Provably Efficient Cooperative Multi-Agent Reinforcement Learning with Function Approximation

Abhimanyu Dubey and Alex Pentland*

Abstract

Reinforcement learning in cooperative multi-agent settings has recently advanced significantly in its scope, with applications in cooperative estimation for advertising, dynamic treatment regimes, distributed control, and federated learning. In this paper, we discuss the problem of cooperative multi-agent RL with function approximation, where a group of agents communicates with each other to jointly solve an episodic MDP. We demonstrate that via careful message-passing and cooperative value iteration, it is possible to achieve near-optimal no-regret learning even with a fixed constant communication budget. Next, we demonstrate that even in heterogeneous cooperative settings, it is possible to achieve Pareto-optimal no-regret learning with limited communication. Our work generalizes several ideas from the multi-agent contextual and multi-armed bandit literature to MDPs and reinforcement learning.

1 Introduction

Cooperative multi-agent reinforcement learning (MARL) systems are widely prevalent in many engineering systems, e.g., robotic systems (Ding et al., 2020), power grids (Yu et al., 2014), traffic control (Bazzan, 2009), as well as team games (Zhao et al., 2019). Increasingly, federated (Yang et al., 2019) and distributed (Peteiro-Barral & Guijarro-Berdiñas, 2013) machine learning is gaining prominence in industrial applications, and reinforcement learning in these large-scale settings is becoming of import in the research community as well (Zhuo et al., 2019; Liu et al., 2019).

Recent research in the statistical learning community has focused on cooperative multi-agent decision-making algorithms with provable guarantees (Zhang et al., 2018b; Wai et al., 2018; Zhang et al., 2018a). However, prior work focuses on algorithms that, while are decentralized, provide guarantees on convergence (e.g., Zhang et al. (2018b)) but no finite-sample guarantees for regret, in contrast to efficient algorithms with function approximation proposed for single-agent RL (e.g., Jin et al. (2018, 2020); Yang et al. (2020)). Moreover, optimization in the decentralized multi-agent setting is also known to be non-convergent without assumptions (Tan, 1993). Developing no-regret multi-agent algorithms is therefore an important problem in RL.

For the (relatively) easier problem of multi-agent multi-armed bandits, there has been significant recent interest in decentralized algorithms involving agents communicating over a network (Landgren et al., 2016a, 2018; Martínez-Rubio et al., 2019; Dubey & Pentland, 2020b), as well as in the distributed settings (Hillel et al., 2013; Wang et al., 2019). Since several application areas for distributed sequential decision-making regularly involve non-stationarity and contextual information (Polydoros & Nalpantidis, 2017), an MDP formulation can potentially provide stronger algorithms for these settings as well. Furthermore, no-regret algorithms in the single-agent RL setting with function approximation (e.g., Jin et al. (2020)) build on analysis techniques for contextual bandits, which leads us to the question – *Can no-regret function approximation be extended to (decentralized) cooperative multi-agent reinforcement learning?*

*Media Lab and Institute for Data, Systems and Society, Massachusetts Institute of Technology. Corresponding email: dubeya@mit.edu.

Contributions. In this paper, we answer the above question affirmatively for *cooperative* multi-agent learning. Specifically, we study cooperative multi-agent reinforcement learning with *linear* function approximation in two practical scenarios - the first being *parallel* reinforcement learning (Kretchmar, 2002), where a group of agents simultaneously solve isolated MDPs that are "similar" to each other, and communicate to facilitate faster learning. This corresponds to *heterogenous* federated learning (Li et al., 2018), and generalizes transfer learning in RL (Taylor & Stone, 2009) to multiple sources (Yao & Doretto, 2010).

The second scenario we study is the heterogeneous multi-agent MDP, where a group of agents interact in an MDP by playing moves simultaneously, with the objective being to recover *Pareto-optimal* multi-agent policies (Tuyls & Nowé, 2005), a more general notion of performance, compared to the standard objective of maximizing cumulative reward (Boutilier, 1996). Multi-agent MDPs are present in cooperative and distributed applications such as multi-agent robotics (Yang & Gu, 2004; Gupta et al., 2017).

For each setting, we propose decentralized algorithms that are provably efficient with limited communication. Existing regret bounds for single-agent episodic settings scale as $\widetilde{\mathcal{O}}(H^2\sqrt{d^3T})$ for T episodes of length H[1], leading to a cumulative regret of $\widetilde{\mathcal{O}}(MH^2\sqrt{d^3T})$ if M agents operate in isolation. Similarly, a *fully centralized* agent running for MT episodes will consequently obtain $\widetilde{\mathcal{O}}(H^2\sqrt{d^3MT})$ regret. In comparison, for the *parallel* setting, we provide a least-squares value iteration (LSVI) algorithm `CoopLSVI` which obtains a group regret of $\widetilde{\mathcal{O}}((d+k)H^2\sqrt{(d+\chi)MT})$, where χ is a measure of heterogeneity between different MDPs, and k is the minimum number of dimensions required to model the heterogeneity. When the MDPs are homogenous, our rate matches the *centralized* single-agent regret. Moreover, our algorithm only requires $\mathcal{O}(HM^3)$ rounds of communication, i.e., independent of the number of episodes T. The algorithm is a multi-agent variant of the popular upper confidence bound (UCB) strategies for reinforcement learning, where our key contributions are to first design an estimator that takes into account the bias introduced via heterogeneity between the different MDPs, and second, to strategically select episodes in which to synchronize statistics across agents without excessive communication.

For multi-agent MDPs, we introduce a variant of `CoopLSVI`, which attempts to recover the set of *cooperative* Pareto-optimal policies, i.e., policies that cannot improve any individual agent's reward without decreasing the reward of the other agents (Desai et al., 2018). `CoopLSVI` obtains a cumulative *Bayes regret* of $\widetilde{\mathcal{O}}(H^2\sqrt{d^3T})$ over T episodes, which is the first no-regret bound on learning Pareto-optimal policies. We use the method of *random scalarizations*, a popular approach in multi-objective decision-making (Van Moffaert & Nowé, 2014) coupled with optimistic least-squares value iteration, to provide a no-regret algorithm. Moreover, a direct corollary of our analysis is the *first* no-regret algorithm for multi-objective RL (Mossalam et al., 2016) with function approximation.

2 Related Work

Our work is related to several areas of multi-agent learning. We discuss connections sequentially.

Multi-Agent Multi-Armed Bandits. The multi-agent bandit literature has seen a lot of interest recently, and given that our techniques build on function approximation techniques initially employed by the bandit community, many parallels can be drawn to our work and the multi-agent bandit literature. Several recent works have proposed no-regret algorithms. One line of work utilizes consensus-based averaging (Martínez-Rubio et al., 2019; Landgren et al., 2016a,b, 2018) which provide regret depending on statistics of the communication graph. For contextual bandits, similar algorithms have been derived that alternatively utilize message-passing algorithms (Dubey & Pentland, 2020a) or server-synchronization (Wang et al., 2019). In the competitive multi-agent bandit setting, where agents must avoid collisions, algorithms have been proposed for distributed (Liu &

[1]The $\widetilde{\mathcal{O}}$ notation ignores logarithmic factors and failure probability, and d is the dimensionality of the ambient feature space. See, e.g., Yang et al. (2020) and Jin et al. (2018) for bounds.

Zhao, 2010a,b; Hillel et al., 2013) and limited-communication (Bistritz & Leshem, 2018) settings. Differentially-private algorithms have also been proposed (Dubey & Pentland, 2020b).

Reinforcement Learning with Function Approximation. Our work builds on the body of recent work in (single-agent) reinforcement learning with function approximation. Classical work in this line of research, e.g., Bradtke & Barto (1996); Melo & Ribeiro (2007) provide algorithms, however, with no polynomial-time sample efficiency guarantees. In the presence of a simulator Yang & Wang (2020) provide a sample-efficient algorithm under linear function approximation. For the linear MDP assumption studied in this paper, our algorithms build on the seminal work of (Jin et al., 2020), that present an efficient (i.e., no-regret) algorithm. This research was further extended to kernel and neural function approximation in the recent work of Yang et al. (2020); Wang et al. (2020). Other approaches in this approximation setting are either computationally intractable (Krishnamurthy et al., 2016; Dann et al., 2018; Dong et al., 2020) or require strong assumptions on the transition model (Wen & Van Roy, 2017).

Cooperative Multi-Agent Reinforcement Learning. Cooperative multi-agent reinforcement learning has a very large body of related work, beginning from classical algorithms in the *fully*-cooperative setting (Boutilier, 1996), i.e., when all agents share identical reward functions. This setting has been explored as multi-agent MDPs in the AI community (Lauer & Riedmiller, 2000; Boutilier, 1996) and as *team Markov games* in the control community (Yoshikawa, 1978; Wang & Sandholm, 2003). However, the more general *heterogeneous* reward setting considered in our work, where each agent may have unique rewards, corresponds to the *team average* games studied previously (Kar et al., 2013; Zhang et al., 2018b,a). While some of these approaches do provide tractable algorithms that are decentralized and convergent, none consider the setting of linear MDPs with polynomial regret. Moreover, as pointed out in prior work (e.g., as studied in Szepesvári & Littman (1999)), a centralized agent controlling each agent can converge to the optimal joint policy, which leaves only the sparse communication theoretically interesting. In our paper, however, we study a more general form of regret in order to discover multiple policies on the *Pareto frontier*, instead of the single policy that maximizes team-average reward. We refer the readers to the illuminating survey paper by Zhang et al. (2019) for a detailed overview of algorithms in this setting.

