a real-world problem in thermal control of building energy systems. Through empirical evaluation, we demonstrate that our randomized exploration strategies outperform existing deep Q-network (DQN)-based baselines. We also show that our random exploration strategies in cooperative MARL can be adapted to the existing federated RL framework when data transitions are not shared.

2 Preliminary

In parallel Markov Decision Processes (MDPs), M agents interact independently with their respective discrete-time MDPs. All agents have the same but independent state and action spaces. Each agent might have its unique reward functions and transition probabilities. Specifically, for agent $m \in \mathcal{M}$, the associated MDP is defined by the tuple $\text{MDP}(\mathcal{S}, \mathcal{A}, \text{H}, \mathbb{P}_m, \mathbf{r}_m)$. Here $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces respectively. The transition probabilities for agent m are denoted by $\mathbb{P}_m = \{\mathbb{P}_{m,h}\}_{h\in[H]}$, where $\mathbb{P}_{m,h} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is the transition probability at step h, and H is the horizon length. $\mathbf{r}_m = \{r_{m,h}\}_{h\in[H]}$ are the reward functions, with $r_{m,h} : \mathcal{S} \times \mathcal{A} \to [0, 1]$. The policy for agent m is represented by $\pi_m = \{\pi_{m,h}\}_{h\in[H]}$, consisting of a set of H functions where $\pi_{m,h} : \mathcal{S} \to \mathcal{A}$.

In each episode $k = 1, 2, \ldots,$ each agent $m \in \mathcal{M}$ adopts a fixed policy $\pi^k_m = \{\pi^k_{m,h}\}_{h\in[H]}$ and begins in an initial state $s^k_{m,1}$ chosen arbitrarily by the environment. During each step $h \in [H]$ in this episode, the agent observes its current state $s^k_{m,h}$, selects an action $a^k_{m,h} \sim \pi^k_{m,h}(\cdot|s^k_{m,h})$, receives a reward $r_{m,h}(s^k_{m,h}, a^k_{m,h})$, and then transitions to the next state $s^k_{m,h+1}$ based on the transition probability $\mathbb{P}_{m,h}(\cdot|s^k_{m,h}, a^k_{m,h})$. The reward defaults to 0 when the episode terminates at step $H + 1$.

The effectiveness of any policy π in the mth MDP is evaluated by the value function $V^\pi_{m,h}(s) : \mathcal{S} \to \mathbb{R}$, for any $s \in \mathcal{S}, h \in [H], m \in \mathcal{M}$, which is defined as follows $V^\pi_{m,h}(s) := \mathbb{E}_\pi[\sum_{h'=h}^H r_{m,h'}(s_{m,h'}, a_{m,h'})|s_{m,h} = s]$. Furthermore, the state-action value func-

tion $Q^{\pi}_{m,h}(s,a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, which calculates the total expected reward from an action-state pair at step h for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ in the mth MDP, is defined as follows $Q^{\pi}_{m,h}(s,a) := \mathbb{E}_{\pi}[\sum_{h'=h}^{H} r_{m,h'}(s_{m,h'}, a_{m,h'}) | s_{m,h} = s, a_{m,h} = a]$. The optimal policy for the mth MDP is denoted as π^*_m, and we denote $V^*_{m,h}(s) = V^{\pi^*_m}_{m,h}(s)$. The cumulative group regret for total K episodes is defined as

$$\text{Regret}(K) = \sum_{m \in \mathcal{M}} \sum_{k=1}^{K} \left[V^*_{m,1}\left(s^k_{m,1}\right) - V^{\pi_{m,k}}_{m,1}\left(s^k_{m,1}\right) \right].$$

3 Algorithm Design

In this section, we first present a unified algorithm framework for conducting randomized exploration in cooperative MARL. Then we introduce two practical randomized exploration strategies.

3.1 Unified Algorithm Framework

A unified algorithm framework is presented in Algorithm 1. The core idea is that each agent executes Least-Square Value Iteration (LSVI) in parallel and makes decisions based on collective data obtained from communication between each agent and the server. Before we describe the details of our algorithm, we first define notations about the datasets stored on each agent's local machine and the server.

Algorithm 1 Unified Algorithm Framework for Randomized Exploration in Parallel MDPs

Initialization: set $U^{\text{ser}}_h(k), U^{\text{loc}}_{m,h}(k) = \emptyset$.
for episode $k = 1, ..., K$ **do**
  for agent $m \in \mathcal{M}$ **do**
    Receive initial state $s^k_{m,1}$.
    $V^k_{m,H+1}(\cdot) \leftarrow 0$.
    $\{Q^k_{m,h}(\cdot,\cdot)\}_{h=1}^{H} \leftarrow$ **Randomized Exploration** ◁ Algorithm 2 or Algorithm 3
    for step $h = 1, ..., H$ **do**
      $a^k_{m,h} \leftarrow \text{argmax}_{a \in \mathcal{A}} Q^k_{m,h}(s^k_{m,h}, a)$.
      Receive $s^k_{m,h+1}$ and r_h.
      $U^{\text{loc}}_{m,h}(k) \leftarrow U^{\text{loc}}_{m,h}(k) \bigcup \left(s^k_{m,h}, a^k_{m,h}, s^k_{m,h+1}\right)$.
      if Condition **then**
        SYNCHRONIZE $\leftarrow$ True.
      end if
    end for
  end for
  if SYNCHRONIZE **then**
    for step $h = H, ..., 1$ **do**
      $\forall$ *AGENT*: Send $U^{\text{loc}}_{m,h}(k)$ to *SERVER*.
      SERVER: $U^{\text{loc}}_h(k) \leftarrow \bigcup_{m \in \mathcal{M}} U^{\text{loc}}_{m,h}(k)$.
      SERVER: $U^{\text{ser}}_h(k) \leftarrow U^{\text{ser}}_h(k) \bigcup U^{\text{loc}}_h(k)$.
      SERVER: Send $U^{\text{ser}}_h(k)$ to each *AGENT*.
      $\forall$ *AGENT*: Set $U^{\text{loc}}_{m,h}(k) \leftarrow \emptyset$.
    end for
  end if
end for

Index notation We define $k_s(k)$ (denoted as k_s when no ambiguity arises) as the last episode before episode k where synchronization happens. For episode k and step h, we define three datasets

as follows.

$$U_h^{\text{ser}}(k) = \left\{\left(s_{n,h}^{\tau}, a_{n,h}^{\tau}, s_{n,h+1}^{\tau}\right)\right\}_{n\in\mathcal{M},\tau\in[k_s]}, \tag{3.1a}$$

$$U_{m,h}^{\text{loc}}(k) = \left\{\left(s_{m,h}^{\tau}, a_{m,h}^{\tau}, s_{m,h+1}^{\tau}\right)\right\}_{\tau=k_s+1}^{k-1}, \tag{3.1b}$$

$$U_{m,h}(k) = U_h^{\text{ser}}(k) \bigcup U_{m,h}^{\text{loc}}(k). \tag{3.1c}$$

By definition, $U_h^{\text{ser}}(k)$ is the dataset that is shared across all agents due to the latest synchronization at episode k_s. $U_{m,h}^{\text{loc}}(k)$ is the unique data collected by agent m since episode k_s. Then $U_{m,h}(k)$ is the total dataset available for agent m at the current time. Let $\mathcal{K}(k) = |U_{m,h}(k)|$ be the total number of data points. For the simplicity of notation, we also re-order the data points in $U_{m,h}(k)$, and rename the tuple $(s_{m,h}^{\tau}, a_{m,h}^{\tau}, s_{m,h+1}^{\tau})$ as (s^l, a^l, s'^l) such that we have $U_{m,h}(k) = \bigcup_{l=1}^{\mathcal{K}(k)}(s^l, a^l, s'^l)$. In fact, this can be done by the following one-to-one mapping

$$l_{m,k}(n,\tau) = \begin{cases} (\tau-1)M + n & \tau \le k_s, \\ (M-1)k_s + \tau & k_s < \tau \le k-1. \end{cases} \tag{3.2}$$

Therefore, we use indices $(s, a, s') \in U_{m,h}(k)$ and $l \in [\mathcal{K}(k)]$ interchangeably for the summation over set $U_{m,h}(k)$.

Algorithm interpretation At a high level, each episode k in Algorithm 1 consists of two stages. The first stage (Lines 3-15) is parallelly executed by all agents and the second stage (Lines 16-24) involves the communication among agents and the server.

In the first stage (Lines 3-15) of Algorithm 1, each agent m operates in two parts. The first part (Line 6) updates estimated Q functions $\{Q_{m,h}^k\}_{h=1}^H$ through LSVI with a randomized exploration strategy (Algorithm 2 or Algorithm 3, which will be introduced in Section 3.2). In particular, given the estimated value functions $V_{m,h+1}^k(\cdot) = \max_{a\in A} Q_{m,h}^k(\cdot, a)$ at step $h+1$, we perform one step robust backward Bellman update to obtain $V_{m,h}^k(\cdot)$ at step h. And we initialize $V_{m,H+1}^k(\cdot)$ to be 0 (Line 5). In the second part (Lines 7-14), after obtaining the estimated Q functions, in each step h we execute the greedy policy with respect to $Q_{m,h}^k$ and collect new data points which are added to the local dataset $U_{m,h}^{\text{loc}}(k)$ (Lines 8-10). Then we verify the synchronization condition (Lines 11-13). In this paper, we mainly use three types of synchronization rules. (1) We can synchronize every c episode where c is a user-defined constant, which is easy to implement in practice. (2) We can also synchronize at the episode of $b^1, b^2, ..., b^n$, with b representing the base of the exponential function. This is guided by the intuition that agents require more transitions urgently at the early learning stages. (3) Additionally, if we have a feature mapping $\boldsymbol{\phi}(s,a) : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$, based on (3.1), we define the following empirical covariance matrices.

$$\begin{aligned} {}^{\text{ser}}\boldsymbol{\Lambda}_h^k &= \textstyle\sum_{(s^l,a^l,s'^l)\in U_h^{\text{ser}}(k)} \boldsymbol{\phi}\left(s^l,a^l\right)\boldsymbol{\phi}\left(s^l,a^l\right)^\top, \\ {}^{\text{loc}}\boldsymbol{\Lambda}_{m,h}^k &= \textstyle\sum_{(s^l,a^l,s'^l)\in U_{m,h}^{\text{loc}}(k)} \boldsymbol{\phi}\left(s^l,a^l\right)\boldsymbol{\phi}\left(s^l,a^l\right)^\top, \\ \boldsymbol{\Lambda}_{m,h}^k &= {}^{\text{ser}}\boldsymbol{\Lambda}_h^k + {}^{\text{loc}}\boldsymbol{\Lambda}_{m,h}^k + \lambda\mathbf{I}. \end{aligned}$$

We synchronize as long as the following condition is met,

$$\log\frac{\det\left({}^{\text{ser}}\boldsymbol{\Lambda}_h^k + {}^{\text{loc}}\boldsymbol{\Lambda}_{m,h}^k + \lambda\mathbf{I}\right)}{\det\left({}^{\text{ser}}\boldsymbol{\Lambda}_h^k + \lambda\mathbf{I}\right)} \ge \frac{\gamma}{(k-k_s)}, \tag{3.3}$$

where γ is a communication control factor. In our experiments, we try all three rules and compare their performance, which is discussed in detail in Appendix M.1.

The second stage (Lines 16-24) is executed only when the synchronization condition is satisfied. First, all the agents upload their local transition set $U_{m,h}^{\text{loc}}(k)$, i.e., the newly collected local data after the last synchronization, to the server. Then, the server gathers all information together in $U_h^{\text{ser}}(k)$ and

sends it back to each agent. Finally, each agent resets the local transition set $U^{\text{loc}}_{m,h}(k) \leftarrow \emptyset$. Now agent m can access the dataset $U_{m,h}(k) = U^{\text{ser}}_h(k) \bigcup U^{\text{loc}}_{m,h}(k)$, which contains the historical data of all agents up to last synchronization and its local dataset.

3.2 Randomized Exploration Strategies

When we update the model parameter and estimate Q functions in Algorithm 1 (Line 6), we use exploration strategies to avoid suboptimal policies. Previous work adopted Upper Confidence Bound (UCB) exploration in the linear function class (Dubey & Pentland, 2021; Min et al., 2023) to estimate the Q function $\{Q^k_{m,h}\}_{h=1}^H$. Although UCB-based methods come with strong theoretical guarantees, they often perform poorly in practice (Chapelle & Li, 2011; Osband et al., 2013; Osband & Van Roy, 2017). Moreover, UCB requires a precise computation of the confidence set, which is usually hard to be implemented beyond the linear structure. In contrast, randomized exploration strategies exhibit more robust performance, are flexible in design and easy to implement, and do not require a linear structure.

We approximate the Q functions with the following function class $\mathcal{F} = \{f_{\mathbf{w}} : \mathcal{S} \times \mathcal{A} \to \mathbb{R} | f_{\mathbf{w}}(s,a) = f(\mathbf{w}; \boldsymbol{\phi}(s,a))\}$, where $\mathbf{w} \in \mathbb{R}^d$ is the parameter and $\boldsymbol{\phi} \in \mathbb{R}^d$ is a feature mapping associated with state-action pairs. Now we define the loss function for estimating the Q functions.

$$L^k_{m,h}(\mathbf{w}) = \sum_{l=1}^{\mathcal{K}(k)} L\big(r^l_h + V^k_{m,h+1}(s'^l), f\big(\mathbf{w}; \boldsymbol{\phi}^l\big)\big) + \lambda \|\mathbf{w}\|^2, \tag{3.4}$$

where $r^l_h = r_h\big(s^l, a^l\big)$, $\boldsymbol{\phi}^l = \boldsymbol{\phi}\big(s^l, a^l\big)$, and L is a user-specified loss function.

Perturbed-History Exploration The first strategy we use in Algorithm 1 is called the perturbed-history exploration (Kveton et al., 2019; 2020b; Ishfaq et al., 2021), displayed in Algorithm 2. We refer to the resulting algorithm as CoopTS-PHE. In particular, we optimize the following randomized loss function, where we add random Gaussian noises to the rewards and regularizer in (3.4).

$$\widetilde{L}^{k,n}_{m,h}(\mathbf{w}) = \sum_{l=1}^{\mathcal{K}(k)} L\big(\big(r^l_h + \epsilon^{k,l,n}_h\big) + V^k_{m,h+1}(s'^l), f\big(\mathbf{w}; \boldsymbol{\phi}^l\big)\big) + \lambda \|\mathbf{w} + \boldsymbol{\xi}^{k,n}_h\|^2, \tag{3.5}$$

where $\epsilon^{k,l,n}_h \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$, $\boldsymbol{\xi}^{k,n}_h \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$, and $n \in [N]$. Then we obtain the following perturbed estimated parameter

$$\widetilde{\mathbf{w}}^{k,n}_{m,h} = \text{argmin}_{\mathbf{w} \in \mathbb{R}^d} \widetilde{L}^{k,n}_{m,h}(\mathbf{w}). \tag{3.6}$$

Note that we repeat the above steps for $n = 1, \ldots, N$ to obtain independent copies of parameters, which is referred to as the multi-sampling process (Ishfaq et al., 2021; 2024). Then we obtain the estimated Q function $Q^k_{m,h}$ based on Line 7 in Algorithm 2. Finally, by maximizing $Q^k_{m,h}$ over action space $\mathcal{A}$, we obtain the estimated value function $V^k_{m,h}$.

Langevin Monte Carlo Exploration Next we introduce the Langevin Monte Carlo exploration strategy (Xu et al., 2022; Ishfaq et al., 2024) (displayed in Algorithm 3), which stems from the Langevin dynamics (Roberts & Tweedie, 1996; Bakry et al., 2014; Dalalyan, 2017; Xu et al., 2018; Zou et al., 2021). Combining it with Algorithm 1 leads to our second proposed algorithm, CoopTS-LMC. Specifically, we update the model parameter iteratively. For iterate $j = 1, \ldots, J_k$, the update is given by

$$\mathbf{w}^{k,j,n}_{m,h} = \mathbf{w}^{k,j-1,n}_{m,h} - \eta_{m,k} \nabla L^k_{m,h}\big(\mathbf{w}^{k,j-1,n}_{m,h}\big) + \sqrt{2\eta_{m,k}\beta^{-1}_{m,k}}\boldsymbol{\epsilon}^{k,j,n}_{m,h}, \tag{3.7}$$

where $L^k_{m,h}$ is defined in (3.4), $\boldsymbol{\epsilon}^{k,j,n}_{m,h} \in \mathbb{R}^d$ is a standard Gaussian noise, $\eta_{m,k}$ is the learning rate, and $\beta_{m,k}$ is the inverse temperature parameter. We again use the multi-sampling trick to obtain N

Algorithm 2 Perturbed-History Exploration

Input: multi-sampling number $N \in \mathbb{N}^+$, function class $\mathcal{F} = \{f_{\mathbf{w}} : \mathcal{S} \times \mathcal{A} \to \mathbb{R} | f_{\mathbf{w}}(s,a) = f(\mathbf{w}; \boldsymbol{\phi}(s,a))\}$.
for step $h = H, ..., 1$ **do**
  for $n = 1, ..., N$ **do**
    Sample $\{\epsilon_h^{k,l,n}\}_{l \in [\mathcal{K}(k)]} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$ and $\boldsymbol{\xi}_h^{k,n} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ independently.
    Solve $\widetilde{\mathbf{w}}_{m,h}^{k,n}$ according to (3.6).
  end for
  $Q_{m,h}^k \leftarrow \min\big\{\max_{n \in [N]} f\big(\widetilde{\mathbf{w}}_{m,h}^{k,n}; \boldsymbol{\phi}\big), H - h + 1\big\}^+$.
  $V_{m,h}^k(\cdot) \leftarrow \max_{a \in \mathcal{A}} Q_{m,h}^k(\cdot, a)$.
end for
Output: $\{Q_{m,h}^k(\cdot,\cdot), V_{m,h}^k(\cdot,\cdot)\}_{h=1}^H$.

independent estimators and similarly obtain the estimated Q function $Q_{m,h}^k$ by truncation based on Line 10 in Algorithm 3.

Algorithm 3 Langevin Monte Carlo Exploration

1: Input: multi-sampling number $N \in \mathbb{N}^+$, function class $\mathcal{F} = \{f_{\mathbf{w}} : \mathcal{S} \times \mathcal{A} \to \mathbb{R} | f_{\mathbf{w}}(s,a) = f(\mathbf{w}; \boldsymbol{\phi}(s,a))\}$, step sizes $\{\eta_{m,k}\}_{m \in \mathcal{M}, k \in [K]}$, inverse temperature parameters $\{\beta_{m,k}\}_{m \in \mathcal{M}, k \in [K]}$.
2: **for** step $h = H, ..., 1$ **do**
3: **for** $n = 1, ..., N$ **do**
4: $\mathbf{w}_{m,h}^{k,0,n} = \mathbf{w}_{m,h}^{k-1,J_{k-1},n}$.
5: **for** $j = 1, ..., J_k$ **do**
6: Sample $\boldsymbol{\epsilon}_{m,h}^{k,j,n} \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I})$.
7: Update $\mathbf{w}_{m,h}^{k,j,n}$ by (3.7).
8: **end for**
9: **end for**
10: $Q_{m,h}^k \leftarrow \min\big\{\max_{n \in [N]} f\big(\mathbf{w}_{m,h}^{k,J_k,n}; \boldsymbol{\phi}\big), H - h + 1\big\}^+$.
11: $V_{m,h}^k(\cdot) \leftarrow \max_{a \in A} Q_{m,h}^k(\cdot, a)$.
12: **end for**
13: Output: $\{Q_{m,h}^k(\cdot,\cdot), V_{m,h}^k(\cdot,\cdot)\}_{h=1}^H$.

4 Theoretical Analysis

In this section, we provide theoretical analyses of our algorithms in the linear structure, which is under the assumption of linear function approximation and linear MDP setting. We first present the definition of linear MDPs.

Definition 4.1 (Linear MDP (Jin et al., 2020))**.** An MDP$(\mathcal{S}, \mathcal{A}, \mathrm{H}, \mathbb{P}_m, \mathbf{r}_m)$ is a linear MDP with feature map $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, if for any $h \in [H]$, there exist d unknown measures $\boldsymbol{\mu}_h = (\mu_h^1, ..., \mu_h^d)$ over $\mathcal{S}$ and an unknown vector $\boldsymbol{\theta}_h \in \mathbb{R}^d$ such that for any $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\mathbb{P}_h(\cdot | s,a) = \big\langle \boldsymbol{\phi}(s,a), \boldsymbol{\mu}_h(\cdot) \big\rangle, \quad r_h(s,a) = \big\langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h \big\rangle.$$

Without loss of generality, we assume that for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, $\|\boldsymbol{\phi}(s,a)\| \leq 1$ and $\max\{\|\boldsymbol{\mu}_h(\mathcal{S})\|, \|\boldsymbol{\theta}_h\|\} \leq \sqrt{d}$.

Throughout the analyses in this section, we assume the homogeneous parallel MDPs setting where all agents share the same linear MDP defined in Definition 4.1. We also provide the results when the MDPs across agents are approximately linear and heterogeneous, which is deferred to Appendix E due to the space limit. Next, we first have the following regret bound for CoopTS-PHE.

Theorem 4.2. In CoopTS-PHE (Algorithm 1+Algorithm 2), let $N = \widetilde{C}\log(\delta)/\log(c_0)$ where $\widetilde{C} = \widetilde{\mathcal{O}}(d)$ and $c_0 = \Phi(1)$, $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution. Let $\lambda = 1$ and $0 < \delta < 1$. Under the determinant synchronization condition (3.3), we obtain the following cumulative regret

$$\text{Regret}(K) = \widetilde{\mathcal{O}}\big(d^{\frac{3}{2}} H^2 \sqrt{M}\big(\sqrt{dM\gamma} + \sqrt{K}\big)\big),$$

with probability at least $1 - \delta$.

Remark 4.3. When we choose $\gamma = \mathcal{O}(K/dM)$ in the synchronization condition (3.3), the cumulative regret of CoopTS-PHE becomes $\widetilde{\mathcal{O}}(d^{3/2}H^2\sqrt{MK})$, which matches the result of UCB exploration (Dubey & Pentland, 2021). When $M = 1$, the regret becomes $\widetilde{\mathcal{O}}(d^{3/2}H^2\sqrt{K})$, which matches the existing best single-agent result (Jin et al., 2020; Ishfaq et al., 2021; 2024). Note that if there is no communication at all and agents act independently, with the same number of learning rounds (or samples), the cumulative regret becomes $\widetilde{\mathcal{O}}(M \cdot d^{3/2}H^2\sqrt{K})$. By incorporating communication, our regret bound in Theorem 4.2 is lower than that of the independent setting by a factor $\sqrt{M}$. A similar strategy called rare-switching update with a determinant synchronization condition has also been adopted in parallel bandit problems (Ruan et al., 2021; Chan et al., 2021).

Similarly, we have the following result for CoopTS-LMC.

Theorem 4.4. In CoopTS-LMC (Algorithm 1+Algorithm 3), let $N = \bar{C}\log(\delta)/\log(c_0')$ where $c_0' = 1 - 1/2\sqrt{2e\pi}$ and $\bar{C} = \widetilde{\mathcal{O}}(d)$. Let $1/\sqrt{\beta_{m,k}} = \widetilde{\mathcal{O}}\big(H\sqrt{d}\big)$ for all $m \in \mathcal{M}$, $\lambda = 1$, and $0 < \delta < 1$. For any episode $k \in [K]$ and agent $m \in \mathcal{M}$, let the learning rate $\eta_{m,k} = 1/\big(4\lambda_{\max}\big(\mathbf{\Lambda}_{m,h}^k\big)\big)$, the update number $J_k = 2\kappa_k \log(4HKMd)$ where $\kappa_k = \lambda_{\max}\big(\mathbf{\Lambda}_{m,h}^k\big)/\lambda_{\min}\big(\mathbf{\Lambda}_{m,h}^k\big)$ is the condition number of $\mathbf{\Lambda}_{m,h}^k$. Under the determinant synchronization condition (3.3), we have

$$\text{Regret}(K) = \widetilde{\mathcal{O}}\big(d^{\frac{3}{2}} H^2 \sqrt{M}\big(\sqrt{dM\gamma} + \sqrt{K}\big)\big),$$

with probability at least $1 - \delta$.

Remark 4.5. Note that CoopTS-PHE and CoopTS-LMC have the same order of regret. Hence the discussion in Remark 4.3 also applies to CoopTS-LMC. We would also like to highlight that our results are the first rigorous regret bounds for randomized MARL algorithms.

From the perspective of technical novelty, our analysis of randomized MARL algorithms is different from that of UCB-based algorithms (Dubey & Pentland, 2021) because the model prediction error here contains randomness, causing a more complex probability analysis and an additional approximation error. We would also like to point out that in proofs for both CoopTS-LMC and CoopTS-PHE we use a new ε-covering technique to prove that the optimism lemma holds for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ instead of just the state-action pairs encountered by the algorithm, which is essential for the regret analysis. This was ignored by previous works (Cai et al., 2020) and its follow-up works (Zhong & Zhang, 2024; Ishfaq et al., 2024) that use the same regret decomposition technique. Furthermore, the multi-agent setting and the communications from synchronization in our algorithms also significantly increase the challenges in our analysis compared to randomized exploration in the single-agent setting (Ishfaq et al., 2021; 2024).

5 Conclusion

We proposed a unified algorithm framework for provably efficient randomized exploration in parallel MDPs. By combining this unified algorithm framework with two TS-type randomized exploration strategies, PHE and LMC, we obtained two algorithms for parallel MDPs: CoopTS-PHE and CoopTS-LMC. These two algorithms are both flexible in design and easy to implement in practice. Under the linear MDP setting, we derived the theoretical regret bounds and communication complexities of CoopTS-PHE and CoopTS-LMC. This is the first result for randomized exploration in cooperative MARL, matching the best existing regret bounds for single-agent RL (Ishfaq et al., 2021;

2024). We also extended our theoretical analysis to the misspecified setting. Our experiments on diverse RL parallel environments verified that randomized exploration improves the balance between exploration and exploitation in both homogeneous and heterogeneous settings. Future research could explore extending our randomized exploration algorithm to a fully decentralized or advanced federated learning setting. Additionally, developing a more communication-efficient algorithm to reduce the substantial communication costs in the general function class setting is another potential direction.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Marc Abeille and Alessandro Lazaric. Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*, pp. 176–184. PMLR, 2017.

Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, 1968.

Priyank Agrawal, Jinglin Chen, and Nan Jiang. Improved worst-case regret bounds for randomized least-squares value iteration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 6566–6573, 2021.

S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pp. 127–135, 2013a.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pp. 127–135. PMLR, 2013b.

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems*, volume 30, pp. 1184–1194, 2017.

Dominique Bakry, Ivan Gentil, Michel Ledoux, et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

Nikhil Angad Bakshi, Tejus Gupta, Ramina Ghods, and Jeff Schneider. Guts: Generalized uncertainty-aware thompson sampling for multi-agent active search. *arXiv preprint arXiv:2304.02075*, 2023.

Ana LC Bazzan. Opportunities for multiagent systems and multiagent reinforcement learning in traffic control. *Autonomous Agents and Multi-Agent Systems*, 18:342–375, 2009.

Daniel S. Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. The complexity of decentralized control of markov decision processes. *Mathematics of Operations Research*, 27(4): 819–840, 2002.

Craig Boutilier. Planning, learning and coordination in multiagent decision processes. In *Theoretical Aspects of Rationality and Knowledge*, 1996.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pp. 1283–1294. PMLR, 2020.

Georgios Chalkiadakis and Craig Boutilier. Coordination in multiagent reinforcement learning: A bayesian approach. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pp. 709–716, 2003.

Jeffrey Chan, Aldo Pacchiano, Nilesh Tripuraneni, Yun S. Song, Peter Bartlett, and Michael I. Jordan. Parallelizing contextual bandits. *arXiv preprint arXiv:2105.10590*, 2021.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214. JMLR Workshop and Conference Proceedings, 2011.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3):651–676, 2017.

Christoph Dann, Mehryar Mohri, Tong Zhang, and Julian Zimmert. A provably efficient model-free posterior sampling method for episodic reinforcement learning. *Advances in neural information processing systems*, 34, 2021.

Guohui Ding, Joewie J Koh, Kelly Merckaert, Bram Vanderborght, Marco M Nicotra, Christoffer Heckman, Alessandro Roncone, and Lijun Chen. Distributed reinforcement learning for cooperative multi-robot object manipulation. *arXiv preprint arXiv:2003.09540*, 2020.

Abhimanyu Dubey and Alex Pentland. Provably efficient cooperative multi-agent reinforcement learning with function approximation. *arXiv preprint arXiv:2103.04972*, 2021.

Yingjie Fei and Ruitu Xu. Cascaded gaps: Towards logarithmic regret for risk-sensitive reinforcement learning. In *International Conference on Machine Learning*, pp. 6392–6417. PMLR, 2022.

Meire Fortunato, Mohammad Gheshlaghi Azar, Bilal Piot, et al. Noisy networks for exploration. *arXiv preprint arXiv:1706.10295*, 2018.

Jianye Hao, Tianpei Yang, Hongyao Tang, Chenjia Bai, Jinyi Liu, Zhaopeng Meng, Peng Liu, and Zhen Wang. Exploration in deep reinforcement learning: From single-agent to multiagent domain. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double qlearning. In *Annual AAAI Conference on Artificial Intelligence (AAAI)*, 2016.

Eshcar Hillel, Zohar S Karnin, Tomer Koren, Ronny Lempel, and Oren Somekh. Distributed exploration in multi-armed bandits. *Advances in Neural Information Processing Systems*, 26, 2013.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Tom Huix, Matthew Zhang, and Alain Durmus. Tight regret and complexity bounds for thompson sampling via langevin monte carlo. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 8749–8770. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/huix23a.html.

Haque Ishfaq, Qiwen Cui, Viet Nguyen, Alex Ayoub, Zhuoran Yang, Zhaoran Wang, Doina Precup, and Lin Yang. Randomized exploration in reinforcement learning with general value function approximation. In *International Conference on Machine Learning*, pp. 4607–4616. PMLR, 2021.

Haque Ishfaq, Qingfeng Lan, Pan Xu, A Rupam Mahmood, Doina Precup, Anima Anandkumar, and Kamyar Azizzadenesheli. Provable and practical: Efficient exploration in reinforcement learning via langevin monte carlo. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=nfIAEJFiBZ.

Mehdi Jafarnia-Jahromi, Rahul Jain, and Ashutosh Nayyar. Learning zero-sum stochastic games with posterior sampling. *https://arxiv.org/abs/2109.03396*, 2023.

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

Hao Jin, Yang Peng, Wenhao Yang, Shusen Wang, and Zhihua Zhang. Federated reinforcement learning with environment heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, pp. 18–37. PMLR, 2022a.

Tianyuan Jin, Pan Xu, Jieming Shi, Xiaokui Xiao, and Quanquan Gu. Mots: Minimax optimal thompson sampling. In *International Conference on Machine Learning*, pp. 5074–5083. PMLR, 2021.

Tianyuan Jin, Pan Xu, Xiaokui Xiao, and Anima Anandkumar. Finite-time regret of thompson sampling algorithms for exponential family multi-armed bandits. *Advances in Neural Information Processing Systems*, 35:38475–38487, 2022b.

Tianyuan Jin, Xianglin Yang, Xiaokui Xiao, and Pan Xu. Thompson sampling with less exploration is fast and optimal. In *International Conference on Machine Learning*, pp. 15239–15261. PMLR, 2023.

Tianyuan Jin, Hao-Lun Hsu, William Chang, and Pan Xu. Finite-time frequentist regret bounds of multi-agent thompson sampling on sparse hypergraphs. In *Annual AAAI Conference on Artificial Intelligence (AAAI)*, 2024.

Amin Karbasi, Nikki Lijing Kuang, Yian Ma, and Siddharth Mitra. Langevin thompson sampling with logarithmic communication: Bandits and reinforcement learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15828–15860. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/karbasi23a.html.

R. Matthew Kretchmar. Parallel reinforcement learning. In *The 6th World Conference on Systemics, Cybernetics, and Informatics*, 2002.

Nikki Kuang, Ming Yin, et al. Posterior sampling with delayed feedback for reinforcement learning with linear function approximation. *Advances in neural information processing systems*, 2023.

Branislav Kveton, Csaba Szepesvari, Mohammad Ghavamzadeh, and Craig Boutilier. Perturbed-history exploration in stochastic multi-armed bandits, 2019.

Branislav Kveton, Manzil Zaheer, Csaba Szepesvari, Lihong Li, Mohammad Ghavamzadeh, and Craig Boutilier. Randomized exploration in generalized linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 2066–2076. PMLR, 2020a.

Branislav Kveton, Manzil Zaheer, Csaba Szepesvari, Lihong Li, Mohammad Ghavamzadeh, and Craig Boutilier. Randomized exploration in generalized linear bandits. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2066–2076. PMLR, 26–28 Aug 2020b. URL https://proceedings.mlr.press/v108/kveton20a.html.

Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. On distributed cooperative decision-making in multiarmed bandits. In *2016 European Control Conference (ECC)*, pp. 243–248. IEEE, 2016.

L. Li, W. Chu, J. Langford, and R. E.r Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pp. 2071–2080. PMLR, 2017.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pp. 429–450, 2020.

Ziniu Li, Yingru Li, Yushun Zhang, Tong Zhang, and Zhi-Quan Luo. Hyperdqn: A randomized exploration method for deep reinforcement learning. In *International Conference on Learning Representations*, 2022.

Justin Lidard, Udari Madhushani, and Naomi Ehrich Leonard. Provably efficient multi-agent reinforcement learning with fully decentralized communication. In *2022 American Control Conference (ACC)*, pp. 3311–3316. IEEE, 2022.

Boyi Liu, Lujia Wang, and Ming Liu. Lifelong federated reinforcement learning: a learning architecture for navigation in cloud robotic systems. *IEEE Robotics and Automation Letters*, 4(4): 4555–4562, 2019.

Zhilong Liu, Jiayi Zhang, Ziheng Liu, Hongyang Du, Zhe Wang, Dusit Niyato, Mohsen Guizani, and Bo Ai. Cell-free xl-mimo meets multi-agent reinforcement learning: Architectures, challenges, and future directions. *arXiv preprint arXiv:2307.02827*, 2023.

Yifei Min, Jiafan He, Tianhao Wang, and Quanquan Gu. Cooperative multi-agent reinforcement learning: asynchronous communication and linear function approximation. In *International Conference on Machine Learning*, pp. 24785–24811. PMLR, 2023.

V. Mnih, K. Kavukcuoglu, D. Silver, et al. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

Alireza Mousavi-Hosseini, Tyler Farghly, Ye He, Krishnakumar Balasubramanian, and Murat A. Erdogdu. Towards a complete analysis of langevin monte carlo: Beyond poincaré inequality, 2023.

Thanh Nguyen-Tang and Raman Arora. On sample-efficient offline reinforcement learning: Data diversity, posterior sampling and beyond. *Advances in neural information processing systems*, 2023.

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *International conference on machine learning*, pp. 2701–2710. PMLR, 2017.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26, 2013.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. *Advances in neural information processing systems*, 29, 2016a.

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pp. 2377–2386. PMLR, 2016b.

Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. *Advances in neural information processing systems*, 31, 2018.

Shuang Qiu, Ziyu Dai, Han Zhong, Zhaoran Wang, Zhuoran Yang, and Tong Zhang. Posterior sampling for competitive rl: Function approximation and partial observation. *Advances in neural information processing systems*, 2023.

Carlos Riquelme, George Tucker, and Jasper Snoek. Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. *arXiv preprint arXiv:1802.09127*, 2018.

Gareth O Roberts and Richard L Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, pp. 341–363, 1996.

Carolina Rojas-Córdova, Amanda J Williamson, Julio A Pertuze, and Gustavo Calvo. Why one strategy does not fit all: a systematic review on exploration–exploitation in different organizational archetypes. *Review of Managerial Science*, 17(7):2251–2295, 2023.

Yufei Ruan, Jiaqi Yang, and Yuan Zhou. Linear bandits with limited adaptivity and learning distributional optimal design. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 74–87, 2021.

Daniel Russo. Worst-case regret bounds for exploration via randomized value functions. *Advances in neural information processing systems*, 32:14410–14420, 2019.

Malcolm Strens. A bayesian framework for reinforcement learning. In *International Conference on Machine Learning*, pp. 943–950. PMLR, 2000.

Matthew E. Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(56):1633–1685, 2009.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

J.-J. Tsay, C.-C. Chen, and J.-J. Hsu. Evolving intelligent mario controller by reinforcement learning. In *International Conference on Technologies and Applications of Artificial Intelligence*, pp. 266–272, 2011.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Ruosong Wang, Russ R Salakhutdinov, and Lin Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33:6123–6135, 2020a.

Yuanhao Wang, Jiachen Hu, Xiaoyu Chen, and Liwei Wang. Distributed bandit learning: Near-optimal regret with efficient communication. In *International Conference on Learning Representations*, 2020b.

Zhendong Wang and Mingyuan Zhou. Thompson sampling via local uncertainty. In *International Conference on Machine Learning*, pp. 10115–10125. PMLR, 2020.

Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pp. 3674–3682. PMLR, 2020.

Wei Xiong, Han Zhong, Chengshuai Shi, Cong Shen, and Tong Zhang. A self-play posterior sampling algorithm for zero-sum markov games. In *International Conference on Machine Learning*, pp. 24496–24523. PMLR, 2022.

Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of langevin dynamics based algorithms for nonconvex optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

Pan Xu, Zheng Wen, Handong Zhao, and Quanquan Gu. Neural contextual bandits with deep representation and shallow exploration. In *International Conference on Learning Representations*, 2021.

Pan Xu, Hongkai Zheng, Eric V Mazumdar, Kamyar Azizzadenesheli, and Animashree Anandkumar. Langevin monte carlo for contextual bandits. In *International Conference on Machine Learning*, pp. 24830–24850. PMLR, 2022.

Deheng Ye, Guibin Chen, Wen Zhang, Sheng Chen, Bo Yuan, Bo Liu, Jia Chen, Zhao Liu, Fuhao Qiu, Hongsheng Yu, et al. Towards playing full moba games with deep reinforcement learning. *Advances in Neural Information Processing Systems*, 33:621–632, 2020.

C. Yeh, V. Li, R. Datta, J. Arroyo, N. Christianson, C. Zhang, Y. Chen, M. Hosseini, A. Golmohammadi, Y. Shi, Y. Yue, and A. Wierman. Sustaingym: A benchmark suite of reinforcement learning for sustainability applications. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. PMLR, 2023.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative multi-agent games. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

Chao Yu, Xinyi Yang, Jiaxuan Gao, Jiayu Chen, Yunfei Li, Jijia Liu, Yunfei Xiang, Ruixin Huang, Huazhong Yang, Yi Wu, and Yu Wang. Asynchronous multi-agent reinforcement learning for efficient real-time multi-robot cooperative exploration. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS*, pp. 1107–1115. ACM, 2023.

Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pp. 1954–1964. PMLR, 2020.

Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Başar. Fully decentralized multi-agent reinforcement learning with networked agents. In *International Conference on Machine Learning*, volume 80, pp. 5872–5881. PMLR, 2018.

W. Zhang, D. Zhou, L. Li, and Q. Gu. Neural thompson sampling. In *International Conference on Learning Representations*, 2021.

Yizhou Zhang, Guannan Qu, Pan Xu, Yiheng Lin, Zaiwei Chen, and Adam Wierman. Global convergence of localized policy iteration in networked multi-agent reinforcement learning. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 7(1):1–51, 2023.

Yunqi Zhao, Igor Borovikov, Jason Rupert, Caedmon Somers, and Ahmad Beirami. On multi-agent learning in team sports games. *arXiv preprint arXiv:1906.10124*, 2019.

Han Zhong and Tong Zhang. A theoretical analysis of optimistic proximal policy optimization in linear markov decision processes. *Advances in Neural Information Processing Systems*, 36, 2024.