Parallel and Federated Reinforcement Learning. Parallel reinforcement learning is a very relevant practical setting for reinforcement learning in large-scale and distributed systems, studied first in (Kretchmar, 2002). A variant of the SARSA was presented for parallel RL in Grounds & Kudenko (2005), that provides an efficient algorithm but with no regret guarantees. Modern deep-learning based approaches (with no regret guarantees) have been studied recently as well (e.g., Clemente et al. (2017); Espeholt et al. (2018); Horgan et al. (2018); Nair et al. (2015)). In the federated setting, which corresponds to a decentralized variant of parallel reinforcement learning, there has been recent interested from application domains as well (Yu et al., 2020; Zhuo et al., 2019).

Multi-Objective Sequential Decision-Making. Our algorithms for multi-agent MDP build on recent work in multi-objective sequential decision-making. We extend the framework of obtaining Pareto-optimal *reinforcement learning* policies from multi-objective optimization, presented in the work of (Paria et al., 2020). We utilize a novel vector-valued noise concentration result from Chowdhury & Gopalan (2020), which is an extension of the self-normalized martingale concentration for scalar noise presented in (Chowdhury & Gopalan, 2017), adapted to multi-objective gaussian process optimization. The framework of scalarizations has been studied both in the context of gaussian process optimization (Knowles, 2006; Zhang & Li, 2007; Zhang et al., 2009) and multi-objective reinforcement learning (Van Moffaert & Nowé, 2014).

Notation. We denote vectors by lowercase solid letters, i.e., $\mathbf{x}$, matrices by uppercase solid letters $\mathbf{X}$, and sets by calligraphic letters, i.e., $\mathcal{X}$. We denote the $\mathbf{S}-$ellipsoid norm of a vector $\mathbf{x}$ as $\|\mathbf{x}\|_{\mathbf{S}} = \sqrt{\mathbf{x}^\top \mathbf{S}\mathbf{x}}$. We denote the interval $a, ..., b$ for $b \geq a$ by $[a, b]$ and by $[b]$ when $a = 1$.

3 CoopLSVI for Parallel MDPs

3.1 Parallel Markov Decision Processes

Parallel MDPs (PMDPs, (Sucar, 2007; Kretchmar, 2002)) are a set of discrete time Markov decision processes that are executed in parallel, where a different agent interacts with any single MDP within the parallel MDP. Each agent interacts with their respective MDPs, each with identical (but disjoint) action and state spaces, but possibly unique reward functions and transition probabilities. We have a group of M agents (denoted by $\mathcal{M}$), where the MDP for any agent $m \in \mathcal{M}$ is given by $\text{MDP}(\mathcal{S}, \mathcal{A}, H, \mathbb{P}_m, \mathbf{r}_m)$, where the state and action spaces are given by $\mathcal{S}$ and $\mathcal{A}$ respectively, the reward functions $\mathbf{r}_m = \{r_{m,h}\}_{h\in[H]}, r_{m,h} : \mathcal{S} \times \mathcal{A} \to [0,1]$[2], and transition probabilities $\mathbb{P}_m = \{\mathbb{P}_{m,h}\}_{h\in[H]}, \mathbb{P}_{m,h} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$, i.e., $\mathbb{P}_{m,h}(x'|x,a)$ denotes the probability of the agent moving to state x' if at step h it selects action a from state x. We assume that $\mathcal{S}$ is measurable with possibly infinite elements, and that $\mathcal{A}$ is finite with some size A. For any agent m, the policy π_m is a set of H functions $\pi_m = \{\pi_{m,h}\}_{m\in[M]}, \pi_{m,h} : \mathcal{S} \to \mathcal{A}$ such that $\sum_{a\in\mathcal{A}} \pi_{m,h}(a|x) = 1 \ \forall \ x \in \mathcal{S}$ and $\pi_{m,h}(a|x)$ is the probability of agent m taking action a from state x at step h.

The problem proceeds as follows. At every episode $t = 1, 2, ...,$ each agent $m \in \mathcal{M}$ fixes a policy $\pi_m^t = \{\pi_{m,h}^t\}_{h\in[H]}$, and starts in an initial state $x_{m,1}^t$ picked arbitrarily by the environment. For each step $h \in [H]$ of the episode, each agent observes its state $x_{m,h}^t$, selects an action $a_{m,h}^t \sim \pi_{m,h}^t(\cdot|x_{m,h}^t)$, obtains a reward $r_{m,h}(x_{m,h}^t, a_{m,h}^t)$, and transitions to state $x_{m,h+1}^t$ sampled according to $\mathbb{P}_{m,h}(\cdot|x_{m,h}^t, a_{m,h}^t)$. The episode terminates at step $H+1$ where agents receive 0 reward. After termination, the agents can communicate among themselves via a server, if required. The performance of any policy π in the m^{th} MDP is measured by the value function $V_{m,h}^\pi(x) : \mathcal{S} \to \mathbb{R}$, defined $\forall \ x \in \mathcal{S}, h \in [H], m \in \mathcal{M}$ as,

$$V_{m,h}^\pi(x) \triangleq \mathbb{E}_\pi \left[\sum_{i=h}^{H} r_{m,i}(x_i, a_i) \ \middle| \ x_{m,h} = x \right].$$

The expectation is taken with respect to the random trajectory followed by the agent in the m^{th} MDP under policy π. A related function $Q_{m,h}^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ determines the total expected reward from any action-state pair at step h for the m^{th} MDP for any state $x \in \mathcal{S}$ and action $a \in \mathcal{A}$:

$$Q_{m,h}^\pi(x,a) \triangleq \mathbb{E}_\pi \left[\sum_{i=h}^{H} r_{m,i}(x_i, a_i) \ \middle| \ (x_{m,h}, a_{m,h}) = (x,a) \right].$$

Let $\pi_m^\star$ denote the optimal policy for the m^{th} MDP, i.e., the policy that gives the maximum value, $V_{m,h}^\star(x) = \sup_\pi V_{m,h}^\pi(x)$, for all $x \in \mathcal{S}, h \in [H]$. We can see that with the current set of assumptions, the optimal policy for each agent is possibly unique. For T episodes, the cumulative *group* regret (in expectation), is defined as,

$$\mathfrak{R}(T) \triangleq \sum_{m\in\mathcal{M}} \sum_{t=1}^{T} \left[V_{m,1}^\star(x_{m,1}^t) - V_{m,1}^{\pi_{m,t}}(x_{m,1}^t) \right].$$

3.2 Cooperative Least-Squares Value Iteration

Our algorithms are a cooperative variant of linear least-squares value iteration with optimism (Jin et al., 2020). The prmary motivation behind our algorithm design is to strategically allow for communication between the agents such that with minimal overhead, we achieve a regret close to

[2]We consider $r_{m,h}$ to be deterministic and bounded for simplicity. Our results can easily be extended to random rewards with sub-Gaussian densities.

the single-agent case. Value iteration proceeds by obtaining the optimal Q-values $\{Q^{\star}_{m,h}\}_{h\in[H],m\in\mathcal{M}}$ by recursively applying the Bellman equation. Specifically, each agent $m \in \mathcal{M}$ constructs a sequence of action-value functions $\{Q_{m,h}\}_{h\in[H]}$ as, for each $x \in \mathcal{S}, a \in \mathcal{A}, m \in \mathcal{M}$,

$$Q_{m,h}(x,a) \leftarrow [r_{m,h} + \mathbb{P}_h V_{m,h+1}](x,a),\ V_{m,h+1}(x,a) \leftarrow \max_{a'\in\mathcal{A}} Q_{m,h+1}(x,a').$$

Where $\mathbb{P}_h V(x,a) = \mathbb{E}_{x'}[V(x')\mathbb{P}(x'|x,a)]$. Our approach is to solve a linear least-squares regression usingproblem based on *multi-agent* historical data. For any function class $\mathcal{F}$, assume that any agent $m \in [M]$ has observed k transition tuples $\{x^{\tau}_h, a^{\tau}_h, x^{\tau}_{h+1}\}_{\tau\in[k]}$ for any step $h \in [H]$. Then, the agent estimates the optimal Q-value for any step by LSVI, solving the regularized least-squares regression:

$$\widehat{Q}^t_{m,h} \leftarrow \arg\min_{f\in\mathcal{F}} \left\{ \sum_{\tau\in[k]} \left[r_h(x^{\tau}_h, a^{\tau}_h) + V^t_{m,h+1}(x^{\tau}_{h+1}) - f(x^{\tau}_h, a^{\tau}_h) \right]^2 + \|f\|^2 \right\}.$$