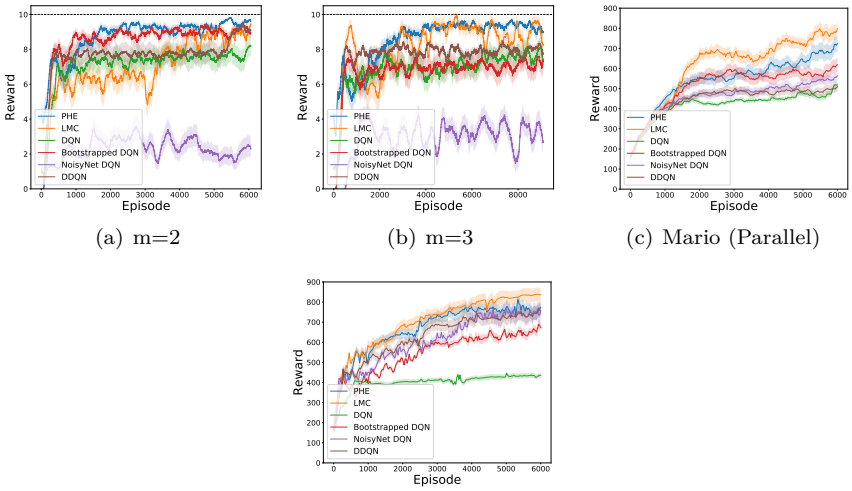
D. Zhou, L. Li, and Q. Gu. Neural contextual bandits with ucb-based exploration. In *International Conference on Machine Learning*, pp. 11492–11502, 2020.

Yichi Zhou, Jialian Li, and Jun Zhu. Posterior sampling for multi-agent reinforcement learning: solving extensive games with imperfect information. In *International Conference on Learning Representations*, 2019.

Difan Zou, Pan Xu, and Quanquan Gu. Faster convergence of stochastic gradient langevin dynamics for non-log-concave sampling. In *Uncertainty in Artificial Intelligence*, pp. 1152–1162. PMLR, 2021.

A Experiments

We present an empirical evaluation of our proposed randomized exploration strategies (*i.e.,* CoopTS-PHE and CoopTS-LMC) with deep Q-networks (DQNs) (Mnih et al., 2015) as the core algorithm on varying tasks under multi-agent settings compared with several baselines: vanilla DQN, Double DQN (Hasselt et al., 2016), Bootstrapped DQN (Osband et al., 2016a), and Noisy-Net (Fortunato et al., 2018)). Given that all experiments are conducted under multi-agent settings unless explicitly specified as a single-agent or centralized scenario, we denote CoopTS-PHE as "PHE" and CoopTS-LMC as "LMC" in both experimental contexts and figures.



(a) m=2 (b) m=3 (c) Mario (Parallel)

(d) Mario (Federated)

Figure 1: Comparison among different exploration strategies in (a)-(b): N-chain with $N = 25$ and (c)-(d): Super Mario Bros. All results are averaged over 10 runs.

A.1 N-chain

The N-chain (Osband et al., 2016a) comprises a sequence of N states denoted as s_l, $\forall 1 \leq l \leq N$. Assuming the existence of m agents, all initiating their trajectories from s_2, this study explores the dynamics of their movement within the chain. At each time step, agents face the decision to move either left or right. Notably, each agent incurs a nominal reward of $r = 0.001$ upon reaching state s_1, while a more substantial reward of $r = 1$ is obtained upon reaching the terminal state s_N. The illustration of N-chain environment is shown in Appendix M.1. With a horizon length of $N + 9$, the optimal return is 10. We consider $N = 25$ with the communication among agents in Figure 1 following the synchronization approach in Algorithm 1. Note that the total training episodes in the x-axis are shared among m agents. In Figure 1(a), we show that PHE and Bootstrapped DQN result in better performance while LMC can also eventually converge to a similar reward. Upon increasing the number of agents to $m = 3$, we show in Figure 1(b) that our randomized exploration methods outperform all other baselines. Notably, the fluctuation in PHE is observed to be less pronounced against LMC. This observation lends support to our theoretical framework regarding performance tolerance in the misspecified setting, as detailed in Appendix E. The complete results for N-chain and ablation studies can be found in Appendix M.1.

A.2 Super Mario Bros

Environmental heterogeneity, arising from various sources, is a prevalent challenge in practical scenarios. In Appendix E, we illustrate the extension of homogeneous parallel MDP to the misspecified setting. In the Super Mario Bros task (Tsay et al., 2011), we examine a scenario where four agents, denoted as $m = 4$, engage in learning within distinct environments. Despite these environments sharing the same state space $\mathcal{S}$, action space $\mathcal{A}$, and reward function, their characteristics are different

described in Appendix M.2. The primary objective of the Super Mario Bros task is to train an agent capable of advancing as far-right and rapidly as possible without collisions or falls. Figure 1(c) visually depicts that both randomized exploration strategies outperform other baselines in cooperative parallel learning. Notably, We observe that the superiority of LMC gets significant against PHE unlike the results in N-chain in Figures 1(a) and 1(b). In the case of PHE, Gaussian noise is introduced to the reward before applying the Bellman update, which can be viewed as a method empirically approximating the posterior distribution of the Q function using a Gaussian distribution. However, it is crucial to note that in practical scenarios, unlike the N-chain setting, Gaussian distributions may not always provide an accurate approximation of the true posterior of the Q function (Ishfaq et al., 2024). Here, transitions are shared among the four agents whenever the synchronization condition in (3.3) is met. We also conducted extra experiments in this task extending our proposed method to federated learning shown in Figure 1(d) with details in Appendix M.2.

A.3 Thermal Control of Building Energy Systems

Finally, we assess the efficacy of our randomized exploration strategies through their application to a practical task within a sustainable energy system: BuildingEnv, as outlined by Yeh et al. (2023). BuildingEnv is designed to manage the heating supply in a multi-zone building, which involves addressing real-world physical constraints and accounting for environmental shifts over time. The objective is to meet user-defined temperature specifications while simultaneously minimizing overall electricity consumption. We defer the environment details to Appendix M.3. With the availability of different cities in varying weather types, we conduct experiments on multiple cities in parallel and share their data following Algorithm 1 for each exploration strategy. We deploy those trained policies to the environment of each city/weather respectively in evaluation. We include all methods and random action (*i.e.,* sample action randomly from action space) in Figure 13 in Appendix M.3 for a fair comparison. We display the distribution of the return with probability density in violin plots, indicating that our PHE and LMC can perform better with a higher mean.

B Related Work

Cooperative Multi-Agent Reinforcement Learning Cooperative MARL is closely intertwined with the domain of multi-agent multi-armed bandits, exemplified by decentralized algorithms featuring communication across a network or hypergraphs (Landgren et al., 2016; Zhang et al., 2023; Jin et al., 2024) and distributed settings (Hillel et al., 2013; Wang et al., 2020b). Cooperative MARL manifests primarily in two categories: multi-agent MDPs (Boutilier, 1996; Zhang et al., 2018; Xie et al., 2020; Dubey & Pentland, 2021) and parallel MDPs (Bernstein et al., 2002; Dubey & Pentland, 2021; Lidard et al., 2022; Bernstein et al., 2002; Min et al., 2023). In the realm of cooperative multi-agent robotics, the former is employed to formulate optimal multi-agent policies across the distributed system (Yu et al., 2022; 2023). On the other hand, homogeneous parallel MDPs leverage inter-agent communication to expedite learning processes (Kretchmar, 2002). Additionally, heterogeneous parallel MDPs establish connections to heterogeneous federated learning (Li et al., 2020) and exhibit improved generalizability in transfer learning scenarios (Taylor & Stone, 2009).

We focus on parallel MDPs in this paper, where agents interact with the environment simultaneously to tackle shared challenges within extensive and distributed systems (Kretchmar, 2002). Recently, Dubey & Pentland (2021) proposed the Coop-LSVI algorithm, extending the LSVI-UCB algorithm (Jin et al., 2020) in single-agent RL to MARL with linear MDPs. In a parallel RL setting with asynchronous communication, Min et al. (2023) builds upon Coop-LSVI while relinquishing compatibility with heterogeneous MDPs. Meanwhile, Lidard et al. (2022) focuses on fully decentralized multi-agent UCB Q-learning in a tabular setting, maintaining polynomial space complexity even as the number of agents increases. However, it is worth noting that neither of the previous works (Dubey & Pentland, 2021; Min et al., 2023) in non-tabular cooperative MRAL provides experimental validation for the efficacy of their proposed communication strategies. The gap arises from their reliance on LSVI-UCB as the core algorithm, wherein optimism is instantiated through

UCB. Empirical evidence suggests that UCB-based approaches tend to underperform in practical scenarios (Osband et al., 2013; Osband & Van Roy, 2017; Ishfaq et al., 2024). Moreover, the computational demands of LSVI-UCB become untenable due to the necessity of recurrently computing the feature covariance matrix for updating the UCB bonus function. Therefore, randomized exploration in this work is critical to make these algorithm designs practical.

Randomized Exploration The roots of randomized exploration, particularly TS, can be traced back to its success in bandit problems (Thompson, 1933). Randomized exploration strategies can typically exhibit superior performance in practical applications due to avoidance of early convergence to suboptimal actions (Jin et al., 2021; 2022b; 2023). Furthermore, these strategies demonstrate robustness in the face of noise and uncertainty, particularly within non-stationary environments (Wang & Zhou, 2020; Bakshi et al., 2023). This success has extended to Langevin Monte Carlo Thompson Sampling (LMCTS), which has been applied to various domains, including linear bandits, generalized linear bandits, and neural contextual bandits (Xu et al., 2022). The exploration of posterior sampling techniques in RL has gained prominence, building upon the foundation laid by TS (Strens, 2000; Agrawal & Jia, 2017). Randomized Least-Square Value Iteration (RLSVI) is an approach that leverages random perturbations to approximate the posterior, with frequentist regret analysis applied under the tabular MDP setting (Osband et al., 2016b), inspiring subsequent works focusing on theoretical analyses aimed at improving worst-case regret under tabular MDPs (Russo, 2019; Agrawal et al., 2021), with extensions to the linear setting (Zanette et al., 2020; Ishfaq et al., 2021; Dann et al., 2021). In addition to theoretical advancements, several practical algorithms have been proposed based on RLSVI to approximate posterior samples of Q functions in deep RL. These approaches involve ensembles of randomly initialized neural networks (Osband et al., 2016a; 2018) and noise injection into the parameters of the neural network (Fortunato et al., 2018; Li et al., 2022). With the success of LMCTS (Xu et al., 2022) in bandit domains, the exploration of randomized methods has expanded to alternative approaches like LMC in tabular RL (Karbasi et al., 2023) and linear MDPs with neural network approximation (Ishfaq et al., 2024). Further works delve into the realm of random exploration from the perspectives of delayed feedback (Kuang et al., 2023) and offline RL (Nguyen-Tang & Arora, 2023).

While posterior sampling demonstrates superiority in various contexts, its theoretical foundations in the multi-agent setting remain underexplored. Existing research predominantly focuses on two-player zero-sum games, considering both Bayesian (Zhou et al., 2019; Jafarnia-Jahromi et al., 2023) and frequentist regrets (Xiong et al., 2022; Qiu et al., 2023). There is no existing work studying randomized exploration for cooperative multi-agent settings.

C Instantiation in the Linear Function Class

In this section, we specifically discuss our TS-related algorithms in the linear structure, which is under the assumption of linear function approximation and linear MDP setting.

Recall from the loss function in (3.4), here we choose L to be l_2 loss and linear function class $f(\mathbf{w};\boldsymbol{\phi}^l) = \mathbf{w}^\top\boldsymbol{\phi}^l$. By solving this least-square regression problem, we obtain the unperturbed regression estimator $\widehat{\mathbf{w}}^k_{m,h}$. In the linear setting, we have the closed-form solution

$$\widehat{\mathbf{w}}^k_{m,h} = (\boldsymbol{\Lambda}^k_{m,h})^{-1}\boldsymbol{b}^k_{m,h}, \tag{C.1}$$

where $\boldsymbol{\Lambda}^k_{m,h}$ and $\boldsymbol{b}^k_{m,h}$ are defined as follows

$$\boldsymbol{\Lambda}^k_{m,h} = \sum_{l=1}^{\mathcal{K}(k)} \boldsymbol{\phi}\big(s^l,a^l\big)\boldsymbol{\phi}\big(s^l,a^l\big)^\top + \lambda\mathbf{I},$$
$$\boldsymbol{b}^k_{m,h} = \sum_{l=1}^{\mathcal{K}(k)} \big[r_h\big(s^l,a^l\big) + V^k_{m,h+1}\big(s'^l\big)\big]\boldsymbol{\phi}\big(s^l,a^l\big).$$

A natural way of doing randomized exploration is to add a noise $\mathcal{N}(\mathbf{0}, \sigma^2(\mathbf{\Lambda}_{m,h}^k)^{-1})$ to $\widehat{\mathbf{w}}_{m,h}^k$ and get the estimated parameter $\bar{\mathbf{w}}_{m,h}^k$. Then we can construct estimated Q function $Q_{m,h}^k(\cdot,\cdot) = \min\{\boldsymbol{\phi}(\cdot,\cdot)^\top \bar{\mathbf{w}}_{m,h}^k, H-h+1\}^+$. We call this method as CoopTS, which is aligned with other linear TS algorithms (Agrawal & Goyal, 2013b; Abeille & Lazaric, 2017). In what follows, we theoretically show that our proposed algorithms are equivalent or approximately converge to the CoopTS algorithm in the linear function approximation setting.

For CoopTS-PHE (Algorithm 1+Algorithm 2), let the function approximation in (3.5) be linear and choose L to be the squared loss. By solving this least-square regression problem, we obtain the perturbed regression estimator $\widetilde{\mathbf{w}}_{m,h}^{k,n}$ in CoopTS-PHE. The following proposition conveys that CoopTS-PHE is actually equivalent to CoopTS.

Proposition C.1 (Equivalent to CoopTS)**.** The output $\widetilde{\mathbf{w}}_{m,h}^{k,n}$ by CoopTS-PHE is equivalent to adding a Gaussian vector to the unperturbed regression estimator $\widehat{\mathbf{w}}_{m,h}^k$, i.e., $\widetilde{\mathbf{w}}_{m,h}^{k,n} = \widehat{\mathbf{w}}_{m,h}^k + \boldsymbol{\zeta}_{m,h}^{k,n}$, where $\boldsymbol{\zeta}_{m,h}^{k,n} \sim \mathcal{N}(\mathbf{0}, \sigma^2(\mathbf{\Lambda}_{m,h}^k)^{-1})$.

For CoopTS-LMC (Algorithm 1+Algorithm 3), let function approximation in (3.4) be linear and choose L to be l_2 loss to get the loss function. Then after finishing the LMC update, we get the estimated parameter $\mathbf{w}_{m,h}^{k,J_k,n}$ and construct the model approximation of Q function. The following proposition conveys that the distribution of $\mathbf{w}_{m,h}^{k,J_k}$ converges to the posterior distribution of Thompson Sampling exploration. The proof of this proposition is given in Xu et al. (2022).

Proposition C.2 (Approximately equivalent to CoopTS (Xu et al., 2022))**.** If the epoch length J_k in Algorithm 3 is sufficiently large, the distribution of $\mathbf{w}_{m,h}^{k,J_k}$ converges to Gaussian distribution $\mathcal{N}(\widehat{\mathbf{w}}_{m,h}^k, \beta_{m,k}^{-1}(\mathbf{\Lambda}_{m,h}^k)^{-1})$.

Propositions C.1 and C.2 indicate that the results of our two randomized exploration strategies are closely related to CoopTS. As we have mentioned above, in CoopTS, the estimated parameter $\bar{\mathbf{w}}_{m,h}^k$ is sampled from the normal distribution $\mathcal{N}(\widehat{\mathbf{w}}_{m,h}^k, \sigma^2(\mathbf{\Lambda}_{m,h}^k)^{-1})$. However, in practice, this sampling is often executed in this way: we sample $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ first, then we calculate $\bar{\mathbf{w}}_{m,h}^k = \widehat{\mathbf{w}}_{m,h}^k + \sigma(\mathbf{\Lambda}_{m,h}^k)^{-\frac{1}{2}}\boldsymbol{\beta}$ and obtain the estimated parameter. Nevertheless, computing $\left(\mathbf{\Lambda}_{m,h}^k\right)^{-\frac{1}{2}}$ can be computationally expensive, often requiring at least $\mathcal{O}(d^3)$ operations with the Cholesky decomposition, making it impractical for high-dimensional machine learning challenges. Additionally, the Gaussian distribution used in Thompson Sampling may not effectively approximate the posterior distribution in more complex bandit models than the linear MDP due to their intricate structures.

Moreover, as pointed out by recent work (Chapelle & Li, 2011; Riquelme et al., 2018; Kveton et al., 2020a; Xu et al., 2022), the Laplace approximation-based Thompson Sampling exhibits a constant approximation error in the estimation of the posterior distribution. Therefore, it necessitates a careful redesign of the covariance matrix to ensure effective performance.

Advantages of PHE and LMC As mentioned above, computing $\left(\mathbf{\Lambda}_{m,h}^k\right)^{-\frac{1}{2}}$ can be computationally expensive. However, Perturbed-History exploration and Langevin Monte Carlo exploration can avoid this. For PHE, by only adding i.i.d random Gaussian noise to perturb reward and regularizer, its performance will be equivalent to TS. For LMC, by only performing noisy gradient descent, we can do the randomized exploration, resulting in similar performance compared with TS. Additionally, these two methods can easily be implemented to general function class while Thompson Sampling usually cannot be generalized except for the linear setting. In summary, these two methods are both flexible in design and easy to implement in practice.

Communication cost We emphasize that agents can just send compressed statistics to the server under the linear setting, which can largely reduce communication cost. In the linear function class, we can calculate the closed-form solution of the regression problem (C.1). In this case, when synchronization process is met, all the agents will only need to send their calculated local statistics ${}^{\text{loc}}\mathbf{\Lambda}_{m,h}^k$

and ${}^{\text{loc}}\boldsymbol{b}_{m,h}^{k}$ to help solve the regression problem. This communication cost is much smaller because $\boldsymbol{\Lambda}$ is only a $d \times d$ matrix and $\boldsymbol{b}$ is only a d-dimensional vector, where d is the feature dimension in linear MDP assumption. This can also avoid privacy disclosure through communications.

Nevertheless, in the general function class setting, our proposed algorithms still require sharing all the collected datasets, which will cause relatively large communication cost. Additionally, in Appendix M.2, we also propose a federated setting algorithm Algorithm 4. In this setting, instead of sharing collected datasets, agents can just share the weight of the collected estimated Q functions, which can largely reduce the communication cost.

D Analysis of the Communication Complexity of Algorithm 1

Next we present the communication complexity of Algorithm 1 with synchronization condition (3.3).

Lemma D.1. The total number of communication rounds between the agents and the server in Algorithm 1 is bounded by $\text{CPX} = \widetilde{\mathcal{O}}((d + K/\gamma)MH)$.

Remark D.2. We provide a refined analysis in Appendix D to get this improved result based on that of Dubey & Pentland (2021), which studied the same communication procedure as ours. When we choose $\gamma = \mathcal{O}(K/dM)$, the communication complexity reduces to $\widetilde{\mathcal{O}}(dHM^2)$. Note that Min et al. (2023) studied the asynchronous setting where only one agent is active in each episode, giving out the regret $\widetilde{\mathcal{O}}\big(d^{3/2}H^2\sqrt{K}\big)$ with the communication complexity $\widetilde{\mathcal{O}}(dHM^2)$. It is interesting to see that our algorithm, though in the synchronous setting, has the same communication complexity as the asynchronous variant. This implies that the asynchronous algorithm can only circumvent current communication by delaying it to the future but does not decrease the communication complexity. In fact, the synchronous setting can learn the policy better in our work, which is indicated by comparison of the average regret (the cumulative regret divided by the total number of samples used by the algorithm) in Table 1. By achieving a matched communication complexity, we find that synchronous and asynchronous settings have their own advantages and cannot replace each other. This phenomenon can help us better understand the properties of these two communication schemes.

The proof of this lemma is largely inspired by that in Dubey & Pentland (2021). However, we provide a refined analysis here, and thus obtain an improved communication complexity $\mathcal{O}(dHM^2)$, in contrast with the $\mathcal{O}(dHM^3)$ complexity in their paper. We also discussed this in Remark D.2 and showed that our result matches that of a recently proposed asynchronous algorithm.

Proof of Lemma D.1. We assume $\sigma = \{\sigma_1, \ldots, \sigma_n\}$ as the synchronization episodes, where $\sigma_i \in [K]$, we also denote $\sigma_0 = 0$. To bound the number of synchronization n, we separate σ into two parts with an undetermined term α

$$I_1 = \{i \in [n] | \sigma_i - \sigma_{i-1} \leq \alpha\},$$
$$I_2 = \{i \in [n] | \sigma_i - \sigma_{i-1} > \alpha\}.$$

Then we have $n = |I_1| + |I_2|$. Note that

$$K \geq \sigma_n = \sum_{i=1}^{n} (\sigma_i - \sigma_{i-1}) \geq \sum_{i \in I_2} (\sigma_i - \sigma_{i-1}) > |I_2|\alpha.$$

Then we have $|I_2| < K/\alpha$. Then note that

$$\begin{aligned}
\sum_{i=1}^{n} \log\left(\frac{\det(\boldsymbol{\Lambda}_{m,h}^{\sigma_i})}{\det(\boldsymbol{\Lambda}_{m,h}^{\sigma_{i-1}})}\right) &\geq \sum_{i \in I_1} \log\left(\frac{\det(\boldsymbol{\Lambda}_{m,h}^{\sigma_i})}{\det(\boldsymbol{\Lambda}_{m,h}^{\sigma_{i-1}})}\right) \\
&\geq \sum_{i \in I_1} \frac{\gamma}{\sigma_i - \sigma_{i-1}} \\
&\geq |I_1| \frac{\gamma}{\alpha}.
\end{aligned} \tag{D.1}$$

Define $\mathbf{\Lambda}_h^K = \sum_{m\in\mathcal{M}}\sum_{k=1}^K \boldsymbol{\phi}(z_{m,h}^k)\boldsymbol{\phi}(z_{m,h}^k)^\top + \lambda\mathbf{I}$ where $z_{m,h}^k = (s_{m,h}^k, a_{m,h}^k)$. On the other hand, we have

$$\begin{aligned}\sum_{i=1}^n \log\left(\frac{\det(\mathbf{\Lambda}_{m,h}^{\sigma_i})}{\det(\mathbf{\Lambda}_{m,h}^{\sigma_{i-1}})}\right) &= \log\left(\frac{\det(\mathbf{\Lambda}_{m,h}^{\sigma_n})}{\det(\mathbf{\Lambda}_{m,h}^{\sigma_0})}\right)\\ &\le \log\left(\frac{\det(\mathbf{\Lambda}_h^K)}{\det(\lambda\mathbf{I})}\right)\\ &\le d\log(1+MK/d), \end{aligned} \tag{D.2}$$

where the first inequality holds due to the trivial fact that $\mathbf{A} \preccurlyeq \mathbf{B} \Rightarrow \det(\mathbf{A}) \le \det(\mathbf{B})$, the second inequality follow from Lemma L.2 and the fact that $\|\boldsymbol{\phi}(\cdot)\|_2 \le 1$. Combine (D.1) and (D.2), then we have $|I_1| \le d\alpha/\gamma \log(1+MK/d)$. Finally, we choose $\alpha = K/d$, then we have

$$n \le \frac{K}{\alpha} + \frac{d\alpha}{\gamma}\log\left(1+\frac{MK}{d}\right) = \left(d+\frac{K}{\gamma}\right)\log\left(1+\frac{MK}{d}\right).$$

When one synchronization occurs, communications between agents and the server will occur M times because we have M agents in total. Recall from Lines 16-24 in Algorithm 1, also note that in one synchronization episode, communications will happen H times between every agent and the server. Finally, the upper bound of communication complexity is

$$\text{CPX} = \widetilde{\mathcal{O}}\big((d+K/\gamma)MH\big).$$

This completes the proof. □

E Misspecified Setting

In this part, we extend our theoretical analysis to the misspecified setting. In this setting, the transition functions $\mathbb{P}_{m,h}$ and the reward functions $r_{m,h}$ are heterogeneous across different MDPs, which is slightly more complicated than the homogeneous setting. Moreover, instead of assuming the transition and reward are linear, we only require each individual MDP is a ζ-approximate linear MDP (Jin et al., 2020) where both the transition and reward are approximately linear up to an controlled error ζ.

Definition E.1 (Misspecified Parallel MDPs)**.** For any $0 < \zeta \le 1$, and for any agent $m \in \mathcal{M}$, the corresponding MDP$(\mathcal{S}, \mathcal{A}, \mathrm{H}, \mathbb{P}_m, \mathbf{r}_m)$ is a ζ-approximate linear MDP with a feature map $\boldsymbol{\phi} : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$, for any $h \in [H]$, there exist d unknown (signed) measures $\boldsymbol{\mu}_h = (\mu_h^{(1)}, \ldots, \mu_h^{(d)})$ over $\mathcal{S}$ and an unknown vector $\boldsymbol{\theta}_h \in \mathbb{R}^d$ such that for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have

$$\begin{aligned}&\big\|\mathbb{P}_{m,h}(\cdot \mid s,a) - \big\langle\boldsymbol{\phi}(s,a), \boldsymbol{\mu}_h(\cdot)\big\rangle\big\|_{\mathrm{TV}} \le \zeta,\\ &\big|r_{m,h}(s,a) - \langle\boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h\rangle\big| \le \zeta,\end{aligned}$$

where $\|\cdot\|_{\mathrm{TV}}$ is the total variation norm, for two distributions P_1 and P_2, we define it as: $\|P_1 - P_2\|_{\mathrm{TV}} = \frac{1}{2}\sum_{\mathbf{x}\in\mathbf{\Omega}} |P_1(\mathbf{x}) - P_2(\mathbf{x})|$. Without loss of generality, we assume that $\|\boldsymbol{\phi}(s,a)\| \le 1$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$, and $\max\{\|\boldsymbol{\mu}_h(\mathcal{S})\|, \|\boldsymbol{\theta}_h\|\} \le \sqrt{d}$ for all $h \in [H]$ and $m \in \mathcal{M}$.

Remark E.2. Note that our misspecified setting defined in Definition E.1 is a generalized notion of misspecification in Jin et al. (2020). Moreover, our misspecified setting is also more general and cover the small heterogeneous setting mentioned in Dubey & Pentland (2021). The triangle inequality can easily be used to derive small heterogeneous setting from our misspecified setting, but not vice versa.

Next we state our regret bound for CoopTS-PHE in the misspecified setting.

Theorem E.3 (Misspecified Regret Bound for CoopTS-PHE)**.** In CoopTS-PHE (Algorithm 1+Algorithm 2), under Definition E.1 and determinant synchronization condition (3.3), with the same initialization with Theorem 4.2, we obtain the following cumulative regret

$$\text{Regret}(K) = \widetilde{\mathcal{O}}\Big(d^{\frac{3}{2}}H^2\sqrt{M}\big(\sqrt{dM\gamma}+\sqrt{K}\big) + dH^2M\sqrt{K}\big(\sqrt{dM\gamma}+\sqrt{K}\big)\zeta\Big),$$

with probability at least $1-\delta$.

Remark E.4. When we choose $\zeta = \mathcal{O}\big(\sqrt{d/MK}\big)$, the cumulative regret becomes $\widetilde{\mathcal{O}}\big(d^{\frac{3}{2}}H^2\sqrt{M}\big(\sqrt{dM\gamma}+\sqrt{K}\big)\big)$. This matches the result of Theorem 4.2 in the linear MDP setting.

Similarly, we can have the following result for CoopTS-LMC.

Theorem E.5 (Misspecified Regret Bound for CoopTS-LMC)**.** In CoopTS-LMC (Algorithm 1+Algorithm 3), under Definition E.1 and determinant synchronization condition (3.3), with the same initialization with Theorem 4.4 except that $1/\sqrt{\beta_{m,k}} = \widetilde{\mathcal{O}}\big(H\sqrt{d}+H\sqrt{MKd}\zeta\big)$, we obtain the following cumulative regret

$$\text{Regret}(K) = \widetilde{\mathcal{O}}\Big(d^{\frac{3}{2}}H^2\sqrt{M}\big(\sqrt{dM\gamma}+\sqrt{K}\big) + d^{\frac{3}{2}}H^2M\sqrt{K}\big(\sqrt{dM\gamma}+\sqrt{K}\big)\zeta\Big),$$

with probability at least $1-\delta$.

Remark E.6. When we choose $\zeta = \mathcal{O}\big(\sqrt{1/MK}\big)$, the cumulative regret becomes $\widetilde{\mathcal{O}}\big(d^{\frac{3}{2}}H^2\sqrt{M}\big(\sqrt{dM\gamma}+\sqrt{K}\big)\big)$. This matches the result of Theorem 4.4 in the linear MDP setting. By comparing Theorems E.3 and E.5, we find the result of CoopTS-LMC has an extra $\sqrt{d}$ factor worse than that of CoopTS-PHE, causing the chosen ζ in CoopTS-PHE has an extra $\sqrt{d}$ order over that in CoopTS-LMC. This indicates that CoopTS-PHE has better performance tolerance for the misspecified setting.

F Proof of the Regret Bound for CoopTS-LMC

The general framework for CoopTS-LMC and CoopTS-PHE is closely similar. To make the article more concise, we first prove CoopTS-LMC completely, which is a bit more complicated. Then we can simplify the following similar proof for CoopTS-PHE in Appendix I.

F.1 Supporting Lemmas

Before deriving the regret bound for CoopTS-LMC, we first provide the necessary technical lemmas for our regret analysis. Note that the loop (Line 3-9) in Algorithm 3 is to do multi-sampling for N times. To simplify the notations, we eliminate the index n before Lemma F.7 because the previous lemmas have nothing to do with multi-sampling.

Definition F.1 (Model prediction error)**.** For any $(m,k,h) \in \mathcal{M}\times[K]\times[H]$, we define the model error associated with the reward r_h,

$$l^k_{m,h}(s,a) = r_h(s,a) + \mathbb{P}_h V^k_{m,h+1}(s,a) - Q^k_{m,h}(s,a).$$

Definition F.2 (Filtration)**.** For any $(m,k,h) \in \mathcal{M}\times[K]\times[H]$, we define the filtration $\mathcal{F}_{m,k,h}$ as

$$\mathcal{F}_{m,k,h} = \sigma\Big(\big\{\big(s^\tau_{n,i}, a^\tau_{n,i}\big)\big\}_{(n,\tau,i)\in\mathcal{M}\times[k-1]\times[H]} \bigcup \big\{\big(s^k_{n,i}, a^k_{n,i}\big)\big\}_{(n,i)\in[m-1]\times[H]} \bigcup \big\{\big(s^k_{m,i}, a^k_{m,i}\big)\big\}_{i\in[h]}\Big).$$

Proposition F.3. In Algorithm 3, the parameter $\mathbf{w}^{k,J_k}_{m,h}$ satisfies the Gaussian distribution $\mathcal{N}\big(\boldsymbol{\mu}^{k,J_k}_{m,h}, \boldsymbol{\Sigma}^{k,J_k}_{m,h}\big)$, where mean vector and the covariance matrix are defined as

$$\boldsymbol{\mu}^{k,J_k}_{m,h} = \mathbf{A}^{J_k}_k \ldots \mathbf{A}^{J_1}_1 \mathbf{w}^{1,0}_{m,h} + \sum_{i=1}^{k} \mathbf{A}^{J_k}_k \ldots \mathbf{A}^{J_{i+1}}_{i+1}\big(\mathbf{I} - \mathbf{A}^{J_i}_i\big)\widehat{\mathbf{w}}^i_{m,h},$$

$$\boldsymbol{\Sigma}^{k,J_k}_{m,h} = \sum_{i=1}^{k} \frac{1}{\beta_{m,i}} \mathbf{A}^{J_k}_k \ldots \mathbf{A}^{J_{i+1}}_{i+1}\big(\mathbf{I} - \mathbf{A}^{2J_i}_i\big)(\boldsymbol{\Lambda}^i_{m,h})^{-1}(\mathbf{I}+\mathbf{A}_i)^{-1}\mathbf{A}^{J_{i+1}}_{i+1}\ldots\mathbf{A}^{J_k}_k,$$

where $\mathbf{A}_i = \mathbf{I} - 2\eta_{m,i}\boldsymbol{\Lambda}^i_{m,h}$ for $i\in[k]$.

Lemma F.4. For any $(m,k,h) \in \mathcal{M} \times [K] \times [H]$, the unperturbed estimated parameter $\widehat{\mathbf{w}}_{m,h}^k$ satisfies

$$\left\|\widehat{\mathbf{w}}_{m,h}^k\right\| \leq 2H\sqrt{Mkd/\lambda}.$$

Lemma F.5. Let $\lambda = 1$ in Algorithm 3. For any fixed $0 < \delta < 1$, with probability at least $1 - \delta^2$, for any $(m,k,h) \in \mathcal{M} \times [K] \times [H]$ and for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\left|\boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^{k,J_k} - \boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}_{m,h}^k\right| \leq \left(5\sqrt{\frac{2d\log(1/\delta)}{3\beta_K}} + \frac{4}{3}\right)\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}.$$

Lemma F.6. Let $\lambda = 1$ in Algorithm 3. For any fixed $0 < \delta < 1$, with probability at least $1 - \delta$, for any $(m,k,h) \in \mathcal{M} \times [K] \times [H]$, we have

$$\left\|\mathbf{w}_{m,h}^{k,J_k}\right\| \leq \frac{16}{3}Hd\sqrt{MK} + \sqrt{\frac{2K}{3\beta_K\delta}}d^{3/2} \stackrel{\text{def}}{=} B_\delta,$$

Lemma F.7. Let $\lambda = 1$ in Algorithm 3. For any fixed $0 < \delta < 1$, with probability at least $1 - \delta$, for all $(m,k,h) \in \mathcal{M} \times [K] \times [H]$, we have

$$\left\|\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \boldsymbol{\phi}(s^l,a^l)\left[\left(V_{m,h+1}^k - \mathbb{P}_h V_{m,h+1}^k\right)(s^l,a^l)\right]\right\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} \leq 3H\sqrt{d}C_\delta,$$

where $C_\delta = \left[\frac{1}{2}\log(K+1) + \log\left(\frac{2\sqrt{2}KB_{\delta/2NMHK}}{H}\right) + \log\frac{3}{\delta}\right]^{1/2}$ and B_δ is defined in Lemma F.6.

Lemma F.8. Let $\lambda = 1$ in Algorithm 3. Under Definition 4.1, for any fixed $0 < \delta < 1$, with probability at least $1 - \delta$, for all $(m,k,h) \in \mathcal{M} \times [K] \times [H]$ and for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\left|\boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}_{m,h}^k - r_h(s,a) - \mathbb{P}_h V_{m,h+1}^k(s,a)\right| \leq 5H\sqrt{d}C_\delta\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}.$$

Lemma F.9 (Error bound)**.** Let $\lambda = 1$ in Algorithm 3. Under Definition 4.1, for any fixed $0 < \delta < 1$, with probability at least $1 - \delta - \delta^2$, for any $(m,k,h) \in \mathcal{M} \times [K] \times [H]$ and for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$-l_{m,h}^k(s,a) \leq \left(5H\sqrt{d}C_\delta + 5\sqrt{\frac{2d\log\left(\sqrt{N}/\delta\right)}{3\beta_K}} + \frac{4}{3}\right)\|\boldsymbol{\phi}(s,a)\|_{\left(\boldsymbol{\Lambda}_{m,h}^k\right)^{-1}},$$

where C_δ is defined in Lemma F.7.

Lemma F.10 (Optimism)**.** Let $\lambda = 1$ in Algorithm 3 and $c_0' = 1 - \frac{1}{2\sqrt{2e\pi}}$. Under Definition 4.1, for any fixed $0 < \delta < 1$, with probability at least $1 - |\mathcal{C}(\varepsilon)|{c_0'}^N - 2\delta$ where $|\mathcal{C}(\varepsilon)| \leq (3/\varepsilon)^d$, for all $(m,h,k) \in \mathcal{M} \times [H] \times [K]$ and for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$l_{m,h}^k(s,a) \leq \alpha_\delta\varepsilon,$$

where $\alpha_\delta = \sqrt{MK}\left(2H\sqrt{d} + B_{\delta/NMHK}\right)$.

Remark F.11. Here we point out that in our proofs for both CoopTS-LMC and CoopTS-PHE, we use a new ε-covering technique to prove that the optimism lemma holds for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ instead of just the state-action pairs encountered by the algorithm, which is essential in applying this lemma to bound the term $\mathbb{E}_{\pi^*}[l_{m,h}^k(s_{m,h},a_{m,h})|s_{m,1} = s_{m,1}^k]$ in (F.2) in the regret analysis. This was ignored by previous works (Cai et al., 2020; Ishfaq et al., 2024) that use the same regret decomposition technique in the single-agent setting.

The following lemma gives the upper bound of self-normalized term summation in the multi-agent setting, which is first introduced by Lemma 9 in Dubey & Pentland (2021). To make our analysis complete, we give out the proof in the Appendix G.9 where we make some necessary modifications compared with Lemma 9 in Dubey & Pentland (2021).

Lemma F.12. Let Algorithm 2 run for any $K > 0$, $M \geq 1$, and γ as the communication control factor. Define $\mathbf{\Lambda}_h^K = \sum_{m\in\mathcal{M}}\sum_{k=1}^K \boldsymbol{\phi}(s_{m,h}^k, a_{m,h}^k)\boldsymbol{\phi}(s_{m,h}^k, a_{m,h}^k)^\top + \lambda\mathbf{I}$, then we have

$$\sum_{m\in\mathcal{M}}\sum_{k=1}^K \left\|\boldsymbol{\phi}(s_{m,h}^k, a_{m,h}^k)\right\|_{(\mathbf{\Lambda}_{m,h}^k)^{-1}} \leq \left(\log\left(\frac{\det(\mathbf{\Lambda}_h^K)}{\det(\lambda\mathbf{I})}\right)+1\right)M\sqrt{\gamma} + 2\sqrt{MK\log\left(\frac{\det(\mathbf{\Lambda}_h^K)}{\det(\lambda\mathbf{I})}\right)}.$$

The following lemma shows that we can decompose the regret of Algorithm 2 into three different components. The proof of this lemma closely resembles Lemma 4.2 in Cai et al. (2020) for the single-agent setting. When we fix the agent $m \in \mathcal{M}$, it is totally same as Lemma 4.2 in Cai et al. (2020).

Lemma F.13 (Lemma 4.2 in Cai et al. (2020))**.** Define the operators and the following terms:

$$\begin{aligned}
(\mathbb{J}_{m,h}f)(s) &= \left\langle f(s,\cdot), \pi_{m,h}^*(\cdot|s)\right\rangle, \quad (\mathbb{J}_{m,k,h}f)(s) = \left\langle f(s,\cdot), \pi_{m,h}^k(\cdot|s)\right\rangle, \\
D_{m,k,h,1} &= \left(\mathbb{J}_{m,k,h}\left(Q_{m,h}^k - Q_{m,h}^{\pi_{m,k}}\right)\right)(s_{m,h}^k) - \left(Q_{m,h}^k - Q_{m,h}^{\pi_{m,k}}\right)(s_{m,h}^k, a_{m,h}^k), \\
D_{m,k,h,2} &= \left(\mathbb{P}_{m,h}\left(V_{m,h+1}^k - V_{m,h+1}^{\pi_{m,k}}\right)\right)(s_{m,h}^k, a_{m,h}^k) - \left(V_{m,h+1}^k - V_{m,h+1}^{\pi_{m,k}}\right)(s_{m,h+1}^k).
\end{aligned} \tag{F.1}$$

Then we can decompose the regret into the following form:

$$\begin{aligned}
\text{Regret}(K) &= \sum_{m\in\mathcal{M}}\sum_{k=1}^K V_{m,1}^*(s_{m,1}^k) - V_{m,1}^{\pi_{m,k}}(s_{m,1}^k) \\
&= \underbrace{\sum_{m\in\mathcal{M}}\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\left\langle Q_{m,h}^k(s_{m,h},\cdot), \pi_{m,h}^*(\cdot,|s_{m,h}) - \pi_{m,h}^k(\cdot|s_{m,h})\right\rangle | s_{m,1} = s_{m,1}^k\right]}_{(i)} \\
&\quad + \underbrace{\sum_{m\in\mathcal{M}}\sum_{k=1}^K\sum_{h=1}^H (D_{m,k,h,1} + D_{m,k,h,2})}_{(ii)} \\
&\quad + \underbrace{\sum_{m\in\mathcal{M}}\sum_{k=1}^K\sum_{h=1}^H \left(\mathbb{E}_{\pi^*}\left[l_{m,h}^k(s_{m,h}, a_{m,h}) | s_{m,1} = s_{m,1}^k\right] - l_{m,h}^k(s_{m,h}^k, a_{m,h}^k)\right)}_{(iii)}.
\end{aligned}$$

F.2 Regret Analysis

In this part, we give out the proof of Theorem 4.4, the regret bound for CoopTS-LMC.

Proof of Theorem 4.4. Based on the result from Lemma F.13, we do the regret decomposition first

$$\begin{aligned}
\text{Regret}(K) &= \sum_{m\in\mathcal{M}}\sum_{k=1}^K V_{m,1}^*(s_{m,1}^k) - V_{m,1}^{\pi_{m,k}}(s_{m,1}^k) \\
&= \underbrace{\sum_{m\in\mathcal{M}}\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\left\langle Q_{m,h}^k(s_{m,h},\cdot), \pi_{m,h}^*(\cdot,|s_{m,h}) - \pi_{m,h}^k(\cdot|s_{m,h})\right\rangle | s_{m,1} = s_{m,1}^k\right]}_{(i)} \\
&\quad + \underbrace{\sum_{m\in\mathcal{M}}\sum_{k=1}^K\sum_{h=1}^H (D_{m,k,h,1} + D_{m,k,h,2})}_{(ii)}
\end{aligned}$$

$$+ \underbrace{\sum_{m\in\mathcal{M}}\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\mathbb{E}_{\pi^*}\left[l_{m,h}^k(s_{m,h},a_{m,h})|s_{m,1}=s_{m,1}^k\right]-l_{m,h}^k\left(s_{m,h}^k,a_{m,h}^k\right)\right)}_{(iii)}. \tag{F.2}$$

Next, we will bound the above three terms respectively.

Bounding Term (i) in (F.2): for the policy $\pi_{m,h}^k$, we have

$$\sum_{m\in\mathcal{M}}\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left\langle Q_{m,h}^k(s_{m,h},\cdot),\pi_{m,h}^*(\cdot,|s_{m,h})-\pi_{m,h}^k(\cdot|s_{m,h})\right\rangle|s_{m,1}=s_{m,1}^k\right]\leq 0. \tag{F.3}$$

This is because by definition $\pi_{m,h}^k$ is the greedy policy for $Q_{m,h}^k$.

Bounding Term (ii) in (F.2): note that $0\leq Q_{m,h}^k\leq H-h+1\leq H$, based on (F.1), for any $(m,k,h)\in\mathcal{M}\times[K]\times[H]$, we have $|D_{m,k,h,1}|\leq 2H$ and $|D_{m,k,h,2}|\leq 2H$. Note that $D_{m,k,h,1}$ is a martingale difference sequence $\mathbb{E}[D_{m,k,h,1}|\mathcal{F}_{m,k,h}]=0$. By applying Azuma-Hoeffding inequality, with probability at least $1-\delta/3$, we have

$$\sum_{m\in\mathcal{M}}\sum_{k=1}^{K}\sum_{h=1}^{H}D_{m,k,h,1}\leq 2\sqrt{2MH^3K\log(6/\delta)}.$$

Note that $D_{m,k,h,2}$ is also a martingale difference sequence. By applying Azuma-Hoeffding inequality, with probability at least $1-\delta/3$, we have

$$\sum_{m\in\mathcal{M}}\sum_{k=1}^{K}\sum_{h=1}^{H}D_{m,k,h,2}\leq 2\sqrt{2MH^3K\log(6/\delta)}.$$

By taking union bound, with probability at least $1-2\delta/3$, we have

$$\sum_{m\in\mathcal{M}}\sum_{k=1}^{K}\sum_{h=1}^{H}D_{m,k,h,1}+\sum_{m\in\mathcal{M}}\sum_{k=1}^{K}\sum_{h=1}^{H}D_{m,k,h,2}\leq 4\sqrt{2MH^3K\log(6/\delta)}. \tag{F.4}$$

Bounding Term (iii) in (F.2): based on Lemma F.9 and Lemma F.10, by taking union bound, with probability at least $1-|\mathcal{C}(\varepsilon)|{c_0'}^N-2\delta'-MHK(\delta'+{\delta'}^2)$, we have

$$\begin{aligned}
&\sum_{m\in\mathcal{M}}\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\mathbb{E}_{\pi^*}\left[l_{m,h}^k(s_{m,h},a_{m,h})|s_{m,1}=s_{m,1}^k\right]-l_{m,h}^k\left(s_{m,h}^k,a_{m,h}^k\right)\right)\\
&\leq\sum_{m\in\mathcal{M}}\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\alpha_{\delta'}\varepsilon-l_{m,h}^k\left(s_{m,h}^k,a_{m,h}^k\right)\right)\\
&\leq HMK\alpha_{\delta'}\varepsilon+\sum_{m\in\mathcal{M}}\sum_{k=1}^{K}\sum_{h=1}^{H}\left(5H\sqrt{d}C_{\delta'}+5\sqrt{\frac{2d\log(\sqrt{N}/\delta')}{3\beta_K}}+\frac{4}{3}\right)\left\|\boldsymbol{\phi}(s_{m,h}^k,a_{m,h}^k)\right\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}\\
&=HMK\alpha_{\delta'}\varepsilon+\left(5H\sqrt{d}C_{\delta'}+5\sqrt{\frac{2d\log(\sqrt{N}/\delta')}{3\beta_K}}+\frac{4}{3}\right)\sum_{h=1}^{H}\sum_{m\in\mathcal{M}}\sum_{k=1}^{K}\left\|\boldsymbol{\phi}(s_{m,h}^k,a_{m,h}^k)\right\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}\\
&\leq HMK\alpha_{\delta'}\varepsilon+\left(5H\sqrt{d}C_{\delta'}+5\sqrt{\frac{2d\log(\sqrt{N}/\delta')}{3\beta_K}}+\frac{4}{3}\right)\\
&\quad\times\sum_{h=1}^{H}\left(\log\left(\frac{\det(\boldsymbol{\Lambda}_h^K)}{\det(\lambda\mathbf{I})}\right)+1\right)M\sqrt{\gamma}+2\sqrt{MK\log\left(\frac{\det(\boldsymbol{\Lambda}_h^K)}{\det(\lambda\mathbf{I})}\right)}
\end{aligned}$$

$$\leq HMK\alpha_{\delta'}\varepsilon + \left(5H\sqrt{d}C_{\delta'} + 5\sqrt{\frac{2d\log(\sqrt{N}/\delta')}{3\beta_K}} + \frac{4}{3}\right)$$
$$\times H\Big(d(\log(1+MK/d)+1)M\sqrt{\gamma} + 2\sqrt{MKd\log(1+MK/d)}\Big).$$

The first inequality follows from Lemma F.10, the second inequality follows from Lemma F.9, the third inequality follows from Lemma F.12, the last inequality holds due to Lemma L.2 and the fact that $\|\boldsymbol{\phi}(\cdot)\|_2 \leq 1$.

Here we choose $\varepsilon = dH\sqrt{d/MK}/\alpha_{\delta'} = \widetilde{\mathcal{O}}(\sqrt{1/dHM^3K^4N})$ and choose $\frac{1}{\sqrt{\beta_K}} = 20H\sqrt{d}C_{\delta'} + \frac{16}{3}$, we have

$$\sum_{m\in\mathcal{M}}\sum_{k=1}^{K}\sum_{h=1}^{H}\big(\mathbb{E}_{\pi^*}\big[l_{m,h}^k(s_{m,h},a_{m,h})|s_{m,1}=s_{m,1}^k\big] - l_{m,h}^k\big(s_{m,h}^k,a_{m,h}^k\big)\big) \leq \widetilde{\mathcal{O}}\big(dH^2\big(dM\sqrt{\gamma}+\sqrt{dMK}\big)\big), \tag{F.5}$$

occurs with probability at least $1 - |\mathcal{C}(\varepsilon)|{c_0'}^N - 2\delta' - MHK(\delta' + {\delta'}^2)$.

We set $\delta' = \delta/12(MHK+1)$ and choose $N = \bar{C}\log(\delta)/\log(c_0')$ where $\bar{C} = \widetilde{\mathcal{O}}(d)$, then we have

$$1 - |\mathcal{C}(\varepsilon)|{c_0'}^N - 2\delta' - MHK(\delta' + {\delta'}^2) \geq 1 - \delta/3.$$

Combining Terms (i)(ii)(iii) together: Based on (F.3), (F.4) and (F.5). By taking union bound, we get that the final regret bound for CoopTS-LMC is $\widetilde{\mathcal{O}}\big(dH^2\big(dM\sqrt{\gamma}+\sqrt{dMK}\big)\big)$ with probability at least $1-\delta$. □

G Proof of Supporting Lemmas in Appendix F

G.1 Proof of Proposition F.3

Recall from Algorithm 3, the LMC update rule is

$$\mathbf{w}_{m,h}^{k,j} = \mathbf{w}_{m,h}^{k,j-1} - \eta_{m,k}\nabla L_{m,h}^k\big(\mathbf{w}_{m,h}^{k,j-1}\big) + \sqrt{2\eta_{m,k}\beta_{m,k}^{-1}}\boldsymbol{\epsilon}_{m,h}^{k,j},$$

where we have $\nabla L_{m,h}^k\big(\mathbf{w}_{m,h}^{k,j-1}\big) = 2\big(\boldsymbol{\Lambda}_{m,h}^k\mathbf{w}_{m,h}^{k,j-1} - \boldsymbol{b}_{m,h}^k\big)$. Plug in the above formula, then we can calculate that

$$\begin{aligned}
\mathbf{w}_{m,h}^{k,J_k} &= \mathbf{w}_{m,h}^{k,J_k-1} - 2\eta_{m,k}\big(\boldsymbol{\Lambda}_{m,h}^k\mathbf{w}_{m,h}^{k,J_k-1} - \boldsymbol{b}_{m,h}^k\big) + \sqrt{2\eta_{m,k}\beta_{m,k}^{-1}}\boldsymbol{\epsilon}_{m,h}^{k,J_k} \\
&= \big(\mathbf{I} - 2\eta_{m,k}\boldsymbol{\Lambda}_{m,h}^k\big)\mathbf{w}_{m,h}^{k,J_k-1} + 2\eta_{m,k}\boldsymbol{b}_{m,h}^k + \sqrt{2\eta_{m,k}\beta_{m,k}^{-1}}\boldsymbol{\epsilon}_{m,h}^{k,J_k} \\
&= \big(\mathbf{I} - 2\eta_{m,k}\boldsymbol{\Lambda}_{m,h}^k\big)^{J_k}\mathbf{w}_{m,h}^{k,0} + \sum_{l=0}^{J_k-1}(\mathbf{I} - 2\eta_{m,k}\boldsymbol{\Lambda}_{m,h}^k)^l\Big(2\eta_{m,k}\boldsymbol{b}_{m,h}^k + \sqrt{2\eta_{m,k}\beta_{m,k}^{-1}}\boldsymbol{\epsilon}_{m,h}^{k,J_k-l}\Big) \\
&= \big(\mathbf{I} - 2\eta_{m,k}\boldsymbol{\Lambda}_{m,h}^k\big)^{J_k}\mathbf{w}_{m,h}^{k,0} + 2\eta_{m,k}\sum_{l=0}^{J_k-1}\big(\mathbf{I} - 2\eta_{m,k}\boldsymbol{\Lambda}_{m,h}^k\big)^l\boldsymbol{b}_{m,h}^k \\
&\quad + \sqrt{2\eta_{m,k}\beta_{m,k}^{-1}}\sum_{l=0}^{J_k-1}\big(\mathbf{I} - 2\eta_{m,k}\boldsymbol{\Lambda}_{m,h}^k\big)^l\boldsymbol{\epsilon}_{m,h}^{k,J_k-l},
\end{aligned}$$

where the third equality follows from iteration. Denote that $\mathbf{A}_i = \mathbf{I} - 2\eta_{m,i}\boldsymbol{\Lambda}_{m,h}^i$. Moreover, we choose the step size such that $0 < \eta_{m,i} < 1/\big(2\lambda_{\max}\big(\boldsymbol{\Lambda}_{m,h}^i\big)\big)$. Thus we have

$$\mathbf{w}_{m,h}^{k,J_k} = \mathbf{A}_k^{J_k}\mathbf{w}_{m,h}^{k-1,J_{k-1}} + 2\eta_{m,k}\sum_{l=0}^{J_k-1}\mathbf{A}_k^l\boldsymbol{\Lambda}_{m,h}^k\widehat{\mathbf{w}}_{m,h}^k + \sqrt{2\eta_{m,k}\beta_{m,k}^{-1}}\sum_{l=0}^{J_k-1}\mathbf{A}_k^l\boldsymbol{\epsilon}_{m,h}^{k,J_k-l}$$

$$
\begin{aligned}
&= \mathbf{A}_k^{J_k}\mathbf{w}_{m,h}^{k-1,J_{k-1}} + (\mathbf{I}-\mathbf{A}_k)\big(\mathbf{I}+\mathbf{A}_k+...+\mathbf{A}_k^{J_k-1}\big)\widehat{\mathbf{w}}_{m,h}^k + \sqrt{2\eta_{m,k}\beta_{m,k}^{-1}}\sum_{l=0}^{J_k-1}\mathbf{A}_k^l\boldsymbol{\epsilon}_{m,h}^{k,J_k-l} \\
&= \mathbf{A}_k^{J_k}\mathbf{w}_{m,h}^{k-1,J_{k-1}} + \big(\mathbf{I}-\mathbf{A}_k^{J_k}\big)\widehat{\mathbf{w}}_{m,h}^k + \sqrt{2\eta_{m,k}\beta_{m,k}^{-1}}\sum_{l=0}^{J_k-1}\mathbf{A}_k^l\boldsymbol{\epsilon}_{m,h}^{k,J_k-l} \\
&= \mathbf{A}_k^{J_k}...\mathbf{A}_1^{J_1}\mathbf{w}_{m,h}^{1,0} + \sum_{i=1}^{k}\mathbf{A}_k^{J_k}...\mathbf{A}_{i+1}^{J_{i+1}}\big(\mathbf{I}-\mathbf{A}_i^{J_i}\big)\widehat{\mathbf{w}}_{m,h}^i + \sum_{i=1}^{k}\sqrt{2\eta_{m,i}\beta_{m,i}^{-1}}\mathbf{A}_k^{J_k}...\mathbf{A}_{i+1}^{J_{i+1}}\sum_{l=0}^{J_i-1}\mathbf{A}_i^l\boldsymbol{\epsilon}_{m,h}^{i,J_i-l},
\end{aligned}
$$

where the first equality holds because $\boldsymbol{b}_{m,h}^k = \boldsymbol{\Lambda}_{m,h}^k\widehat{\mathbf{w}}_{m,h}^k$ and $\mathbf{w}_{m,h}^{k-1,J_{k-1}} = \mathbf{w}_{m,h}^{k,0}$, the third equality follows from the fact that $\mathbf{I}+\mathbf{A}+...+\mathbf{A}^{n-1} = (\mathbf{I}-\mathbf{A}^n)(\mathbf{I}-\mathbf{A})^{-1}$, and the fourth equality holds because of iteration.

Note that $\boldsymbol{\epsilon}_{m,h}^{i,J_i-l} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$, based on the property of multivariate Gaussian distribution, we have $\mathbf{w}_{m,h}^{k,J_k} \sim \mathcal{N}\big(\boldsymbol{\mu}_{m,h}^{k,J_k}, \boldsymbol{\Sigma}_{m,h}^{k,J_k}\big)$. Then we can directly get the mean vector

$$
\boldsymbol{\mu}_{m,h}^{k,J_k} = \mathbf{A}_k^{J_k}...\mathbf{A}_1^{J_1}\mathbf{w}_{m,h}^{1,0} + \sum_{i=1}^{k}\mathbf{A}_k^{J_k}...\mathbf{A}_{i+1}^{J_{i+1}}\big(\mathbf{I}-\mathbf{A}_i^{J_i}\big)\widehat{\mathbf{w}}_{m,h}^i.
$$

Next we will calculate the covariance matrix $\boldsymbol{\Sigma}_{m,h}^{k,J_k}$. For simplicity, we define $\mathbf{M}_i = \sqrt{2\eta_{m,i}\beta_{m,i}^{-1}}\mathbf{A}_k^{J_k}...\mathbf{A}_{i+1}^{J_{i+1}}$, thus we have

$$
\mathbf{M}_i\sum_{l=0}^{J_i-1}\mathbf{A}_i^l\boldsymbol{\epsilon}_{m,h}^{i,J_i-l} \sim \mathcal{N}\left(\mathbf{0}, \sum_{l=0}^{J_i-1}\mathbf{M}_i\mathbf{A}_i^l\big(\mathbf{M}_i\mathbf{A}_i^l\big)^\top\right) \sim \mathcal{N}\left(\mathbf{0}, \mathbf{M}_i\left(\sum_{l=0}^{J_i-1}\mathbf{A}_i^{2l}\right)\mathbf{M}_i^\top\right).
$$

Thus we get the covariance matrix $\boldsymbol{\Sigma}_{m,h}^{k,J_k}$,

$$
\begin{aligned}
\boldsymbol{\Sigma}_{m,h}^{k,J_k} &= \sum_{i=1}^{k}\mathbf{M}_i\left(\sum_{l=0}^{J_i-1}\mathbf{A}_i^{2l}\right)\mathbf{M}_i^\top \\
&= \sum_{i=1}^{k}2\eta_{m,i}\beta_{m,i}^{-1}\mathbf{A}_k^{J_k}...\mathbf{A}_{i+1}^{J_{i+1}}\left(\sum_{l=0}^{J_i-1}\mathbf{A}_i^{2l}\right)\mathbf{A}_{i+1}^{J_{i+1}}...\mathbf{A}_k^{J_k} \\
&= \sum_{i=1}^{k}2\eta_{m,i}\beta_{m,i}^{-1}\mathbf{A}_k^{J_k}...\mathbf{A}_{i+1}^{J_{i+1}}\big(\mathbf{I}-\mathbf{A}_i^{2J_i}\big)(\mathbf{I}-\mathbf{A}_i^2)^{-1}\mathbf{A}_{i+1}^{J_{i+1}}...\mathbf{A}_k^{J_k} \\
&= \sum_{i=1}^{k}\frac{1}{\beta_{m,i}}\mathbf{A}_k^{J_k}...\mathbf{A}_{i+1}^{J_{i+1}}\big(\mathbf{I}-\mathbf{A}_i^{2J_i}\big)\big(\boldsymbol{\Lambda}_{m,h}^i\big)^{-1}(\mathbf{I}+\mathbf{A}_i)^{-1}\mathbf{A}_{i+1}^{J_{i+1}}...\mathbf{A}_k^{J_k},
\end{aligned}
$$

where the third equality follows from the fact that $\mathbf{I}+\mathbf{A}+...+\mathbf{A}^{n-1} = (\mathbf{I}-\mathbf{A}^n)(\mathbf{I}-\mathbf{A})^{-1}$. Here we complete the proof.

G.2 Proof of Lemma F.4

Proof. Note that $\widehat{\mathbf{w}}_{m,h}^k = \big(\boldsymbol{\Lambda}_{m,k}^k\big)^{-1}\boldsymbol{b}_{m,h}^k$, we can calculate that

$$
\begin{aligned}
\big\|\widehat{\mathbf{w}}_{m,h}^k\big\| &= \left\|\big(\boldsymbol{\Lambda}_{m,h}^k\big)^{-1}\boldsymbol{b}_{m,h}^k\right\| \\
&= \left\|\big(\boldsymbol{\Lambda}_{m,h}^k\big)^{-1}\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)}\big[r_h\big(s^l,a^l\big)+V_{m,h+1}^k(s'^l)\big]\boldsymbol{\phi}\big(s^l,a^l\big)\right\| \\
&\le \frac{1}{\sqrt{\lambda}}\sqrt{\mathcal{K}(k)}\left(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)}\big\|\big[r_h\big(s^l,a^l\big)+V_{m,h+1}^k(s'^l)\big]\boldsymbol{\phi}\big(s^l,a^l\big)\big\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}^2\right)^{1/2}
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{2H}{\sqrt{\lambda}}\sqrt{\mathcal{K}(k)}\left(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \left\|\boldsymbol{\phi}(s^l,a^l)\right\|^2_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}}\right)^{1/2} \\
&\leq 2H\sqrt{\mathcal{K}(k)d/\lambda} \\
&\leq 2H\sqrt{Mkd/\lambda},
\end{aligned}
$$

where the first inequality follows from Lemma L.3, the second inequality is due to $0 \leq V^k_{m,h} \leq H-h+1$, $0 \leq r_h \leq 1$ and $\|\boldsymbol{\phi}(s,a)\| \leq 1$, the third inequality follows from Lemma L.4, and the last inequality holds because $\mathcal{K}(k) = (M-1)k_s + k - 1 \leq Mk$. □

G.3 Proof of Lemma F.5

Proof. We separate the error into two terms and bound them respectively,

$$
\left|\boldsymbol{\phi}(s,a)^\top \mathbf{w}^{k,J_k}_{m,h} - \boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}^k_{m,h}\right| \leq \underbrace{\left|\boldsymbol{\phi}(s,a)^\top\left(\mathbf{w}^{k,J_k}_{m,h} - \boldsymbol{\mu}^{k,J_k}_{m,h}\right)\right|}_{I_1} + \underbrace{\left|\boldsymbol{\phi}(s,a)^\top\left(\boldsymbol{\mu}^{k,J_k}_{m,h} - \widehat{\mathbf{w}}^k_{m,h}\right)\right|}_{I_2}. \tag{G.1}
$$

Bounding Term I_1 in (G.1): by Cauchy-Schwarz inequality, we have

$$
\left|\boldsymbol{\phi}(s,a)^\top\left(\mathbf{w}^{k,J_k}_{m,h} - \boldsymbol{\mu}^{k,J_k}_{m,h}\right)\right| \leq \left\|\boldsymbol{\phi}(s,a)\right\|_{\boldsymbol{\Sigma}^{k,J_k}_{m,h}} \cdot \left\|\mathbf{w}^{k,J_k}_{m,h} - \boldsymbol{\mu}^{k,J_k}_{m,h}\right\|_{(\boldsymbol{\Sigma}^{k,J_k}_{m,h})^{-1}}.
$$

By choosing $\eta_{m,k} \leq 1/(4\lambda_{\max}(\boldsymbol{\Lambda}^k_{m,h}))$ for all k and m, then we have

$$
\begin{aligned}
&\frac{1}{2}\mathbf{I} \preccurlyeq \mathbf{A}_k = \mathbf{I} - 2\eta_{m,k}\boldsymbol{\Lambda}^k_{m,h} \preccurlyeq (1 - 2\eta_{m,k}\lambda_{\min}(\boldsymbol{\Lambda}^k_{m,h}))\mathbf{I}, \\
&\frac{3}{2}\mathbf{I} \preccurlyeq \mathbf{I} + \mathbf{A}_k = 2\mathbf{I} - 2\eta_{m,k}\boldsymbol{\Lambda}^k_{m,h} \preccurlyeq 2\mathbf{I}.
\end{aligned} \tag{G.2}
$$

Recall the definition of $\boldsymbol{\Sigma}^{k,J_k}_{m,h}$ in Proposition F.3. By choosing $\beta_{m,i} = \beta_K$ for all $i \in [k]$ and $m \in \mathcal{M}$, then we have

$$
\begin{aligned}
\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}^{k,J_k}_{m,h}\boldsymbol{\phi}(s,a) &= \sum_{i=1}^{k}\frac{1}{\beta_{m,i}}\boldsymbol{\phi}(s,a)^\top \mathbf{A}_k^{J_k}\ldots\mathbf{A}_{i+1}^{J_{i+1}}\left(\mathbf{I} - \mathbf{A}_i^{2J_i}\right)\left(\boldsymbol{\Lambda}^i_{m,h}\right)^{-1}(\mathbf{I}+\mathbf{A}_i)^{-1}\mathbf{A}_{i+1}^{J_{i+1}}\ldots\mathbf{A}_k^{J_k}\boldsymbol{\phi}(s,a) \\
&\leq \frac{2}{3\beta_{m,i}}\sum_{i=1}^{k}\boldsymbol{\phi}(s,a)^\top \mathbf{A}_k^{J_k}\ldots\mathbf{A}_{i+1}^{J_{i+1}}\left(\left(\boldsymbol{\Lambda}^i_{m,h}\right)^{-1} - \mathbf{A}_i^{J_i}\left(\boldsymbol{\Lambda}^i_{m,h}\right)^{-1}\mathbf{A}_i^{J_i}\right)\mathbf{A}_{i+1}^{J_{i+1}}\ldots\mathbf{A}_k^{J_k}\boldsymbol{\phi}(s,a) \\
&= \frac{2}{3\beta_K}\sum_{i=1}^{k-1}\boldsymbol{\phi}(s,a)^\top \mathbf{A}_k^{J_k}\ldots\mathbf{A}_{i+1}^{J_{i+1}}\left(\left(\boldsymbol{\Lambda}^i_{m,h}\right)^{-1} - \left(\boldsymbol{\Lambda}^{i+1}_{m,h}\right)^{-1}\right)\mathbf{A}_{i+1}^{J_{i+1}}\ldots\mathbf{A}_k^{J_k}\boldsymbol{\phi}(s,a) \\
&\quad - \frac{2}{3\beta_K}\boldsymbol{\phi}(s,a)^\top \mathbf{A}_k^{J_k}\ldots\mathbf{A}_1^{J_1}(\boldsymbol{\Lambda}^1_{m,h})^{-1}\mathbf{A}_1^{J_1}\ldots\mathbf{A}_k^{J_k}\boldsymbol{\phi}(s,a) \\
&\quad + \frac{2}{3\beta_K}\boldsymbol{\phi}(s,a)^\top(\boldsymbol{\Lambda}^k_{m,h})^{-1}\boldsymbol{\phi}(s,a),
\end{aligned}
$$

where the first inequality follows from (G.2). By the definition of $\boldsymbol{\Lambda}^i_{m,h}$ and Woodbury formula, we have

$$
\begin{aligned}
\left(\boldsymbol{\Lambda}^i_{m,h}\right)^{-1} - \left(\boldsymbol{\Lambda}^{i+1}_{m,h}\right)^{-1} &= \left(\boldsymbol{\Lambda}^i_{m,h}\right)^{-1} - \left(\boldsymbol{\Lambda}^i_{m,h} + \sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)}\boldsymbol{\phi}\left(s^l,a^l\right)\boldsymbol{\phi}\left(s^l,a^l\right)^\top\right)^{-1} \\
&= (\boldsymbol{\Lambda}^i_{m,h})^{-1}\boldsymbol{\varphi}(\mathbf{I}_n + \boldsymbol{\varphi}^\top(\boldsymbol{\Lambda}^i_{m,h})^{-1}\boldsymbol{\varphi})^{-1}\boldsymbol{\varphi}^\top(\boldsymbol{\Lambda}^i_{m,h})^{-1},
\end{aligned}
$$

where $\boldsymbol{\varphi}$ is a matrix with the dimension of $d \times n$, n is the number difference of $\boldsymbol{\phi}\left(s^l,a^l\right)$ between $\left(\boldsymbol{\Lambda}^i_{m,h}\right)^{-1}$ and $\left(\boldsymbol{\Lambda}^{i+1}_{m,h}\right)^{-1}$ (*i.e.* we concatenate all $\boldsymbol{\phi}\left(s^l,a^l\right)$ into the matrix $\boldsymbol{\varphi}$). Note that $n \leq M$, then we have

$$
\boldsymbol{\phi}(s,a)^\top \mathbf{A}_k^{J_k}\ldots\mathbf{A}_{i+1}^{J_{i+1}}\left(\left(\boldsymbol{\Lambda}^i_{m,h}\right)^{-1} - \left(\boldsymbol{\Lambda}^{i+1}_{m,h}\right)^{-1}\right)\mathbf{A}_{i+1}^{J_{i+1}}\ldots\mathbf{A}_k^{J_k}\boldsymbol{\phi}(s,a)
$$

$$
\begin{aligned}
&= \boldsymbol{\phi}(s,a)^\top \mathbf{A}_k^{J_k} ... \mathbf{A}_{i+1}^{J_{i+1}} \Big(\big(\boldsymbol{\Lambda}_{m,h}^i\big)^{-1} \boldsymbol{\varphi} (\mathbf{I}_n + \boldsymbol{\varphi}^\top \big(\boldsymbol{\Lambda}_{m,h}^i\big)^{-1} \boldsymbol{\varphi})^{-1} \boldsymbol{\varphi}^\top \big(\boldsymbol{\Lambda}_{m,h}^i\big)^{-1} \Big) \mathbf{A}_{i+1}^{J_{i+1}} ... \mathbf{A}_k^{J_k} \boldsymbol{\phi}(s,a) \\
&\le \boldsymbol{\phi}(s,a)^\top \mathbf{A}_k^{J_k} ... \mathbf{A}_{i+1}^{J_{i+1}} \big(\boldsymbol{\Lambda}_{m,h}^i\big)^{-1} \boldsymbol{\varphi}\boldsymbol{\varphi}^\top \big(\boldsymbol{\Lambda}_{m,h}^i\big)^{-1} \mathbf{A}_{i+1}^{J_{i+1}} ... \mathbf{A}_k^{J_k} \boldsymbol{\phi}(s,a) \\
&= \big\| \boldsymbol{\phi}(s,a)^\top \mathbf{A}_k^{J_k} ... \mathbf{A}_{i+1}^{J_i+1} \big(\boldsymbol{\Lambda}_{m,h}^i\big)^{-1} \boldsymbol{\varphi} \big\|_2^2 \\
&\le \big\| \mathbf{A}_k^{J_k} ... \mathbf{A}_{i+1}^{J_i+1} (\boldsymbol{\Lambda}_{m,h}^i)^{-1/2} \boldsymbol{\phi}(s,a) \big\|_2^2 \cdot \big\| (\boldsymbol{\Lambda}_{m,h}^i)^{-1/2} \boldsymbol{\varphi} \big\|_F^2 \\
&\le \prod_{j=i+1}^{k} \Big(1 - 2\eta_{m,j} \lambda_{\min}\big(\boldsymbol{\Lambda}_{m,h}^j\big) \Big)^{2J_j} \operatorname{tr}\big(\boldsymbol{\varphi}^\top \big(\boldsymbol{\Lambda}_{m,h}^i\big)^{-1} \boldsymbol{\varphi}\big) \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^i)^{-1}}^2,
\end{aligned}
$$

where $\|\cdot\|_F$ is Frobenius norm and the last inequality is due to $\|\boldsymbol{\Lambda}^{-\frac{1}{2}}\mathbf{X}\|_F^2 = \operatorname{tr}(\mathbf{X}^\top \boldsymbol{\Lambda}^{-1}\mathbf{X})$ and (G.2). Thus we have

$$
\begin{aligned}
\big\|\boldsymbol{\phi}(s,a)\big\|_{\boldsymbol{\Sigma}_{m,h}^{k,J_k}}^2 \le \frac{2}{3\beta_K} \sum_{i=1}^{k} \prod_{j=i+1}^{k} \Big(1 - 2\eta_{m,j} \lambda_{\min}\big(\boldsymbol{\Lambda}_{m,h}^j\big) \Big)^{2J_j} \operatorname{tr}\big(\boldsymbol{\varphi}^\top \big(\boldsymbol{\Lambda}_{m,h}^i\big)^{-1} \boldsymbol{\varphi}\big) \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^i)^{-1}}^2 \\
+ \frac{2}{3\beta_K} \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}^2.
\end{aligned}
$$

Using the inequality $\sqrt{a^2+b^2} \le a+b$ for $a,b>0$, we get

$$
\begin{aligned}
\big\|\boldsymbol{\phi}(s,a)\big\|_{\boldsymbol{\Sigma}_{m,h}^{k,J_k}} &\le \sqrt{\frac{2}{3\beta_K}} \Bigg(\sum_{i=1}^{k} \prod_{j=i+1}^{k} \Big(1 - 2\eta_{m,j} \lambda_{\min}(\boldsymbol{\Lambda}_{m,h}^j) \Big)^{J_j} \operatorname{tr}(\boldsymbol{\varphi}^\top (\boldsymbol{\Lambda}_{m,h}^i)^{-1} \boldsymbol{\varphi})^{\frac{1}{2}} \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^i)^{-1}} \\
&\qquad + \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} \Bigg) \\
&\stackrel{\text{def}}{=} \widehat{g}_{m,h}^k(\boldsymbol{\phi}(s,a)).
\end{aligned}
$$

Note that $\Big(\boldsymbol{\Sigma}_{m,h}^{k,J_k}\Big)^{-1/2} \Big(\mathbf{w}_{m,h}^{k,J_k} - \boldsymbol{\mu}_{m,h}^{k,J_k}\Big) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. By the Gaussian concentration property, we have

$$
\mathbb{P}\Bigg(\Bigg\| \Big(\boldsymbol{\Sigma}_{m,h}^{k,J_k}\Big)^{-1/2} \Big(\mathbf{w}_{m,h}^{k,J_k} - \boldsymbol{\mu}_{m,h}^{k,J_k}\Big) \Bigg\| \ge \sqrt{4d\log(1/\delta)} \Bigg) \le \delta^2.
$$

Then we have

$$
\begin{aligned}
&\mathbb{P}\Big(\Big| \boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^{k,J_k} - \boldsymbol{\phi}(s,a)^\top \boldsymbol{\mu}_{m,h}^{k,J_k} \Big| \ge 2\widehat{g}_{m,h}^k(\boldsymbol{\phi}(s,a)) \sqrt{d\log(1/\delta)} \Big) \\
&\le \mathbb{P}\Big(\Big| \boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^{k,J_k} - \boldsymbol{\phi}(s,a)^\top \boldsymbol{\mu}_{m,h}^{k,J_k} \Big| \ge 2\sqrt{d\log(1/\delta)} \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Sigma}_{m,h}^{k,J_k}} \Big) \\
&\le \mathbb{P}\Big(\Big\| \boldsymbol{\phi}(s,a) \Big\|_{\boldsymbol{\Sigma}_{m,h}^{k,J_k}} \cdot \Big\| \mathbf{w}_{m,h}^{k,J_k} - \boldsymbol{\mu}_{m,h}^{k,J_k} \Big\|_{(\boldsymbol{\Sigma}_{m,h}^{k,J_k})^{-1}} \ge 2\sqrt{d\log(1/\delta)} \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Sigma}_{m,h}^{k,J_k}} \Big) \\
&= \mathbb{P}\Big(\Big\| \big(\boldsymbol{\Sigma}_{m,h}^{k,J_k}\big)^{-1/2} \big(\mathbf{w}_{m,h}^{k,J_k} - \boldsymbol{\mu}_{m,h}^{k,J_k}\big) \Big\| \ge 2\sqrt{d\log(1/\delta)} \Big) \\
&\le \delta^2. \qquad \text{(G.3)}
\end{aligned}
$$

Bounding Term I_2 in (G.1): Recall from Proposition F.3, we have

$$
\begin{aligned}
\boldsymbol{\mu}_{m,h}^{k,J_k} &= \mathbf{A}_k^{J_k} ... \mathbf{A}_1^{J_1} \mathbf{w}_{m,h}^{1,0} + \sum_{i=1}^{k} \mathbf{A}_k^{J_k} ... \mathbf{A}_{i+1}^{J_{i+1}} \big(\mathbf{I} - \mathbf{A}_i^{J_i}\big) \widehat{\mathbf{w}}_{m,h}^i \\
&= \mathbf{A}_k^{J_k} ... \mathbf{A}_1^{J_1} \mathbf{w}_{m,h}^{1,0} + \sum_{i=1}^{k-1} \mathbf{A}_k^{J_k} ... \mathbf{A}_{i+1}^{J_{i+1}} (\widehat{\mathbf{w}}_{m,h}^i - \widehat{\mathbf{w}}_{m,h}^{i+1}) - \mathbf{A}_k^{J_k} ... \mathbf{A}_1^{J_1} \widehat{\mathbf{w}}_{m,h}^1 + \widehat{\mathbf{w}}_{m,h}^k \\
&= \mathbf{A}_k^{J_k} ... \mathbf{A}_1^{J_1} (\mathbf{w}_{m,h}^{1,0} - \widehat{\mathbf{w}}_{m,h}^1) + \sum_{i=1}^{k-1} \mathbf{A}_k^{J_k} ... \mathbf{A}_{i+1}^{J_{i+1}} (\widehat{\mathbf{w}}_{m,h}^i - \widehat{\mathbf{w}}_{m,h}^{i+1}) + \widehat{\mathbf{w}}_{m,h}^k.
\end{aligned}
$$

Then we can get

$$\boldsymbol{\phi}(s,a)^\top(\boldsymbol{\mu}_{m,h}^{k,J_k}-\widehat{\mathbf{w}}_{m,h}^k)=\underbrace{\boldsymbol{\phi}(s,a)^\top\mathbf{A}_k^{J_k}...\mathbf{A}_1^{J_1}(\mathbf{w}_{m,h}^{1,0}-\widehat{\mathbf{w}}_{m,h}^1)}_{I_{21}}+\underbrace{\boldsymbol{\phi}(s,a)^\top\sum_{i=1}^{k-1}\mathbf{A}_k^{J_k}...\mathbf{A}_{i+1}^{J_{i+1}}(\widehat{\mathbf{w}}_{m,h}^i-\widehat{\mathbf{w}}_{m,h}^{i+1})}_{I_{22}}.$$

In Algorithm 3, we choose $\mathbf{w}_{m,h}^{1,0}=\mathbf{0}$ and $\widehat{\mathbf{w}}_{m,h}^1=(\boldsymbol{\Lambda}_{m,h}^1)^{-1}\boldsymbol{b}_{m,h}^1=\mathbf{0}$. Thus we have $I_{21}=0$. To bound term I_{22}, we use the inequalities in (G.2) and Lemma F.4, we have

$$\begin{aligned}
I_{22}&\le\Big|\sum_{i=1}^{k-1}\boldsymbol{\phi}(s,a)^\top\mathbf{A}_k^{J_k}...\mathbf{A}_{i+1}^{J_{i+1}}(\widehat{\mathbf{w}}_{m,h}^i-\widehat{\mathbf{w}}_{m,h}^{i+1})\Big|\\
&\le\sum_{i=1}^{k-1}\prod_{j=i+1}^{k}\Big(1-2\eta_{m,j}\lambda_{\min}(\boldsymbol{\Lambda}_{m,h}^j)\Big)^{J_j}\|\boldsymbol{\phi}(s,a)\|(\|\widehat{\mathbf{w}}_{m,h}^i\|+\|\widehat{\mathbf{w}}_{m,h}^{i+1}\|)\\
&\le\sum_{i=1}^{k-1}\prod_{j=i+1}^{k}\Big(1-2\eta_{m,j}\lambda_{\min}(\boldsymbol{\Lambda}_{m,h}^j)\Big)^{J_j}\|\boldsymbol{\phi}(s,a)\|\big(2H\sqrt{Mid/\lambda}+2H\sqrt{M(i+1)d/\lambda}\big)\\
&\le 4H\sqrt{MKd/\lambda}\sum_{i=1}^{k-1}\prod_{j=i+1}^{k}\Big(1-2\eta_{m,j}\lambda_{\min}(\boldsymbol{\Lambda}_{m,h}^j)\Big)^{J_j}\|\boldsymbol{\phi}(s,a)\|.
\end{aligned}$$

Thus we get

$$\boldsymbol{\phi}(s,a)^\top\big(\boldsymbol{\mu}_{m,h}^{k,J_k}-\widehat{\mathbf{w}}_{m,h}^k\big)\le 4H\sqrt{MKd/\lambda}\sum_{i=1}^{k-1}\prod_{j=i+1}^{k}\Big(1-2\eta_{m,j}\lambda_{\min}(\boldsymbol{\Lambda}_{m,h}^j)\Big)^{J_j}\|\boldsymbol{\phi}(s,a)\|. \tag{G.4}$$

Substituting (G.3) and (G.4) into (G.1), with probability at least $1-\delta^2$, we have

$$\begin{aligned}
&\Big|\boldsymbol{\phi}(s,a)^\top\mathbf{w}_{m,h}^{k,J_k}-\boldsymbol{\phi}(s,a)^\top\widehat{\mathbf{w}}_{m,h}^k\Big|\\
&\le 4H\sqrt{MKd/\lambda}\sum_{i=1}^{k-1}\prod_{j=i+1}^{k}\Big(1-2\eta_{m,j}\lambda_{\min}(\boldsymbol{\Lambda}_{m,h}^j)\Big)^{J_j}\|\boldsymbol{\phi}(s,a)\|+2\sqrt{\frac{2d\log(1/\delta)}{3\beta_K}}\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}\\
&\quad+2\sqrt{\frac{2d\log(1/\delta)}{3\beta_K}}\sum_{i=1}^{k}\prod_{j=i+1}^{k}\Big(1-2\eta_{m,j}\lambda_{\min}(\boldsymbol{\Lambda}_{m,h}^j)\Big)^{J_j}\operatorname{tr}\big(\boldsymbol{\varphi}^\top\big(\boldsymbol{\Lambda}_{m,h}^i\big)^{-1}\boldsymbol{\varphi}\big)^{\frac12}\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^i)^{-1}}\\
&\overset{\text{def}}{=}W.
\end{aligned}\tag{G.5}$$

Here we choose $\eta_{m,j}=1/(4\lambda_{\max}(\boldsymbol{\Lambda}_{m,h}^j))$ and set $\kappa_j=\lambda_{\max}\big(\boldsymbol{\Lambda}_{m,h}^j\big)/\lambda_{\min}\big(\boldsymbol{\Lambda}_{m,h}^j\big)$, then we have

$$\Big(1-2\eta_{m,j}\lambda_{\min}\big(\boldsymbol{\Lambda}_{m,h}^j\big)\Big)^{J_j}=(1-1/2\kappa_j)^{J_j}.$$

We want to have $(1-1/2\kappa_j)^{J_j}<\epsilon$, it suffices to choose J_j such that

$$J_j\ge\frac{\log(1/\epsilon)}{\log\big(\frac{1}{1-1/2\kappa_j}\big)}.$$

Note that $1/2\kappa_j\le 1/2$, we have $\log(1/(1-1/2\kappa_j))\ge 1/2\kappa_j$ because $e^{-x}>1-x$ for $0<x<1$. Therefore, we only need to pick $J_j\ge 2\kappa_j\log(1/\epsilon)$.