Here, the targets $y^{\tau}_h = r_h(x^{\tau}_h, a^{\tau}_h) + V^t_{m,h+1}(x^{\tau}_{h+1})$ denote the empirical value from specific transitions possessed by the agent, and $\|f\|$ denotes an appropriate regularization term based on the capacity of f and the class $\mathcal{F}$. To foster exploration, an additional bonus $\sigma^t_{m,h} : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ term is added that is inspired by the principle of optimism in the face of uncertainty, giving the final Q-value as,

$$Q^t_{m,h}(\cdot,\cdot) = \min\left\{ \left[\widehat{Q}^t_{m,h} + \beta^t_{m,h}\sigma^t_{m,h} \right](\cdot,\cdot), H-h+1 \right\}, \tag{1}$$

$$V^t_{m,h}(\cdot) = \max_{a\in\mathcal{A}} Q^t_{m,h}(\cdot,a). \tag{2}$$

Here $\{\beta^t_{m,h}\}_{t\in[T],m\in\mathcal{M}}$ is an appropriately selected sequence. For episode t, we denote $\boldsymbol{\pi}^t = \{\pi^t_m\}_{m\in\mathcal{M}}$ as the (joint) greedy policy with respect to the Q-values $\{Q^t_{m,h}\}_{h\in[H]}$ for each agent. While this describes the multiagent LSVI algorithm abstractly for any general function class $\mathcal{F}$, we first describe the *homogeneous* parallel setting, where we assume $\mathcal{F}$ to be linear in d dimensions, called the *linear* MDP assumption (Bradtke & Barto, 1996; Jin et al., 2020; Melo & Ribeiro, 2007).

Definition 1 (Linear MDP, Jin et al. (2020))**.** *An MDP$(\mathcal{S},\mathcal{A},H,\mathbb{P},R)$ 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 (signed) measures $\boldsymbol{\mu}_h = (\mu^1_h, ..., \mu^d_h)$ over $\mathcal{S}$ and an unknown vector $\boldsymbol{\theta}_h \in \mathbb{R}^d$ such that for any $(x,a) \in \mathcal{S}\times\mathcal{A}$,*

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

We assume, without loss of generality, $\|\boldsymbol{\phi}(x,a)\| \le 1\ \forall\ (x,a) \in \mathcal{S}\times\mathcal{A}$; $\max\{\|\boldsymbol{\mu}_h(\mathcal{S})\|, \|\boldsymbol{\theta}_h\|\} \le \sqrt{d}$.

3.3 Warm Up: Homogenous Parallel MDPs

As a warm up, we describe `CoopLSVI` in the homogenous setting for simplicity first. In this case, the transition functions $\mathbb{P}_{m,h}$ and reward functions $r_{m,h}$ are identical for all agents and can be given by $\mathbb{P}_h$ and r_h respectively for any episode $h \in [H]$. This environment corresponds to distributed applications, e.g., federated and concurrent reinforcement learning. Corresponding to Eq. 3.2, we assume $\mathcal{F}$ to be the class of linear functions in d dimensions over the feature map $\boldsymbol{\phi}$, i.e., $f(\cdot) = \mathbf{w}^{\top}\boldsymbol{\phi}(\cdot), \mathbf{w} \in \mathbb{R}^d$, and set the ridge norm $\|\mathbf{w}\|^2_2$ as the regularizer. Furthermore, we fix a threshold constant S that determines the amount of communication between the agents. The algorithm is summarized in Algorithm 1. In a nutshell, the algorithm operates by each agent executing a local linear LSVI and then synchronizing observations between other agents if the threshold condition is met every episode. Specifically, for each $t \in [T]$, each agent $m \in \mathcal{M}$ obtains a sequence of value functions $\{Q^t_{m,h}\}_{h\in[H]}$ by iteratively performing linear least-squares ridge regression from the *multi-agent* history available from the previous $t-1$ episodes. Assume that for any step h, the

Algorithm 1 `Coop-LSVI`

Input: $T, \boldsymbol{\phi}, H, S$, sequence $\beta_h = \{(\beta^t_{m,h})_{m,t}\}$.
Initialize: $\mathbf{S}^t_{m,h}, \delta\mathbf{S}^t_{m,h} = \mathbf{0}, \mathcal{U}^m_h, \mathcal{W}^m_h = \emptyset$.
for episode $t = 1, 2, ..., T$ **do**
  for agent $m \in \mathcal{M}$ **do**
    Receive initial state $x^t_{m,1}$.
    Set $V^t_{m,H+1}(\cdot) \leftarrow 0$.
    for step $h = H, ..., 1$ **do**
      Compute $\boldsymbol{\Lambda}^t_{m,h} \leftarrow \mathbf{S}^t_{m,h} + \delta\mathbf{S}^t_{m,h}$.
      Compute $\widehat{Q}^t_{m,h}$ and $\sigma^t_{m,h}$ (Eqns. 5 and 6).
      Compute $Q^t_{m,h}(\cdot,\cdot)$ (Eqn. 1)
      Set $V^t_{m,h}(\cdot) \leftarrow \max_{a\in\mathcal{A}} Q^t_{m,h}(\cdot, a)$.
    end for
    for step $h = 1, ..., H$ **do**
      Take action $a^t_{m,h} \leftarrow \arg\max_{a\in\mathcal{A}} Q^t_{m,h}(x^t_{m,h}, a)$.
      Observe $r^t_{m,h}, x^t_{m,h+1}$.
      Update $\delta\mathbf{S}^t_{m,h} \leftarrow \delta\mathbf{S}^t_{m,h} + \boldsymbol{\phi}(z^t_{m,h})\boldsymbol{\phi}(z^t_{m,h})^\top$.
      Update $\mathcal{W}^m_h \leftarrow \mathcal{W}^m_h \cup (m, x, a, x')$.
      if $\log \frac{\det(\mathbf{S}^t_{m,h}+\delta\mathbf{S}^t_{m,h}+\lambda\mathbf{I})}{\det(\mathbf{S}^t_{m,h}+\lambda\mathbf{I})} > \frac{S}{\Delta t_{m,h}}$ **then**
        SYNCHRONIZE$\leftarrow$ TRUE.
      end if
    end for
  end for
  if SYNCHRONIZE **then**
    for step $h = H, ..., 1$ **do**
      [$\forall$ AGENTS] Send $\mathcal{W}^h_m \rightarrow$SERVER.
      [SERVER] Aggregate $\mathcal{W}^h \rightarrow \cup_{m\in\mathcal{M}}\mathcal{W}^m_h$.
      [SERVER] Communicate $\mathcal{W}^h$ to each agent.
      [$\forall$ AGENTS] Set $\delta\mathbf{S}^t_h \leftarrow 0, \mathcal{W}^m_h \leftarrow \emptyset$.
      [$\forall$ AGENTS] Set $\mathbf{S}^t_h \leftarrow \mathbf{S}^t_h + \sum_{z\in\mathcal{W}^h} \boldsymbol{\phi}(z)\boldsymbol{\phi}(z)^\top$.
      [$\forall$ AGENTS] Set $\mathcal{U}^m_h \leftarrow \mathcal{U}^m_h \cup \mathcal{W}^m_h$
    end for
  end if
end for

previous synchronization round occured after episode k_t. Then, the set of transitions available to agent $m \in \mathcal{M}$ for any step h before episode t can be given by,

$$\mathcal{U}^m_h(t) = \left\{\cup\left(x^\tau_{n,h}, a^\tau_{n,h}, x^\tau_{n,h+1}\right)_{n\in\mathcal{M},\tau\in[k_t]}\right\} \bigcup \left\{\cup\left(x^\tau_{m,h}, a^\tau_{m,h}, x^\tau_{m,h+1}\right)_{\tau=k_t+1}^{t-1}\right\}.$$

Let $\psi^m_h(t)$ be an ordering of $\mathcal{U}^m_h(t)$, and $U^m_h(t) = |\mathcal{U}^m_h(t)|$. We have that $\mathcal{U}^m_h$ is a set of $U^m_h(t)$ elements, where each element is a set of the form (n, x, a, x'), where $n \in \mathcal{M}$ specifies the agent, and (x, a, x') specifies a transition occuring at step h. Each agent m first sets $Q^t_{m,H+1}$ to be $\mathbf{0}_d$, and for $h = H, ..., 1$, iteratively solves H regressions:

$$\widehat{Q}^t_{m,h} \leftarrow \arg\min_{\mathbf{w}} \left\{ \sum_{(n,x,a,x')\in\mathcal{U}^m_h(t)} \left[r_h(x,a) + V^t_{m,h+1}(x') - \mathbf{w}^\top\boldsymbol{\phi}(x,a)\right]^2 + \lambda\|\mathbf{w}\|_2^2 \right\}. \tag{3}$$

Here $\lambda > 0$ is a regularizer. Next, $Q^t_{m,h}$ and $V^t_{m,h}$ are obtained via Equations 1 and 2. We denote the targets $y_\tau = y_{m,h}(x_\tau, a_\tau, a'_\tau)$, and the features $\boldsymbol{\phi}_\tau = \boldsymbol{\phi}(x_\tau, a_\tau), \forall\ \tau \in \psi^m_h(t)$. Next, we denote

the covariance $\mathbf{\Lambda}^t_{m,h} \in \mathbb{R}^{d\times d}$ and bias $\mathbf{u}^t_{m,h} \in \mathbb{R}^d$ as,

$$\mathbf{\Lambda}^t_{m,h} = \sum_{\tau\in\psi^m_h(t)} \left[\boldsymbol{\phi}_\tau \boldsymbol{\phi}_\tau^\top\right] + \lambda \mathbf{I}_d, \mathbf{u}^t_{m,h} = \sum_{\tau\in\psi^m_h(t)} \left[\boldsymbol{\phi}_\tau y_\tau\right] \tag{4}$$