Also note that $1\ge\|\boldsymbol{\phi}(s,a)\|\ge\sqrt{\lambda}\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^i)^{-1}}$ and $\operatorname{tr}\big(\boldsymbol{\varphi}^\top\big(\boldsymbol{\Lambda}_{m,h}^i\big)^{-1}\boldsymbol{\varphi}\big)\le M$ due to the fact that $n\le M$. By setting $\epsilon=1/(4HMKd)$ and $\lambda=1$, we obtain

$$W\le\sum_{i=1}^{k-1}\epsilon^{k-i}4H\sqrt{MKd/\lambda}\|\boldsymbol{\phi}(s,a)\|+2\sqrt{\frac{2d\log(1/\delta)}{3\beta_K}}\Big(\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}+\sum_{i=1}^{k-1}\epsilon^{k-i}\sqrt{M}\|\boldsymbol{\phi}(s,a)\|\Big)$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{k-1} \epsilon^{k-i} 4H\sqrt{MKd/\lambda}\sqrt{MK}\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} \\
&\quad + 2\sqrt{\frac{2d\log(1/\delta)}{3\beta_K}}\left(\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} + \sum_{i=1}^{k-1}\epsilon^{k-i}M\sqrt{K}\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}\right) \\
&\leq \sum_{i=1}^{k-1}\epsilon^{k-i-1}\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} + 2\sqrt{\frac{2d\log(1/\delta)}{3\beta_K}}\left(\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} + \sum_{i=1}^{k-1}\epsilon^{k-i-1}\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}\right) \\
&\leq \left(5\sqrt{\frac{2d\log(1/\delta)}{3\beta_K}} + \frac{4}{3}\right)\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}},
\end{aligned}
$$

where the second inequality follows from $\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} \geq 1/\sqrt{\mathcal{K}(k)+1}\|\boldsymbol{\phi}(s,a)\| \geq 1/\sqrt{MK}\|\boldsymbol{\phi}(s,a)\|$, the fourth inequality follows from $\sum_{i=1}^{k-1}\epsilon^{k-i-1} = \sum_{i=0}^{k-2}\epsilon^i < 1/(1-\epsilon) \leq 4/3$. Finally we have

$$
\begin{aligned}
&\mathbb{P}\left(\left|\boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^{k,J_k} - \boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}_{m,h}^k\right| \leq \left(5\sqrt{\frac{2d\log(1/\delta)}{3\beta_K}} + \frac{4}{3}\right)\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}\right) \\
&\geq \mathbb{P}\left(\left|\boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^{k,J_k} - \boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}_{m,h}^k\right| \leq W\right) \\
&\geq 1-\delta^2.
\end{aligned}
$$

This completes the proof. □

G.4 Proof of Lemma F.6

Proof. Recall that $\mathbf{w}_{m,h}^{k,J_k} \sim \mathcal{N}(\boldsymbol{\mu}_{m,h}^{k,J_k}, \boldsymbol{\Sigma}_{m,h}^{k,J_k})$. Let $\boldsymbol{\xi}_{m,h}^{k,J_k} = \mathbf{w}_{m,h}^{k,J_k} - \boldsymbol{\mu}_{m,h}^{k,J_k} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{m,h}^{k,J_k})$, thus we have

$$
\left\|\mathbf{w}_{m,h}^{k,J_k}\right\| = \left\|\boldsymbol{\mu}_{m,h}^{k,J_k} + \boldsymbol{\xi}_{m,h}^{k,J_k}\right\| \leq \left\|\boldsymbol{\mu}_{m,h}^{k,J_k}\right\| + \left\|\boldsymbol{\xi}_{m,h}^{k,J_k}\right\|. \tag{G.6}
$$

Bounding $\left\|\boldsymbol{\mu}_{m,h}^{k,J_k}\right\|$ in (G.6): Based on Proposition F.3, we have

$$
\begin{aligned}
\left\|\boldsymbol{\mu}_{m,h}^{k,J_k}\right\| &= \left\|\mathbf{A}_k^{J_k}\dots\mathbf{A}_1^{J_1}\mathbf{w}_{m,h}^{1,0} + \sum_{i=1}^{k}\mathbf{A}_k^{J_k}\dots\mathbf{A}_{i+1}^{J_{i+1}}\left(\mathbf{I}-\mathbf{A}_i^{J_i}\right)\widehat{\mathbf{w}}_{m,h}^i\right\| \\
&\leq \sum_{i=1}^{k}\left\|\mathbf{A}_k^{J_k}\dots\mathbf{A}_{i+1}^{J_{i+1}}\left(\mathbf{I}-\mathbf{A}_i^{J_i}\right)\right\|_F \cdot \left\|\widehat{\mathbf{w}}_{m,h}^i\right\| \\
&\leq 2H\sqrt{MKd/\lambda}\sum_{i=1}^{k}\left\|\mathbf{A}_k^{J_k}\dots\mathbf{A}_{i+1}^{J_{i+1}}\left(\mathbf{I}-\mathbf{A}_i^{J_i}\right)\right\|_F \\
&\leq 2Hd\sqrt{MK/\lambda}\sum_{i=1}^{k}\|\mathbf{A}_k\|_2^{J_k}\dots\|\mathbf{A}_{i+1}\|_2^{J_{i+1}}\left\|\left(\mathbf{I}-\mathbf{A}_i^{J_i}\right)\right\|_2 \\
&\leq 2Hd\sqrt{MK/\lambda}\sum_{i=1}^{k}\prod_{j=i+1}^{k}\left(1-2\eta_{m,j}\lambda_{\min}\left(\boldsymbol{\Lambda}_{m,h}^j\right)\right)^{J_j}\left(\|\mathbf{I}\|_2 + \left\|\mathbf{A}_i\right\|_2^{J_i}\right) \\
&\leq 2Hd\sqrt{MK/\lambda}\sum_{i=1}^{k}\prod_{j=i+1}^{k}\left(1-2\eta_{m,j}\lambda_{\min}\left(\boldsymbol{\Lambda}_{m,h}^j\right)\right)^{J_j}\left(1+\left(1-2\eta_{m,i}\lambda_{\min}\left(\boldsymbol{\Lambda}_{m,h}^i\right)\right)^{J_j}\right),
\end{aligned}
$$

where the second inequality holds from Lemma F.4, the third inequality follows from the fact that $\operatorname{rank}\left(\mathbf{A}_k^{J_k}\dots\mathbf{A}_{i+1}^{J_{i+1}}\left(\mathbf{I}-\mathbf{A}_i^{J_i}\right)\right) \leq d$ and $\|\mathbf{X}\|_2 \leq \|\mathbf{X}\|_F \leq \operatorname{rank}(\mathbf{X})\|\mathbf{X}\|_2$ where $\|\mathbf{X}\|_2 = \sigma_{\max}(\mathbf{X})$.

Recall that in Lemma F.5, we set $J_j \geq 2\kappa_j \log(1/\epsilon)$ where $\kappa_j = \lambda_{\max}(\mathbf{\Lambda}_{m,h}^j)/\lambda_{\min}(\mathbf{\Lambda}_{m,h}^j)$, $\epsilon = 1/(4HMKd)$ and $\lambda = 1$, thus we get

$$\begin{aligned}\left\|\boldsymbol{\mu}_{m,h}^{k,J_k}\right\| &\leq 2Hd\sqrt{MK/\lambda}\sum_{i=1}^{k}(\epsilon^{k-i}+\epsilon^{k-i+1}) \\ &\leq 4Hd\sqrt{MK/\lambda}\sum_{i=0}^{\infty}\epsilon^{i} \\ &\leq \frac{16}{3}Hd\sqrt{MK}.\end{aligned}$$

Bounding $\left\|\boldsymbol{\xi}_{m,h}^{k,J_k}\right\|$ in (G.6): Note that $\boldsymbol{\xi}_{m,h}^{k,J_k} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{m,h}^{k,J_k})$, using Gaussian concentration Lemma L.5, we have

$$\mathbb{P}\left(\left\|\boldsymbol{\xi}_{m,h}^{k,J_k}\right\| \leq \sqrt{\frac{1}{\delta}\operatorname{tr}\left(\mathbf{\Sigma}_{m,h}^{k,J_k}\right)}\right) \geq 1-\delta.$$

Recall from Proposition F.3, we have

$$\begin{aligned}\operatorname{tr}\left(\mathbf{\Sigma}_{m,h}^{k,J_k}\right) &= \sum_{i=1}^{k}\frac{1}{\beta_{m,i}}\operatorname{tr}\left(\mathbf{A}_k^{J_k}\dots\mathbf{A}_{i+1}^{J_{i+1}}(\mathbf{I}-\mathbf{A}_i^{2J_i})(\mathbf{\Lambda}_{m,h}^i)^{-1}(\mathbf{I}+\mathbf{A}_i)^{-1}\mathbf{A}_{i+1}^{J_{i+1}}\dots\mathbf{A}_k^{J_k}\right) \\ &\leq \sum_{i=1}^{k}\frac{1}{\beta_{m,i}}\operatorname{tr}\left(\mathbf{A}_k^{J_k}\right)\dots\operatorname{tr}\left(\mathbf{A}_{i+1}^{J_{i+1}}\right)\operatorname{tr}\left(\mathbf{I}-\mathbf{A}_i^{2J_i}\right)\operatorname{tr}\left(\left(\mathbf{\Lambda}_{m,h}^i\right)^{-1}\right)\operatorname{tr}\left(\left(\mathbf{I}+\mathbf{A}_i\right)^{-1}\right) \\ &\quad\times\operatorname{tr}\left(\mathbf{A}_{i+1}^{J_{i+1}}\right)\dots\operatorname{tr}\left(\mathbf{A}_k^{J_k}\right),\end{aligned}$$

where the inequality holds due to Lemma L.6. Recall from (G.2) that, when $\eta_{m,k} \leq 1/(4\lambda_{\max}(\mathbf{\Lambda}_{m,h}^k))$ for all k and m, we have $\mathbf{A}_i^{J_i} \preccurlyeq (1-2\eta_{m,k}\lambda_{\min}(\mathbf{\Lambda}_{m,h}^k))^{J_j}\mathbf{I}$, set $\lambda = 1$, then we obtain

$$\operatorname{tr}(\mathbf{A}_i^{J_i}) \leq \operatorname{tr}\left(\left(1-2\eta_{m,k}\lambda_{\min}\left(\mathbf{\Lambda}_{m,h}^k\right)\right)^{J_j}\mathbf{I}\right) \leq d\left(1-2\eta_{m,k}\lambda_{\min}\left(\mathbf{\Lambda}_{m,h}^k\right)\right)^{J_j} \leq d\epsilon \leq 1.$$

Similarly, we have $\mathbf{I}-\mathbf{A}_i^{2J_i} \preccurlyeq \left(1-\frac{1}{2^{2J_i}}\right)\mathbf{I}$, then we get

$$\operatorname{tr}(\mathbf{I}-\mathbf{A}_i^{2J_i}) \leq \left(1-\frac{1}{2^{2J_i}}\right)d < d.$$

Also, based on $(\mathbf{I}+\mathbf{A}_i)^{-1} \preccurlyeq \frac{2}{3}\mathbf{I}$, we have

$$\operatorname{tr}\left((\mathbf{I}+\mathbf{A}_i)^{-1}\right) \leq \frac{2}{3}d.$$

Note that $\lambda_{\max}\left(\left(\mathbf{\Lambda}_{m,h}^i\right)^{-1}\right) \leq 1$, we have

$$\operatorname{tr}\left(\left(\mathbf{\Lambda}_{m,h}^i\right)^{-1}\right) \leq \sum\lambda\left(\left(\mathbf{\Lambda}_{m,h}^i\right)^{-1}\right) \leq d.$$

Combine the above results together and choose $\beta_{m,i} = \beta_K$ for all $i \in [K]$ and $m \in \mathcal{M}$, we have

$$\operatorname{tr}\left(\mathbf{\Sigma}_{m,h}^{k,J_k}\right) \leq \sum_{i=1}^{K}\frac{1}{\beta_{m,i}}\cdot\frac{2}{3}\cdot d^3 = \frac{2}{3\beta_K}Kd^3.$$

Then we have

$$\mathbb{P}\left(\left\|\boldsymbol{\xi}_{m,h}^{k,J_k}\right\| \leq \sqrt{\frac{1}{\delta}\cdot\frac{2}{3\beta_K}Kd^3}\right) \geq \mathbb{P}\left(\left\|\boldsymbol{\xi}_{m,h}^{k,J_k}\right\| \leq \sqrt{\frac{1}{\delta}\operatorname{tr}\left(\mathbf{\Sigma}_{m,h}^{k,J_k}\right)}\right) \geq 1-\delta.$$

Combine above results together: with probability at least $1-\delta$, we have

$$\left\|\mathbf{w}_{m,h}^{k,J_k}\right\| \leq \frac{16}{3}Hd\sqrt{MK} + \sqrt{\frac{2K}{3\beta_K\delta}}d^{3/2}.$$

This completes the proof. □

G.5 Proof of Lemma F.7

Proof. Based on Lemma F.6, for any fixed $n \in [N]$, with probability at least $1-\delta$, for any $(m,k,h) \in \mathcal{M} \times [K] \times [H]$, we have

$$\left\|\mathbf{w}_{m,h}^{k,J_k,n}\right\| \leq \frac{16}{3} H d \sqrt{MK} + \sqrt{\frac{2K}{3\beta_K \delta}} d^{3/2}.$$

By taking union over n, m, k, h, we have for all $(m,k,h) \in \mathcal{M} \times [K] \times [H]$ and for all $n \in [N]$, with probability $1-\delta/2$, we have

$$\left\|\mathbf{w}_{m,h}^{k,J_k,n}\right\| \leq \frac{16}{3} H d \sqrt{MK} + \sqrt{\frac{4NMHK^2}{3\beta_K \delta}} d^{3/2} = B_{\delta/2NMHK}. \tag{G.7}$$

Based on Lemma L.7 and Lemma L.9, we have that for any $\varepsilon > 0$ and $\delta > 0$, with probability at least $1-\delta/2$,

$$\begin{aligned}
&\left\| \sum_{(s^l,a^l,s'^l) \in U_{m,h}(k)} \boldsymbol{\phi}(s^l,a^l) \left[\left(V_{m,h+1}^k - \mathbb{P}_h V_{m,h+1}^k\right)(s^l,a^l) \right] \right\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} \\
&\leq \left(4H^2 \left[\frac{d}{2} \log\left(\frac{k+\lambda}{\lambda}\right) + d \log\left(\frac{B_{\delta/2NMHK}}{\varepsilon}\right) + \log\frac{3}{\delta} \right] + \frac{8k^2\varepsilon^2}{\lambda} \right)^{1/2} \\
&\leq 2H \left[\frac{d}{2} \log\left(\frac{k+\lambda}{\lambda}\right) + d \log\left(\frac{B_{\delta/2NMHK}}{\varepsilon}\right) + \log\frac{3}{\delta} \right]^{1/2} + \frac{2\sqrt{2}k\varepsilon}{\sqrt{\lambda}}.
\end{aligned}$$

Here we set $\lambda = 1, \varepsilon = \frac{H}{2\sqrt{2}k}$, with probability at least $1-\delta/2$, we have

$$\begin{aligned}
&\left\| \sum_{(s^l,a^l,s'^l) \in U_{m,h}(k)} \boldsymbol{\phi}(s^l,a^l) \left[\left(V_{m,h+1}^k - \mathbb{P}_h V_{m,h+1}^k\right)(s^l,a^l) \right] \right\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} \\
&\leq 2H\sqrt{d} \left[\frac{1}{2} \log(k+1) + \log\left(\frac{B_{\delta/2NMHK}}{\frac{H}{2\sqrt{2}k}}\right) + \log\frac{3}{\delta} \right]^{1/2} + H \\
&\leq 3H\sqrt{d} \left[\frac{1}{2} \log(K+1) + \log\left(\frac{2\sqrt{2}KB_{\delta/2NMHK}}{H}\right) + \log\frac{3}{\delta} \right]^{1/2}.
\end{aligned} \tag{G.8}$$

By applying union bound between (G.7) and (G.8), and define that $C_\delta = \left[\frac{1}{2}\log(K+1) + \log\frac{3}{\delta} + \log\left(\frac{2\sqrt{2}KB_{\delta/2NMHK}}{H}\right)\right]^{1/2}$, finally we obtain that for all $(m,k,h) \in \mathcal{M} \times [K] \times [H]$,

$$\left\| \sum_{(s^l,a^l,s'^l) \in U_{m,h}(k)} \boldsymbol{\phi}(s^l,a^l) \left[\left(V_{m,h+1}^k - \mathbb{P}_h V_{m,h+1}^k\right)(s^l,a^l) \right] \right\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} \leq 3H\sqrt{d}C_\delta,$$

with probability at least $1-\delta$. □

G.6 Proof of Lemma F.8

Proof. We denote the inner product over $\mathcal{S}$ by $\langle \cdot, \cdot \rangle_{\mathcal{S}}$. Based on $\mathbb{P}_h(\cdot|s,a) = \left\langle \boldsymbol{\phi}(s,a), \mu_h(\cdot) \right\rangle_{\mathcal{S}}$ in Definition 4.1, we have

$$\begin{aligned}
\mathbb{P}_h V_{m,h+1}^k(s,a) &= \boldsymbol{\phi}(s,a)^\top \left\langle \boldsymbol{\mu}_h, V_{m,h+1}^k \right\rangle_{\mathcal{S}} \\
&= \boldsymbol{\phi}(s,a)^\top \left(\boldsymbol{\Lambda}_{m,h}^k\right)^{-1} \left(\boldsymbol{\Lambda}_{m,h}^k\right) \left\langle \boldsymbol{\mu}_h, V_{m,h+1}^k \right\rangle_{\mathcal{S}}
\end{aligned}$$

$$= \boldsymbol{\phi}(s,a)^\top \big(\boldsymbol{\Lambda}^k_{m,h}\big)^{-1} \Bigg(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \boldsymbol{\phi}\big(s^l,a^l\big)\boldsymbol{\phi}\big(s^l,a^l\big)^\top + \lambda \mathbf{I} \Bigg) \big\langle \boldsymbol{\mu}_h, V^k_{m,h+1} \big\rangle_{\mathcal{S}}$$

$$= \boldsymbol{\phi}(s,a)^\top \big(\boldsymbol{\Lambda}^k_{m,h}\big)^{-1} \Bigg(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \boldsymbol{\phi}\big(s^l,a^l\big)\big(\mathbb{P}_h V^k_{m,h+1}\big)\big(s^l,a^l\big) + \lambda \mathbf{I} \big\langle \boldsymbol{\mu}_h, V^k_{m,h+1} \big\rangle_{\mathcal{S}} \Bigg). \tag{G.9}$$

Here the last equality uses $\mathbb{P}_h(\cdot|s,a) = \big\langle \boldsymbol{\phi}(s,a), \boldsymbol{\mu}_h(\cdot) \big\rangle_{\mathcal{S}}$ again. Then we can separate the following error into three parts,

$$\begin{aligned}
&\boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}^k_{m,h} - r_h(s,a) - \mathbb{P}_h V^k_{m,h+1}(s,a) \\
&= \boldsymbol{\phi}(s,a)^\top \big(\boldsymbol{\Lambda}^k_{m,h}\big)^{-1} \sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \big[r_h\big(s^l,a^l\big) + V^k_{m,h+1}(s'^l) \big] \boldsymbol{\phi}\big(s^l,a^l\big) - r_h(s,a) \\
&\quad - \boldsymbol{\phi}(s,a)^\top \big(\boldsymbol{\Lambda}^k_{m,h}\big)^{-1} \Bigg(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \boldsymbol{\phi}\big(s^l,a^l\big)\big(\mathbb{P}_h V^k_{m,h+1}\big)\big(s^l,a^l\big) + \lambda \mathbf{I} \big\langle \boldsymbol{\mu}_h, V^k_{m,h+1} \big\rangle_{\mathcal{S}} \Bigg) \\
&= \underbrace{\boldsymbol{\phi}(s,a)^\top \big(\boldsymbol{\Lambda}^k_{m,h}\big)^{-1} \Bigg(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \boldsymbol{\phi}\big(s^l,a^l\big) \big[\big(V^k_{m,h+1} - \mathbb{P}_h V^k_{m,h+1}\big)\big(s^l,a^l\big) \big] \Bigg)}_{\text{(i)}} \\
&\quad + \underbrace{\boldsymbol{\phi}(s,a)^\top \big(\boldsymbol{\Lambda}^k_{m,h}\big)^{-1} \Bigg(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} r_h\big(s^l,a^l\big)\boldsymbol{\phi}\big(s^l,a^l\big) \Bigg) - r_h(s,a)}_{\text{(ii)}} \\
&\quad - \underbrace{\lambda \boldsymbol{\phi}(s,a)^\top \big(\boldsymbol{\Lambda}^k_{m,h}\big)^{-1} \big\langle \boldsymbol{\mu}_h, V^k_{m,h+1} \big\rangle_{\mathcal{S}}}_{\text{(iii)}}.
\end{aligned} \tag{G.10}$$

Here the first equality holds due to (G.9). We now provide an upper bound for each of the terms in (G.10).

Bounding Term (i) in (G.10): using Cauchy-Schwarz inequality and Lemma F.7, with probability at least $1-\delta$, for all $(m,k,h) \in \mathcal{M} \times [K] \times [H]$ and for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\begin{aligned}
&\boldsymbol{\phi}(s,a)^\top \big(\boldsymbol{\Lambda}^k_{m,h}\big)^{-1} \Bigg(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \boldsymbol{\phi}\big(s^l,a^l\big) \big[\big(V^k_{m,h+1} - \mathbb{P}_h V^k_{m,h+1}\big)\big(s^l,a^l\big) \big] \Bigg) \\
&\le \Bigg\| \sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \boldsymbol{\phi}\big(s^l,a^l\big) \big[\big(V^k_{m,h+1} - \mathbb{P}_h V^k_{m,h+1}\big)\big(s^l,a^l\big) \big] \Bigg\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}} \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}} \\
&\le 3H\sqrt{d} C_\delta \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}}.
\end{aligned} \tag{G.11}$$

Bounding Term (ii) in (G.10): we first note that

$$\begin{aligned}
&\boldsymbol{\phi}(s,a)^\top \big(\boldsymbol{\Lambda}^k_{m,h}\big)^{-1} \Bigg(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} r_h\big(s^l,a^l\big)\boldsymbol{\phi}\big(s^l,a^l\big) \Bigg) - r_h(s,a) \\
&= \boldsymbol{\phi}(s,a)^\top \big(\boldsymbol{\Lambda}^k_{m,h}\big)^{-1} \Bigg(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} r_h\big(s^l,a^l\big)\boldsymbol{\phi}\big(s^l,a^l\big) \Bigg) - \boldsymbol{\phi}(s,a)^\top \boldsymbol{\theta}_h \\
&= \boldsymbol{\phi}(s,a)^\top \big(\boldsymbol{\Lambda}^k_{m,h}\big)^{-1} \Bigg(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} r_h\big(s^l,a^l\big)\boldsymbol{\phi}\big(s^l,a^l\big) - \boldsymbol{\Lambda}^k_{m,h}\boldsymbol{\theta}_h \Bigg)
\end{aligned}$$

$$
\begin{aligned}
&= \boldsymbol{\phi}(s,a)^\top \big(\boldsymbol{\Lambda}_{m,h}^k\big)^{-1} \Bigg(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} r_h\big(s^l,a^l\big)\boldsymbol{\phi}\big(s^l,a^l\big) - \sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \boldsymbol{\phi}\big(s^l,a^l\big)\boldsymbol{\phi}\big(s^l,a^l\big)^\top \boldsymbol{\theta}_h - \lambda \mathbf{I}\boldsymbol{\theta}_h \Bigg) \\
&= \boldsymbol{\phi}(s,a)^\top \big(\boldsymbol{\Lambda}_{m,h}^k\big)^{-1} \Bigg(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} r_h\big(s^l,a^l\big)\boldsymbol{\phi}\big(s^l,a^l\big) - \sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \boldsymbol{\phi}\big(s^l,a^l\big) r_h\big(s^l,a^l\big) - \lambda \boldsymbol{\theta}_h \Bigg) \\
&= -\lambda \boldsymbol{\phi}(s,a)^\top \big(\boldsymbol{\Lambda}_{m,h}^k\big)^{-1} \boldsymbol{\theta}_h, \qquad \text{(G.12)}
\end{aligned}
$$

where the first and fourth equality holds due to the definition $r_h(s,a) = \big\langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h \big\rangle$ from Definition 4.1, the third equality uses the definition of $\boldsymbol{\Lambda}_{m,h}^k$. Next we can obtain that

$$
\begin{aligned}
-\lambda \boldsymbol{\phi}(s,a)^\top \big(\boldsymbol{\Lambda}_{m,h}^k\big)^{-1} \boldsymbol{\theta}_h &\le \lambda \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} \|\boldsymbol{\theta}_h\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} \\
&\le \sqrt{\lambda} \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} \|\boldsymbol{\theta}_h\| \\
&\le \sqrt{\lambda d} \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}, \qquad \text{(G.13)}
\end{aligned}
$$

where we use the fact that $\lambda_{\max}\big(\big(\boldsymbol{\Lambda}_{m,h}^k\big)^{-1}\big) \le 1/\lambda$ and $\|\boldsymbol{\theta}_h\| \le \sqrt{d}$ from Definition 4.1. By Combining (G.12) and (G.13), we obtain

$$
\boldsymbol{\phi}(s,a)^\top \big(\boldsymbol{\Lambda}_{m,h}^k\big)^{-1} \Bigg(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} r_h\big(s^l,a^l\big)\boldsymbol{\phi}\big(s^l,a^l\big) \Bigg) - r_h(s,a) \le \sqrt{\lambda d} \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}. \qquad \text{(G.14)}
$$

Bounding Term (iii) in (G.10): we have

$$
\begin{aligned}
\lambda \boldsymbol{\phi}(s,a)^\top \big(\boldsymbol{\Lambda}_{m,h}^k\big)^{-1} \big\langle \boldsymbol{\mu}_h, V_{m,h+1}^k \big\rangle_{\mathcal{S}} &\le \lambda \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} \big\| \big\langle \boldsymbol{\mu}_h, V_{m,h+1}^k \big\rangle_{\mathcal{S}} \big\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} \\
&\le \sqrt{\lambda} \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} \big\| \big\langle \boldsymbol{\mu}_h, V_{m,h+1}^k \big\rangle_{\mathcal{S}} \big\| \\
&\le H\sqrt{\lambda} \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} \|\boldsymbol{\mu}_h\| \\
&\le H\sqrt{\lambda d} \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}, \qquad \text{(G.15)}
\end{aligned}
$$

where the second inequality holds due to the fact that $\lambda_{\max}\big(\big(\boldsymbol{\Lambda}_{m,h}^k\big)^{-1}\big) \le 1/\lambda$, the third inequality uses the fact that $V_{m,h+1}^k \le H$ and the last inequality follows from $\|\boldsymbol{\mu}_h\| \le \sqrt{d}$ in Definition 4.1.

Combine Terms (i)(ii)(iii) together: combine (G.11), (G.14) and (G.15), then set $\lambda = 1$, with probability at least $1-\delta$, we get

$$
\begin{aligned}
\Big| \boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}_{m,h}^k - r_h(s,a) - \mathbb{P}_h V_{m,h+1}^k(s,a) \Big| &\le \Big(3HC_\delta + \sqrt{\lambda d} + H\sqrt{\lambda d} \Big) \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} \\
&\le 5H\sqrt{d} C_\delta \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}},
\end{aligned}
$$

This completes the proof. □

G.7 Proof of Lemma F.9

Proof. Recall from Definition F.1,

$$
\begin{aligned}
-l_{m,h}^k(s,a) &= Q_{m,h}^k(s,a) - r_h(s,a) - \mathbb{P}_h V_{m,h+1}^k(s,a) \\
&= \min\Big\{ \max_{n\in[N]} \boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^{k,J_k,n}, H-h+1 \Big\}^+ - r_h(s,a) - \mathbb{P}_h V_{m,h+1}^k(s,a) \\
&\le \max_{n\in[N]} \boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^{k,J_k,n} - r_h(s,a) - \mathbb{P}_h V_{m,h+1}^k(s,a)
\end{aligned}
$$

$$= \max_{n\in[N]} \boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^{k,J_k,n} - \boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}_{m,h}^k + \boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}_{m,h}^k - r_h(s,a) - \mathbb{P}_h V_{m,h+1}^k(s,a)$$
$$\leq \underbrace{\max_{n\in[N]} \left|\boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^{k,J_k,n} - \boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}_{m,h}^k\right|}_{I_1} + \underbrace{\left|\boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}_{m,h}^k - r_h(s,a) - \mathbb{P}_h V_{m,h+1}^k(s,a)\right|}_{I_2}.$$

Bounding Term I_1: based on Lemma F.5, for any fixed $n \in [N]$, for any $(m,h,k) \in \mathcal{M}\times[H]\times[K]$ and for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, with probability at least $1-\delta^2$, we have

$$\left|\boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^{k,J_k,n} - \boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}_{m,h}^k\right| \leq \left(5\sqrt{\frac{2d\log(1/\delta)}{3\beta_K}} + \frac{4}{3}\right)\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}.$$

By taking union bound over n, we have for all $n \in [N]$, with probability $1-\delta^2$, we have

$$\left|\boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^{k,J_k,n} - \boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}_{m,h}^k\right| \leq \left(5\sqrt{\frac{2d\log\left(\sqrt{N}/\delta\right)}{3\beta_K}} + \frac{4}{3}\right)\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}.$$

This indicates, for any $(m,h,k) \in \mathcal{M}\times[H]\times[K]$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$, with probability at least $1-\delta^2$, we have

$$\textbf{Term } I_1 = \max_{n\in[N]} \left|\boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^{k,J_k,n} - \boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}_{m,h}^k\right| \leq \left(5\sqrt{\frac{2d\log\left(\sqrt{N}/\delta\right)}{3\beta_K}} + \frac{4}{3}\right)\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}. \tag{G.16}$$

Bounding Term I_2: based on Lemma F.8, with probability at least $1-\delta$, for any $(m,h,k) \in \mathcal{M}\times[H]\times[K]$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have

$$\left|\boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}_{m,h}^k - r_h^k(s,a) - \mathbb{P}_h V_{m,h+1}^k(s,a)\right| \leq 5H\sqrt{d}C_\delta\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}.$$

Combine the two result above, by taking union bound, with probability at least $1-\delta-\delta^2$, for any $(m,h,k) \in \mathcal{M}\times[H]\times[K]$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have

$$-l_{m,h}^k(s,a) \leq \left(5H\sqrt{d}C_\delta + 5\sqrt{\frac{2d\log\left(\sqrt{N}/\delta\right)}{3\beta_K}} + \frac{4}{3}\right)\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}.$$

This completes the proof. □

G.8 Proof of Lemma F.10

Proof. Recall from Definition F.1,

$$l_{m,h}^k(s,a) = r_h(s,a) + \mathbb{P}_h V_{m,h+1}^k(s,a) - Q_{m,h}^k(s,a).$$

Note that

$$Q_{m,h}^k(s,a) = \min\left\{\max_{n\in[N]} \boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^{k,J_k,n}, H-h+1\right\}^+ \leq \max_{n\in[N]} \boldsymbol{\phi}(x,a)^\top \mathbf{w}_{m,h}^{k,J_k,n}.$$

Note that $\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} \leq \sqrt{1/\lambda}\|\boldsymbol{\phi}(s,a)\| \leq 1$ for all $\boldsymbol{\phi}(s,a)$. Define $\mathcal{C}(\varepsilon)$ to be a ε-cover of $\left\{\boldsymbol{\phi} \mid \|\boldsymbol{\phi}\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} \leq 1\right\}$. Based on Lemma L.8, we have $|\mathcal{C}(\varepsilon)| \leq (3/\varepsilon)^d$.

First, for any fixed $\boldsymbol{\phi}(s,a) \in \mathcal{C}(\varepsilon)$, based on the results in Proposition F.3, we have that $\boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^{k,J_k,n} \sim \mathcal{N}\left(\boldsymbol{\phi}(s,a)^\top \boldsymbol{\mu}_{m,h}^{k,J_k}, \boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{m,h}^{k,J_k}\boldsymbol{\phi}(s,a)\right)$ for any fixed $n \in [N]$. Now we define

$$Z_k = \frac{r_h(s,a) + \mathbb{P}_h V_{m,h+1}^k(s,a) - \boldsymbol{\phi}(s,a)^\top \boldsymbol{\mu}_{m,h}^{k,J_k}}{\sqrt{\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{m,h}^{k,J_k}\boldsymbol{\phi}(s,a)}}.$$

When $|Z_k| < 1$, by Gaussian concentration Lemma L.10, we have

$$\begin{aligned}
&\mathbb{P}\Big(\boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^{k,J_k,n} \geq r_h(s,a) + \mathbb{P}_h V_{m,h+1}^k(s,a)\Big) \\
&= \mathbb{P}\left(\frac{\boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^{k,J_k,n} - \boldsymbol{\phi}(s,a)^\top \boldsymbol{\mu}_{m,h}^{k,J_k}}{\sqrt{\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{m,h}^{k,J_k} \boldsymbol{\phi}(s,a)}} \geq \frac{r_h(s,a) + \mathbb{P}_h V_{m,h+1}^k(s,a) - \boldsymbol{\phi}(s,a)^\top \boldsymbol{\mu}_{m,h}^{k,J_k}}{\sqrt{\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{m,h}^{k,J_k} \boldsymbol{\phi}(s,a)}}\right) \\
&= \mathbb{P}\left(\frac{\boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^{k,J_k,n} - \boldsymbol{\phi}(s,a)^\top \boldsymbol{\mu}_{m,h}^{k,J_k}}{\sqrt{\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{m,h}^{k,J_k} \boldsymbol{\phi}(s,a)}} \geq Z_k\right) \\
&\geq \frac{1}{2\sqrt{2\pi}} \exp(-Z_k^2/2) \\
&\geq \frac{1}{2\sqrt{2e\pi}}.
\end{aligned}$$

Consider the numerator of Z_k:

$$\begin{aligned}
&\left|r_h(s,a) + \mathbb{P}_h V_{m,h+1}^k(s,a) - \boldsymbol{\phi}(s,a)^\top \boldsymbol{\mu}_{m,h}^{k,J_k}\right| \\
&\leq \left|r_h(s,a) + \mathbb{P}_h V_{m,h+1}^k(s,a) - \boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}_{m,h}^k\right| + \left|\boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}_{m,h}^k - \boldsymbol{\phi}(s,a)^\top \boldsymbol{\mu}_{m,h}^{k,J_k}\right|.
\end{aligned}$$

Based on Lemma F.8, with probablity at least $1-\delta$, we have

$$|r_h(s,a) + \mathbb{P}_h V_{m,h+1}^k(s,a) - \boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}_{m,h}^k| \leq 5H\sqrt{d}C_\delta \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}},$$

From (G.4), we have

$$\boldsymbol{\phi}(s,a)^\top \left(\boldsymbol{\mu}_{m,h}^{k,J_k} - \widehat{\mathbf{w}}_{m,h}^k\right) \leq 4H\sqrt{MKd/\lambda} \sum_{i=1}^{k-1} \prod_{j=i+1}^{k} \left(1 - 2\eta_{m,j}\lambda_{\min}(\boldsymbol{\Lambda}_{m,h}^j)\right)^{J_j} \|\boldsymbol{\phi}(s,a)\|.$$

Recall the proof of Lemma F.5, we set $\eta_{m,j} = 1/(4\lambda_{\max}(\boldsymbol{\Lambda}_{m,h}^j))$, $J_j \geq 2\kappa_j \log(1/\epsilon)$, then we have for all $j \in [K]$, $(1 - 2\eta_{m,j}\lambda_{\min}(\boldsymbol{\Lambda}_{m,h}^j))^{J_j} \leq \epsilon$, set $\epsilon = 1/4HMKd$ and $\lambda = 1$, we have

$$\begin{aligned}
\left|\boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}_{m,h}^k - \boldsymbol{\phi}(s,a)^\top \boldsymbol{\mu}_{m,h}^{k,J_k}\right| &\leq 4H\sqrt{MKd} \sum_{i=1}^{k-1} \epsilon^{k-i} \|\boldsymbol{\phi}(s,a)\| \\
&\leq \sum_{i=1}^{k-1} \epsilon^{k-i-1} \frac{1}{4MHKd} 4H\sqrt{MKd}\sqrt{MK} \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} \\
&\leq \frac{4}{3} \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}.
\end{aligned}$$

So, with probablity at least $1-\delta$, we have

$$\left|r_h(s,a) + \mathbb{P}_h V_{m,h+1}^k(s,a) - \boldsymbol{\phi}(s,a)^\top \boldsymbol{\mu}_{m,h}^{k,J_k}\right| \leq \left(5H\sqrt{d}C_\delta + \frac{4}{3}\right) \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}. \tag{G.17}$$

Consider the denominator of Z_k: recall from the definition of $\boldsymbol{\Sigma}_{m,h}^{k,J_k}$ from Proposition F.3, then we have

$$\begin{aligned}
\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{m,h}^{k,J_k} \boldsymbol{\phi}(s,a) &= \sum_{i=1}^{k} \frac{1}{\beta_{m,i}} \boldsymbol{\phi}(s,a)^\top \mathbf{A}_k^{J_k} \dots \mathbf{A}_{i+1}^{J_{i+1}} \left(\mathbf{I} - \mathbf{A}^{2J_i}\right) \left(\boldsymbol{\Lambda}_{m,h}^i\right)^{-1} (\mathbf{I} + \mathbf{A}_i)^{-1} \mathbf{A}_{i+1}^{J_{i+1}} \dots \mathbf{A}_k^{J_k} \boldsymbol{\phi}(s,a) \\
&\geq \sum_{i=1}^{k} \frac{1}{2\beta_{m,i}} \boldsymbol{\phi}(s,a)^\top \mathbf{A}_k^{J_k} \dots \mathbf{A}_{i+1}^{J_{i+1}} \left(\mathbf{I} - \mathbf{A}^{2J_i}\right) \left(\boldsymbol{\Lambda}_{m,h}^i\right)^{-1} \mathbf{A}_{i+1}^{J_{i+1}} \dots \mathbf{A}_k^{J_k} \boldsymbol{\phi}(s,a),
\end{aligned}$$

where we used the fact that $\frac{1}{2}\mathbf{I} \preccurlyeq (\mathbf{I}+\mathbf{A}_k)^{-1}$. Then we have

$$\begin{aligned}
\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{m,h}^{k,J_k} \boldsymbol{\phi}(s,a) &\geq \sum_{i=1}^{k} \frac{1}{2\beta_{m,i}} \boldsymbol{\phi}(s,a)^\top \mathbf{A}_k^{J_k} \dots \mathbf{A}_{i+1}^{J_{i+1}} \big(\big(\boldsymbol{\Lambda}_{m,h}^i\big)^{-1} - \mathbf{A}_i^{J_i} \big(\boldsymbol{\Lambda}_{m,h}^i\big)^{-1} \mathbf{A}_i^{J_i} \big) \mathbf{A}_{i+1}^{J_{i+1}} \dots \mathbf{A}_k^{J_k} \boldsymbol{\phi}(s,a) \\
&= \frac{1}{2\beta_K} \sum_{i=1}^{k-1} \boldsymbol{\phi}(s,a)^\top \mathbf{A}_k^{J_k} \dots \mathbf{A}_{i+1}^{J_{i+1}} \big(\big(\boldsymbol{\Lambda}_{m,h}^i\big)^{-1} - \big(\boldsymbol{\Lambda}_{m,h}^{i+1}\big)^{-1} \big) \mathbf{A}_{i+1}^{J_{i+1}} \dots \mathbf{A}_k^{J_k} \boldsymbol{\phi}(s,a) \\
&\quad - \frac{1}{2\beta_K} \boldsymbol{\phi}(s,a)^\top \mathbf{A}_k^{J_k} \dots \mathbf{A}_1^{J_1} \big(\boldsymbol{\Lambda}_{m,h}^1\big)^{-1} \mathbf{A}_1^{J_1} \dots \mathbf{A}_k^{J_k} \boldsymbol{\phi}(s,a) \\
&\quad + \frac{1}{2\beta_K} \boldsymbol{\phi}(s,a)^\top \big(\boldsymbol{\Lambda}_{m,h}^k\big)^{-1} \boldsymbol{\phi}(s,a).
\end{aligned}$$

By the definition of $\boldsymbol{\Lambda}_{m,h}^i$ and Woodbury formula, we have

$$\begin{aligned}
\big(\boldsymbol{\Lambda}_{m,h}^i\big)^{-1} - \big(\boldsymbol{\Lambda}_{m,h}^{i+1}\big)^{-1} &= \big(\boldsymbol{\Lambda}_{m,h}^i\big)^{-1} - \Bigg(\boldsymbol{\Lambda}_{m,h}^i + \sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \boldsymbol{\phi}\big(s^l,a^l\big) \boldsymbol{\phi}\big(s^l,a^l\big)^\top \Bigg)^{-1} \\
&= (\boldsymbol{\Lambda}_{m,h}^i)^{-1} \boldsymbol{\varphi} \big(\mathbf{I}_n + \boldsymbol{\varphi}^\top (\boldsymbol{\Lambda}_{m,h}^i)^{-1} \boldsymbol{\varphi} \big)^{-1} \boldsymbol{\varphi}^\top (\boldsymbol{\Lambda}_{m,h}^i)^{-1},
\end{aligned}$$

where $\boldsymbol{\varphi}$ is a matrix with the dimension of $d \times n$, n is the number difference of $\boldsymbol{\phi}\big(s^l,a^l\big)$ between $\big(\boldsymbol{\Lambda}_{m,h}^i\big)^{-1}$ and $\big(\boldsymbol{\Lambda}_{m,h}^{i+1}\big)^{-1}$ (i.e. we concatenate all $\boldsymbol{\phi}\big(s^l,a^l\big)$ in to the matrix $\boldsymbol{\varphi}$). Note that $n \leq M$, we have

$$\begin{aligned}
&\boldsymbol{\phi}(s,a)^\top \mathbf{A}_k^{J_k} \dots \mathbf{A}_{i+1}^{J_{i+1}} \Big(\big(\boldsymbol{\Lambda}_{m,h}^i\big)^{-1} - \big(\boldsymbol{\Lambda}_{m,h}^{i+1}\big)^{-1} \Big) \mathbf{A}_{i+1}^{J_{i+1}} \dots \mathbf{A}_k^{J_k} \boldsymbol{\phi}(s,a) \\
&= \boldsymbol{\phi}(s,a)^\top \mathbf{A}_k^{J_k} \dots \mathbf{A}_{i+1}^{J_{i+1}} \Big(\big(\boldsymbol{\Lambda}_{m,h}^i\big)^{-1} \boldsymbol{\varphi} (\mathbf{I}_n + \boldsymbol{\varphi}^\top \big(\boldsymbol{\Lambda}_{m,h}^i\big)^{-1} \boldsymbol{\varphi})^{-1} \boldsymbol{\varphi}^\top \big(\boldsymbol{\Lambda}_{m,h}^i\big)^{-1} \Big) \mathbf{A}_{i+1}^{J_{i+1}} \dots \mathbf{A}_k^{J_k} \boldsymbol{\phi}(s,a) \\
&\leq \boldsymbol{\phi}(s,a)^\top \mathbf{A}_k^{J_k} \dots \mathbf{A}_{i+1}^{J_{i+1}} \big(\boldsymbol{\Lambda}_{m,h}^i\big)^{-1} \boldsymbol{\varphi} \boldsymbol{\varphi}^\top \big(\boldsymbol{\Lambda}_{m,h}^i\big)^{-1} \mathbf{A}_{i+1}^{J_{i+1}} \dots \mathbf{A}_k^{J_k} \boldsymbol{\phi}(s,a) \\
&= \big\| \boldsymbol{\phi}(s,a)^\top \mathbf{A}_k^{J_k} \dots \mathbf{A}_{i+1}^{J_{i+1}} \big(\boldsymbol{\Lambda}_{m,h}^i\big)^{-1} \boldsymbol{\varphi} \big\|_2^2 \\
&\leq \big\| \mathbf{A}_k^{J_k} \dots \mathbf{A}_{i+1}^{J_i+1} (\boldsymbol{\Lambda}_{m,h}^i)^{-1/2} \boldsymbol{\phi}(s,a) \big\|_2^2 \cdot \big\| (\boldsymbol{\Lambda}_{m,h}^i)^{-1/2} \boldsymbol{\varphi} \big\|_F^2 \\
&\leq \prod_{j=i+1}^{k} \Big(1 - 2\eta_{m,j} \lambda_{\min}\big(\boldsymbol{\Lambda}_{m,h}^j\big) \Big)^{2J_j} \operatorname{tr}\big(\boldsymbol{\varphi}^\top \big(\boldsymbol{\Lambda}_{m,h}^i\big)^{-1} \boldsymbol{\varphi} \big) \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^i)^{-1}}^2,
\end{aligned}$$

where $\|\cdot\|_F$ is Frobenius norm and the last inequality is due to $\|\boldsymbol{\Lambda}^{-\frac{1}{2}}\mathbf{X}\|_F^2 = \operatorname{tr}(\mathbf{X}^\top \boldsymbol{\Lambda}^{-1} \mathbf{X})$ and (G.2). Therefore, we have

$$\begin{aligned}
\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{m,h}^{k,J_k} \boldsymbol{\phi}(s,a) &\geq \frac{1}{2\beta_K} \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}^2 - \frac{1}{2\beta_K} \prod_{i=1}^{k} \big(1 - 2\eta_{m,i} \lambda_{\min}\big(\boldsymbol{\Lambda}_{m,h}^i\big) \big)^{2J_i} \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^1)^{-1}}^2 \\
&\quad - \frac{1}{2\beta_K} \sum_{i=1}^{k-1} \prod_{j=i+1}^{k} \big(1 - 2\eta_{m,j} \lambda_{\min}\big(\boldsymbol{\Lambda}_{m,h}^j\big) \big)^{2J_j} \operatorname{tr}\big(\boldsymbol{\varphi}^\top \big(\boldsymbol{\Lambda}_{m,h}^i\big)^{-1} \boldsymbol{\varphi} \big) \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^i)^{-1}}^2.
\end{aligned}$$

Similar to the proof of Lemma F.5, note that $\operatorname{tr}\big(\boldsymbol{\varphi}^\top (\boldsymbol{\Lambda}_{m,h}^i)^{-1} \boldsymbol{\varphi}\big) \leq M$, when we choose $J_j \geq 2\kappa_j \log(3kM)$, we have

$$\begin{aligned}
\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Sigma}_{m,h}^{k,J_k}} &\geq \frac{1}{2\sqrt{\beta_K}} \Bigg(\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} - \frac{\|\boldsymbol{\phi}(s,a)\|}{(3KM)^k} - \sum_{i=1}^{k-1} \frac{\sqrt{M}}{(3kM)^{k-i}} \|\boldsymbol{\phi}(s,a)\| \Bigg) \\
&\geq \frac{1}{2\sqrt{\beta_K}} \Bigg(\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} - \frac{1}{3\sqrt{kM}} \|\boldsymbol{\phi}(s,a)\| - \frac{1}{6\sqrt{kM}} \|\boldsymbol{\phi}(s,a)\| \Bigg) \\
&\geq \frac{1}{4\sqrt{\beta_K}} \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}},
\end{aligned} \tag{G.18}$$

where we used the fact that $\lambda_{\min}\big(\big(\mathbf{\Lambda}_{m,h}^k\big)^{-1}\big) \geq 1/kM$ and $\|\boldsymbol{\phi}(s,a)\|_{(\mathbf{\Lambda}_{m,h}^k)^{-1}} \geq 1/\sqrt{kM}\|\boldsymbol{\phi}(s,a)\|$. Therefore, according to (G.17) and (G.18), with probablity at least $1-\delta$, it holds that

$$
\begin{aligned}
|Z_k| &= \left|\frac{r_h(s,a) + \mathbb{P}_h V_{m,h+1}^k(s,a) - \boldsymbol{\phi}(s,a)^\top \boldsymbol{\mu}_{m,h}^{k,J_k}}{\sqrt{\boldsymbol{\phi}(s,a)^\top \mathbf{\Sigma}_{m,h}^{k,J_k} \boldsymbol{\phi}(s,a)}}\right| \\
&\leq \frac{5H\sqrt{d}C_\delta + \frac{4}{3}}{\frac{1}{4\sqrt{\beta_K}}},
\end{aligned}
$$

which implies $|Z_k| < 1$ when $\frac{1}{\sqrt{\beta_K}} = 20H\sqrt{d}C_\delta + \frac{16}{3}$.

Till now we have proved that for any fixed $\boldsymbol{\phi}(s,a) \in \mathcal{C}(\varepsilon)$ and for all $(m,h,k) \in \mathcal{M} \times [H] \times [K]$, for any fixed $n \in [N]$, with probablity at least $1-\delta$, we have

$$
\mathbb{P}\Big(\boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^{k,J_k,n} - r_h(s,a) - \mathbb{P}_h V_{m,h+1}^k(s,a) \geq 0\Big) \geq \frac{1}{2\sqrt{2e\pi}}.
$$

By taking union bound over $n \in [N]$, with probablity at least $1-\delta$, we have

$$
\mathbb{P}\Big(\max_{n\in[N]} \big\{\boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^{k,J_k,n} - r_h(s,a) - \mathbb{P}_h V_{m,h+1}^k(s,a)\big\} \geq 0\Big) \geq 1 - \Big(1 - \frac{1}{2\sqrt{2e\pi}}\Big)^N = 1 - {c_0'}^N,
$$

where $c_0' = 1 - \frac{1}{2\sqrt{2e\pi}}$. Therefore, for any fixed $\boldsymbol{\phi}(s,a) \in \mathcal{C}(\varepsilon)$ and for all $(m,h,k) \in \mathcal{M} \times [H] \times [K]$, with probability at least $(1-\delta)\big(1-{c_0'}^N\big) > 1 - \delta - {c_0'}^N$, we have

$$
\max_{n\in[N]} \big\{\boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^{k,J_k,n} - r_h(s,a) - \mathbb{P}_h V_{m,h+1}^k(s,a)\big\} \geq 0. \tag{G.19}
$$

Next for any $\boldsymbol{\phi} = \boldsymbol{\phi}(s,a)$, we can find $\boldsymbol{\phi}' \in \mathcal{C}(\varepsilon)$ such that $\|\boldsymbol{\phi} - \boldsymbol{\phi}'\|_{(\mathbf{\Lambda}_{m,h}^k)^{-1}} \leq \varepsilon$. We define $\Delta\boldsymbol{\phi} = \boldsymbol{\phi} - \boldsymbol{\phi}'$. Recall from Definition 4.1, we have

$$
r_h(s,a) + \mathbb{P}_h V_{m,h+1}^k(s,a) = \boldsymbol{\phi}(s,a)^\top \boldsymbol{\theta}_h + \boldsymbol{\phi}(s,a)^\top \big\langle \boldsymbol{\mu}_h, V_{m,h+1}^k \big\rangle_{\mathcal{S}} \stackrel{\text{def}}{=} \boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^k,
$$

where $\mathbf{w}_{m,h}^k = \boldsymbol{\theta}_h + \big\langle \boldsymbol{\mu}_h, V_{m,h+1}^k \big\rangle_{\mathcal{S}}$. Note that $\max\{\|\boldsymbol{\mu}_h(\mathcal{S})\|, \|\boldsymbol{\theta}_h\|\} \leq \sqrt{d}$ and $V_{m,h+1}^k \leq H - h \leq H$, thus we have

$$
\big\|\mathbf{w}_{m,h}^k\big\| \leq \|\boldsymbol{\theta}_h\| + \big\|\big\langle \boldsymbol{\mu}_h, V_{m,h+1}^k \big\rangle_{\mathcal{S}}\big\| \leq \sqrt{d} + H\sqrt{d} \leq 2H\sqrt{d}.
$$

Then we define the regression error $\Delta\mathbf{w}_{m,h}^k = \mathbf{w}_{m,h}^k - \mathbf{w}_{m,h}^{k,J_k,n}$. Thus we have

$$
\max_{n\in[N]} \big\{\boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^{k,J_k,n} - r_h(s,a) - \mathbb{P}_h V_{m,h+1}^k(s,a)\big\} = \max_{n\in[N]} \big\{-\boldsymbol{\phi}(s,a)^\top \Delta\mathbf{w}_{m,h}^k\big\}.
$$

Then by Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
\boldsymbol{\phi}^\top \Delta\mathbf{w}_{m,h}^k &= {\boldsymbol{\phi}'}^\top \Delta\mathbf{w}_{m,h}^k + \Delta\boldsymbol{\phi}^\top \Delta\mathbf{w}_{m,h}^k \\
&\geq {\boldsymbol{\phi}'}^\top \Delta\mathbf{w}_{m,h}^k - \|\Delta\boldsymbol{\phi}\| \cdot \big\|\Delta\mathbf{w}_{m,h}^k\big\| \\
&\geq {\boldsymbol{\phi}'}^\top \Delta\mathbf{w}_{m,h}^k - \sqrt{MK}\varepsilon\big\|\Delta\mathbf{w}_{m,h}^k\big\|.
\end{aligned}
$$

By triangle inequality, with probability at least $1-\delta$, we have

$$
\big\|\Delta\mathbf{w}_{m,h}^k\big\| \leq \big\|\mathbf{w}_{m,h}^k\big\| + \big\|\mathbf{w}_{m,h}^{k,J_k,n}\big\| \leq 2H\sqrt{d} + B_{\delta/NMHK}
$$

Denote $\alpha_\delta = \sqrt{MK}\big(2H\sqrt{d} + B_{\delta/NMHK}\big)$. Then, for all $(m,h,k) \in \mathcal{M} \times [H] \times [K]$, with probability at least $1-\delta$, we have

$$
\max_{n\in[N]} \big\{\boldsymbol{\phi}^\top \Delta\mathbf{w}_{m,h}^k\big\} \geq \max_{n\in[N]} \big\{{\boldsymbol{\phi}'}^\top \Delta\mathbf{w}_{m,h}^k\big\} - \alpha_\delta\varepsilon.
$$

Recall from (G.19), by taking union bound, with probability at least $1 - |\mathcal{C}(\varepsilon)|{c_0'}^N - 2\delta$, for all $(m, h, k) \in \mathcal{M} \times [H] \times [K]$ and for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\max_{n\in[N]} \left\{\boldsymbol{\phi}^\top \Delta \mathbf{w}_{m,h}^k\right\} \geq -\alpha_\delta \varepsilon.$$

Finally, with probability at least $1 - |\mathcal{C}(\varepsilon)|{c_0'}^N - 2\delta$, for all $(m, h, k) \in \mathcal{M} \times [H] \times [K]$ and for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have

$$l_{m,h}^k(s, a) \leq \alpha_\delta \varepsilon.$$

This completes the proof. □

G.9 Proof of Lemma F.12

Proof. For simplicity, we denote $(s_{m,h}^k, a_{m,h}^k)$ as $z_{m,h}^k$. Then we consider the following mappings $(\nu_M, \nu_K) : [MK] \to [M] \times [K]$,

$$\nu_M(\tau) = \tau(\mathrm{mod} M), \qquad \nu_K = \left\lceil \frac{\tau}{M} \right\rceil,$$

where we set $\nu_M(\tau) = M$ if $M|\tau$. Next, for any $\tau \geq 0$, we define

$$\bar{\boldsymbol{\Lambda}}_h^\tau = \lambda \mathbf{I} + \sum_{u=1}^{\tau M} \boldsymbol{\phi}\Big(z_{\nu_M(u),h}^{\nu_K(u)}\Big)\boldsymbol{\phi}\Big(z_{\nu_M(u),h}^{\nu_K(u)}\Big)^\top, \quad \text{for } \tau > 0,$$
$$\bar{\boldsymbol{\Lambda}}_h^0 = \lambda \mathbf{I}, \quad \text{for } \tau = 0.$$

We denote $\sigma = \{\sigma_1, \ldots, \sigma_n\}$ as the synchronization episodes, where $\sigma_i \in [K]$, we also denote $\sigma_0 = 0$. Then we separate the episodes $k = 1, \ldots, K$ into two groups based on the following condition,

$$1 \leq \frac{\det(\bar{\boldsymbol{\Lambda}}_h^{\sigma_i})}{\det(\bar{\boldsymbol{\Lambda}}_h^{\sigma_{i-1}})} \leq 3. \tag{G.20}$$

Note that the left inequality always holds due to $\bar{\boldsymbol{\Lambda}}_h^{\sigma_{i-1}} \preccurlyeq \bar{\boldsymbol{\Lambda}}_h^{\sigma_i}$ and the trivial fact that $\mathbf{A} \preccurlyeq \mathbf{B} \Rightarrow \det(\mathbf{A}) \leq \det(\mathbf{B})$. Then we define that $I_1 = \{k \in \mathbb{N}^+, k \in [\sigma_{i-1}, \sigma_i), \forall i \in [n] | \text{(G.20) is true}\}$ and $I_2 = \{k \in \mathbb{N}^+, k \in [\sigma_{i-1}, \sigma_i), \forall i \in [n] | \text{(G.20) is false}\}$, then $[K] = I_1 \cup I_2 \cup \{K\}$. For any $k \in [\sigma_{i-1}, \sigma_i)$ and $k \in I_1$, note that $\bar{\boldsymbol{\Lambda}}_h^{\sigma_{i-1}} \preccurlyeq \boldsymbol{\Lambda}_{m,h}^k \preccurlyeq \bar{\boldsymbol{\Lambda}}_h^k \preccurlyeq \bar{\boldsymbol{\Lambda}}_h^{\sigma_i}$, thus for any $m \in \mathcal{M}$, we have

$$\begin{aligned}
\big\|\boldsymbol{\phi}\big(z_{m,h}^k\big)\big\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} &\leq \big\|\boldsymbol{\phi}\big(z_{m,h}^k\big)\big\|_{(\bar{\boldsymbol{\Lambda}}_h^k)^{-1}} \sqrt{\frac{\det(\bar{\boldsymbol{\Lambda}}_h^k)}{\det(\boldsymbol{\Lambda}_{m,h}^k)}} \\
&\leq \big\|\boldsymbol{\phi}\big(z_{m,h}^k\big)\big\|_{(\bar{\boldsymbol{\Lambda}}_h^k)^{-1}} \sqrt{\frac{\det(\bar{\boldsymbol{\Lambda}}_h^{\sigma_i})}{\det(\bar{\boldsymbol{\Lambda}}_h^{\sigma_{i-1}})}} \\
&\leq 2\big\|\boldsymbol{\phi}\big(z_{m,h}^k\big)\big\|_{(\bar{\boldsymbol{\Lambda}}_h^k)^{-1}},
\end{aligned} \tag{G.21}$$

where the first inequality follows from Lemma L.12, the second inequality follows from the trivial fact that $\mathbf{A} \preccurlyeq \mathbf{B} \Rightarrow \det(\mathbf{A}) \leq \det(\mathbf{B})$, and the final inequality holds because $k \in I_1$. Then we will bound the summation for $k \in I_1$ and $k \in I_2$ respectively.

$$\begin{aligned}
\sum_{k\in I_1\cup\{K\}} \sum_{m\in\mathcal{M}} \|\boldsymbol{\phi}(z_{m,h}^k)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} &\leq \sqrt{MK \sum_{m\in\mathcal{M}} \sum_{k\in I_1\cup\{K\}} \|\boldsymbol{\phi}(z_{m,h}^k)\|^2_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}} \\
&\leq 2\sqrt{MK \sum_{m\in\mathcal{M}} \sum_{k\in I_1\cup\{K\}} \|\boldsymbol{\phi}(z_{m,h}^k)\|^2_{(\bar{\boldsymbol{\Lambda}}_h^k)^{-1}}}
\end{aligned}$$

$$\leq 2\sqrt{MK\sum_{m\in\mathcal{M}}\sum_{k=1}^{K}\|\boldsymbol{\phi}(z_{m,h}^k)\|^2_{(\bar{\boldsymbol{\Lambda}}_h^k)^{-1}}}$$
$$\leq 2\sqrt{MK\log\left(\frac{\det(\boldsymbol{\Lambda}_h^K)}{\det(\lambda\mathbf{I})}\right)},$$

where the first inequality follows from Cauchy-Schwarz inequality, the second inequality holds due to (G.21), the final equality follows from Lemma L.1 and $\boldsymbol{\Lambda}_h^K = \sum_{m\in\mathcal{M}}\sum_{k=1}^K \boldsymbol{\phi}\big(s_{m,h}^k, a_{m,h}^k\big)\boldsymbol{\phi}\big(s_{m,h}^k, a_{m,h}^k\big)^\top + \lambda\mathbf{I}$.

For any interval $[\sigma_{i-1}, \sigma_i)$, define $\Delta_i = \sigma_i - \sigma_{i-1} - 1$, we calculate that

$$\sum_{k=\sigma_{i-1}}^{\sigma_i-1}\big\|\boldsymbol{\phi}(z_{m,h}^k)\big\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} \leq \sqrt{\Delta_i\sum_{k=\sigma_{i-1}}^{\sigma_i-1}\big\|\boldsymbol{\phi}(z_{m,h}^k)\big\|^2_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}}$$
$$\leq \sqrt{\Delta_i\log\left(\frac{\det(\boldsymbol{\Lambda}_{m,h}^{\sigma_i-1})}{\det(\boldsymbol{\Lambda}_{m,h}^{\sigma_{i-1}})}\right)}$$
$$\leq \sqrt{\gamma},$$

where the last inequality follows from the synchronization condition (3.3).

Define $R_h = \left\lceil \log\left(\frac{\det(\boldsymbol{\Lambda}_h^K)}{\det(\lambda\mathbf{I})}\right)\right\rceil$, note that $\sigma_n \leq K$, then we can find that

$$R_h \geq \log\left(\frac{\det(\bar{\boldsymbol{\Lambda}}_h^{\sigma_n})}{\det(\bar{\boldsymbol{\Lambda}}_h^{\sigma_0})}\right) = \sum_{i=1}^n \log\left(\frac{\det(\bar{\boldsymbol{\Lambda}}_h^{\sigma_i})}{\det(\bar{\boldsymbol{\Lambda}}_h^{\sigma_{i-1}})}\right).$$

We can claim that I_2 has at most R_h synchronization episodes, otherwise

$$R_h \geq \sum_{i=1}^n \log\left(\frac{\det(\bar{\boldsymbol{\Lambda}}_h^{\sigma_i})}{\det(\bar{\boldsymbol{\Lambda}}_h^{\sigma_{i-1}})}\right) \geq \sum_{i\in\{i|\sigma_{i-1}\in I_2\}} \log\left(\frac{\det(\bar{\boldsymbol{\Lambda}}_h^{\sigma_i})}{\det(\bar{\boldsymbol{\Lambda}}_h^{\sigma_{i-1}})}\right) \geq R_h\log 3,$$

which causes the contradiction. Thus I_2 has at most R_h intervals, then we get

$$\sum_{k\in I_2}\sum_{m\in\mathcal{M}}\big\|\boldsymbol{\phi}(z_{m,h}^k)\big\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} \leq R_h M\sqrt{\gamma} \leq \left(\log\left(\frac{\det(\boldsymbol{\Lambda}_h^K)}{\det(\lambda\mathbf{I})}\right)+1\right)M\sqrt{\gamma}.$$

Finally, we can bound the total summation,

$$\sum_{m\in\mathcal{M}}\sum_{k=1}^K\big\|\boldsymbol{\phi}(z_{m,h}^k)\big\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} \leq \sum_{m\in\mathcal{M}}\sum_{k\in I_2}\big\|\boldsymbol{\phi}(z_{m,h}^k)\big\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} + \sum_{m\in\mathcal{M}}\sum_{k\in I_1\cup\{K\}}\|\boldsymbol{\phi}(z_{m,h}^k)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}$$
$$\leq \left(\log\left(\frac{\det(\boldsymbol{\Lambda}_h^K)}{\det(\lambda\mathbf{I})}\right)+1\right)M\sqrt{\gamma} + 2\sqrt{MK\log\left(\frac{\det(\boldsymbol{\Lambda}_h^K)}{\det(\lambda\mathbf{I})}\right)}.$$

This completes the proof. □

H Proof of the Regret Bound for CoopTS-LMC in Misspecified Setting

In this section, we prove the regret bound for CoopTS-LMC in the misspecified setting. The regret analysis, the essential supporting lemmas and their corresponding proofs are almost same as what we have presented in Appendix F and Appendix G. Here we mainly point out the differences of proof between these two settings.

H.1 Supporting Lemmas

Definition H.1 (Model prediction error)**.** For any $(m,k,h) \in \mathcal{M} \times [K] \times [H]$, we define the model error associated with the reward $r_{m,h}$,

$$l^k_{m,h}(s,a) = r_{m,h}(s,a) + \mathbb{P}_{m,h} V^k_{m,h+1}(s,a) - Q^k_{m,h}(s,a).$$

Lemma H.2. Let $\lambda = 1$ in Algorithm 3. Under Definition E.1, for any fixed $0 < \delta < 1$, with probability at least $1-\delta$, for all $(m,k,h) \in \mathcal{M} \times [K] \times [H]$ and for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\left|\boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}^k_{m,h} - r_{m,h}(s,a) - \mathbb{P}_{m,h} V^k_{m,h+1}(s,a)\right| \le \big(5H\sqrt{d}C_\delta + 3H\zeta\sqrt{MKd}\big)\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}} + 3H\zeta,$$

where C_δ is defined in Lemma F.7.

Proof of Lemma H.2. Recall from Definition E.1, we have

$$\begin{aligned}
\left|\mathbb{P}_{m,h} V^k_{m,h+1}(s,a) - \boldsymbol{\phi}(s,a)^\top \big\langle \boldsymbol{\mu}_h, V^k_{m,h+1}\big\rangle_{\mathcal{S}}\right| &\le \left\|\mathbb{P}_{m,h}(\cdot \mid s,a) - \big\langle \boldsymbol{\phi}(s,a), \boldsymbol{\mu}_h(\cdot)\big\rangle\right\|_1 \|V^k_{m,h+1}\|_\infty \\
&\le 2H\left\|\mathbb{P}_{m,h}(\cdot \mid s,a) - \big\langle \boldsymbol{\phi}(s,a), \boldsymbol{\mu}_h(\cdot)\big\rangle\right\|_{\mathrm{TV}} \\
&\le 2H\zeta,
\end{aligned}$$

where the first inequality follows from Cauchy-Schwarz inequality, the second inequality follows from the fact that $\|V^k_{m,h+1}\|_\infty \le H$ and P_2, $\|P_1 - P_2\|_{\mathrm{TV}} = \frac{1}{2}\sum_{\mathbf{x}\in\boldsymbol{\omega}} |P_1(\mathbf{x}) - P_2(\mathbf{x})| = \frac{1}{2}\|P_1 - P_2\|_1$ for two distributions P_1 and P_2, note that here we regard distribution as infinite dimensional vector, the third inequality follows from Definition E.1. Define $\Delta_{m,1} = \mathbb{P}_{m,h} V^k_{m,h+1}(s,a) - \boldsymbol{\phi}(s,a)^\top \big\langle \boldsymbol{\mu}_h, V^k_{m,h+1}\big\rangle_{\mathcal{S}}$, thus $|\Delta_{m,1}| \le 2H\zeta$. Then we have

$$\begin{aligned}
\mathbb{P}_{m,h} V^k_{m,h+1}(s,a) &= \boldsymbol{\phi}(s,a)^\top \big\langle \boldsymbol{\mu}_h, V^k_{m,h+1}\big\rangle_{\mathcal{S}} + \Delta_{m,1} \\
&= \boldsymbol{\phi}(s,a)^\top \big(\boldsymbol{\Lambda}^k_{m,h}\big)^{-1} \Bigg(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \boldsymbol{\phi}\big(s^l,a^l\big)\boldsymbol{\phi}\big(s^l,a^l\big)^\top + \lambda \mathbf{I} \Bigg) \big\langle \boldsymbol{\mu}_h, V^k_{m,h+1}\big\rangle_{\mathcal{S}} + \Delta_{m,1} \\
&= \boldsymbol{\phi}(s,a)^\top \big(\boldsymbol{\Lambda}^k_{m,h}\big)^{-1} \Bigg(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \boldsymbol{\phi}\big(s^l,a^l\big)\boldsymbol{\phi}\big(s^l,a^l\big)^\top \big\langle \boldsymbol{\mu}_h, V^k_{m,h+1}\big\rangle_{\mathcal{S}} \Bigg) \\
&\quad + \lambda \boldsymbol{\phi}(s,a)^\top \big(\boldsymbol{\Lambda}^k_{m,h}\big)^{-1} \big\langle \boldsymbol{\mu}_h, V^k_{m,h+1}\big\rangle_{\mathcal{S}} + \Delta_{m,1} \\
&= \boldsymbol{\phi}(s,a)^\top \big(\boldsymbol{\Lambda}^k_{m,h}\big)^{-1} \Bigg(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \boldsymbol{\phi}\big(s^l,a^l\big)\big(\mathbb{P}_{m,h} V^k_{m,h+1}\big)\big(s^l,a^l\big) \Bigg) \\
&\quad - \boldsymbol{\phi}(s,a)^\top \big(\boldsymbol{\Lambda}^k_{m,h}\big)^{-1} \Bigg(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \Delta_{m,1}\boldsymbol{\phi}\big(s^l,a^l\big) \Bigg) \\
&\quad + \lambda \boldsymbol{\phi}(s,a)^\top \big(\boldsymbol{\Lambda}^k_{m,h}\big)^{-1} \big\langle \boldsymbol{\mu}_h, V^k_{m,h+1}\big\rangle_{\mathcal{S}} + \Delta_{m,1}.
\end{aligned} \tag{H.1}$$

Based on (H.1), we can separate the following error into four parts,

$$\begin{aligned}
&\boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}^k_{m,h} - r_{m,h}(s,a) - \mathbb{P}_{m,h} V^k_{m,h+1}(s,a) \\
&= \boldsymbol{\phi}(s,a)^\top \big(\boldsymbol{\Lambda}^k_{m,h}\big)^{-1} \sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \big[r_{m,h}\big(s^l,a^l\big) + V^k_{m,h+1}(s'^l)\big]\boldsymbol{\phi}\big(s^l,a^l\big) - r_{m,h}(s,a) \\
&\quad - \boldsymbol{\phi}(s,a)^\top \big(\boldsymbol{\Lambda}^k_{m,h}\big)^{-1} \Bigg(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \boldsymbol{\phi}\big(s^l,a^l\big)\big(\mathbb{P}_{m,h} V^k_{m,h+1}\big)\big(s^l,a^l\big) \Bigg) \\
&\quad + \Delta_{m,1}\boldsymbol{\phi}(s,a)^\top \big(\boldsymbol{\Lambda}^k_{m,h}\big)^{-1} \Bigg(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \boldsymbol{\phi}\big(s^l,a^l\big) \Bigg) - \lambda \boldsymbol{\phi}(s,a)^\top \big(\boldsymbol{\Lambda}^k_{m,h}\big)^{-1} \big\langle \boldsymbol{\mu}_h, V^k_{m,h+1}\big\rangle_{\mathcal{S}} - \Delta_{m,1}
\end{aligned}$$

$$
\begin{aligned}
&= \underbrace{\boldsymbol{\phi}(s,a)^\top \left(\boldsymbol{\Lambda}_{m,h}^k\right)^{-1} \left(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \boldsymbol{\phi}\left(s^l,a^l\right) \left[\left(V_{m,h+1}^k - \mathbb{P}_{m,h} V_{m,h+1}^k\right)\left(s^l,a^l\right) \right] \right)}_{\text{(i)}} \\
&\quad + \underbrace{\boldsymbol{\phi}(s,a)^\top \left(\boldsymbol{\Lambda}_{m,h}^k\right)^{-1} \left(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} r_{m,h}\left(s^l,a^l\right) \boldsymbol{\phi}\left(s^l,a^l\right) \right) - r_{m,h}(s,a)}_{\text{(ii)}} \\
&\quad - \underbrace{\lambda \boldsymbol{\phi}(s,a)^\top \left(\boldsymbol{\Lambda}_{m,h}^k\right)^{-1} \left\langle \boldsymbol{\mu}_h, V_{m,h+1}^k \right\rangle_{\mathcal{S}}}_{\text{(iii)}} \\
&\quad + \underbrace{\Delta_{m,1} \boldsymbol{\phi}(s,a)^\top \left(\boldsymbol{\Lambda}_{m,h}^k\right)^{-1} \left(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \boldsymbol{\phi}\left(s^l,a^l\right) \right) - \Delta_{m,1}}_{\text{(iv)}}
\end{aligned} \tag{H.2}
$$

We now provide an upper bound for each of the terms in (H.2).

Bounding Term (i) in (H.2): same as (G.11) in Appendix G.6, with probability at least $1-\delta$, we have

$$
|\textbf{Term (i)}| \le 3H\sqrt{d} C_\delta \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}. \tag{H.3}
$$

Bounding Term (ii) + Term (iv) in (H.2): define $\Delta_{m,2} = r_{m,h}(s,a) - \boldsymbol{\phi}(s,a)^\top \boldsymbol{\theta}_h$, then we have $|\Delta_{m,2}| \le \zeta$ due to Definition E.1. Next we have

$$
\begin{aligned}
&\boldsymbol{\phi}(s,a)^\top \left(\boldsymbol{\Lambda}_{m,h}^k\right)^{-1} \left(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} r_{m,h}\left(s^l,a^l\right) \boldsymbol{\phi}\left(s^l,a^l\right) \right) - r_{m,h}(s,a) \\
&= \boldsymbol{\phi}(s,a)^\top \left(\boldsymbol{\Lambda}_{m,h}^k\right)^{-1} \left(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} r_{m,h}\left(s^l,a^l\right) \boldsymbol{\phi}\left(s^l,a^l\right) \right) - \boldsymbol{\phi}(s,a)^\top \boldsymbol{\theta}_h - \Delta_{m,2} \\
&= \boldsymbol{\phi}(s,a)^\top \left(\boldsymbol{\Lambda}_{m,h}^k\right)^{-1} \left(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} r_{m,h}\left(s^l,a^l\right) \boldsymbol{\phi}\left(s^l,a^l\right) - \boldsymbol{\Lambda}_{m,h}^k \boldsymbol{\theta}_h \right) - \Delta_{m,2} \\
&= \boldsymbol{\phi}(s,a)^\top \left(\boldsymbol{\Lambda}_{m,h}^k\right)^{-1} \left(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \boldsymbol{\phi}\left(s^l,a^l\right) r_{m,h}\left(s^l,a^l\right) - \sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \boldsymbol{\phi}\left(s^l,a^l\right) \boldsymbol{\phi}\left(s^l,a^l\right)^\top \boldsymbol{\theta}_h - \lambda \mathbf{I} \boldsymbol{\theta}_h \right) - \Delta_{m,2} \\
&= \boldsymbol{\phi}(s,a)^\top \left(\boldsymbol{\Lambda}_{m,h}^k\right)^{-1} \left(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \boldsymbol{\phi}\left(s^l,a^l\right) \Delta_{m,2} - \lambda \mathbf{I} \boldsymbol{\theta}_h \right) - \Delta_{m,2} \\
&= -\lambda \boldsymbol{\phi}(s,a)^\top \left(\boldsymbol{\Lambda}_{m,h}^k\right)^{-1} \boldsymbol{\theta}_h + \underbrace{\Delta_{m,2} \boldsymbol{\phi}(s,a)^\top \left(\boldsymbol{\Lambda}_{m,h}^k\right)^{-1} \left(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \boldsymbol{\phi}\left(s^l,a^l\right) \right) - \Delta_{m,2}}_{\text{(v)}},
\end{aligned} \tag{H.4}
$$

where the third equality uses the definition of $\boldsymbol{\Lambda}_{m,h}^k$. By Combining (H.4) and (G.13) in Appendix G.6, we obtain

$$
|\textbf{Term (ii)} + \textbf{Term (iv)}| \le \sqrt{\lambda d} \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} + |\textbf{Term (iv)} + \textbf{Term (v)}|. \tag{H.5}
$$

Then we calculate that

$$
|\textbf{Term (iv)} + \textbf{Term (v)}| = \left| (\Delta_{m,1} + \Delta_{m,2}) \boldsymbol{\phi}(s,a)^\top \left(\boldsymbol{\Lambda}_{m,h}^k\right)^{-1} \left(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \boldsymbol{\phi}\left(s^l,a^l\right) \right) - (\Delta_{m,1} + \Delta_{m,2}) \right|
$$

$$
\begin{aligned}
&\leq |\Delta_{m,1}+\Delta_{m,2}| \cdot \left| \boldsymbol{\phi}(s,a)^\top \left(\boldsymbol{\Lambda}_{m,h}^k\right)^{-1} \left(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \boldsymbol{\phi}\left(s^l,a^l\right) \right) \right| + |\Delta_{m,1}+\Delta_{m,2}| \\
&\leq 3H\zeta \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} \sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \left\|\boldsymbol{\phi}(s^l,a^l)\right\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} + 3H\zeta \\
&\leq 3H\zeta \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} \left(\mathcal{K}(k) \sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \left\|\boldsymbol{\phi}(s^l,a^l)\right\|^2_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} \right)^{\frac{1}{2}} + 3H\zeta \\
&\leq 3H\zeta\sqrt{MKd} \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} + 3H\zeta, \qquad \text{(H.6)}
\end{aligned}
$$

where the second inequality follows from Cauchy-Schwarz inequality and the fact that $|\Delta_{m,1}+\Delta_{m,2}| \leq |\Delta_{m,1}|+|\Delta_{m,2}| \leq 2H\zeta+\zeta \leq 3H\zeta$, the third inequality holds because of Cauchy-Schwarz inequality, and the last inequality holds because $\mathcal{K}(k) \leq MK$ and Lemma L.4. Substitute (H.6) into (H.5), we have

$$
|\textbf{Term (ii)} + \textbf{Term (iv)}| \leq \left(3H\zeta\sqrt{MKd}+\sqrt{\lambda d}\right) \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} + 3H\zeta. \qquad \text{(H.7)}
$$

Bounding Term (iii) in (H.2): same as (G.15) in Appendix G.6, we have

$$
|\textbf{Term (iii)}| \leq H\sqrt{\lambda d} \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}. \qquad \text{(H.8)}
$$

Combine all the terms in (H.2) together: by using triangle inequality in (H.2), we combine (H.3), (H.7) and (H.8), then set $\lambda = 1$, with probability at least $1-\delta$, we get

$$
\begin{aligned}
&\left| \boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}_{m,h}^k - r_{m,h}(s,a) - \mathbb{P}_{m,h} V_{m,h+1}^k(s,a) \right| \\
&\leq \left(3H\sqrt{d}C_\delta + \sqrt{d} + H\sqrt{d} + 3H\zeta\sqrt{MKd} \right) \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} + 3H\zeta \\
&\leq \left(5H\sqrt{d}C_\delta + 3H\zeta\sqrt{MKd}\right) \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} + 3H\zeta.
\end{aligned}
$$

This completes the proof. □

Lemma H.3 (Error bound)**.** Let $\lambda = 1$ in Algorithm 3. Under Definition E.1, for any fixed $0 < \delta < 1$, with probability at least $1-\delta-\delta^2$, for any $(m,k,h) \in \mathcal{M} \times [K] \times [H]$ and for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$
-l_{m,h}^k(s,a) \leq \left(5H\sqrt{d}C_\delta + 3H\zeta\sqrt{MKd} + 5\sqrt{\frac{2d\log\left(\sqrt{N}/\delta\right)}{3\beta_K}} + \frac{4}{3} \right) \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} + 3H\zeta,
$$

where C_δ is defined in Lemma F.7.