Denote $\mathcal{Z} = \mathcal{S}\times\mathcal{A}$ and $z = (x, a)$. We have $\forall\ z \in \mathcal{Z}$,

$$\widehat{Q}^t_{m,h}(z) = \boldsymbol{\phi}(z)^\top \left(\mathbf{\Lambda}^t_{m,h}\right)^{-1} \mathbf{u}^t_{m,h}. \tag{5}$$

And correspondingly, the UCB bonus is given as,

$$\sigma^t_{m,h}(z) = \|\boldsymbol{\phi}(z)\|_{(\mathbf{\Lambda}^t_{m,h})^{-1}} = \sqrt{\boldsymbol{\phi}(z)^\top (\mathbf{\Lambda}^t_{m,h})^{-1} \boldsymbol{\phi}(z)}. \tag{6}$$

This exploration bonus is similar to that of Gaussian process (GP) optimization (Srinivas et al., 2009) and linear bandit (Abbasi-Yadkori et al., 2011) algorithms, and it can be interpreted as the posterior variance of a Gaussian process regression. The motivation for adding the UCB term is similar to that in the bandit and GP case, to adequately overestimate the uncertainty in the ridge regression solution. When $\beta^t_{m,h}$ is appropriately selected, the $Q-$values overestimate the optimal Q-values with high probability, which is the foundation to bounding the regret.

The algorithm essentially involves this synchronization of transitions at carefully chosen instances. To achieve this, each agent maintains two sets of parameters. The first is $\mathbf{S}^t_{m,h}$ which refers to the parameters that have been updated with the other agents via the synchronization, and the second is $\delta\mathbf{S}^t_{m,h}$, which refers to the parameters that are not synchronized. Now, in each step of any episode t, after computing $Q-$values (Eq. 1), each agent executes the greedy policy with respect to the $Q-$values, i.e., $a^t_{m,h} = \arg\max_{a\in\mathcal{A}} Q^t_{m,h}(x^t_{m,h}, a)$, and updates the unsyncrhonized parameters $\delta\mathbf{S}^t_{m,h}$. If any agent's new unsynchronized parameters satisfy the determinant condition with threshold S, i.e., if

$$\log \frac{\det\left(\mathbf{S}^t_{m,h} + \delta\mathbf{S}^t_{m,h} + \lambda\mathbf{I}_d\right)}{\det\left(\mathbf{S}^t_{m,h} + \lambda\mathbf{I}_d\right)} \geq \frac{S}{(t-k_t)}, \tag{7}$$

then the agent signals a synchronization with the server, and messages are exchanged. Here k_t denotes the episode after which the previous round of synchronization took place (step 19 of Algorithm 1). We can demonstrate the complexity of communication as a function of S.

Lemma 1 (Communication Complexity)**.** *If Algorithm 1 is run with threshold S, then the total number of episodes with communication $n \leq 2H\sqrt{d(T/S)\log(MT)} + 4H$.*

The above result demonstrates that when $S = o(T)$, it is possible to ensure that the agents communicate only in a constant number of episodes, regardless of T. We now present the regret guarantee for the homogenous setting.

Theorem 1 (Homogenous Regret)**.** *Algorithm 1 when run on M agents with communication threshold S, $\beta_t = \mathcal{O}(H\sqrt{d\log(tMH)})$ and $\lambda = 1$ obtains the following cumulative regret after T episodes, with probability at least $1-\alpha$,*

$$\mathfrak{R}(T) = \widetilde{\mathcal{O}}\left(dH^2\left(dM\sqrt{S} + \sqrt{dMT}\right)\sqrt{\log\left(\frac{1}{\alpha}\right)}\right).$$

Remark 1 (Regret Optimality)**.** Theorem 1 claims that appropriately chosen β and λ ensures sublinear group regret. Similar to the single-agent analysis in linear (Jin et al., 2020) and kernel (Yang

et al., 2020) function approximation settings, our analysis admits a dependence on the (linear) function class via the $\ell_\infty-$covering number, which we simplify in Theorem 1 by selecting appropriate values of the parameters. Generally, the regret scales as $\mathcal{O}\left(H^2\left(M\sqrt{S}+\sqrt{MT\log\mathcal{N}_\infty(\epsilon^\star)}\right)\sqrt{\log\left(\frac{1}{\alpha}\right)}\right)$, where $\mathcal{N}_\infty(\epsilon)$ is the ϵ-covering number of the set of linear value functions under the ℓ_∞ norm, and $\epsilon^\star = \mathcal{O}(dH/T)$. We elaborate on this connection in the full proof.

Remark 2 (Multi-Agent Analysis)**.** The aspect central to the multi-agent analysis is the dependence on the communication parameter S. If the agents communicate every round, i.e., $S = \mathcal{O}(1)$, we observe that the cumulative regret is $\widetilde{\mathcal{O}}(d^{\frac{3}{2}}H^2\sqrt{MT})$, matching the centralized setting. With no communication, the agents simply operate independently and the regret incurred is $\widetilde{\mathcal{O}}(M\sqrt{T})$, matching the group regret incurred by isolated agents. Furthermore, with $S = \mathcal{O}(T\log(MT)/dM^2)$ we observe that with a total of $\mathcal{O}(dHM^3)$ episodes with communication, we recover a group regret of $\mathcal{O}(d^{\frac{3}{2}}H^2\sqrt{MT}(\log MT))$, which matches the optimal rate (in terms of T) up to logarithmic factors.

3.4 Heterogeneous Parallel MDPs

The above algorithm assumes that the underlying MDPs each agent interacts with are identical. We now present variants of this algorithm that generalizes to the case when the underlying MDPs are different. Note that even for heterogeneous settings, our algorithms require assumptions on the nature of heterogeneity in order to benefit from cooperative estimation.

3.4.1 Small Heterogeneity

The first heterogeneous setting we consider is when the deviations between MDPs are much smaller than the horizon T, which allows Algorithm 1 to be no-regret as long as an upper bound on the heterogeneity is known. We first present this "small deviation" assumption.

Assumption 1. *For any $\xi = o(T^{-\alpha}) < 1, \alpha > 0$, a parallel MDP setting demonstrates "small deviations" if for any $m, m' \in \mathcal{M}$, the corresponding linear MDPs defined in Definition 1 obey the following for all $(x, a) \in \mathcal{S} \times \mathcal{A}$:*

$$\|(\mathbb{P}_{m,h} - \mathbb{P}_{m',h})(\cdot|x,a)\|_{\mathrm{TV}} \leq \xi, \text{ and } |(r_{m,h} - r_{m',h})(x,a)| \leq \xi.$$

Under this assumption, when an upper bound on ξ is known, we do not have to modify Algorithm 1 as the confidence intervals employed by `CoopLSVI` are robust to small deviations. We formalize this with the following regret bound.

Theorem 2. *Algorithm 1 when run on M agents with parameter S in the small deviation setting (Assumption 1), with $\beta_t = \mathcal{O}(H\sqrt{d\log(tMH)}+\xi\sqrt{dMT})$ and $\lambda = 1$ obtains the following cumulative regret after T episodes, with probability at least $1-\alpha$,*

$$\mathfrak{R}(T) = \widetilde{\mathcal{O}}\left(dH^2\left(dM\sqrt{S}+\sqrt{dMT}\right)\left(\sqrt{\log\left(\frac{1}{\alpha}\right)}+2\xi\sqrt{dMT}\right)\right).$$

Remark 3 (Comparison with Misspecification)**.** While this demonstrates that `CoopLSVI` is robust to small deviations in the different MDPs, the analysis can be extended to the case when the MDPs are "approximately" linear, as done in Jin et al. (2020) (Theorem 3.2). A key distinction in the above result and the standard bound in the misspecification setting is that in the general misspecified linear MDP there are two aspects to the anlaysis - the first being the (adversarial) error introduced from the linear approximation, and the second being the error introduced by executing a policy following the misspecified linear approximation. In our case, the second term does not exist as the policy is valid within each agents' own MDP, but the misspecification error still remains.

3.4.2 Large Heterogeneity

For the large heterogeneity case, in order to transfer knowledge from a different agents' MDP, we assume that each agent $m \in \mathcal{M}$ possesses an additional contextual description $\boldsymbol{\kappa}(m) \in \mathbb{R}^k$ that describes the heterogeneity linearly. We state this formally below.