Proof of Lemma H.3. We do the same process as that in Appendix G.7, and we have

$$
-l_{m,h}^k(s,a) \leq \underbrace{\max_{n\in[N]} \left| \boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^{k,J_k,n} - \boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}_{m,h}^k \right|}_{(i)} + \underbrace{\left| \boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}_{m,h}^k - r_{m,h}(s,a) - \mathbb{P}_{m,h} V_{m,h+1}^k(s,a) \right|}_{(ii)}.
$$

Bounding Term (i): based on (G.16), for any $(m,h,k) \in \mathcal{M} \times [H] \times [K]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$, with probability at least $1-\delta^2$, we have

$$
\max_{n\in[N]} \left| \boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^{k,J_k,n} - \boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}_{m,h}^k \right| \leq \left(5\sqrt{\frac{2d\log\left(\sqrt{N}/\delta\right)}{3\beta_K}} + \frac{4}{3} \right) \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}.,
$$

Bounding Term (ii): based on Lemma H.2, for all $(m,h,k) \in \mathcal{M} \times [H] \times [K]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\left|\boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}_{m,h}^k - r_h^k(s,a) - \mathbb{P}_h V_{m,h+1}^k(s,a)\right| \leq \left(5H\sqrt{d}C_\delta + 3H\zeta\sqrt{MKd}\right)\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} + 3H\zeta,$$

Combine the two result above, by taking union bound, with probability at least $1-\delta-\delta^2$, we have

$$-l_{m,h}^k(s,a) \leq \left(5H\sqrt{d}C_\delta + 3H\zeta\sqrt{MKd} + 5\sqrt{\frac{2d\log\left(\sqrt{N}/\delta\right)}{3\beta_K}} + \frac{4}{3}\right)\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} + 3H\zeta.$$

This completes the proof. □

Lemma H.4 (Optimism)**.** Let $\lambda = 1$ in Algorithm 3 and $c_0' = 1 - \frac{1}{2\sqrt{2e\pi}}$. Under Definition E.1, for any fixed $0 < \delta < 1$, with probability at least $1 - |\mathcal{C}(\varepsilon)|{c_0'}^N - 2\delta$ where $|\mathcal{C}(\varepsilon)| \leq (3/\varepsilon)^d$, for all $(m,h,k) \in \mathcal{M} \times [H] \times [K]$ and for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$l_{m,h}^k(s,a) \leq \alpha_\delta \varepsilon + 3H\zeta,$$

where $\alpha_\delta = \sqrt{MK}\left(2H\sqrt{d} + B_{\delta/NMHK}\right)$.

Proof of Lemma H.4. This proof is similar to the proof in Appendix G.8, we just prove the part that for fixed $\boldsymbol{\phi} \in \mathcal{C}(\varepsilon)$. Recall from Definition H.1,

$$l_{m,h}^k(s,a) = r_{m,h}(s,a) + \mathbb{P}_{m,h}V_{m,h+1}^k(s,a) - Q_{m,h}^k(s,a).$$

Note that

$$Q_{m,h}^k(s,a) = \min\left\{\max_{n\in[N]} \boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^{k,J_k,n}, H-h+1\right\}^+ \leq \max_{n\in[N]} \boldsymbol{\phi}(x,a)^\top \mathbf{w}_{m,h}^{k,J_k,n}.$$

Here we define

$$Z_k = \frac{r_{m,h}(s,a) + \mathbb{P}_{m,h}V_{m,h+1}^k(s,a) - \boldsymbol{\phi}(s,a)^\top \boldsymbol{\mu}_{m,h}^{k,J_k} - (\Delta_{m,1} + \Delta_{m,2})}{\sqrt{\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{m,h}^{k,J_k}\boldsymbol{\phi}(s,a)}},$$

where $\Delta_{m,1} = \mathbb{P}_{m,h}V_{m,h+1}^k(s,a) - \boldsymbol{\phi}(s,a)^\top\langle \boldsymbol{\mu}_h, V_{m,h+1}^k\rangle_{\mathcal{S}}, \Delta_{m,2} = r_{m,h}(s,a) - \boldsymbol{\phi}(s,a)^\top\boldsymbol{\theta}_h$. Based on the results in Proposition F.3, we have that $\boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^{k,J_k,n} \sim \mathcal{N}\left(\boldsymbol{\phi}(s,a)^\top \boldsymbol{\mu}_{m,h}^{k,J_k}, \boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{m,h}^{k,J_k}\boldsymbol{\phi}(s,a)\right)$, for any fixed $n \in [N]$. When $|Z_k| < 1$, by Gaussian concentration Lemma L.10, we have

$$\begin{aligned}
&\mathbb{P}\left(r_{m,h}(s,a) + \mathbb{P}_{m,h}V_{m,h+1}^k(s,a) - \boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^{k,J_k,n} \leq (\Delta_{m,1} + \Delta_{m,2})\right)\\
&= \mathbb{P}\left(\boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^{k,J_k,n} \geq r_{m,h}(s,a) + \mathbb{P}_{m,h}V_{m,h+1}^k(s,a) - (\Delta_{m,1} + \Delta_{m,2})\right)\\
&= \mathbb{P}\left(\frac{\boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^{k,J_k,n} - \boldsymbol{\phi}(s,a)^\top \boldsymbol{\mu}_{m,h}^{k,J_k}}{\sqrt{\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{m,h}^{k,J_k}\boldsymbol{\phi}(s,a)}} \geq \frac{r_{m,h}(s,a) + \mathbb{P}_{m,h}V_{m,h+1}^k(s,a) - (\Delta_{m,1} + \Delta_{m,2}) - \boldsymbol{\phi}(s,a)^\top \boldsymbol{\mu}_{m,h}^{k,J_k}}{\sqrt{\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{m,h}^{k,J_k}\boldsymbol{\phi}(s,a)}}\right)\\
&= \mathbb{P}\left(\frac{\boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^{k,J_k,n} - \boldsymbol{\phi}(s,a)^\top \boldsymbol{\mu}_{m,h}^{k,J_k}}{\sqrt{\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{m,h}^{k,J_k}\boldsymbol{\phi}(s,a)}} \geq Z_k\right)\\
&\geq \frac{1}{2\sqrt{2\pi}}\exp(-Z_k^2/2)\\
&\geq \frac{1}{2\sqrt{2e\pi}}.
\end{aligned}$$

Consider the numerator of Z_k:

$$\begin{aligned}&\left|r_{m,h}(s,a)+\mathbb{P}_{m,h}V^k_{m,h+1}(s,a)-\boldsymbol{\phi}(s,a)^\top\boldsymbol{\mu}^{k,J_k}_{m,h}-(\Delta_{m,1}+\Delta_{m,2})\right|\\&\leq\underbrace{\left|r_{m,h}(s,a)+\mathbb{P}_{m,h}V^k_{m,h+1}(s,a)-\boldsymbol{\phi}(s,a)^\top\widehat{\mathbf{w}}^k_{m,h}-(\Delta_{m,1}+\Delta_{m,2})\right|}_{I_1}+\underbrace{\left|\boldsymbol{\phi}(s,a)^\top\widehat{\mathbf{w}}^k_{m,h}-\boldsymbol{\phi}(s,a)^\top\boldsymbol{\mu}^{k,J_k}_{m,h}\right|}_{I_2}.\end{aligned}\tag{H.9}$$

Bounding Term I_1 in (H.9): recall the proof of Lemma H.2, we do the almost same error decomposition as (H.2) with the only difference of adding term $(\Delta_{m,1}+\Delta_{m,2})$

$$\begin{aligned}&\boldsymbol{\phi}(s,a)^\top\widehat{\mathbf{w}}^k_{m,h}-r_{m,h}(s,a)-\mathbb{P}_{m,h}V^k_{m,h+1}(s,a)+(\Delta_{m,1}+\Delta_{m,2})\\&=\underbrace{\boldsymbol{\phi}(s,a)^\top\left(\boldsymbol{\Lambda}^k_{m,h}\right)^{-1}\left(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)}\boldsymbol{\phi}\left(s^l,a^l\right)\left[\left(V^k_{m,h+1}-\mathbb{P}_{m,h}V^k_{m,h+1}\right)\left(s^l,a^l\right)\right]\right)}_{\text{(i)}}\\&\quad+\underbrace{\boldsymbol{\phi}(s,a)^\top\left(\boldsymbol{\Lambda}^k_{m,h}\right)^{-1}\left(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)}r_{m,h}\left(s^l,a^l\right)\boldsymbol{\phi}\left(s^l,a^l\right)\right)-r_{m,h}(s,a)}_{\text{(ii)}}\\&\quad-\underbrace{\lambda\boldsymbol{\phi}(s,a)^\top\left(\boldsymbol{\Lambda}^k_{m,h}\right)^{-1}\left\langle\boldsymbol{\mu}_h,V^k_{m,h+1}\right\rangle_{\mathcal{S}}}_{\text{(iii)}}\\&\quad+\underbrace{\Delta_{m,1}\boldsymbol{\phi}(s,a)^\top\left(\boldsymbol{\Lambda}^k_{m,h}\right)^{-1}\left(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)}\boldsymbol{\phi}\left(s^l,a^l\right)\right)+\Delta_{m,2}}_{\text{(iv)}}\end{aligned}\tag{H.10}$$

We now provide an upper bound for each of the terms in (H.10).

Bounding Term (i) in (H.10): almost same as (G.11) in Appendix G.6 with the only difference between $\mathbb{P}_h$ and $\mathbb{P}_{m,h}$, with probability at least $1-\delta$, we have

$$|\textbf{Term (i)}|\leq 3H\sqrt{d}C_\delta\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}}.\tag{H.11}$$

Bounding Term (ii) + Term (iv) in (H.10): we do the same calculation as that in the proof of Lemma H.2, based on (H.4), we have

$$\begin{aligned}\textbf{Term (ii)}&=\boldsymbol{\phi}(s,a)^\top\left(\boldsymbol{\Lambda}^k_{m,h}\right)^{-1}\left(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)}r_{m,h}\left(s^l,a^l\right)\boldsymbol{\phi}\left(s^l,a^l\right)\right)-r_{m,h}(s,a)\\&=-\lambda\boldsymbol{\phi}(s,a)^\top\left(\boldsymbol{\Lambda}^k_{m,h}\right)^{-1}\boldsymbol{\theta}_h+\underbrace{\Delta_{m,2}\boldsymbol{\phi}(s,a)^\top\left(\boldsymbol{\Lambda}^k_{m,h}\right)^{-1}\left(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)}\boldsymbol{\phi}\left(s^l,a^l\right)\right)-\Delta_{m,2}}_{\text{(v)}}.\end{aligned}\tag{H.12}$$

By Combining (H.12) and (G.13) in Appendix G.6, we obtain

$$|\textbf{Term (ii)}+\textbf{Term (iv)}|\leq\sqrt{\lambda d}\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}}+|\textbf{Term (iv)}+\textbf{Term (v)}|.\tag{H.13}$$

Then we calculate that

$$|\textbf{Term (iv)}+\textbf{Term (v)}|=|\Delta_{m,1}+\Delta_{m,2}|\cdot\left|\boldsymbol{\phi}(s,a)^\top\left(\boldsymbol{\Lambda}^k_{m,h}\right)^{-1}\left(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)}\boldsymbol{\phi}\left(s^l,a^l\right)\right)\right|$$

$$
\begin{aligned}
&\leq 3H\zeta\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}} \sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \left\|\boldsymbol{\phi}(s^l,a^l)\right\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}} \\
&\leq 3H\zeta\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}} \left(\mathcal{K}(k) \sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \left\|\boldsymbol{\phi}(s^l,a^l)\right\|^2_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}}\right)^{\frac{1}{2}} \\
&\leq 3H\zeta\sqrt{MKd}\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}}, \qquad \text{(H.14)}
\end{aligned}
$$

where the first inequality follows from Cauchy-Schwarz inequality and the fact that $|\Delta_{m,1}+\Delta_{m,2}| \leq 3H\zeta$, the second inequality holds because of Cauchy-Schwarz inequality, and the last inequality holds because $\mathcal{K}(k) \leq MK$ and Lemma L.4. Substitute (H.14) into (H.13), we have

$$
|\textbf{Term (ii)} + \textbf{Term (iv)}| \leq \left(3H\zeta\sqrt{MKd}+\sqrt{\lambda d}\right)\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}}. \qquad \text{(H.15)}
$$

Bounding Term (iii) in (H.10): same as (G.15) in Appendix G.6, we have

$$
|\textbf{Term (iii)}| \leq H\sqrt{\lambda d}\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}}. \qquad \text{(H.16)}
$$

Combine all the terms in (H.10) together: by using triangle inequality in (H.10), we combine (H.11), (H.15) and (H.16), then set $\lambda = 1$, with probability at least $1-\delta$, we get

$$
\begin{aligned}
&\left|\boldsymbol{\phi}(s,a)^\top\widehat{\mathbf{w}}^k_{m,h} - r_{m,h}(s,a) - \mathbb{P}_{m,h}V^k_{m,h+1}(s,a) + (\Delta_{m,1}+\Delta_{m,2})\right| \\
&\leq \left(5H\sqrt{d}C_\delta + 3H\zeta\sqrt{MKd}\right)\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}}.
\end{aligned}
$$

Bounding Term I_2 in (H.9): same as the proof in Appendix G.8, we have

$$
\left|\boldsymbol{\phi}(s,a)^\top\widehat{\mathbf{w}}^k_{m,h} - \boldsymbol{\phi}(s,a)^\top\boldsymbol{\mu}^{k,J_k}_{m,h}\right| \leq \frac{4}{3}\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}}.
$$

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

$$
\left|r_{m,h}(s,a) + \mathbb{P}_{m,h}V^k_{m,h+1}(s,a) - \boldsymbol{\phi}(s,a)^\top\boldsymbol{\mu}^{k,J_k}_{m,h}\right| \leq \left(5H\sqrt{d}C_\delta + 3H\zeta\sqrt{MKd} + \frac{4}{3}\right)\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}}. \qquad \text{(H.17)}
$$

Consider the denominator of Z_k: same as the proof in Appendix G.8, with (G.18), we have

$$
\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Sigma}^{k,J_k}_{m,h}} \geq \frac{1}{4\sqrt{\beta_K}}\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}}, \qquad \text{(H.18)}
$$

where we used the fact that $\lambda_{\min}\left(\left(\boldsymbol{\Lambda}^k_{m,h}\right)^{-1}\right) \geq 1/k$ and $\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}} \geq 1/\sqrt{k}\|\boldsymbol{\phi}(s,a)\|_2$. Therefore, according to (H.17) and (H.18), with probability at least $1-\delta$, it holds that

$$
\begin{aligned}
|Z_k| &= \left|\frac{r_{m,h}(s,a) + \mathbb{P}_{m,h}V^k_{m,h+1}(s,a) - \boldsymbol{\phi}(s,a)^\top\boldsymbol{\mu}^{k,J_k}_{m,h}}{\sqrt{\boldsymbol{\phi}(s,a)^\top\boldsymbol{\Sigma}^{k,J_k}_{m,h}\boldsymbol{\phi}(s,a)}}\right| \\
&\leq \frac{\left(5H\sqrt{d}C_\delta + 3H\zeta\sqrt{MKd} + \frac{4}{3}\right)\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}}}{\frac{1}{4\sqrt{\beta_K}}\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}}} \\
&= \frac{5H\sqrt{d}C_\delta + 3H\zeta\sqrt{MKd} + \frac{4}{3}}{\frac{1}{4\sqrt{\beta_K}}},
\end{aligned}
$$

which implies $|Z_k| < 1$ when $\frac{1}{\sqrt{\beta_K}} = 20H\sqrt{d}C_\delta + 12H\zeta\sqrt{MKd} + \frac{16}{3}$.

Now we have already proved that, for any fixed $n \in [N]$, with probability at least $1-\delta$, we have

$$\mathbb{P}\Big(r_{m,h}(s,a) + \mathbb{P}_{m,h}V^k_{m,h+1}(s,a) - \boldsymbol{\phi}(s,a)^\top \mathbf{w}^{k,J_k,n}_{m,h} \le (\Delta_{m,1}+\Delta_{m,2})\Big) \ge \frac{1}{2\sqrt{2e\pi}}.$$

By taking union bound over $n \in [N]$, with probablity at least $1-\delta$, we have

$$\begin{aligned}
&\mathbb{P}\Big(\max_{n\in[N]} \big\{\boldsymbol{\phi}(s,a)^\top \mathbf{w}^{k,J_k,n}_{m,h} - r_{m,h}(s,a) - \mathbb{P}_{m,h}V^k_{m,h+1}(s,a)\big\} \ge -(\Delta_{m,1}+\Delta_{m,2})\Big)\\
&\ge 1 - \Big(1 - \frac{1}{2\sqrt{2e\pi}}\Big)^N\\
&= 1 - {c_0'}^N,
\end{aligned}$$

where $c_0' = 1 - \frac{1}{2\sqrt{2e\pi}}$. Finally, with probability at least $(1-\delta)\big(1-{c_0'}^N\big)$, for all $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have

$$l^k_{m,h}(s,a) \le 3H\zeta.$$

Till now we have completed the proof of fixed $\boldsymbol{\phi} \in \mathcal{C}(\varepsilon)$. Follow the proof in Appendix G.8, we can get the final result. □

H.2 Regret Analysis

In this part, we give out the proof of Theorem E.5, the regret bound for CoopTS-LMC in the misspecified setting.

Proof of Theorem E.5. This proof is almost same as the proof in Appendix F.2. We do the same regret decomposition (F.2) and obtain the same bound for **Term(i)** (F.3) and **Term(ii)** (F.4). Next we bound **Term (iii)** with new lemmas in the misspecified setting.

Bounding Term (iii) in (F.2): based on Lemma H.3 and Lemma H.4, by taking union bound, with probability at least $1 - |\mathcal{C}(\varepsilon)|{c_0'}^N - 2\delta' - MHK(\delta' + {\delta'}^2)$, we have

$$\begin{aligned}
&\sum_{m\in\mathcal{M}}\sum_{k=1}^K\sum_{h=1}^H \big(\mathbb{E}_{\pi^*}\big[l^k_{m,h}(s_{m,h},a_{m,h})|s_{m,1}=s^k_{m,1}\big] - l^k_{m,h}\big(s^k_{m,h},a^k_{m,h}\big)\big)\\
&\le \sum_{m\in\mathcal{M}}\sum_{k=1}^K\sum_{h=1}^H \Big(-l^k_{m,h}\big(s^k_{m,h},a^k_{m,h}\big) + \alpha_{\delta'}\varepsilon + 3H\zeta\Big)\\
&\le \sum_{m\in\mathcal{M}}\sum_{k=1}^K\sum_{h=1}^H \Bigg(\Bigg(5H\sqrt{d}C_{\delta'} + 3H\zeta\sqrt{MKd} + 5\sqrt{\frac{2d\log\big(\sqrt{N}/\delta'\big)}{3\beta_K}} + \frac{4}{3}\Bigg)\big\|\boldsymbol{\phi}(s^k_{m,h},a^k_{m,h})\big\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}}\\
&\qquad + \alpha_{\delta'}\varepsilon + 6H\zeta\Bigg)\\
&= HMK\alpha_{\delta'}\varepsilon + 6H^2MK\zeta + \Bigg(5H\sqrt{d}C_{\delta'} + 3H\zeta\sqrt{MKd} + 5\sqrt{\frac{2d\log\big(\sqrt{N}/\delta'\big)}{3\beta_K}} + \frac{4}{3}\Bigg)\\
&\qquad \times \sum_{h=1}^H\sum_{m\in\mathcal{M}}\sum_{k=1}^K \big\|\boldsymbol{\phi}(s^k_{m,h},a^k_{m,h})\big\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}}\\
&\le HMK\alpha_{\delta'}\varepsilon + 6H^2MK\zeta + \Bigg(5H\sqrt{d}C_{\delta'} + 3H\zeta\sqrt{MKd} + 5\sqrt{\frac{2d\log\big(\sqrt{N}/\delta'\big)}{3\beta_K}} + \frac{4}{3}\Bigg)\\
&\qquad \times \sum_{h=1}^H\Bigg(\log\Bigg(\frac{\det(\boldsymbol{\Lambda}^K_h)}{\det(\lambda\mathbf{I})}\Bigg) + 1\Bigg)M\sqrt{\gamma} + 2\sqrt{MK\log\Bigg(\frac{\det(\boldsymbol{\Lambda}^K_h)}{\det(\lambda\mathbf{I})}\Bigg)}
\end{aligned}$$

$$
\begin{aligned}
&\leq HMK\alpha_{\delta'}\varepsilon + 6H^2MK\zeta + \left(5H\sqrt{d}C_{\delta'} + 3H\zeta\sqrt{MKd} + 5\sqrt{\frac{2d\log\left(\sqrt{N}/\delta'\right)}{3\beta_K}} + \frac{4}{3}\right) \\
&\qquad \times H\Big(d(\log(1+MK/d)+1)M\sqrt{\gamma} + 2\sqrt{MKd\log(1+MK/d)}\Big) \\
&= \widetilde{\mathcal{O}}\Big(d^{\frac{3}{2}}H^2\sqrt{M}\big(\sqrt{dM\gamma}+\sqrt{K}\big) + d^{\frac{3}{2}}H^2M\sqrt{K}\big(\sqrt{dM\gamma}+\sqrt{K}\big)\zeta\Big). \qquad \text{(H.19)}
\end{aligned}
$$

The first inequality follows from Lemma H.4, the second inequality follows from Lemma H.3, the third inequality follows from Lemma F.12, the last inequality holds due to Lemma L.2 and the fact that $\|\boldsymbol{\phi}(\cdot)\|_2 \leq 1$, the last equality follows from $\frac{1}{\sqrt{\beta_K}} = 20H\sqrt{d}C_{\delta'} + 12H\zeta\sqrt{MKd} + \frac{16}{3}$, which we define in Lemma H.4.

The probability calculation is same as that in Appendix F.2. By combining **Terms (i)(ii)(iii)** together, we get that the final regret bound for CoopTS-LMC in misspecified setting is

$$
\widetilde{\mathcal{O}}\Big(d^{\frac{3}{2}}H^2\sqrt{M}\big(\sqrt{dM\gamma}+\sqrt{K}\big) + d^{\frac{3}{2}}H^2M\sqrt{K}\big(\sqrt{dM\gamma}+\sqrt{K}\big)\zeta\Big),
$$

with probability at least $1-\delta$. Here we finish the proof. □

I Proof of the Regret Bound for CoopTS-PHE

Before getting the regret bound for CoopTS-PHE, we first present some essential technical lemmas required for our analysis.

I.1 Supporting Lemmas

Proposition I.1. The difference between the perturbed estimated parameter $\widetilde{\mathbf{w}}_{m,h}^{k,n}$ and unperturbed estimated parameter $\widehat{\mathbf{w}}_{m,h}^{k}$ satisfies the Gaussian distribution,

$$
\boldsymbol{\zeta}_{m,h}^{k,n} = \widetilde{\mathbf{w}}_{m,h}^{k,n} - \widehat{\mathbf{w}}_{m,h}^{k} \sim \mathcal{N}\Big(\mathbf{0}, \sigma^2\big(\boldsymbol{\Lambda}_{m,h}^{k}\big)^{-1}\Big),
$$

where $\widehat{\mathbf{w}}_{m,h}^{k} = \big(\boldsymbol{\Lambda}_{m,h}^{k}\big)^{-1}\Big(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)}\big[r_h + V_{m,h+1}^{k}\big(s'^l\big)\big]\boldsymbol{\phi}\big(s^l,a^l\big)\Big)$ is the unperturbed estimated parameter.

Next we will define some good events that hold with high probability to help prove the critical lemmas in this section.

Lemma I.2 (Good events)**.** For any fixed $0 < \delta < 1$, with some constant $c > 0$, we define the following random events

$$
\begin{aligned}
\mathcal{G}_{m,h}^{k}(\boldsymbol{\zeta},\delta) &\stackrel{\text{def}}{=} \Big\{\max_{n\in[N]}\big\|\boldsymbol{\zeta}_{m,h}^{k,n}\big\|_{\boldsymbol{\Lambda}_{m,h}^{k}} \leq c_1\sigma\sqrt{d}\Big\}, \\
\mathcal{G}(M,K,H,\delta) &\stackrel{\text{def}}{=} \bigcap_{m\in\mathcal{M}}\bigcap_{k\leq K}\bigcap_{h\leq H}\mathcal{G}_{m,h}^{k}(\boldsymbol{\zeta},\delta),
\end{aligned}
$$

where $c_1 = c\sqrt{\log(dNMKH/\delta)}$. Then the event $\mathcal{G}(M,K,H,\delta)$ occurs with probability at least $1-\delta$.

Lemma I.3. Let $\lambda = 1$ in Algorithm 2. For any fixed $0 < \delta < 1$, conditioned on the event $\mathcal{G}(M,K,H,\delta)$, with probability $1-\delta$, for all $(m,k,h) \in \mathcal{M}\times[K]\times[H]$, we have

$$
\Bigg\|\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)}\boldsymbol{\phi}\big(s^l,a^l\big)\big[\big(V_{m,h+1}^{k} - \mathbb{P}_hV_{m,h+1}^{k}\big)\big(s^l,a^l\big)\big]\Bigg\|_{(\boldsymbol{\Lambda}_{m,h}^{k})^{-1}} \leq 3H\sqrt{d}D_\delta,
$$

where we define $D_\delta = \Big[\frac{1}{2}\log(K+1) + \log\Big(\frac{6\sqrt{2}K(2H\sqrt{MKd}+c_1\sigma\sqrt{d})}{H}\Big) + \log\frac{1}{\delta}\Big]^{1/2}$.

Lemma I.4. Let $\lambda = 1$ in Algorithm 2. Under Definition 4.1, for any fixed $0 < \delta < 1$, conditioned on the event $\mathcal{G}(M, K, H, \delta)$, with probability $1 - \delta$, for all $(m, k, h) \in \mathcal{M} \times [K] \times [H]$ and for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\left|\boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}_{m,h}^k - r_h(s,a) - \mathbb{P}_h V_{m,h+1}^k(s,a)\right| \leq 5H\sqrt{d}D_\delta \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}. \tag{I.1}$$

Lemma I.5 (Optimism)**.** Let $\lambda = 1$ in Algorithm 2 and set $c_0 = \Phi(1)$. Under Definition 4.1, conditioned on the event $\mathcal{G}(M, K, H, \delta)$, with probability at least $1 - |\mathcal{C}(\varepsilon)|c_0^N - \delta$ where $|\mathcal{C}(\varepsilon)| \leq (3/\varepsilon)^d$, for all $(m, k, h) \in \mathcal{M} \times [K] \times [H]$ and for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have

$$l_{m,h}^k(s,a) \leq A_\delta \varepsilon,$$

where $A_\delta = c_1\sigma\sqrt{d} + 5H\sqrt{d}D_\delta = \widetilde{\mathcal{O}}(Hd)$.

Lemma I.6 (Error bound)**.** Let $\lambda = 1$ in Algorithm 2. Under Definition 4.1, for any fixed $0 < \delta < 1$, conditioned on the event $\mathcal{G}(M, K, H, \delta)$, with probability $1 - \delta$, for all $(m, k, h) \in \mathcal{M} \times [K] \times [H]$ and for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have

$$-l_{m,h}^k(s,a) \leq c_2 Hd \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}},$$

where $c_2 = \widetilde{\mathcal{O}}(1)$.

I.2 Regret Analysis

In this part, we give out the proof of Theorem 4.2, the regret bound for CoopTS-PHE.

Proof of Theorem 4.2. Based on the result from Lemma F.13, we do the regret decomposition first

$$\begin{aligned}
\text{Regret}(K) &= \sum_{m\in\mathcal{M}} \sum_{k=1}^K V_{m,1}^*\left(s_{m,1}^k\right) - V_{m,1}^{\pi_{m,k}}\left(s_{m,1}^k\right) \\
&= \underbrace{\sum_{m\in\mathcal{M}} \sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\left\langle Q_{m,h}^k(s_{m,h},\cdot), \pi_{m,h}^*(\cdot|s_{m,h}) - \pi_{m,h}^k(\cdot|s_{m,h})\right\rangle | s_{m,1} = s_{m,1}^k\right]}_{(i)} \\
&\quad + \underbrace{\sum_{m\in\mathcal{M}} \sum_{k=1}^K \sum_{h=1}^H (D_{m,k,h,1} + D_{m,k,h,2})}_{(ii)} \\
&\quad + \underbrace{\sum_{m\in\mathcal{M}} \sum_{k=1}^K \sum_{h=1}^H \left(\mathbb{E}_{\pi^*}\left[l_{m,h}^k(s_{m,h}, a_{m,h}) | s_{m,1} = s_{m,1}^k\right] - l_{m,h}^k\left(s_{m,h}^k, a_{m,h}^k\right)\right)}_{(iii)}.
\end{aligned} \tag{I.2}$$

Next, we will bound the above three terms respectively.

Bounding Term (i) in (I.2): for the policy $\pi_{m,h}^k$, we have

$$\sum_{m\in\mathcal{M}} \sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\left\langle Q_{m,h}^k(s_{m,h},\cdot), \pi_{m,h}^*(\cdot|s_{m,h}) - \pi_{m,h}^k(\cdot|s_{m,h})\right\rangle | s_{m,1} = s_{m,1}^k\right] \leq 0. \tag{I.3}$$

This is because by definition $\pi_{m,h}^k$ is the greedy policy for $Q_{m,h}^k$.

Bounding Term (ii) in (I.2): note that $0 \leq Q_{m,h}^k \leq H - h + 1 \leq H$, based on (F.1), for any $(m, k, h) \in \mathcal{M} \times [K] \times [H]$, we have $|D_{m,k,h,1}| \leq 2H$ and $|D_{m,k,h,2}| \leq 2H$. Note that $D_{m,k,h,1}$ is

a martingale difference sequence $\mathbb{E}[D_{m,k,h,1}|\mathcal{F}_{m,k,h}] = 0$. By applying Azuma-Hoeffding inequality, with probability at least $1-\delta/3$, we have

$$\sum_{m\in\mathcal{M}}\sum_{k=1}^{K}\sum_{h=1}^{H} D_{m,k,h,1} \le 2\sqrt{2MH^3K\log(6/\delta)}.$$

Note that $D_{m,k,h,2}$ is also a martingale difference sequence. By applying Azuma-Hoeffding inequality, with probability at least $1-\delta/3$, we have

$$\sum_{m\in\mathcal{M}}\sum_{k=1}^{K}\sum_{h=1}^{H} D_{m,k,h,2} \le 2\sqrt{2MH^3K\log(6/\delta)}.$$

By taking union bound, with probability at least $1-2\delta/3$, we have

$$\sum_{m\in\mathcal{M}}\sum_{k=1}^{K}\sum_{h=1}^{H} D_{m,k,h,1} + \sum_{m\in\mathcal{M}}\sum_{k=1}^{K}\sum_{h=1}^{H} D_{m,k,h,2} \le 4\sqrt{2MH^3K\log(6/\delta)}. \tag{I.4}$$

Bounding Term (iii) in (I.2): conditioned on the event $\mathcal{G}(M,K,H,\delta')$, based on Lemma I.6 and Lemma I.5, by taking union bound, with probability at least $1-|\mathcal{C}(\varepsilon)|c_0^N-\delta'-MHK\delta'$, we have

$$\begin{aligned}
&\sum_{m\in\mathcal{M}}\sum_{k=1}^{K}\sum_{h=1}^{H}\big(\mathbb{E}_{\pi^*}\big[l_{m,h}^k(s_{m,h},a_{m,h})|s_{m,1}=s_{m,1}^k\big]-l_{m,h}^k\big(s_{m,h}^k,a_{m,h}^k\big)\big)\\
&\le \sum_{m\in\mathcal{M}}\sum_{k=1}^{K}\sum_{h=1}^{H}\big(A_{\delta'}\varepsilon-l_{m,h}^k\big(s_{m,h}^k,a_{m,h}^k\big)\big)\\
&\le HMKA_{\delta'}\varepsilon+\sum_{m\in\mathcal{M}}\sum_{k=1}^{K}\sum_{h=1}^{H}c_2dH\big\|\boldsymbol{\phi}(s_{m,h}^k,a_{m,h}^k)\big\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}\\
&\le HMKA_{\delta'}\varepsilon+c_2dH\sum_{h=1}^{H}\left(\log\left(\frac{\det(\boldsymbol{\Lambda}_h^K)}{\det(\lambda\mathbf{I})}\right)+1\right)M\sqrt{\gamma}+2\sqrt{MK\log\left(\frac{\det(\boldsymbol{\Lambda}_h^K)}{\det(\lambda\mathbf{I})}\right)}\\
&\le HMKA_{\delta'}\varepsilon+c_2dH\cdot H\Big(d(\log(1+MK/d)+1)M\sqrt{\gamma}+2\sqrt{MKd\log(1+MK/d)}\Big).
\end{aligned}$$

The first inequality follows from Lemma I.5, the second inequality holds due to Lemma I.6, the third inequality follows from Lemma F.12, the last inequality holds due to Lemma L.2 and the fact that $\|\boldsymbol{\phi}(\cdot)\|_2 \le 1$.