Definition 2 (Heterogenous Linear MDP)**.** *A heterogeneous parallel MDP$(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, R)$ is a set of linear MDPs with two feature maps $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and $\boldsymbol{\kappa} : \mathcal{M} \to \mathbb{R}^k$, if for any $h \in [H]$, there exist d unknown (signed) measures $\boldsymbol{\mu}_h = (\mu_h^1, ..., \mu_h^d)$ over $\mathcal{S}$, an unknown vector $\boldsymbol{\theta}_h \in \mathbb{R}^d$, k unknown (signed) measures $\boldsymbol{\nu}_h = (\nu_h^1, ..., \nu_h^k)$ over $\mathcal{S}$ and an unknown vector $\boldsymbol{\alpha}_h \in \mathbb{R}^k$ such that for any $(x, a) \in \mathcal{S} \times \mathcal{A}$ and $m \in \mathcal{M}$,*

$$\mathbb{P}_{m,h}(\cdot|x, a) = \begin{bmatrix} \boldsymbol{\phi}(x, a) \\ \boldsymbol{\kappa}(m) \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{\mu}_h(\cdot) \\ \boldsymbol{\nu}_h(\cdot) \end{bmatrix}, r_{m,h}(x, a) = \begin{bmatrix} \boldsymbol{\phi}(x, a) \\ \boldsymbol{\kappa}(m) \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{\theta}_h \\ \boldsymbol{\alpha}_h \end{bmatrix}.$$

We denote the combined features via the shorthand:

$$\widetilde{\boldsymbol{\phi}}(m, x, a) = \left[\boldsymbol{\phi}(x, a)^\top, \boldsymbol{\kappa}(m)^\top\right]^\top, \widetilde{\boldsymbol{\mu}}_h(\cdot) = \left[\boldsymbol{\mu}_h(\cdot)^\top, \boldsymbol{\nu}_h(\cdot)^\top\right]^\top, \widetilde{\boldsymbol{\theta}}_h = \left[\boldsymbol{\theta}_h^\top, \boldsymbol{\alpha}_h^\top\right]^\top.$$

We assume, WLOG, that $\|\widetilde{\boldsymbol{\phi}}(m, x, a)\| \leq 1 \ \forall \ (m, x, a) \in \mathcal{M} \times \mathcal{S} \times \mathcal{A}$, $\max\{\|\boldsymbol{\mu}_h(\mathcal{S})\|, \|\boldsymbol{\theta}_h\|\} \leq \sqrt{d}$, and $\max\{\|\boldsymbol{\nu}_h(\mathcal{S})\|, \|\boldsymbol{\alpha}_h\|\} \leq \sqrt{k}$.

The above assumption is identical to the linear MDP assumption (Definition 1) except that the transition and reward functions are also a function of contextual information possessed by the agent about its environment. Such information is usually present in many applications, e.g., as demonstrated in Krause & Ong (2011) for the contextual bandit. Concretely, we assume that for each MDP, the discrepancies between both the transition and reward functions can be explained as a linear function of the underlying agent-specific contexts, i.e., $\boldsymbol{\kappa}$, which is independent of the state and action pair (x, a)[3]. Then, each agent predicts an *agent-specific* Q and value function.

Specifically, for each $t \in [T]$, each agent $m \in \mathcal{M}$ obtains a sequence of value functions $\{Q_{m,h}^t\}_{h \in [H]}$ by iteratively performing linear least-squares ridge regression from the *multi-agent* history available from the previous $t-1$ episodes, but in contrast to the homogenous case, it now learns a $Q-$function over $\mathcal{M} \times \mathcal{S} \times \mathcal{A}$ and value function over $\mathcal{M} \times \mathcal{A}$. Each agent m first sets $Q_{m,H+1}^t$ to be a zero function, and for any $h \in [H]$, solves the regression problem in $\mathbb{R}^{d+k}$ to obtain $Q-$values.

$$\widehat{Q}_{m,h}^t \leftarrow \underset{\mathbf{w} \in \mathbb{R}^{d+k}}{\arg\min} \left\{ \sum_{(n,x,a,x') \in \mathcal{U}_h^m(t)} \left[r_{n,h}(x, a) + V_{m,h+1}^t(n, x') - \mathbf{w}^\top \widetilde{\boldsymbol{\phi}}(n, x, a) \right]^2 + \lambda \|\mathbf{w}\|_2^2 \right\}. \tag{8}$$

Here, $\lambda > 0$ is a regularizer, and $n \in \mathcal{M}$ denotes the agent whose Q function the agent m is estimating, and $V_{m,h+1}^t(n, x) = \max_{a \in \mathcal{A}} Q(n, x, a)$, where the Q values are given by, for any n, x, a,

$$Q(n, x, a) = \widehat{Q}_{m,h}^t(n, x, a) + \beta_{m,h}^t \|\widetilde{\boldsymbol{\phi}}(n, x, a)\|_{\widetilde{\boldsymbol{\Lambda}}_{m,h}^{t}}.$$

Here as well, we have the least-squares solution described as follows. Let $\psi_h^m(t)$ be an ordering of $\mathcal{U}_h^m(t)$, and $U_h^m(t) = |\mathcal{U}_h^m(t)|$. Then, we denote the targets $y_\tau = y_{m,h}(n_\tau, x_\tau, a_\tau, a'_\tau)$, and the features $\widetilde{\boldsymbol{\phi}}_\tau = \widetilde{\boldsymbol{\phi}}(n_\tau, x_\tau, a_\tau), \forall \ \tau \in \psi_h^m(t)$. Next, we denote the covariance $\widetilde{\boldsymbol{\Lambda}}_{m,h}^t \in \mathbb{R}^{d+k \times d+k}$ and bias $\widetilde{\mathbf{u}}_{m,h}^t \in \mathbb{R}^{d+k}$ as,

$$\widetilde{\boldsymbol{\Lambda}}_{m,h}^t = \sum_{\tau \in \psi_h^m(t)} \left[\widetilde{\boldsymbol{\phi}}_\tau \widetilde{\boldsymbol{\phi}}_\tau^\top \right] + \lambda \mathbf{I}_{d+k}, \widetilde{\mathbf{u}}_{m,h}^t = \sum_{\tau \in \psi_h^m(t)} \left[\widetilde{\boldsymbol{\phi}}_\tau y_\tau \right]. \tag{9}$$

[3]Intricate models can be assumed that exploit interdependence in a sophisticated manner (see, e.g., Dubey & Pentland (2020a); Deshmukh et al. (2017)), however, we leave that for future work.

Let $\widetilde{\mathcal{Z}} = \mathcal{M} \times \mathcal{S} \times \mathcal{A}$ and $\tilde{z} = (n, x, a)$. We have $\forall\ \tilde{z} \in \widetilde{\mathcal{Z}}$,

$$\widehat{Q}^t_{m,h}(\tilde{z}) = \widetilde{\boldsymbol{\phi}}(\tilde{z})^\top \left(\widetilde{\boldsymbol{\Lambda}}^t_{m,h}\right)^{-1} \widetilde{\mathbf{u}}^t_{m,h}. \tag{10}$$

At any episode t and step h, each agent m then follows the greedy policy with respect to $Q(m, x^t_{m,h})$. The remainder of the algorithm is identical to Algorithm 1, and is presented in Algorithm 3 in the appendix. To present the regret bound, we first define coefficient of heterogeneity χ, and then present the regret bound in terms of this coefficient.

Definition 3. *For the parallel MDP defined in Definition 2, let* $\mathbf{K}^\kappa_h = \left[\boldsymbol{\nu}_h(m)^\top \boldsymbol{\nu}_h(m')\right]_{m,m' \in \mathcal{M}}$ *be the Gram matrix of agent-specific contexts. The coefficient of heterogeneity is defined as* $\chi = \max_{h \in [H]} \mathit{rank}(\mathbf{K}^\kappa_h) \leq k$.

Theorem 3. *Algorithm 3 when run on M agents with parameter S in the heterogeneous setting (Definition 2), with $\beta_t = \mathcal{O}(H\sqrt{(d+k)\log(tMH)})$ and $\lambda = 1$ obtains the following cumulative regret after T episodes, with probability at least $1-\alpha$,*

$$\mathfrak{R}(T) = \widetilde{\mathcal{O}}\left((d+k)H^2\left(M(d+\chi)\sqrt{S} + \sqrt{(d+\chi)MT}\right)\sqrt{\log\left(\frac{1}{\alpha}\right)}\right).$$

Remark 4 (Optimality Discussion)**.** Heterogeneous `CoopLSVI` regret is bounded by the similarity in the agents' MDPs. In the case when the agents' have identical MDPs, $\chi = 1$, which implies that the heterogenous variant has a worse regret by a factor of $(1 + \frac{k}{d})\sqrt{1 + \frac{1}{d}}$, which arises from the fact that we use a model that lies in $\mathbb{R}^{d+k}$. Nevertheless, this suboptimality is indeed an artifact of our regret analysis, particularly introduced by the covering number of linear functions in $\mathbb{R}^{d+k}$, and future work can address modifications to ensure tightness. Alternatively, in the worst case, $\chi = k$, which matches the linear parallel MDP in $d + k$ dimensions, which ensures that no suboptimality has been introduced by the heterogeneous analysis. Under this model however, one can only observe improvements when $k = o(\sqrt{M})$ suffices for modeling the heterogeneity within the parallel MDP.

4 `CoopLSVI` for Multiagent MDPs

The next environment we consider is the simultaneous-move multi-agent MDP (Boutilier, 1996), which is an extension of an MDP to multiple agents (Xie et al., 2020).