Here we choose $\varepsilon = dH\sqrt{d/MK}/A_{\delta'} = \widetilde{\mathcal{O}}(\sqrt{d/MK})$. Conditioned on the event $\mathcal{G}(M,K,H,\delta')$, we have

$$\sum_{m\in\mathcal{M}}\sum_{k=1}^{K}\sum_{h=1}^{H}\big(\mathbb{E}_{\pi^*}\big[l_{m,h}^k(s_{m,h},a_{m,h})|s_{m,1}=s_{m,1}^k\big]-l_{m,h}^k\big(s_{m,h}^k,a_{m,h}^k\big)\big) \le \widetilde{\mathcal{O}}\big(dH^2\big(dM\sqrt{\gamma}+\sqrt{dMK}\big)\big), \tag{I.5}$$

with probability at least $1-|\mathcal{C}(\varepsilon)|c_0^N-\delta'-MHK\delta'$. Based on Lemma I.2, the event $\mathcal{G}(M,K,H,\delta')$ occurs with probability at least $1-\delta'$. Therefore, (I.5) occurs with probability at least

$$\big(1-\delta'\big)\big(1-|\mathcal{C}(\varepsilon)|c_0^N-\delta'-MHK\delta'\big).$$

We set $\delta' = \delta/6(MHK+2)$ and choose $N = \widetilde{C}\log(\delta)/\log(c_0)$ where $\widetilde{C} = \widetilde{\mathcal{O}}(d)$, then we have

$$\big(1-\delta'\big)\big(1-|\mathcal{C}(\varepsilon)|c_0^N-\delta'-MHK\delta'\big) \ge 1-\delta/3.$$

Combining Terms (i)(ii)(iii) together: Based on (I.3), (I.4) and (I.5). By taking union bound, we get that the final regret bound for CoopTS-PHE is $\widetilde{\mathcal{O}}\big(dH^2\big(dM\sqrt{\gamma}+\sqrt{dMK}\big)\big)$ with probability at least $1-\delta$. □

J Proof of Supporting Lemmas in Appendix I

J.1 Proof of Proposition I.1

Proof. Based on (3.5), we can calculate that

$$\begin{aligned}\widetilde{\mathbf{w}}_{m,h}^{k,n} &= \big(\mathbf{\Lambda}_{m,h}^{k}\big)^{-1}\Bigg(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)}\big[\big(r_h\big(s^l,a^l\big)+\epsilon_h^{k,l,n}\big)+V_{m,h+1}^{k}\big(s'^l\big)\big]\boldsymbol{\phi}\big(s^l,a^l\big)-\lambda\boldsymbol{\xi}_h^{k,n}\Bigg)\\ &= \widehat{\mathbf{w}}_{m,h}^{k}+\big(\mathbf{\Lambda}_{m,h}^{k}\big)^{-1}\Bigg(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)}\epsilon_h^{k,l,n}\boldsymbol{\phi}\big(s^l,a^l\big)-\lambda\boldsymbol{\xi}_h^{k,n}\Bigg),\end{aligned}\tag{J.1}$$

where $\widehat{\mathbf{w}}_{m,h}^{k} = \big(\mathbf{\Lambda}_{m,h}^{k}\big)^{-1}\Big(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)}\big[r_h+V_{m,h+1}^{k}\big(s'^l\big)\big]\boldsymbol{\phi}\big(s^l,a^l\big)\Big)$ is the unperturbed estimated parameter. Since $\epsilon_h^{k,l,n}\sim\mathcal{N}(0,\sigma^2)$, for $l\in[\mathcal{K}(k)]$, based on the property of Gaussian distribution, we have

$$\epsilon_h^{k,l,n}\boldsymbol{\phi}\big(s^l,a^l\big)\sim\mathcal{N}\Big(0,\sigma^2\boldsymbol{\phi}\big(s^l,a^l\big)\boldsymbol{\phi}\big(s^l,a^l\big)^\top\Big),$$

Since $\boldsymbol{\xi}_h^{k,n}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})$, we can calculate the covariance matrix of the second term in (J.1),

$$\begin{aligned}&\big(\mathbf{\Lambda}_{m,h}^{k}\big)^{-1}\operatorname{Cov}\Bigg(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)}\epsilon_h^{k,l,n}\boldsymbol{\phi}\big(s^l,a^l\big)-\lambda\boldsymbol{\xi}_h^{k,n}\Bigg)\big(\mathbf{\Lambda}_{m,h}^{k}\big)^{-1}\\ &=\big(\mathbf{\Lambda}_{m,h}^{k}\big)^{-1}\sigma^2\Bigg(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)}\boldsymbol{\phi}\big(s^l,a^l\big)\boldsymbol{\phi}\big(s^l,a^l\big)^\top+\lambda\mathbf{I}\Bigg)\big(\mathbf{\Lambda}_{m,h}^{k}\big)^{-1}\\ &=\sigma^2\big(\mathbf{\Lambda}_{m,h}^{k}\big)^{-1}\mathbf{\Lambda}_{m,h}^{k}\big(\mathbf{\Lambda}_{m,h}^{k}\big)^{-1}\\ &=\sigma^2\big(\mathbf{\Lambda}_{m,h}^{k}\big)^{-1}.\end{aligned}$$

It is obvious that the mean of the second term in (J.1) is $\mathbf{0}$. Thus, we have

$$\boldsymbol{\zeta}_{m,h}^{k,n}=\widetilde{\mathbf{w}}_{m,h}^{k,n}-\widehat{\mathbf{w}}_{m,h}^{k}\sim\mathcal{N}\Big(0,\sigma^2\big(\mathbf{\Lambda}_{m,h}^{k}\big)^{-1}\Big).$$

This completes the proof. □

J.2 Proof of Lemma I.2

Proof. Recall that in Proposition I.1, we have

$$\big\{\boldsymbol{\zeta}_{m,h}^{k,n}\big\}\sim\mathcal{N}\Big(\mathbf{0},\sigma^2\big(\mathbf{\Lambda}_{m,h}^{k}\big)^{-1}\Big).$$

By Lemma L.10, for fixed $n\in[N]$, with probability at least $1-\delta$, we have

$$\big\|\boldsymbol{\zeta}_{m,h}^{k,n}\big\|_{\mathbf{\Lambda}_{m,h}^{k}}\le c\sqrt{d\sigma^2\log(d/\delta)}.$$

By applying union bound over N samples, we have

$$\mathbb{P}\Big(\max_{n\in[N]}\big\|\boldsymbol{\zeta}_{m,h}^{k,n}\big\|_{\mathbf{\Lambda}_{m,h}^{k}}\le c\sqrt{d\sigma^2\log(d/\delta)}\Big)\ge 1-N\delta.$$

Now we define $c_1=c\sqrt{\log(dNMKH/\delta)}$, and we define the event

$$\mathcal{G}_{m,h}^{k}(\boldsymbol{\zeta},\delta)\overset{\text{def}}{=}\Big\{\max_{n\in[N]}\big\|\boldsymbol{\zeta}_{m,h}^{k,n}\big\|_{\mathbf{\Lambda}_{m,h}^{k}}\le c_1\sigma\sqrt{d}\Big\}.$$

Thus for any fixed m, h and k, the event $\mathcal{G}^k_{m,h}(\boldsymbol{\zeta},\delta)$ occurs with a probability of at least $1-\delta/MKH$. By taking union bound over all $(m,h,k)\in\mathcal{M}\times[H]\times[K]$, we have

$$\mathbb{P}\big(\mathcal{G}(M,K,H,\delta)\big)=\mathbb{P}\left(\bigcap_{m\in\mathcal{M}}\bigcap_{k\le K}\bigcap_{h\le H}\mathcal{G}^k_{m,h}(\boldsymbol{\zeta},\delta)\right)\ge 1-\delta.$$

This completes the proof. □

J.3 Proof of Lemma I.3

Proof. Based on the result in Lemma F.4, for any $(m,h,k)\in\mathcal{M}\times[H]\times[K]$, we have

$$\big\|\widehat{\mathbf{w}}^k_{m,h}\big\|\le 2H\sqrt{Mkd/\lambda}.$$

By recalling the construction of $\boldsymbol{\Lambda}^k_{m,h}$, it is trivial to find that $\lambda_{\min}\big(\boldsymbol{\Lambda}^k_{m,h}\big)\ge\lambda$. Conditioned on the event $\mathcal{G}(M,K,H,\delta)$, we have

$$\sqrt{\lambda}\big\|\boldsymbol{\zeta}^{k,n}_{m,h}\big\|\le\big\|\boldsymbol{\zeta}^{k,n}_{m,h}\big\|_{\boldsymbol{\Lambda}^k_{m,h}}\le c_1\sigma\sqrt{d}.$$

Then by triangle inequality, for all $n\in[N]$, we obtain the upper bound

$$\big\|\widetilde{\mathbf{w}}^{k,n}_{m,h}\big\|=\big\|\widehat{\mathbf{w}}^k_{m,h}+\boldsymbol{\zeta}^{k,n}_{m,h}\big\|\le 2H\sqrt{Mkd/\lambda}+c_1\sigma\sqrt{d/\lambda}.$$

Based on the result from Lemma L.7 and Lemma L.9, we have that, for any $\varepsilon>0$, and for all $(m,k,h)\in\mathcal{M}\times[K]\times[H]$, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\left\|\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)}\boldsymbol{\phi}\big(s^l,a^l\big)\big[\big(V^k_{m,h+1}-\mathbb{P}_hV^k_{m,h+1}\big)\big(s^l,a^l\big)\big]\right\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}}\\
&\le\left(4H^2\left[\frac{d}{2}\log\left(\frac{k+\lambda}{\lambda}\right)+d\log\left(\frac{3\big(2H\sqrt{Mkd/\lambda}+c_1\sigma\sqrt{d/\lambda}\big)}{\varepsilon}\right)+\log\frac{1}{\delta}\right]+\frac{8k^2\varepsilon^2}{\lambda}\right)^{1/2}\\
&\le 2H\left[\frac{d}{2}\log\left(\frac{k+\lambda}{\lambda}\right)+d\log\left(\frac{3\big(2H\sqrt{Mkd/\lambda}+c_1\sigma\sqrt{d/\lambda}\big)}{\varepsilon}\right)+\log\frac{1}{\delta}\right]^{1/2}+\frac{2\sqrt{2}k\varepsilon}{\sqrt{\lambda}}.
\end{aligned}$$

Here we set $\lambda=1,\varepsilon=\frac{H}{2\sqrt{2}k}$, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\left\|\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)}\boldsymbol{\phi}\big(s^l,a^l\big)\big[\big(V^k_{m,h+1}-\mathbb{P}_hV^k_{m,h+1}\big)\big(s^l,a^l\big)\big]\right\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}}\\
&\le 2H\sqrt{d}\left[\frac{1}{2}\log(K+1)+\log\left(\frac{6\sqrt{2}K\big(2H\sqrt{MKd}+c_1\sigma\sqrt{d}\big)}{H}\right)+\log\frac{1}{\delta}\right]^{1/2}+H\\
&\le 3H\sqrt{d}D_\delta,
\end{aligned}$$

where we define $D_\delta=\left[\frac{1}{2}\log(K+1)+\log\left(\frac{6\sqrt{2}K(2H\sqrt{MKd}+c_1\sigma\sqrt{d})}{H}\right)+\log\frac{1}{\delta}\right]^{1/2}$. Here we finish the proof. □

J.4 Proof of Lemma I.4

Proof. This proof is almost same as the proof of Lemma F.8 in Appendix G.6. The only difference is the **Term (i)** in (G.10). Here based on Lemma I.6, conditioned on the event $\mathcal{G}(M,K,H,\delta)$, with probability $1-\delta$, we have

$$\textbf{Term (i)}\le 3H\sqrt{d}D_\delta\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}}.$$

Finally, conditioned on the event $\mathcal{G}(M, K, H, \delta)$, with probability $1-\delta$, for all $(m, k, h) \in \mathcal{M} \times [K] \times [H]$ and for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\begin{aligned}\left|\boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}_{m,h}^k - r_h(s,a) - \mathbb{P}_h V_{m,h+1}^k(s,a)\right| &\leq \big(3H\sqrt{d}D_\delta + H\sqrt{d} + \sqrt{d}\big)\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} \\ &\leq 5H\sqrt{d}D_\delta \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,k}^k)^{-1}}.\end{aligned}$$

Here we finish the proof. □

J.5 Proof of Lemma I.5

Proof. Recall from Definition 4.1, we have

$$r_h(s,a) + \mathbb{P}_h V_{m,h+1}^k(s,a) = \boldsymbol{\phi}(s,a)^\top \boldsymbol{\theta}_h + \boldsymbol{\phi}(s,a)^\top \big\langle \boldsymbol{\mu}_h, V_{m,h+1}^k \big\rangle_{\mathcal{S}} \stackrel{\text{def}}{=} \boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^k,$$

where $\mathbf{w}_{m,h}^k = \boldsymbol{\theta}_h + \big\langle \boldsymbol{\mu}_h, V_{m,h+1}^k \big\rangle_{\mathcal{S}}$. Note that $\max\{\|\boldsymbol{\mu}_h(\mathcal{S})\|, \|\boldsymbol{\theta}_h\|\} \leq \sqrt{d}$ and $V_{m,h+1}^k \leq H - h \leq H$, thus we have

$$\begin{aligned}\big\|\mathbf{w}_{m,h}^k\big\| &\leq \|\boldsymbol{\theta}_h\| + \big\|\big\langle \boldsymbol{\mu}_h, V_{m,h+1}^k \big\rangle_{\mathcal{S}}\big\| \\ &\leq \sqrt{d} + H\sqrt{d} \\ &\leq 2H\sqrt{d}.\end{aligned}$$

Then we define the regression error $\Delta \mathbf{w}_{m,h}^k = \mathbf{w}_{m,h}^k - \widehat{\mathbf{w}}_{m,h}^k$. For any $(m, h, k) \in \mathcal{M} \times [H] \times [K]$ and any $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\begin{aligned}l_{m,h}^k(s,a) &= r_h(s,a) + \mathbb{P}_h V_{m,h+1}^k(s,a) - Q_{m,h}^k(s,a) \\ &= r_h(s,a) + \mathbb{P}_h V_{m,h+1}^k(s,a) - \min\Big\{H - h + 1, \max_{n\in[N]} \boldsymbol{\phi}(s,a)^\top \Big(\widehat{\mathbf{w}}_{m,h}^k + \boldsymbol{\zeta}_{m,h}^{k,n}\Big)\Big\}^+ \\ &\leq \max\Big\{\boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^k - (H-h+1), \boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^k - \max_{n\in[N]} \boldsymbol{\phi}(s,a)^\top \Big(\widehat{\mathbf{w}}_{m,h}^k + \boldsymbol{\zeta}_{m,h}^{k,n}\Big)\Big\} \\ &\leq \max\Big\{0, \boldsymbol{\phi}(s,a)^\top \Delta \mathbf{w}_{m,h}^k - \max_{n\in[N]} \boldsymbol{\phi}(s,a)^\top \boldsymbol{\zeta}_{m,h}^{k,n}\Big\}, \qquad \text{(J.2)}\end{aligned}$$

where the last inequality holds because $|r_h| \leq 1$ and $V_{m,h+1}^k \leq H - h$, this indicates $r_h(s,a) + \mathbb{P}_h V_{m,h+1}^k(s,a) = \boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^k \leq H - h + 1$. Note that $\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} \leq \sqrt{1/\lambda}\|\boldsymbol{\phi}(s,a)\| \leq 1$ for all $\boldsymbol{\phi}(s,a)$. Define $\mathcal{C}(\varepsilon)$ to be a ε-cover of $\big\{\boldsymbol{\phi} \mid \|\boldsymbol{\phi}\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} \leq 1\big\}$. Based on Lemma L.8, we have $|\mathcal{C}(\varepsilon)| \leq (3/\varepsilon)^d$.

First, for any fixed $\boldsymbol{\phi}(s,a) \in \mathcal{C}(\varepsilon)$, we have

$$\big\{\boldsymbol{\phi}^\top \boldsymbol{\zeta}_{m,h}^{k,n}\big\} \sim \mathcal{N}\Big(\mathbf{0}, \sigma^2 \|\boldsymbol{\phi}\|^2_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}\Big).$$

Use the property of Gaussian distribution, we obtain

$$\mathbb{P}\Big(\boldsymbol{\phi}^\top \boldsymbol{\zeta}_{m,h}^{k,n} - \sigma\|\boldsymbol{\phi}\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} \geq 0\Big) = \Phi(-1).$$

By taking union bound over $n \in [N]$, we obtain

$$\mathbb{P}\Big(\max_{n\in[N]} \big\{\boldsymbol{\phi}^\top \boldsymbol{\zeta}_{m,h}^{k,n} - \sigma\|\boldsymbol{\phi}\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}\big\} \geq 0\Big) \geq 1 - (1 - \Phi(-1))^N = 1 - \Phi(1)^N = 1 - c_0^N.$$

By applying union bound over $\mathcal{C}(\varepsilon)$, with probability $1 - |\mathcal{C}(\varepsilon)|c_0^N$, for all $\boldsymbol{\phi} \in \mathcal{C}(\varepsilon)$, we have

$$\max_{n\in[N]} \big\{\boldsymbol{\phi}^\top \boldsymbol{\zeta}_{m,h}^{k,n} - \sigma\|\boldsymbol{\phi}\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}\big\} \geq 0. \qquad \text{(J.3)}$$

Then, for any $\boldsymbol{\phi} = \boldsymbol{\phi}(s,a)$, we can find $\boldsymbol{\phi}' \in \mathcal{C}(\varepsilon)$ such that $\|\boldsymbol{\phi}-\boldsymbol{\phi}'\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}} \le \varepsilon$. Define $\Delta\boldsymbol{\phi} = \boldsymbol{\phi}-\boldsymbol{\phi}'$, we have

$$\begin{aligned}
\boldsymbol{\phi}^\top \boldsymbol{\zeta}^{k,n}_{m,h} - \boldsymbol{\phi}^\top \Delta\mathbf{w}^k_{m,h} &= \boldsymbol{\phi}'^\top \boldsymbol{\zeta}^{k,n}_{m,h} - \boldsymbol{\phi}'^\top \Delta\mathbf{w}^k_{m,h} + \Delta\boldsymbol{\phi}^\top \boldsymbol{\zeta}^{k,n}_{m,h} - \Delta\boldsymbol{\phi}^\top \Delta\mathbf{w}^k_{m,h} \\
&\ge \boldsymbol{\phi}'^\top \boldsymbol{\zeta}^{k,n}_{m,h} - \|\boldsymbol{\phi}'\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}} \big\|\Delta\mathbf{w}^k_{m,h}\big\|_{\boldsymbol{\Lambda}^k_{m,h}} - \|\Delta\boldsymbol{\phi}\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}} \big\|\boldsymbol{\zeta}^{k,n}_{m,h}\big\|_{\boldsymbol{\Lambda}^k_{m,h}} \\
&\quad - \|\Delta\boldsymbol{\phi}\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}} \big\|\Delta\mathbf{w}^k_{m,h}\big\|_{\boldsymbol{\Lambda}^k_{m,h}} \\
&\ge \boldsymbol{\phi}'^\top \boldsymbol{\zeta}^{k,n}_{m,h} - \|\boldsymbol{\phi}'\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}} \big\|\Delta\mathbf{w}^k_{m,h}\big\|_{\boldsymbol{\Lambda}^k_{m,h}} - \varepsilon\big(\big\|\boldsymbol{\zeta}^{k,n}_{m,h}\big\|_{\boldsymbol{\Lambda}^k_{m,h}} + \big\|\Delta\mathbf{w}^k_{m,h}\big\|_{\boldsymbol{\Lambda}^k_{m,h}}\big).
\end{aligned} \tag{J.4}$$

Conditioned on the event $\mathcal{G}(M,K,H,\delta)$, we have

$$\big\|\boldsymbol{\zeta}^{k,n}_{m,h}\big\|_{\boldsymbol{\Lambda}^k_{m,h}} \le c_1\sigma\sqrt{d}.$$

For any vector $\mathbf{x} \in \mathbb{R}^d$, we have

$$\begin{aligned}
\mathbf{x}^\top \Delta\mathbf{w}^k_{m,h} &= \mathbf{x}^\top\big(\mathbf{w}^k_{m,h} - \widehat{\mathbf{w}}^k_{m,h}\big) \\
&= \mathbf{x}^\top\big(\boldsymbol{\Lambda}^k_{m,h}\big)^{-1}\bigg(\boldsymbol{\Lambda}^k_{m,h}\mathbf{w}^k_{m,h} - \bigg(\sum_{(s^l,a^l,s'^l)\in U_{m,h}(k)} \big[r_h + V^k_{m,h+1}\big(s'^l\big)\big]\boldsymbol{\phi}\big(s^l,a^l\big)\bigg)\bigg) \\
&= \mathbf{x}^\top\big(\boldsymbol{\Lambda}^k_{m,h}\big)^{-1}\bigg(\sum_{l=1}^{\mathcal{K}(k)} \boldsymbol{\phi}\big(s^l,a^l\big)\boldsymbol{\phi}\big(s^l,a^l\big)^\top\mathbf{w}^k_{m,h} + \lambda\mathbf{w}^k_{m,h} - \bigg(\sum_{l=1}^{\mathcal{K}(k)} \big[r_h + V^k_{m,h+1}\big(s'^l\big)\big]\boldsymbol{\phi}\big(s^l,a^l\big)\bigg)\bigg) \\
&= \mathbf{x}^\top\big(\boldsymbol{\Lambda}^k_{m,h}\big)^{-1}\bigg(\mathbf{w}^k_{m,h} + \bigg(\sum_{l=1}^{\mathcal{K}(k)} \big[\mathbb{P}_h V^k_{m,h+1} - V^k_{m,h+1}\big(s'^l\big)\big]\boldsymbol{\phi}\big(s^l,a^l\big)\bigg)\bigg),
\end{aligned}$$

where the third equality holds due to the definition of $\boldsymbol{\Lambda}^k_{m,h}$. We set $\mathbf{x} = \boldsymbol{\Lambda}^k_{m,h}\Delta\mathbf{w}^k_{m,h}$. By using Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\big\|\Delta\mathbf{w}^k_{m,h}\big\|^2_{\boldsymbol{\Lambda}^k_{m,h}} &= \Delta\mathbf{w}^k_{m,h}{}^\top\bigg(\mathbf{w}^k_{m,h} + \bigg(\sum_{l=1}^{\mathcal{K}(k)} \big[\mathbb{P}_h V^k_{m,h+1} - V^k_{m,h+1}\big(s'^l\big)\big]\boldsymbol{\phi}\big(s^l,a^l\big)\bigg)\bigg) \\
&\le \big\|\Delta\mathbf{w}^k_{m,h}\big\|_{\boldsymbol{\Lambda}^k_{m,h}} \cdot \bigg\|\mathbf{w}^k_{m,h} + \bigg(\sum_{l=1}^{\mathcal{K}(k)} \big[\mathbb{P}_h V^k_{m,h+1} - V^k_{m,h+1}\big(s'^l\big)\big]\boldsymbol{\phi}\big(s^l,a^l\big)\bigg)\bigg\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}}.
\end{aligned}$$

This indicates that with probability at least $1-\delta$, for all $(m,h,k) \in \mathcal{M}\times[H]\times[K]$, we have

$$\begin{aligned}
\big\|\Delta\mathbf{w}^k_{m,h}\big\|_{\boldsymbol{\Lambda}^k_{m,h}} &\le \big\|\mathbf{w}^k_{m,h}\big\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}} + \bigg\|\sum_{l=1}^{\mathcal{K}(k)} \big[\mathbb{P}_h V^k_{m,h+1} - V^k_{m,h+1}\big(s'^l\big)\big]\boldsymbol{\phi}\big(s^l,a^l\big)\bigg\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}} \\
&\le \big\|\mathbf{w}^k_{m,h}\big\| + 3H\sqrt{d}D_\delta \\
&\le 5H\sqrt{d}D_\delta,
\end{aligned}$$

where the second inequality holds because of Lemma I.3. Then for all $(m,h,k) \in \mathcal{M}\times[H]\times[K]$, with probability at least $1-\delta$, (J.4) becomes

$$\max_{n\in[N]}\big\{\boldsymbol{\phi}^\top\boldsymbol{\zeta}^{k,n}_{m,h} - \boldsymbol{\phi}^\top\Delta\mathbf{w}^k_{m,h}\big\} \ge \max_{n\in[N]}\big\{\boldsymbol{\phi}'^\top\boldsymbol{\zeta}^{k,n}_{m,h} - \|\boldsymbol{\phi}'\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}}\big\|\Delta\mathbf{w}^k_{m,h}\big\|_{\boldsymbol{\Lambda}^k_{m,h}}\big\} - \varepsilon\big(c_1\sigma\sqrt{d} + 5H\sqrt{d}D_\delta\big).$$

Now we choose $\sigma = \widetilde{\mathcal{O}}(H\sqrt{d})$ and guarantee that $\sigma > 5H\sqrt{d}D_\delta \ge \big\|\Delta\mathbf{w}^k_{m,h}\big\|_{\boldsymbol{\Lambda}^k_{m,h}}$, this is achievable through calculation. Define $A_\delta = c_1\sigma\sqrt{d} + 5H\sqrt{d}D_\delta = \widetilde{\mathcal{O}}(Hd)$. Then, for all $(m,h,k) \in \mathcal{M}\times[H]\times[K]$, with probability at least $1-\delta$, we have

$$\max_{n\in[N]}\big\{\boldsymbol{\phi}^\top\boldsymbol{\zeta}^{k,n}_{m,h} - \boldsymbol{\phi}^\top\Delta\mathbf{w}^k_{m,h}\big\} \ge \max_{n\in[N]}\big\{\boldsymbol{\phi}'^\top\boldsymbol{\zeta}^{k,n}_{m,h} - \sigma\|\boldsymbol{\phi}'\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}}\big\} - A_\delta\varepsilon.$$

Recall from (J.3), by taking union bound, with probability at least $1-|\mathcal{C}(\varepsilon)|c_0^N-\delta$, for all $(m,h,k)\in\mathcal{M}\times[H]\times[K]$ and for all $(s,a)\in\mathcal{S}\times\mathcal{A}$, we have

$$\max_{n\in[N]}\left\{\boldsymbol{\phi}^\top\boldsymbol{\zeta}_{m,h}^{k,n}-\boldsymbol{\phi}^\top\Delta\mathbf{w}_{m,h}^k\right\}\geq -A_\delta\varepsilon.$$

Finally, recall from (J.2), we have, with probability at least $1-|\mathcal{C}(\varepsilon)|c_0^N-\delta$, for all $(m,h,k)\in\mathcal{M}\times[H]\times[K]$ and for all $(s,a)\in\mathcal{S}\times\mathcal{A}$, we have

$$l_{m,h}^k(s,a)\leq A_\delta\varepsilon.$$

This completes the proof. □

J.6 Proof of Lemma I.6

Proof. Recall the definition of model prediction error in Definition F.1, we get

$$\begin{aligned}
-l_{m,h}^k(s,a) &= Q_{m,h}^k(s,a)-r_h(s,a)-\mathbb{P}_hV_{m,h+1}^k(s,a)\\
&= \min\left\{\max_{n\in[N]}\boldsymbol{\phi}(s,a)^\top\left(\widehat{\mathbf{w}}_{m,h}^k+\boldsymbol{\zeta}_{m,h}^{k,n}\right),H-h+1\right\}^+-r_h(s,a)-\mathbb{P}_hV_{m,h+1}^k(s,a)\\
&\leq \max_{n\in[N]}\boldsymbol{\phi}(s,a)^\top\left(\widehat{\mathbf{w}}_{m,h}^k+\boldsymbol{\zeta}_{m,h}^{k,n}\right)-r_h(s,a)-\mathbb{P}_hV_{m,h+1}^k(s,a)\\
&= \max_{n\in[N]}\boldsymbol{\phi}(s,a)^\top\boldsymbol{\zeta}_{m,h}^{k,n}-\left(r_h(s,a)+\mathbb{P}_hV_{m,h+1}^k(s,a)-\boldsymbol{\phi}(s,a)^\top\widehat{\mathbf{w}}_{m,h}^k\right)\\
&\leq \left|r_h(s,a)+\mathbb{P}_hV_{m,h+1}^k(s,a)-\boldsymbol{\phi}(s,a)^\top\widehat{\mathbf{w}}_{m,h}^k\right|+\max_{n\in[N]}\left|\boldsymbol{\phi}(s,a)^\top\boldsymbol{\zeta}_{m,h}^{k,n}\right|.
\end{aligned}$$

Based on Lemma I.4, conditioned on the event $\mathcal{G}(M,K,H,\delta)$, with probability $1-\delta$, for all $(m,k,h)\in\mathcal{M}\times[K]\times[H]$ and for any $(s,a)\in\mathcal{S}\times\mathcal{A}$, we have

$$\left|\boldsymbol{\phi}(s,a)^\top\widehat{\mathbf{w}}_{m,h}^k-r_h(s,a)-\mathbb{P}_hV_{m,h+1}^k(s,a)\right|\leq 5H\sqrt{d}D_\delta\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}\tag{J.5}$$

Conditioned on the event $\mathcal{G}(M,K,H,\delta)$, for all $(m,h,k)\in\mathcal{M}\times[H]\times[K]$ and for any $(s,a)\in\mathcal{S}\times\mathcal{A}$, we have

$$\max_{n\in[N]}\left|\boldsymbol{\phi}(s,a)^\top\boldsymbol{\zeta}_{m,h}^{k,n}\right|\leq c_1\sigma\sqrt{d}\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}.\tag{J.6}$$

Combine (J.5) and (J.6), then use σ defined in Lemma I.5. Conditioned on the event $\mathcal{G}(M,K,H,\delta)$, with probability $1-\delta$, for all $(m,h,k)\in\mathcal{M}\times[H]\times[K]$ and for any $(s,a)\in\mathcal{S}\times\mathcal{A}$, we get

$$\begin{aligned}
-l_{m,h}^k(s,a) &\leq \left(5H\sqrt{d}D_\delta+c_1\sigma\sqrt{d}\right)\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}\\
&\leq c_2Hd\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}},
\end{aligned}$$

where $c_2=\widetilde{\mathcal{O}}(1)$. Here we completes the proof. □

K Proof of the Regret Bound for CoopTS-PHE in Misspecified Setting

In this section, we prove the regret bound for CoopTS-PHE in the misspecified setting. The regret analysis, the essential supporting lemmas and their corresponding proofs are very similar to what we have presented in Appendix I and Appendix J. Here we mainly point out the differences of proof between these two settings.

K.1 Supporting Lemmas

Lemma K.1. Let $\lambda = 1$ in Algorithm 2. Under Definition E.1, for any fixed $0 < \delta < 1$, conditioned on the event $\mathcal{G}(M, K, H, \delta)$, with probability $1 - \delta$, for all $(m, k, h) \in \mathcal{M} \times [K] \times [H]$ and for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\left|\boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}_{m,h}^k - r_h(s,a) - \mathbb{P}_h V_{m,h+1}^k(s,a)\right| \leq \left(5H\sqrt{d}D_\delta + 3H\zeta\sqrt{MKd}\right)\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}} + 3H\zeta, \tag{K.1}$$

where D_δ is defined in Lemma I.3.

Proof of Lemma K.1. This proof is almost same as the proof of Lemma H.2, with the only difference in bounding **Term(i)** in (H.2). Here (H.3) becomes

$$|\mathbf{Term(i)}| \leq 3H\sqrt{d}D_\delta \|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}_{m,h}^k)^{-1}}.$$

Finally we can get the desired result. □

Lemma K.2 (Optimism)**.** Let $\lambda = 1$ in Algorithm 2 and set $c_0 = \Phi(1)$. Under Definition E.1, conditioned on the event $\mathcal{G}(M, K, H, \delta)$, with probability at least $1 - |\mathcal{C}(\varepsilon)|c_0^N - \delta$ where $|\mathcal{C}(\varepsilon)| \leq (3/\varepsilon)^d$, for all $(m, k, h) \in \mathcal{M} \times [K] \times [H]$ and for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have

$$l_{m,h}^k \leq A_\delta \varepsilon + 3H\zeta,$$

where $A_\delta = c_1\sigma\sqrt{d} + 5H\sqrt{d}D_\delta = \widetilde{\mathcal{O}}(Hd)$.

Proof of Lemma K.2. This proof is similar to the proof in Appendix J.5. In the previous part, we have defined

$$\begin{aligned}
\Delta_{m,1} &= \mathbb{P}_{m,h} V_{m,h+1}^k(s,a) - \boldsymbol{\phi}(s,a)^\top \langle \boldsymbol{\mu}_h, V_{m,h+1}^k \rangle_{\mathcal{S}}, \\
\Delta_{m,2} &= r_{m,h}(s,a) - \boldsymbol{\phi}(s,a)^\top \boldsymbol{\theta}_h,
\end{aligned}$$

where $|\Delta_{m,1}| \leq 2H\zeta$ and $|\Delta_{m,2}| \leq \zeta$. Thus we have

$$r_{m,h}(s,a) + \mathbb{P}_{m,h} V_{m,h+1}^k(s,a) = \boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^k + \Delta_{m,1} + \Delta_{m,2},$$

where $\mathbf{w}_{m,h}^k = \langle \boldsymbol{\mu}_h, V_{m,h+1}^k \rangle_{\mathcal{S}} + \boldsymbol{\theta}_h$. Then we define $\Delta \mathbf{w}_{m,h}^k = \mathbf{w}_{m,h}^k - \widehat{\mathbf{w}}_{m,h}^k$. For any $(m, h, k) \in \mathcal{M} \times [H] \times [K]$ and any $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\begin{aligned}
l_{m,h}^k(s,a) &= r_{m,h}(s,a) + \mathbb{P}_{m,h} V_{m,h+1}^k(s,a) - Q_{m,h}^k(s,a) \\
&= r_{m,h}(s,a) + \mathbb{P}_{m,h} V_{m,h+1}^k(s,a) - \min\left\{H - h + 1, \max_{n\in[N]} \boldsymbol{\phi}(s,a)^\top \left(\widehat{\mathbf{w}}_{m,h}^k + \boldsymbol{\zeta}_{m,h}^{k,n}\right)\right\}^+ \\
&\leq \max\left\{\boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^k - (H - h + 1), \boldsymbol{\phi}(s,a)^\top \mathbf{w}_{m,h}^k - \max_{n\in[N]} \boldsymbol{\phi}(s,a)^\top \left(\widehat{\mathbf{w}}_{m,h}^k + \boldsymbol{\zeta}_{m,h}^{k,n}\right)\right\} \\
&\qquad + \Delta_{m,1} + \Delta_{m,2} \\
&\leq \max\left\{0, \boldsymbol{\phi}(s,a)^\top \Delta \mathbf{w}_{m,h}^k - \max_{n\in[N]} \boldsymbol{\phi}(s,a)^\top \boldsymbol{\zeta}_{m,h}^{k,n}\right\} + 3H\zeta.
\end{aligned} \tag{K.2}$$

In Appendix J.5, we have proved that with probability at least $1 - |\mathcal{C}(\varepsilon)|c_0^N - \delta$, for all $(m, h, k) \in \mathcal{M} \times [H] \times [K]$ and for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\max_{n\in[N]} \left\{\boldsymbol{\phi}^\top \boldsymbol{\zeta}_{m,h}^{k,n} - \boldsymbol{\phi}^\top \Delta \mathbf{w}_{m,h}^k\right\} \geq -A_\delta \varepsilon.$$

Substitute it into (K.2), we can get the final result. □

Lemma K.3 (Error bound)**.** Let $\lambda = 1$ in Algorithm 2. Under Definition E.1, for any fixed $0 < \delta < 1$, conditioned on the event $\mathcal{G}(M, K, H, \delta)$, with probability $1 - \delta$, for all $(m, k, h) \in \mathcal{M} \times [K] \times [H]$ and for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have

$$-l^k_{m,h}(s,a) \leq \big(c_2 Hd + 3H\zeta\sqrt{MKd}\big)\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}^k_h)^{-1}} + 3H\zeta,$$

where $c_2 = \widetilde{\mathcal{O}}(1)$ is same as that in Lemma I.6.

Proof of Lemma K.3. Similar to the proof in Appendix J.6, using (J.6) in Appendix J.6 and (K.1), we have

$$\begin{aligned}
-l^k_{m,h}(s,a) &\leq \Big|r_h(s,a) + \mathbb{P}_h V^k_{m,h+1}(s,a) - \boldsymbol{\phi}(s,a)^\top \widehat{\mathbf{w}}^k_{m,h}\Big| + \max_{n\in[N]} \big|\boldsymbol{\phi}(s,a)^\top \boldsymbol{\zeta}^{k,n}_{m,h}\big| \\
&\leq \big(5H\sqrt{d}D_\delta + 3H\zeta\sqrt{MKd} + c_1\sigma\sqrt{d}\big)\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}^k_h)^{-1}} + 3H\zeta \\
&\leq \big(c_2 Hd + 3H\zeta\sqrt{MKd}\big)\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}^k_h)^{-1}} + 3H\zeta,
\end{aligned}$$

where $c_2 = \widetilde{\mathcal{O}}(1)$ is same as that in Lemma I.6. Here we completes the proof. □

K.2 Regret Analysis

In this part, we give out the proof of Theorem E.3, the regret bound for CoopTS-PHE in the misspecified setting.

Proof of Theorem E.3. This proof is almost same as the proof in Appendix I.2. We do the same regret decomposition (I.2) and obtain the same bound for **Term (i)** (I.3) and **Term (ii)** (I.4). Next we bound **Term (iii)** with new lemmas in misspecified setting.

Bounding Term (iii) in (I.2): conditioned on the event $\mathcal{G}(M, K, H, \delta')$, based on Lemma K.3 and Lemma K.2, by taking union bound, with probability at least $1 - |\mathcal{C}(\varepsilon)|{c'_0}^N - \delta' - MHK\delta'$, we have

$$\begin{aligned}
&\sum_{m\in\mathcal{M}}\sum_{k=1}^K\sum_{h=1}^H \big(\mathbb{E}_{\pi^*}\big[l^k_{m,h}(s_{m,h}, a_{m,h})|s_{m,1} = s^k_{m,1}\big] - l^k_{m,h}\big(s^k_{m,h}, a^k_{m,h}\big)\big) \\
&\leq \sum_{m\in\mathcal{M}}\sum_{k=1}^K\sum_{h=1}^H \Big(-l^k_{m,h}\big(s^k_{m,h}, a^k_{m,h}\big) + A_{\delta'}\varepsilon + 3H\zeta\Big) \\
&\leq \sum_{m\in\mathcal{M}}\sum_{k=1}^K\sum_{h=1}^H \Big(\big(c_2 dH + 3H\zeta\sqrt{MKd}\big)\|\boldsymbol{\phi}(s,a)\|_{(\boldsymbol{\Lambda}^k_h)^{-1}} + 3H\zeta + A_{\delta'}\varepsilon + 3H\zeta\Big) \\
&= HMKA_{\delta'}\varepsilon + 6H^2MK\zeta + \big(c_2 dH + 3H\zeta\sqrt{MKd}\big)\sum_{h=1}^H\sum_{m\in\mathcal{M}}\sum_{k=1}^K \big\|\boldsymbol{\phi}(s^k_{m,h}, a^k_{m,h})\big\|_{(\boldsymbol{\Lambda}^k_{m,h})^{-1}} \\
&\leq HMKA_{\delta'}\varepsilon + 6H^2MK\zeta + \big(c_2 dH + 3H\zeta\sqrt{MKd}\big) \\
&\qquad \times \sum_{h=1}^H \left(\log\left(\frac{\det(\boldsymbol{\Lambda}^K_h)}{\det(\lambda\mathbf{I})}\right) + 1\right) M\sqrt{\gamma} + 2\sqrt{MK\log\left(\frac{\det(\boldsymbol{\Lambda}^K_h)}{\det(\lambda\mathbf{I})}\right)} \\
&\leq HMKA_{\delta'}\varepsilon + 6H^2MK\zeta + \big(c_2 dH + 3H\zeta\sqrt{MKd}\big) \\
&\qquad \times H\Big(d(\log(1 + MK/d) + 1)M\sqrt{\gamma} + 2\sqrt{MKd\log(1 + MK/d)}\Big) \\
&= \widetilde{\mathcal{O}}\Big(d^{\frac{3}{2}}H^2\sqrt{M}\big(\sqrt{dM\gamma} + \sqrt{K}\big) + dH^2M\sqrt{K}\big(\sqrt{dM\gamma} + \sqrt{K}\big)\zeta\Big).
\end{aligned}$$

The first inequality follows from Lemma K.2, the second inequality holds due to Lemma K.3, the third inequality follows from Lemma F.12, the last inequality holds due to Lemma L.2 and the fact that $\|\boldsymbol{\phi}(\cdot)\|_2 \leq 1$, and again we choose $\varepsilon = dH\sqrt{d/MK}/A_{\delta'} = \widetilde{\mathcal{O}}(\sqrt{d/MK})$.

The probability calculation is same as that in Appendix I.2. By combining **Terms (i)(ii)(iii)** together, we get that the final regret bound for CoopTS-PHE in misspecified setting is

$$\text{Regret}(K) = \widetilde{\mathcal{O}}\Big(d^{\frac{3}{2}} H^2 \sqrt{M}\big(\sqrt{dM\gamma} + \sqrt{K}\big) + dH^2 M\sqrt{K}\big(\sqrt{dM\gamma} + \sqrt{K}\big)\zeta\Big),$$

with probability at least $1-\delta$. Here we finish the proof. □

L Auxiliary Lemmas

Lemma L.1. (Abbasi-Yadkori et al., 2011, Lemma 11) Let $\{\mathbf{X}_t\}_{t=1}^{\infty}$ be a sequence in $\mathbb{R}^d$, $\mathbf{V}$ is $d \times d$ positive definite matrix and define $\bar{\mathbf{V}}_t = \mathbf{V} + \sum_{s=1}^{t} \mathbf{X}_s \mathbf{X}_s^\top$. Then, we have that

$$\log\left(\frac{\det(\bar{\mathbf{V}}_n)}{\det(\mathbf{V})}\right) \leq \sum_{t=1}^{n} \|\mathbf{X}_t\|^2_{\bar{\mathbf{V}}_{t-1}^{-1}}.$$

Further, if $\|\mathbf{X}_t\|_2 \leq L$ for all t, then

$$\sum_{t=1}^{n} \min\left\{1, \|\mathbf{X}_t\|^2_{\bar{\mathbf{V}}_{t-1}^{-1}}\right\} \leq 2\big(\log\det(\bar{\mathbf{V}}_n) - \log\det\mathbf{V}\big) \leq 2\big(d\log\big(\big(\text{trace}(\mathbf{V}) + nL^2\big)/d\big) - \log\det\mathbf{V}\big),$$

and finally, if $\lambda_{\min}(\mathbf{V}) \geq \max\big(1, L^2\big)$ then

$$\sum_{t=1}^{n} \|\mathbf{X}_t\|^2_{\bar{\mathbf{V}}_{t-1}^{-1}} \leq 2\log\frac{\det(\bar{\mathbf{V}}_n)}{\det(\mathbf{V})}.$$

Lemma L.2. (Abbasi-Yadkori et al., 2011, Lemma 10) Suppose $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_t \in \mathbb{R}^d$ and for any $1 \leq s \leq t$, $\|\mathbf{X}_s\|_2 \leq L$. Let $\bar{\mathbf{V}}_t = \lambda\mathbf{I} + \sum_{s=1}^{t} \mathbf{X}_s\mathbf{X}_s^\top$ for some $\lambda > 0$. Then,

$$\det\left(\bar{\mathbf{V}}_t\right) \leq \left(\lambda + tL^2/d\right)^d.$$

Lemma L.3. (Ishfaq et al., 2021, Lemma D.5) Let $\mathbf{A} \in \mathbb{R}^{d\times d}$ be a positive definite matrix where its largest eigenvalue $\lambda_{\max}(\mathbf{A}) \leq \lambda$. Let $\mathbf{x}_1, ..., \mathbf{x}_k$ be k vectors in $\mathbb{R}^d$. Then it holds that

$$\left\|\mathbf{A}\sum_{i=1}^{k}\mathbf{x}_i\right\| \leq \sqrt{\lambda k}\left(\sum_{i=1}^{k}\|\mathbf{x}_i\|^2_{\mathbf{A}}\right)^{1/2}.$$

Lemma L.4. (Jin et al., 2020, Lemma D.1) Let $\mathbf{\Lambda}_t = \lambda\mathbf{I} + \sum_{i=1}^{t}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top$, where $\boldsymbol{\phi}_i \in \mathbb{R}^d$ and $\lambda > 0$. Then it holds that

$$\sum_{i=1}^{t}\boldsymbol{\phi}_i^\top(\mathbf{\Lambda}_t)^{-1}\boldsymbol{\phi}_i \leq d.$$

Lemma L.5. (Ishfaq et al., 2024, Lemma D.1) Given a multivariate normal distribution $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, we have,

$$\mathbb{P}\left(\|\mathbf{X}\| \leq \sqrt{\frac{1}{\delta}\,\text{tr}(\mathbf{\Sigma})}\right) \geq 1-\delta.$$

Lemma L.6. (Horn & Johnson, 2012) If $\mathbf{A}$ and $\mathbf{B}$ are positive semi-definite square matrices of the same size, then

$$0 \leq [\text{tr}(\mathbf{AB})]^2 \leq \text{tr}\left(\mathbf{A}^2\right)\text{tr}\left(\mathbf{B}^2\right) \leq [\text{tr}(\mathbf{A})]^2[\text{tr}(\mathbf{B})]^2.$$

Lemma L.7. (Jin et al., 2020, Lemma D.4) Let $\{s_i\}_{i=1}^{\infty}$ be a stochastic process on state space $\mathcal{S}$ with corresponding filtration $\{\mathcal{F}_i\}_{i=1}^{\infty}$. Let $\{\boldsymbol{\phi}_i\}_{i=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process where $\boldsymbol{\phi}_i \in \mathcal{F}_{i-1}$, and $\|\boldsymbol{\phi}_i\| \leq 1$. Let $\boldsymbol{\Lambda}_k = \lambda \mathbf{I} + \sum_{i=1}^{k} \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top$. Then for any $\delta > 0$, with probability at least $1-\delta$, for all $k \geq 0$, and any $V \in \mathcal{V}$ with $\sup_{s\in\mathcal{S}} |V(s)| \leq H$, we have

$$\left\| \sum_{i=1}^{k} \boldsymbol{\phi}_i \{V(s_i) - \mathbb{E}[V(s_i) \mid \mathcal{F}_{i-1}]\} \right\|_{\boldsymbol{\Lambda}_k^{-1}}^2 \leq 4H^2 \left[\frac{d}{2} \log\left(\frac{k+\lambda}{\lambda}\right) + \log \frac{\mathcal{N}_\varepsilon}{\delta} \right] + \frac{8k^2\varepsilon^2}{\lambda},$$

where $\mathcal{N}_\varepsilon$ is the ε-covering number of $\mathcal{V}$ with respect to the distance $\operatorname{dist}(V, V') = \sup_{s\in\mathcal{S}} |V(s) - V'(s)|$.