4.1 Heterogeneous Multi-agent MDPs

A multi-agent MDP (MMDP) can be formally described as MMDP$(\mathcal{S}, \mathcal{M}, \mathcal{A}, H, \mathbb{P}, \mathbf{R})$, where the set of agents $\mathcal{M}$ is finite and countable with cardinality M, the state and action spaces are factorized as $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2 \times \ldots \mathcal{S}_M$ and $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times \ldots \mathcal{A}_M$, where $\mathcal{S}_i$ and $\mathcal{A}_i$ denote the individual state and action space for agent i respectively. Furthermore, the transition matrix $\mathbb{P} = \{\mathbb{P}_h\}_{h \in [H]}$ depends on the joint action-state configuration for all agents, i.e., $\mathbb{P}_h : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$, and so does the (vector-valued) reward function $\mathbf{R} = \{\mathbf{r}_h\}_{h \in [H]}, \mathbf{r}_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^M$.

The MMDP proceeds as follows. In each episode $t \in [T]$ each agent fixes a policy $\pi^t_m = \{\pi^t_{m,h}\}_{h \in [H]}$ in a common initial state $\mathbf{x}^t_1 = \{x^t_{m,1}\}_{h \in [H]}$ picked arbitrarily by the environment. For each step $h \in [H]$ of the episode, each agent observes the overall state $\mathbf{x}^t_h$, selects an individual action $a^t_{m,h} \sim \pi^t_{m,h}(\cdot|\mathbf{x}^t_h)$ (denoted collectively as the joint action $\mathbf{a}^t_h = \{a^t_{m,h}\}_{m \in \mathcal{M}}$), and obtains a reward $r_{m,h}(\mathbf{x}^t_h, \mathbf{a}^t_h)$ (denoted collectively as the joint reward $\mathbf{r}_h(\mathbf{x}^t_h, \mathbf{a}^t_h) = \{r_{m,h}(\mathbf{x}^t_h, \mathbf{a}^t_h)\}_{m \in \mathcal{M}}$). All agents transition subsequently to a new joint state $\mathbf{x}^t_{h+1} = \{x^t_{m,h+1}\}_{m \in \mathcal{M}}$ sampled according to $\mathbb{P}_h(\cdot|\mathbf{x}^t_h, \mathbf{a}^t_h)$. The episode terminates at step $H + 1$ where all agents receive no reward. After termination, the agents communicate by exchanging messages with a server, prior to the next episode.

Let $\boldsymbol{\pi} = \{\boldsymbol{\pi}_h\}_{h\in[H]}, \boldsymbol{\pi}_h = \{\pi_{m,h}\}_{m\in\mathcal{M}}$ denote the joint policy for all M agents. We can define the vector-valued value function over all states $\mathbf{x} \in \mathcal{S}$ for a policy $\boldsymbol{\pi}$ as,

$$\mathbf{V}_h^{\boldsymbol{\pi}}(\mathbf{x}) \triangleq \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{i=h}^{H} \mathbf{r}_i(\mathbf{x}_i, \mathbf{a}_i) \;\middle|\; \mathbf{x}_h = \mathbf{x}\right].$$

One can define the analogous vector-valued Q-function for a policy $\boldsymbol{\pi}$ and any $\mathbf{x} \in \mathcal{S}, \mathbf{a} \in \mathcal{A}, h \in [H]$,

$$\mathbf{Q}_h^{\boldsymbol{\pi}}(\mathbf{x}, \mathbf{a}) \triangleq \mathbf{r}_h(\mathbf{x}, \mathbf{a}) + \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{i=h+1}^{H} \mathbf{r}_i(\mathbf{x}_i, \mathbf{a}_i) \;\middle|\; \mathbf{x}_h = \mathbf{x}, \mathbf{a}_h = \mathbf{a}\right].$$

Without assumptions, a general MMDP represents a wide variety of environments, including competitive stochastic games as well as cooperative games (Shapley, 1953). In this paper, our focus is to consider cooperative learning in the *fully-observable* setting, where the complete state and actions (but *not* rewards) are visible to all agents.

Remark 5 (Multi-agent Environments)**.** MMDPs have been used to model a variety of decision processes, and they are an instance of *stochastic games* (Shapley, 1953), and are most closely related to the general framework for *repeated games* (Myerson, 1982). Repeated games are in turn generalizations of partially observable MDPs (POMDPs, Åström (1965)), and involve a variety of distinct challenges beyond communication and non-stationarity. For these multi-agent environments it is not possible to characterize optimal behavior without global objectives, as, unlike the single-agent setting, achieving a large cumulative reward is typically at odds with individually optimal behavior.

We focus on recovering the set of *cooperative* Pareto-optimal policies, i.e., policies that cannot improve any individual agent's reward without decreasing the reward of the other agents. Formally,

Definition 4 (Pareto domination, Paria et al. (2020))**.** *A (multi-agent) policy $\boldsymbol{\pi}$ Pareto-dominates another policy $\boldsymbol{\pi}'$ if and only if $\mathbf{V}_1^{\boldsymbol{\pi}}(\mathbf{x}) \succeq \mathbf{V}_1^{\boldsymbol{\pi}'}(\mathbf{x}) \forall\ \mathbf{x} \in \mathcal{S}$. Consequently, a policy is Pareto-optimal if it is not Pareto-dominated by any policy. We denote the set of possible policies by $\boldsymbol{\Pi}$, and the set of Pareto-optimal policies by $\boldsymbol{\Pi}^\star$. As an example, it is evident that the policies that maximize any agent's individual reward as well as the average reward are all elements of $\boldsymbol{\Pi}^\star$.*

Our objective is to design an algorithm that recovers $\boldsymbol{\Pi}^\star$ with high probability.

4.2 LSVI with Random Scalarizations

Our approach for multiagent MDPs, inspired by multi-objective optimization, is to utilize the method of *random scalarizations* (Knowles, 2006; Paria et al., 2020), where we define a probability distribution over the Pareto-optimal policies by converting the vector-valued function $\mathbf{V}$ to a parameterized scalar measure. Consider a *scalarization function* $\mathfrak{s}_{\boldsymbol{\upsilon}}(\mathbf{x}) = \boldsymbol{\upsilon}^\top \mathbf{x} : \mathbb{R}^M \to \mathbb{R}$ parameterized by $\boldsymbol{\upsilon}$ belonging to $\boldsymbol{\Upsilon} \subseteq \Delta^M$ (unit simplex in M dimensions). We then have the *scalarized* value function $V_{\boldsymbol{\upsilon},h}^{\boldsymbol{\pi}}(x) : \mathcal{S} \to \mathbb{R}$ and $Q-$function $Q_{\boldsymbol{\upsilon},h}^{\boldsymbol{\pi}} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ for some joint policy $\boldsymbol{\pi}$ as

$$V_{\boldsymbol{\upsilon},h}^{\boldsymbol{\pi}}(\mathbf{x}) \triangleq \mathfrak{s}_{\boldsymbol{\upsilon}}(\mathbf{V}_h^{\boldsymbol{\pi}}(\mathbf{x})) = \boldsymbol{\upsilon}^\top \mathbf{V}_h^{\boldsymbol{\pi}}(\mathbf{x}) \text{ , and } Q_{\boldsymbol{\upsilon},h}^{\boldsymbol{\pi}}(\mathbf{x}, \mathbf{a}) \triangleq \mathfrak{s}_{\boldsymbol{\upsilon}}(\mathbf{Q}_h^{\boldsymbol{\pi}}(\mathbf{x}, \mathbf{a})) = \boldsymbol{\upsilon}^\top \mathbf{Q}_h^{\boldsymbol{\pi}}(\mathbf{x}, \mathbf{a}). \tag{11}$$

Since both $\mathcal{A} = \prod_i \mathcal{A}_i$ and H are finite, there exists an optimal multi-agent policy for any fixed scalarization $\boldsymbol{\upsilon}$, which gives the optimal value $V_{\boldsymbol{\upsilon},h}^\star = \sup_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} V_{\boldsymbol{\upsilon},h}^{\boldsymbol{\pi}}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{S}$ and $h \in [H]$. The primary advantage of considering a scalarized formulation is that for any fixed scalarization parameter, an optimal *joint* policy exists, which coincides with the optimal policy for an MDP defined over the joint action space $\mathcal{S} \times \mathcal{A}$, given as follows.

Proposition 1. *For the scalarized value function given in Equation 11, the Bellman optimality conditions are given as, for all* $h \in [H], \mathbf{x} \in \mathcal{S}, \mathbf{a} \in \mathcal{A}, \boldsymbol{\upsilon} \in \boldsymbol{\Upsilon}$,

$$Q^{\star}_{\boldsymbol{\upsilon},h}(\mathbf{x},\mathbf{a}) = \mathfrak{s}_{\boldsymbol{\upsilon}}\mathbf{r}_h(\mathbf{x},\mathbf{a}) + \mathbb{P}_h V^{\star}_{\boldsymbol{\upsilon},h}(\mathbf{x},\mathbf{a}), V^{\star}_{\boldsymbol{\upsilon},h}(\mathbf{x}) = \max_{\mathbf{a}\in\mathcal{A}} Q^{\star}_{\boldsymbol{\upsilon},h}(\mathbf{x},\mathbf{a}), \text{ and } V^{\star}_{\boldsymbol{\upsilon},H+1}(\mathbf{x}) = 0.$$

The optimal policy for any fixed $\boldsymbol{\upsilon}$ is given by the greedy policy with respect to the Bellman-optimal scalarized Q values. We denote this (unique) optimal policy by $\pi^{\star}_{\boldsymbol{\upsilon}}$. We now demonstrate that each $\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}$ lies in the Pareto frontier.

Proposition 2. *For any parameter* $\boldsymbol{\upsilon} \in \boldsymbol{\Upsilon}$, *the optimal greedy policy* $\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}$ *with respect to the scalarized Q-value that satisfies Proposition 1 lies in the Pareto frontier* $\boldsymbol{\Pi}^{\star}$.

The above result claims that by "projecting" an MMDP to an MDP via scalarization, one can recover a policy on the Pareto frontier. Indeed, if the set of policies $\boldsymbol{\Pi}^{\star}_{\boldsymbol{\Upsilon}} = \{\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}} | \boldsymbol{\upsilon} \in \boldsymbol{\Upsilon}\}$ spans $\boldsymbol{\Pi}^{\star}$, one can recover $\boldsymbol{\Pi}^{\star}$ by simply learning $\boldsymbol{\Pi}^{\star}_{\boldsymbol{\Upsilon}}$. The success of this approach, is however, limited by the scalarization technique as well as $\boldsymbol{\Pi}^{\star}$ may not be convex.

Remark 6 (Limits of Scalarization)**.** Using scalarizations to recover $\boldsymbol{\Pi}^{\star}$ suffers from the drawback that convexity assumptions on the scalarization function limit algorithms to only recover policies within the convex regions of $\boldsymbol{\Pi}^{\star}$ (Vamplew et al., 2008). Subsequently, our algorithm is limited in this sense as it relies on convex scalarizations, however, we leave the extension to non-convex regions as future work, and assume $\boldsymbol{\Pi}^{\star}$ to be convex for simplicity.

4.3 Bayes Regret

Generally, algorithms for cooperative multi-agent RL consider maximizing the cumulative reward of all agents (Littman, 1994). In the fully-observable scenario (i.e., all agent observe the complete $(\mathbf{x},\mathbf{a})$), the problem reduces to that of an MDP with fixed value and reward functions (given as the sum of individual value and rewards). Indeed by considering the scalarization $\boldsymbol{\upsilon}' = \frac{1}{M} \cdot \mathbf{1}_M$ we can observe that by Proposition 1, the optimal policy $\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}'}$ corresponds to the optimal policy for the MPD defined over $\mathcal{S} \times \mathcal{A}$ with rewards given by the average rewards obtained by each agent. It is therefore straightforward to recover a no-regret policy using a single-agent algorithm (by making a linear MDP assumption over the joint state and action space $\mathcal{S} \times \mathcal{A}$, as in Jin et al. (2020)), as it is similar to the parallel setting. Moreover, in distributed applications, additional constraints in the environment (e.g., load balancing, Schaerf et al. (1994)), may require learning policies that prioritize an agent over others, and hence we consider a general notion of *Bayes regret*. Our objective is to approximate $\boldsymbol{\Pi}^{\star}$ by learning a set of T policies $\widehat{\boldsymbol{\Pi}}_T$ that minimize the Bayes regret, given by,

$$\mathfrak{R}_B(T) \triangleq \mathop{\mathbb{E}}_{\boldsymbol{\upsilon}\sim p_{\boldsymbol{\Upsilon}}} \left[\max_{\mathbf{x}\in\mathcal{S}} \left[V^{\star}_{\boldsymbol{\upsilon},1}(\mathbf{x}) - \max_{\boldsymbol{\pi}\in\widehat{\boldsymbol{\Pi}}_T} V^{\boldsymbol{\pi}}_{\boldsymbol{\upsilon},1}(\mathbf{x}) \right] \right]. \tag{12}$$

Here $p_{\boldsymbol{\Upsilon}}$ is a distribution over $\boldsymbol{\Upsilon}$ that characterizes the nature of policies we wish to recover. For example, if we set $p_{\boldsymbol{\Upsilon}}$ as the uniform distribution over Δ^M then we can expect the policies recovered to prioritize all agents equally[4]. The advantage of minimizing Bayes regret can be understood as follows. For any $\boldsymbol{\upsilon} \in \boldsymbol{\Upsilon}$, if $\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}} \in \widehat{\boldsymbol{\Pi}}_T$, then the regret incurred is 0. Hence, by collecting policies that minimize Bayes regret, we are effectively searching for policies that span dense regions of $\boldsymbol{\Pi}^{\star}$ (assuming convexity, see Remark 6). Consider now the *cumulative regret*:

$$\mathfrak{R}_C(T) \triangleq \sum_{t\in[T]} \mathop{\mathbb{E}}_{\boldsymbol{\upsilon}_t\sim p_{\boldsymbol{\Upsilon}}} \left[\max_{\mathbf{x}\in\mathcal{S}} \left[V^{\star}_{\boldsymbol{\upsilon}_t,1}(\mathbf{x}) - V^{\boldsymbol{\pi}_t}_{\boldsymbol{\upsilon}_t,1}(\mathbf{x}) \right] \right]. \tag{13}$$

Where $\boldsymbol{\upsilon}_1, ..., \boldsymbol{\upsilon}_T \sim p_{\boldsymbol{\Upsilon}}$ are sampled i.i.d. from $p_{\boldsymbol{\Upsilon}}$, and $\boldsymbol{\pi}_t$ refers to the joint policy at episode t. Under suitable conditions on $\mathfrak{s}$ and $\boldsymbol{\Upsilon}$, we can bound the two quantities.

[4] One may consider minimizing the regret for a fixed scalarization $\boldsymbol{\upsilon}' = \mathbb{E}_{p_{\boldsymbol{\Upsilon}}} \boldsymbol{\upsilon}$, however, that will also recover only one policy in $\boldsymbol{\Pi}^{\star}$, whereas we desire to capture regions of $\boldsymbol{\Pi}^{\star}$.

Proposition 3. *For $\mathfrak{s}$ that is Lipschitz and bounded $\boldsymbol{\Upsilon}$, we have that $\mathfrak{R}_B(T) \leq \frac{1}{T}\mathfrak{R}_C(T) + o(1)$.*

We focus on minimizing the regret $\mathfrak{R}_C(T)$, as any no-regret algorithm for $\mathfrak{R}_C$ bounds $\mathfrak{R}_B$.

4.4 `Coop-LSVI` for Linear Multiagent MDPs

The key observation for algorithm design in this setting is that for any MMDP$(\mathcal{S}, \mathcal{M}, \mathcal{A}, H, \mathbb{P}, \mathbf{R})$, the optimal policy with respect to a fixed scalarization parameter $\boldsymbol{\upsilon}$ coincides with the optimal policy for the MDP$(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, R')$ where $R'(\mathbf{x}, \mathbf{a}) = \boldsymbol{\upsilon}^\top \mathbf{r}(\mathbf{x}, \mathbf{a})\ \forall\ \mathcal{S} \times \mathcal{A}$. Based on this observation, we extend the notion of a linear MDP to the multi-agent setting under random scalarizations. We first provide a natural regularity assumption of linearity.

Definition 5 (Linear Multiagent MDP)**.** *A Multiagent MDP MMDP$(\mathcal{S}, \mathcal{M}, \mathcal{A}, H, \mathbb{P}, \mathbf{R})$ is a linear multi-agent MDP with $M+1$ features $\{\boldsymbol{\phi}_i\}_{i=1}^M, \boldsymbol{\phi}_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_1}$ and $\boldsymbol{\phi}_c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_2}$ where $d_1 + d_2 = d$, if for any $h \in [H]$ there exist an unknown vector $\boldsymbol{\theta}_h \in \mathbb{R}^{d_1}$ and there exist d_2 unknown (signed) measures $\boldsymbol{\mu}_h = (\mu_h^1, ..., \mu_h^{d_2})$ over $\mathcal{S}$ such that for any $m \in \mathcal{M}, \mathbf{x} \in \mathcal{S}, \mathbf{a} \in \mathcal{A}, h \in [H]$,*

$$r_{m,h} = \langle \boldsymbol{\phi}_m(\mathbf{x}, \mathbf{a}), \boldsymbol{\theta}_h \rangle\ ,\ \text{and}\ \mathbb{P}_h(\cdot|\mathbf{x}, \mathbf{a}) = \langle \boldsymbol{\phi}_c(\mathbf{x}, \mathbf{a}), \boldsymbol{\mu}_h(\cdot) \rangle.$$