Lemma L.8. (Vershynin, 2018, Covering number of Euclidean ball) For any $\varepsilon > 0$, $\mathcal{N}_\varepsilon$, the ε-covering number of the Euclidean ball of radius $B > 0$ in $\mathbb{R}^d$ satisfies

$$\mathcal{N}_\varepsilon \leq \left(1 + \frac{2B}{\varepsilon}\right)^d \leq \left(\frac{3B}{\varepsilon}\right)^d.$$

Lemma L.9. Let $\mathcal{V}$ denote a class of functions mapping from $\mathcal{S}$ to $\mathbb{R}$ with the following parametric form

$$V(\cdot) = \max_{a\in\mathcal{A}} \left\{ \min \left\{ \max_{n\in[N]} \boldsymbol{\phi}(\cdot, a)^\top \mathbf{w}^n, H - h + 1 \right\}^+ \right\},$$

where the parameter $\mathbf{w}^n$ satisifies $\|\mathbf{w}^n\| \leq B$ for all $n \in [N]$ and for all $(x, a) \in \mathcal{S} \times \mathcal{A}$, we have $\|\boldsymbol{\phi}(x, a)\| \leq 1$. Let $N_{\mathcal{V},\varepsilon}$ be the ε-covering number of $\mathcal{V}$ with respect to the distance dist $(V, V') = \sup_{s\in\mathcal{S}} \left|V(s) - V'(s)\right|$. Then

$$\mathcal{N}_{\mathcal{V},\varepsilon} \leq \left(\frac{3B}{\varepsilon}\right)^d.$$

Proof. Consider any two functions $V_1, V_2 \in \mathcal{V}$ with parameters $\{\mathbf{w}_1^n\}_{n\in[N]}$ and $\{\mathbf{w}_2^n\}_{n\in[N]}$ respectively. Then we have

$$\begin{aligned}
\operatorname{dist}(V_1, V_2) &\leq \sup_{s,a} \left| \max_{n\in[N]} \boldsymbol{\phi}(s,a)^\top \mathbf{w}_1^n - \max_{n\in[N]} \boldsymbol{\phi}(s,a)^\top \mathbf{w}_2^n \right| \\
&\leq \sup_{s,a} \left| \max_{n\in[N]} \left(\boldsymbol{\phi}(s,a)^\top \mathbf{w}_1^n - \boldsymbol{\phi}(s,a)^\top \mathbf{w}_2^n \right) \right| \\
&\leq \sup_{\|\boldsymbol{\phi}\|\leq 1} \max_{n\in[N]} \left| \boldsymbol{\phi}^\top \mathbf{w}_1^n - \boldsymbol{\phi}^\top \mathbf{w}_2^n \right| \\
&= \max_{n\in[N]} \sup_{\|\boldsymbol{\phi}\|\leq 1} \left| \boldsymbol{\phi}^\top (\mathbf{w}_1^n - \mathbf{w}_2^n) \right| \\
&\leq \max_{n\in[N]} \sup_{\|\boldsymbol{\phi}\|\leq 1} \|\boldsymbol{\phi}\| \left\| \mathbf{w}_1^n - \mathbf{w}_2^n \right\| \\
&\leq \max_{n\in[N]} \left\| \mathbf{w}_1^n - \mathbf{w}_2^n \right\|.
\end{aligned}$$

Let $\mathcal{N}_{\mathbf{w},\varepsilon}$ denote the ε-covering number of $\{\mathbf{w} \in \mathbb{R}^d \mid \|\mathbf{w}\| \leq B\}$. Then, Lemma L.8 implies

$$\mathcal{N}_{\mathbf{w},\varepsilon} \leq \left(1 + \frac{2B}{\varepsilon}\right)^d \leq \left(\frac{3B}{\varepsilon}\right)^d.$$

For any $V_1 \in \mathcal{V}$, we consider its corresponding parameters $\{\mathbf{w}_1^n\}_{n\in[N]}$. For any $n \in [N]$, we can find $\mathbf{w}_2^n$ such that $\|\mathbf{w}_1^n - \mathbf{w}_2^n\| \leq \varepsilon$, then we can get $V_2 \in \mathcal{V}$ with parameters $\{\mathbf{w}_2^n\}_{n\in[N]}$. Then we have $\operatorname{dist}(V_1, V_2) \leq \max_{n\in[N]} \|\mathbf{w}_1^n - \mathbf{w}_2^n\| \leq \varepsilon$. Thus, we have,

$$\mathcal{N}_{\mathcal{V},\varepsilon} \leq \mathcal{N}_{\mathbf{w},\varepsilon} \leq \left(1 + \frac{2B}{\varepsilon}\right)^d \leq \left(\frac{3B}{\varepsilon}\right)^d.$$

This completes the proof. □

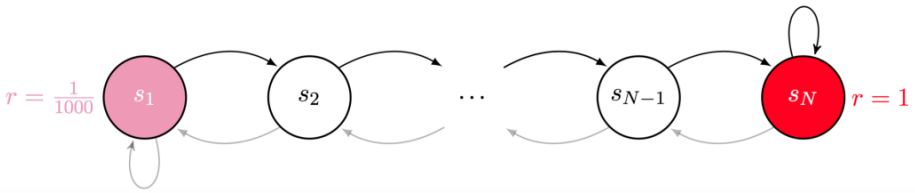


Figure 2: The N-Chain environment (Osband et al., 2016a).

Lemma L.10. (Abramowitz & Stegun, 1968) Suppose Z is a Gaussian random variable $Z \sim \mathcal{N}(\mu, \sigma^2)$, where $\sigma > 0$. For $0 \leq z \leq 1$, we have

$$\mathbb{P}(Z > \mu + z\sigma) \geq \frac{1}{\sqrt{8\pi}} e^{\frac{-z^2}{2}}, \quad \mathbb{P}(Z < \mu - z\sigma) \geq \frac{1}{\sqrt{8\pi}} e^{\frac{-z^2}{2}}.$$

And for $z \geq 1$, we have

$$\frac{e^{-z^2/2}}{2z\sqrt{\pi}} \leq \mathbb{P}(|Z - \mu| > z\sigma) \leq \frac{e^{-z^2/2}}{z\sqrt{\pi}}.$$

Lemma L.11. (Ishfaq et al., 2021, Lemma D.2) Consider a d-dimensional multivariate normal distribution $\mathcal{N}(\mathbf{0}, A\mathbf{\Lambda}^{-1})$ where A is a scalar. Let $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_N$ be N independent samples from the distribution. Then for any $\delta > 0$

$$\mathbb{P}\left(\max_{j \in [M]} \|\boldsymbol{\eta}_j\|_{\mathbf{\Lambda}} \leq c\sqrt{dA\log(d/\delta)}\right) \geq 1 - M\delta,$$

where c is some absolute constant.

Lemma L.12. (Abbasi-Yadkori et al., 2011, Lemma 12) Let $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ be positive semi-definite matrices such that $\mathbf{A} = \mathbf{B} + \mathbf{C}$. Then we have that

$$\sup_{\mathbf{x} \neq 0} \frac{\mathbf{x}^\top \mathbf{A} \mathbf{x}}{\mathbf{x}^\top \mathbf{B} \mathbf{x}} \leq \frac{\det(\mathbf{A})}{\det(\mathbf{B})}.$$

M Additional Experimental Details

We conduct comprehensive experiments investigating the exploration strategies for DQN under a multi-agent setting. For all the Q networks in our experiments, we use ReLU as our activation function. Given that all experiments are conducted under multi-agent settings unless explicitly specified as a single-agent or centralized scenario, we denote our methods: CoopTS-PHE as "PHE" and CoopTS-LMC as "LMC" in experimental contexts and figures. In addition to our methods, the baselines we selected are either commonly used (DQN (Mnih et al., 2015), DDQN (Hasselt et al., 2016)) or with competitive empirical performance (Bootstrapped DQN (Osband et al., 2016a), NoisyNet DQN (Fortunato et al., 2018)). Both Bootstrapped DQN and NoisyNet DQN are randomized exploration methods. Bootstrapped DQN uses finite ensembles to generate the randomized value functions and views them as approximate posterior samples of Q-value functions. NoisyNet DQN injects noise into the parameters of neural networks to aid efficient exploration. Note that we run all our experiments on Nvidia RTX A5000 with 24GB RAM.

M.1 N-chain

We commence by presenting the comprehensive results for $N = 25$ in Figure 3, illustrating that our randomized exploration methods exhibit greater suitability in realistic scenarios characterized by an increasing number of agents. This superiority is particularly evident under two potential

Table 2: The swept hyper-parameters in N-Chain for PHE

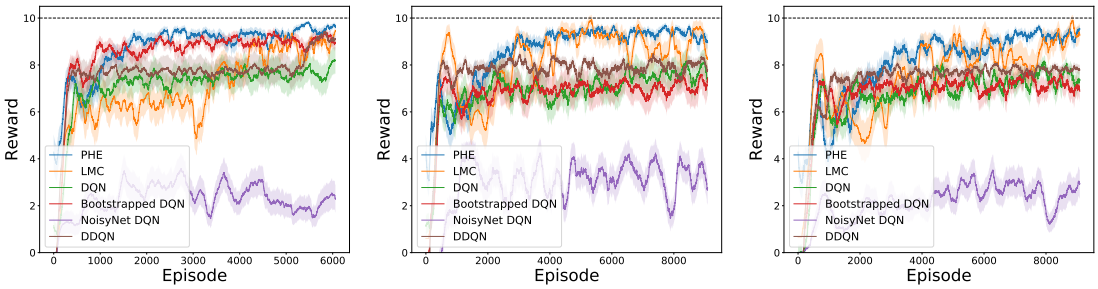
Hyper-parameter	Values
Learning Rate η_k	$\{10^{-1}, 3\times10^{-2}, 10^{-2}, 3\times10^{-3}, 10^{-3}, 3\times10^{-4}, 10^{-4}\}$
No Target Networks	$\{1, 2, 4, 8\}$
Noise std	$\{0, 10^{-4}10^{-3}, 10^{-2}, 10^{-1}, 1.0\}$
Noise Regularization	$\{0, 10^{-4}10^{-3}, 10^{-2}, 10^{-1}, 1.0\}$

Table 3: The swept hyper-parameters in N-Chain for LMC

Hyper-parameter	Values
Learning Rate η_k	$\{10^{-1}, 3\times10^{-2}, 10^{-2}, 3\times10^{-3}, 10^{-3}, 3\times10^{-4}, 10^{-4}\}$
Bias Factor α	$\{1.0, 0.1, 0.01\}$
Inverse Temperature β_k	$\{10^2, 10^4, 10^6, 10^8, 10^{10}\}$
No Update J_k	$\{1, 4, 16, 32\}$

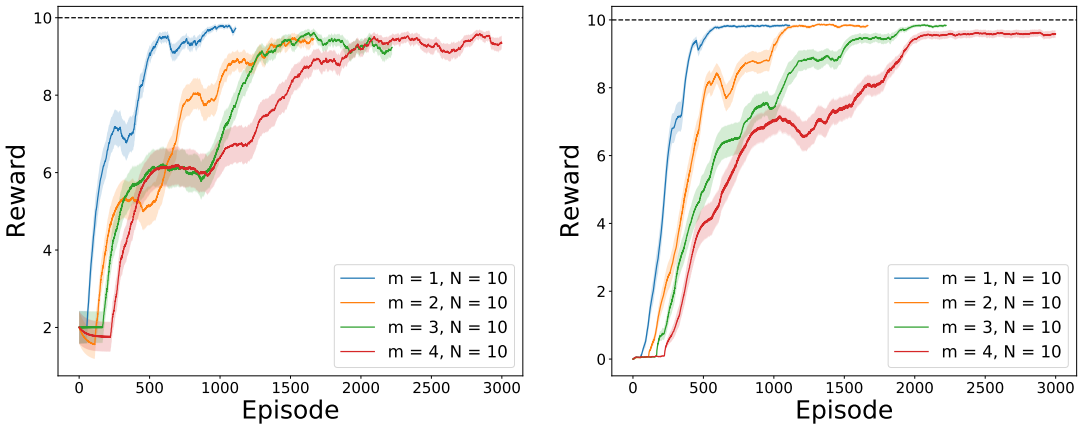
Table 4: Hyper-parameters used in the N-chain

Hyper-parameter	PHE	LMC	DQN	Bootstrapped DQN	Noisy DQN	DDQN
NN size	32×32	32×32	32×32	32×32	32×32	32×32
Task Horizon	19/34	19/34	19/34	19/34	19/34	19/34
Discount Factor λ	0.99	0.99	0.99	0.99	0.99	0.99
Learning Rate η_k	3×10^{-2}	10^{-4}	3×10^{-2}	3×10^{-2}	3×10^{-2}	3×10^{-2}
Hidden Activation	Relu	Relu	Relu	Relu	Relu	Relu
Output Activation	Linear	Linear	Linear	Linear	Linear	Linear
No Update J_k	1	4	1	1	1	1
No Target Networks	2	1	1	4	1	1



(a) m=2 (b) m=3 (c) m=4

Figure 3: Comparison among different exploration strategies in N-chain with $N = 25$. All results are averaged over 10 runs.



(a) PHE (b) LMC

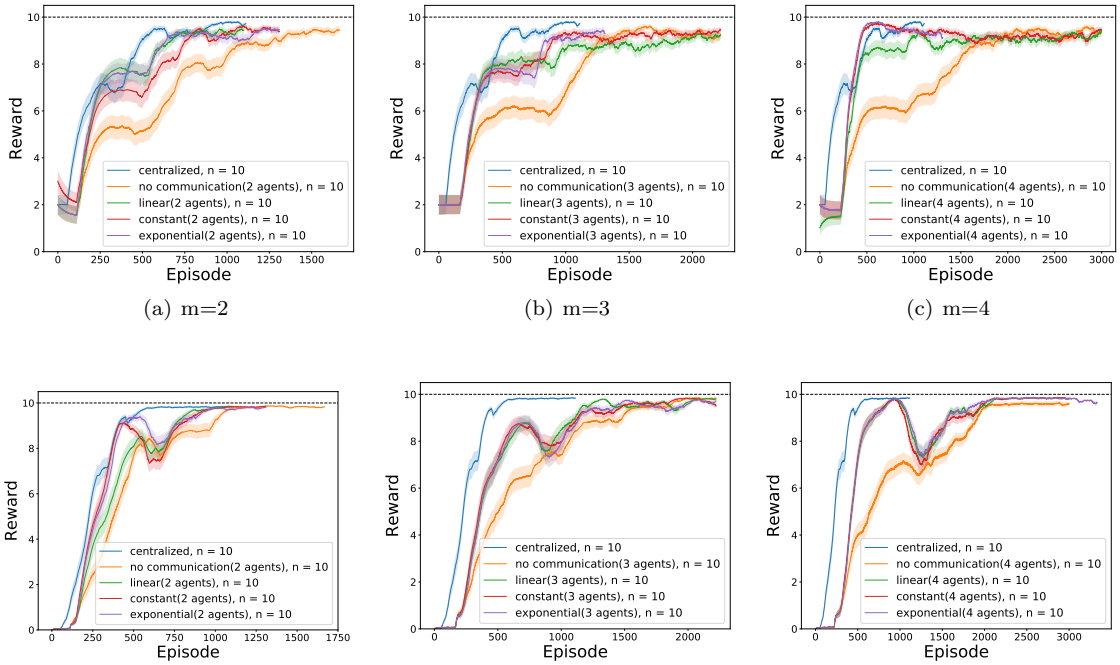
Figure 4: Rewards with averaged over 10 independent runs for different numbers of agents among algorithms without communication. Note that when $m = 1$, one agent indicates a centralized setting.

circumstances: (1) where there are more limitations on computation or data access from each source in the real world, and (2) when parallel learning from multiple sources can significantly enhance runtime efficiency.

Subsequently, we provide a more comprehensive study to investigate the exploration capabilities facilitated by parallel training. Preliminary experiments are conducted with a reduced state space, specifically considering $N = 10$. The study aims to investigate exploration capabilities across varying agent counts, specifically within the set $m \in \{1, 2, 3, 4\}$.

Performance Consistency with Varying m In the investigation detailed in Figure 4, we explore parallel learning without inter-agent communication. Consequently, while multiple agents engage in simultaneous policy learning, each agent independently formulates its policies without the exchange of transition information. The discernible trend in this scenario is that an increase in the number of agents sharing the total episodes results in a slower rate of policy learning. Notably, despite this temporal discrepancy, all learning trajectories eventually approximate convergence towards the optimal dashed line.

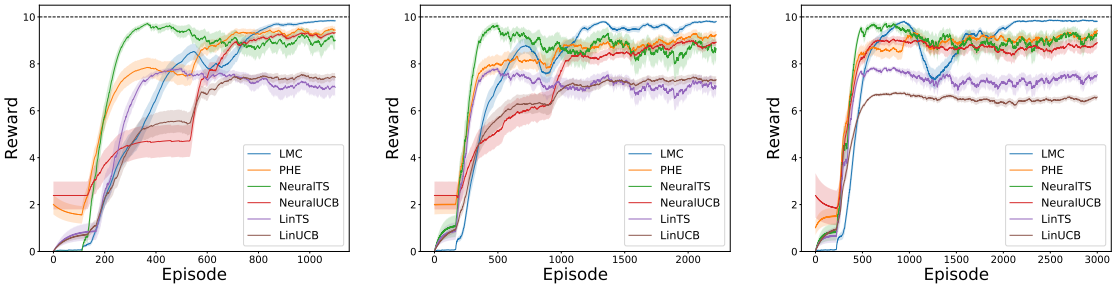
Different Synchronization Conditions To further demonstrate the efficiency of parallel learning with communication, we compare different synchronization conditions in Section 3.1. Specifically, we denote synchronization (1) in every constant step as *constant*, (2) following exponential function as *exponential*, and (3) based on (3.3) as *linear*. To have a fair comparison among different synchro-



(a) m=2 (b) m=3 (c) m=4

(d) m=2 (e) m=3 (f) m=4

Figure 5: Different number of agents m with different synchronization strategies as well as the single-agent and no communication settings in $N = 10$. **Top:** PHE, **Bottom:** LMC



(a) m=2 (b) m=3 (c) m=4

Figure 6: Performance with different number of agents m compared with bandit-inspired exploration in $N = 10$.

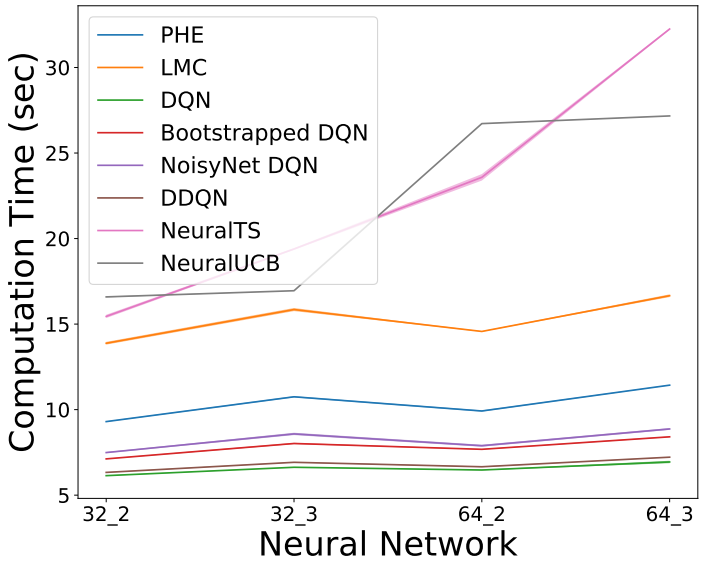
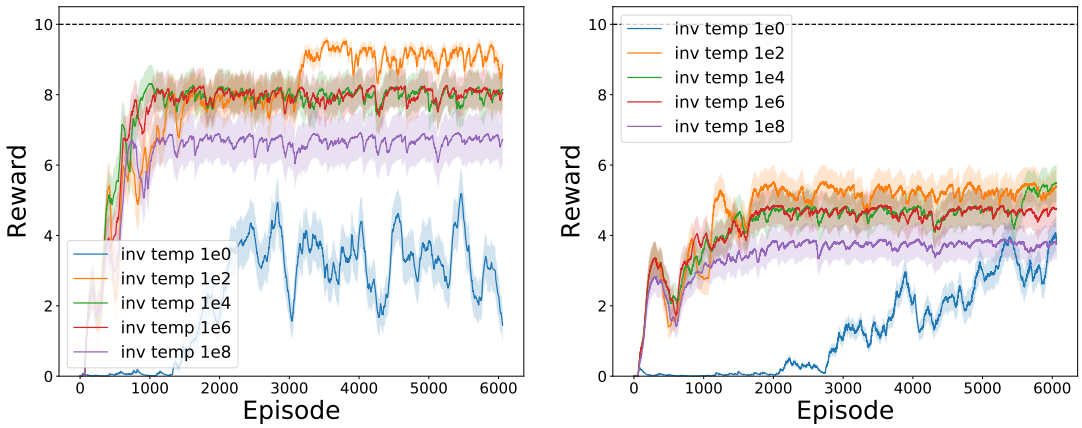


Figure 7: Computation time with different exploration strategies.



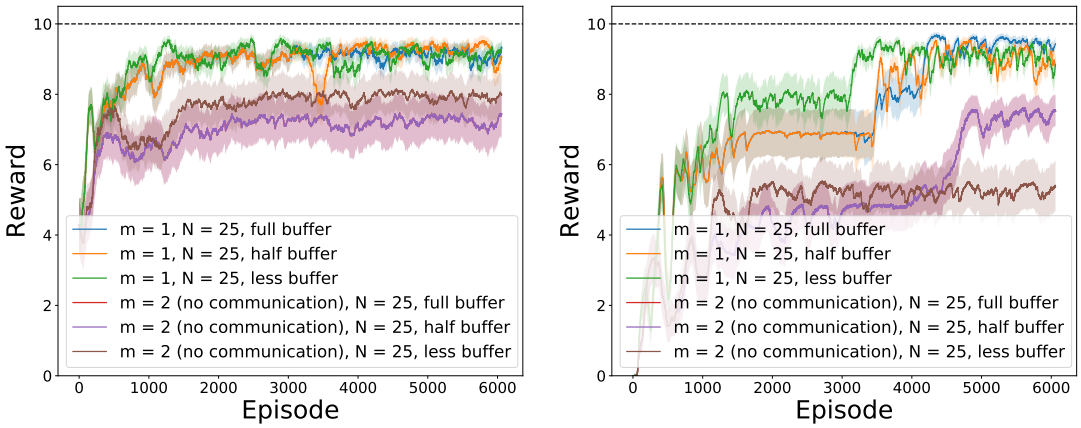
(a) m=1 (centralized), N=25

(b) m=2 (no communication), N=25

Figure 8: Hyper-parameter tuning of inverse temperature (inv temp) $\beta_{m,k}$ for LMC with $N = 25$: (a) centralized setting $m = 1$ (b) 2 agents without communication $m = 2$.

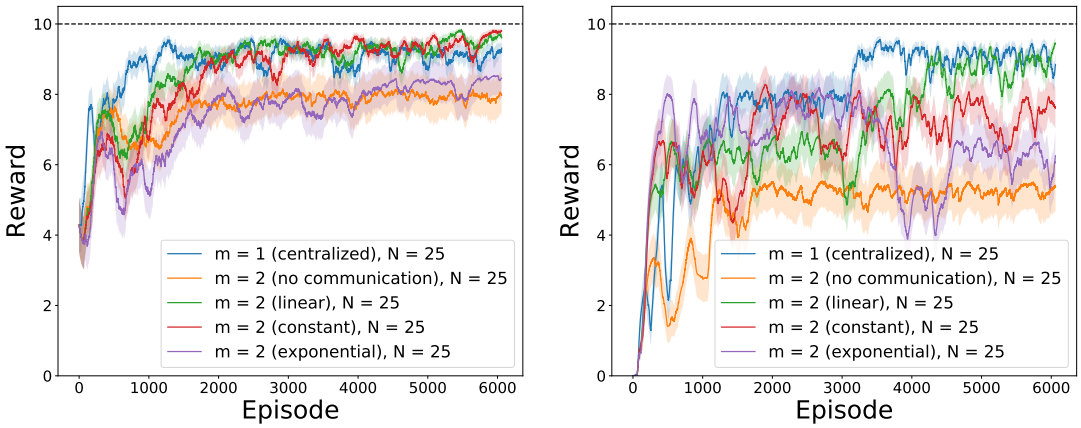
nization conditions, we firstly record the empirical number of synchronization via *linear* condition in average, and then we consider constant value for *constant* condition and select proper base b for *exponential* condition with a similar number of synchronization. Figure 5 illustrates that any synchronization condition can improve learning efficiency but still with centralized learning as an upper bound.

Performance Compared with Bandit-inspired Methods Since one of our proposed random exploration strategies, PHE is a variant of approximation TS, it is fair for us to investigate the performance of other exploration methods from bandit algorithms with the integration of DQN. We mainly compare both TS and UCB under neural network (*i.e.,* NeuralTS (Zhang et al., 2021) and NeuralUCB (Zhou et al., 2020)) and linear (*i.e.,* LinTS (Agrawal & Goyal, 2013a) and LinUCB (Li et al., 2010)) settings. We show that a performance gap exists between linear approaches and other neural-based methods even in a small-scale exploration problem with $N = 10$ in Figure 6.



(a) PHE (b) LMC

Figure 9: Different buffer size with $N = 25$ between single agent (centralized) and 2 agents (no communication). Note that the full buffer indicates the size of the total episodes. Each agent in no communication setting only occupies half of the total episodes. Therefore, two curves (full buffer, half buffer) in no communication are consistent.



(a) PHE (b) LMC

Figure 10: Different synchronization strategies as well as the single-agent and no communication settings in $N = 25$.

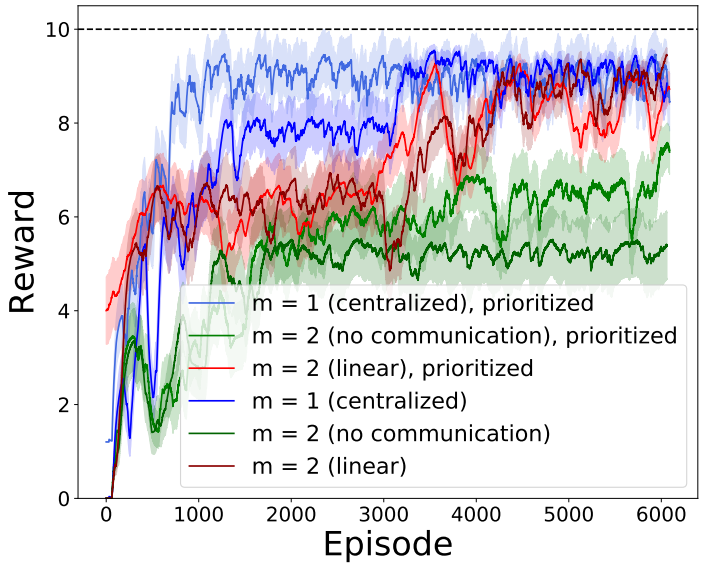


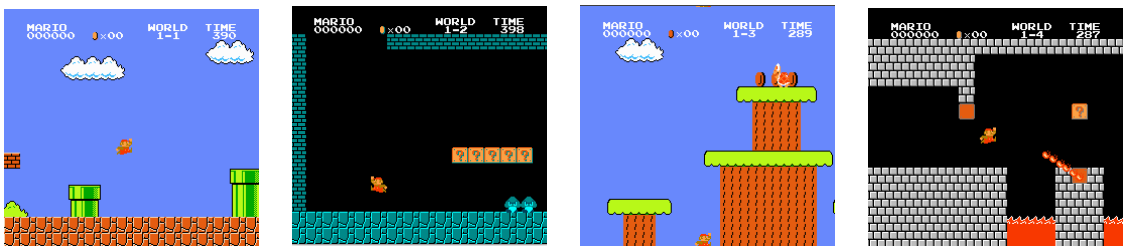
Figure 11: Gap reduction improvement with prioritized experience replay for parallel learning without communication. Note that the same settings with standard and prioritized experience replay are in the same-ish color.

Computational Time We have demonstrated that both NeuralTS and NeuralUCB exhibit convergence to performance levels comparable to our proposed randomized exploration strategies (*i.e.,* PHE and LMC) when considering the case of $N = 10$ with $m = 4$ under the synchronization condition (*linear*), as outlined in (3.3). However, we argue that the scalability of both methods is limited due to their associated computational costs. To substantiate this assertion, we conduct experiments across all methods including DQN baselines with $N = 10$ and $m = 4$ over 10^4 steps with varying neural network sizes, such as $[32, 32, 32]$, which signifies three layers with 32 neurons in each layer. Importantly, the length of the chain N has no bearing on the running time.

In Figure 7, we show the computational time of all methods under different neural network sizes. The solid lines represent the average computational time over 10 random seeds and the shaded area represents the standard deviation. We observe that NeuralTS and NeuralUCB have heavy running time consistently with varying network sizes. Although the computation time of LMC is still higher than other remaining approaches, we observe that it maintains a similar computation time with different neural network sizes, which can still be scaled up to more complex problems with larger neural networks.

Hyper-parameter Tuning of Inverse Temperature $\beta_{m,k}$ Subsequently, we scale the problem to $N = 25$. Given the extended horizon, the demand for exploration intensifies, leading us to conduct hyper-parameter tuning for the inverse temperature parameter $\beta_{m,k}$ in LMC, as illustrated in Figure 8. It is crucial to note that the efficacy of learning is significantly influenced by the exploration capacity in both centralized learning and parallel learning without communication. Our observations reveal a discernible gap between centralized and parallel learning, a departure from the pattern observed in Figure 4. We posit that the disparity may stem from issues associated with the replay buffer size in off-policy RL algorithms. Specifically, when the replay buffer exhausts its capacity for new transitions, the incoming transition replaces the oldest one.

Hyper-parameter Tuning of Buffer Size Therefore, we present a performance comparison between a solitary agent ($m = 1$) and a scenario involving two agents ($m = 2$) in Figure 9 with different buffer sizes. Full buffer and half buffer indicate the replay buffer's capacity to store the complete set and half of the transitions during training, respectively. We observe that the learning



(a) SuperMarioBros-1-1-v0 (b) SuperMarioBros-1-2-v0 (c) SuperMarioBros-1-3-v0 (d) SuperMarioBros-1-4-v0

Figure 12: Illustrations of 4 different environments in Super Mario Bros task.

process is more efficient with less buffer size in a centralized setting because having an excessively large replay buffer may potentially impede the efficiency of the learning process. Furthermore, the gap between centralized setting and paralleling learning still exists among different buffer sizes. Therefore, we focus on the setting of less buffer size with different synchronization conditions in Figure 10. We conclude that *linear* condition results in competitive performance in both PHE and LMC in the N-chain problem and we report all exploration strategies with *linear* condition in Appendix A.1.

Ablation Study of Sampling Mechanism To reduce the reward gap, we adopt a better sampling mechanism in the replay buffer with prioritized experience replay (PER). In Figure 11, parallel learning without inter-agent communication can increase reward with PER. However, centralized learning with PER improves faster convergence with similar performance and the trends for *linear* condition curves are similar. Therefore, the gap between centralized and parallel learning without communication is reduced with PER. Note that the main experimental results in Figure 1 are based on standard experience replay because standard sampling in *linear* condition has similar performance against PER with faster training time.

M.2 Super Mario Bros

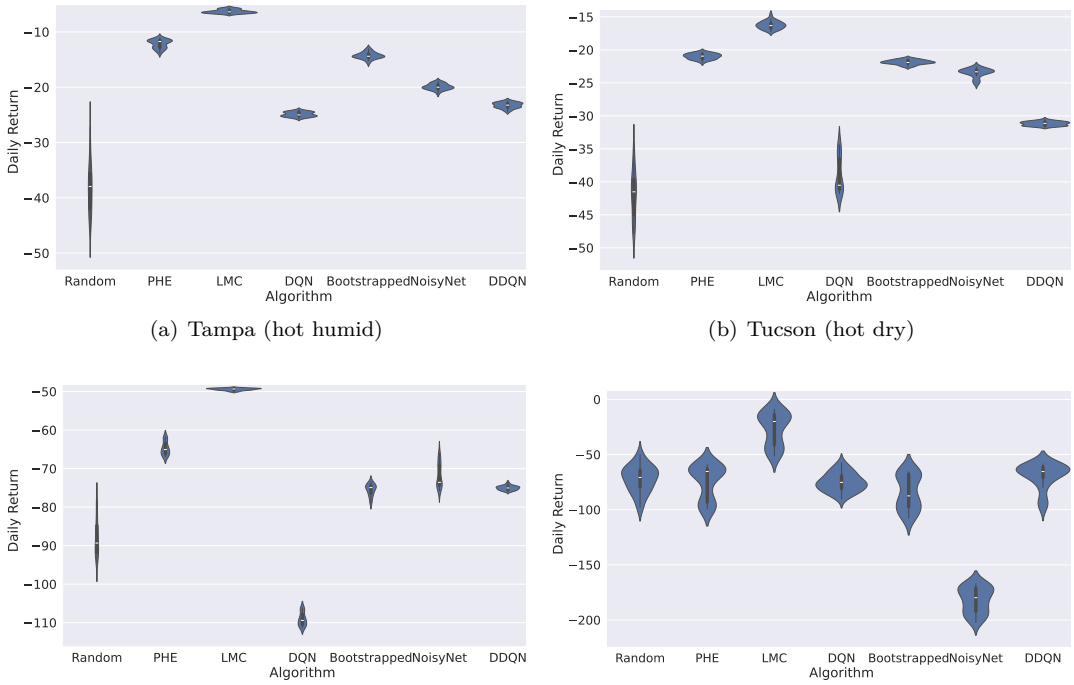
While cooperative parallel learning enhances training efficiency through data sharing, challenges emerge when handling data from devices capturing images or audio due to privacy concerns in real-world applications. In response, our approach extends randomized exploration strategies to a federated reinforcement learning framework as shown in Algorithm 4, from Algorithm 1, which incorporates parameter synchronization among Q neural networks rather than relying on the conventional practice of sharing agents' transitions in Line 14 in Algorithm 4. Note that the synchronization follows the format as in Algorithm 1 to update Q functions with horizon $h \in H$. However, in practice, we can directly update the weight of the neural network to reduce the communication cost.

The training process unfolds within a federated reinforcement learning framework, wherein local updates and global aggregations are iteratively executed (Jin et al., 2022a). Specifically, each agent iterates through multiple local updates of its value function, followed by server-mediated averaging of these functions across all agents, constituting a form of parameter sharing. Note that the transitions are not accessible among agents, leading us to directly synchronize all agents with parameter sharing every constant local iteration instead of synchronization condition in (3.3). We use the same architecture for all the experiments in the Super Mario Bros task with the preprocessed images as the input states and 7 discrete actions in action space.

Particularly, we construct 3 convolutional neural network layers with width [32, 64, 32], followed by 2 linear layers with the output of action space in the Q network.

Algorithm 4 Unified Algorithm Framework for Randomized Exploration in Federated Learning

for episode $k = 1, ..., K$ **do**
  for agent $m \in \mathcal{M}$ **do**
    Receive initial state $s^k_{m,1}$.
    $V^k_{m,H+1}(\cdot) \leftarrow 0$.
    $\{Q^k_{m,h}(\cdot,\cdot)\}_{h=1}^H \leftarrow$ **Randomized Exploration** ◁ Algorithm 2 or Algorithm 3
    for step $h = 1, ..., H$ **do**
      $a^k_{m,h} \leftarrow \text{argmax}_{a\in\mathcal{A}} Q^k_{m,h}(s^k_{m,h}, a)$.
      Receive $s^k_{m,h+1}$ and r_h.
      if Condition **then**
        SYNCHRONIZE $\leftarrow$ True.
      end if
    end for
  end for
  if SYNCHRONIZE **then**
    for step $h = H, ..., 1$ **do**
      $\bar{Q}^k_m \leftarrow \frac{1}{M}\sum_{m=1}^M Q^k_{m,h}$
      $Q^k_{m,h} \leftarrow \bar{Q}^k_{m,h}, \forall m$
    end for
  end if
end for



(a) Tampa (hot humid)

(b) Tucson (hot dry)

(c) Rochester (cold humid)

(d) Great Falls (cold dry)

Figure 13: Evaluation performance at different cities in building energy systems

M.3 Thermal Control of Building Energy Systems

BuildingEnv encompasses the regulation of heat flow in a multi-zone building to sustain a desired temperature setpoint. We focus on one pre-defined building called "office small" in different cities with varying weather types, *i.e.,* Tampa (Hot Humid), Tucson (Hot Dry), Rochester (Cold Humid), and Great Falls (Cold Dry). Each episode is designed to span a single day, comprising 5-minute time intervals (H = 288, $\tau = 5/60$ hours).

Observation Space The state at time step t, denoted as $s(t) \in \mathbb{R}^{M+4}$, encompasses the temperatures $T_i(t)$ of each zone, where $i \in M$, along with four additional properties: $Q^{GHI}(t)$, $\bar{Q}^p(t)$, $T_G(t)$, and $T_E(t)$. Specifically, $Q^{GHI}(t)$ represents the heat gain from solar irradiance, $\bar{Q}^p(t)$ denotes the heat acquired from occupant activities, while $T_G(t)$ and $T_E(t)$ signify the ground and outdoor environment temperatures, respectively.

Action Space The continuous version of the action $a(t) \in [-1, 1]^M$ controls the heating of M zones. However, since our randomized exploration strategies use DQN (Mnih et al., 2015) as the backbone, we adopt the multi-discrete action space defined in Yeh et al. (2023), which is a vector of action spaces. Then we convert the multi-discrete action space to a single discrete action space with action mapping.

Reward Function The primary objective is to minimize energy consumption while ensuring the maintenance of temperature within a specified comfort range. Therefore, the reward is penalized with both temperature deviations and HVAC energy consumption as follows:

$$r(t) = -(1 - \beta)\|a(t)\|_2 - \beta\|T^{target}(t) - T(t)\|_2,$$

where $T^{target}(t) = [T_1^{target}(t), T_2^{target}(t), ..., T_M^{target}(t)]$ are the target temperatures and $T^(t) = [T_1(t), T_2(t), ..., T_M(t)]$ are the actual zonal temperatures. The parameter β is the trade-off between the energy consumption and temperature deviation penalties.

We execute experiments following the united framework in Algorithm 1, synchronizing every constant number of steps across diverse weather conditions in varying cities. We choose a 2-hidden layer neural network with width $[64, 64]$ for the Q network. Subsequently, we evaluate the performance of all methods in distinct cities respectively, as illustrated in Figure 13. Notably, our proposed random exploration strategies demonstrate a consistently higher mean return across all cities. However, it is worth highlighting that DQN in Figure 13(c) and Noisy-Net in Figure 13(d) exhibit lower returns compared to random actions. This outcome can be attributed to the discrete action space configuration (Yeh et al., 2023). In addition, we observe that maintaining thermal control of buildings is more challenging in cold weather conditions compared to hot weather conditions.