We denote the overall feature vector as $\boldsymbol{\Phi}(\cdot) \in \mathbb{R}^{d \times M}$, where, for any $\mathbf{x} \in \mathcal{S}, \mathbf{a} \in \mathcal{A}$,

$$\boldsymbol{\Phi}(\mathbf{x}, \mathbf{a}) = \begin{bmatrix} \boldsymbol{\phi}_1(\mathbf{x}, \mathbf{a})^\top, & \boldsymbol{\phi}_c(\mathbf{x}, \mathbf{a})^\top \\ \vdots & \vdots \\ \boldsymbol{\phi}_M(\mathbf{x}, \mathbf{a})^\top, & \boldsymbol{\phi}_c(\mathbf{x}, \mathbf{a})^\top \end{bmatrix}^\top.$$

Under this representation, we have that for any $\mathbf{x} \in \mathcal{S}, \mathbf{a} \in \mathcal{A}, h \in [H]$,

$$\mathbf{r}_h(\mathbf{x}, \mathbf{a}) = \boldsymbol{\Phi}(\mathbf{x}, \mathbf{a})^\top \begin{bmatrix} \boldsymbol{\theta}_h \\ \mathbf{0}_{d_2} \end{bmatrix},\ \text{and,}\ \mathbf{1}_M \cdot \mathbb{P}_h(\cdot|\mathbf{x}, \mathbf{a}) = \boldsymbol{\Phi}(\mathbf{x}, \mathbf{a})^\top \begin{bmatrix} \mathbf{0}_{d_1} \\ \boldsymbol{\mu}_h(\cdot) \end{bmatrix}.$$

We assume without loss of generality that $\|\boldsymbol{\Phi}(\mathbf{x}, \mathbf{a})\| \leq 1\ \forall\ (\mathbf{x}, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$, $\|\boldsymbol{\theta}_h\| \leq \sqrt{d_1} \leq \sqrt{d}$ and $\|\boldsymbol{\mu}_h(\mathcal{S})\| \leq \sqrt{d_2} \leq \sqrt{d}$.

Observe that this assumption implies that there exist a set of weights such that the scalarized $Q-$values for any sclarization parameter $\boldsymbol{\upsilon}$ are linear projections of the combined features $\boldsymbol{\Phi}(\cdot)$.

Lemma 2 (Linearity of weights in MMDP)**.** *Under the linear MMDP Assumption (Definition 5), for any fixed policy π and $\boldsymbol{\upsilon} \in \boldsymbol{\Upsilon}$, there exist weights $\{\mathbf{w}^\pi_{\boldsymbol{\upsilon},h}\}_{h \in [H]}$ such that $Q^\pi_{\boldsymbol{\upsilon},h}(\mathbf{x}, \mathbf{a}) = \boldsymbol{\upsilon}^\top \boldsymbol{\Phi}(\mathbf{x}, \mathbf{a})^\top \mathbf{w}^\pi_{\boldsymbol{\upsilon},h}$ for all $(\mathbf{x}, \mathbf{a}, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, where $\|\mathbf{w}^\pi_{\boldsymbol{\upsilon},h}\|_2 \leq 2H\sqrt{d}$.*

Remark 7 (Multi-agent modeling assumptions)**.** We now discuss how this assumption differs from the typical modeling assumption in the single-agent function approximation setting (Jin et al., 2020; Wang et al., 2020). In contrast to the typical assumption (see Assumption A in Jin et al. (2020), also in Bradtke & Barto (1996); Melo & Ribeiro (2007)), here we model, for any agent, the reward $r_{m,h}$ and dynamics $\mathbb{P}_h$ for all agents separately, each with d_1 and d_2 dimensions respectively such that $d_1 + d_2 = d$. In the single-agent setting, identical assumptions on the reward and transition kernels will lead to a model with complexity $\max\{d_1, d_2\}$, whereas in our formulation we have a complexity of $d_1 + d_2$. Given that both $d_1, d_2 \leq d$ this implies that our fomulation incurs a maximum overhead of $2\sqrt{2}$ in the regret if applied to the LSVI algorithm presented in Jin et al. (2020). Furthermore, observe that in the *fully-cooperative* setting, where $r_{1,h} = ... = r_{M,h} \forall h \in [H]$, we have that $\boldsymbol{\phi}_1 = \boldsymbol{\phi}_2 = ... = \boldsymbol{\phi}_M$ satisfies the modeling requirement. We also have that the features $\boldsymbol{\Phi}$ are functions of the *joint* action, which differentiates the MMDP from a parallel MDP setting.

Algorithm 2 `Coop-LSVI` for Multiagent MDPs

Input: $T, \mathbf{\Phi}, H, S$, sequence $\beta_h = \{(\beta_h^t)_t\}$.
Initialize: $\mathbf{\Lambda}_h^t = \lambda \mathbf{I}_d, \delta\mathbf{\Lambda}_h^t = \mathbf{0}, \mathcal{U}_h^m, \mathcal{W}_h^m = \emptyset$.
for episode $t = 1, 2, ..., T$ **do**
  for agent $m \in \mathcal{M}$ **do**
    Receive initial state $\mathbf{x}_1^t, \boldsymbol{\upsilon}_t \sim p_{\boldsymbol{\Upsilon}}$.
    Set $V_{\boldsymbol{\upsilon}_t, H+1}^t(\cdot) \leftarrow 0$.
    for step $h = H, ..., 1$ **do**
      Compute $Q_{\boldsymbol{\upsilon}_t, h}^t(\cdot, \cdot)$ (Eqn. 14)
      Set $V_{\boldsymbol{\upsilon}_t, h}^t(\cdot) \leftarrow \max_{\mathbf{a} \in \mathcal{A}} Q_{\boldsymbol{\upsilon}_t, h}^t(\cdot, \mathbf{a})$.
    end for
    for step $h = 1, ..., H$ **do**
      Take action $a_{m,h}^t \leftarrow [\arg\max_{a \in \mathcal{A}} Q_{\boldsymbol{\upsilon}_t, h}^t(\mathbf{x}_h^t, \mathbf{a})]_m$.
      Observe $r_{m,h}^t, \mathbf{x}_{h+1}^t$.
      Update $\delta\mathbf{\Lambda}_h^t \leftarrow \delta\mathbf{\Lambda}_h^t + \mathbf{\Phi}(\mathbf{z}_h^t)\mathbf{\Phi}(\mathbf{z}_h^t)^\top$.
      Update $\mathcal{W}_h^m \leftarrow \mathcal{W}_h^m \cup (m, r_{m,h}^t)$.
      if $\log \frac{\det(\mathbf{\Lambda}_h^t + \delta\mathbf{\Lambda}_h^t + \lambda\mathbf{I})}{\det(\mathbf{\Lambda}_h^t + \lambda\mathbf{I})} > S$ **then**
        SYNCHRONIZE$\leftarrow$ TRUE.
      end if
    end for
  end for
  if SYNCHRONIZE **then**
    for step $h = H, ..., 1$ **do**
      [$\forall$ AGENTS] Send $\mathcal{W}_m^h \rightarrow$SERVER.
      [SERVER] Aggregate $\mathcal{W}^h \rightarrow \cup_{m \in \mathcal{M}} \mathcal{W}_h^m$.
      [SERVER] Communicate $\mathcal{W}^h$ to each agent.
      [$\forall$ AGENTS] Set $\delta\mathbf{\Lambda}_h^t \leftarrow 0, \mathcal{W}_h^m \leftarrow \emptyset$.
      [$\forall$ AGENTS] Set $\mathbf{\Lambda}_h^t \leftarrow \mathbf{\Lambda}_h^t + \sum_{(\mathbf{x}, \mathbf{a}) \in \mathcal{W}^h} \mathbf{\Phi}(\mathbf{x}, \mathbf{a})\mathbf{\Phi}(\mathbf{x}, \mathbf{a})^\top$.
      [$\forall$ AGENTS] Set $\mathcal{U}_h^m \leftarrow \mathcal{U}_h^m \cup \mathcal{W}_h^m$
    end for
  end if
end for

4.5 Algorithm Design

The motivation behind our design is to learn a function that simultaneously can recover a policy close to $\boldsymbol{\pi}_{\boldsymbol{\upsilon}}^\star$ for each MDP corresponding to $\boldsymbol{\upsilon} \in \boldsymbol{\Upsilon}$, recovering $\mathbf{\Pi}^\star$. The algorithm is once again a distributed variant of least-squares value iteration with UCB exploration. Following Proposition 1, the central idea is to scalarize the MMDP with a randomly sampled parameter $\boldsymbol{\upsilon}$, and each of the M agents will execute the *joint* policy that coincides with (an approximation of) $\boldsymbol{\pi}_{\boldsymbol{\upsilon}}^\star$. Now, the key idea is to make sure that each agent acts according to the joint policy that is aiming to mimic $\boldsymbol{\pi}_{\boldsymbol{\upsilon}}^\star$. Therefore, we must ensure that the local estimate for the *joint* policy obtained by any agent must be identical, such that the *joint* action is in accordance with $\boldsymbol{\pi}_{\boldsymbol{\upsilon}}^\star$. To achieve this we will utilize a rare-switching technique similar to Algorithm 1. In any episode $t \in [T]$, the objective would be to obtain the optimal (scalar) $Q-$values $Q_{\boldsymbol{\upsilon}_t, h}^\star$ by recursively applying the Bellman equation and solving the resulting equations via a *vector-valued* regression. Since the policy variables are designed to be identical across agents at all times, we drop the m subscript.

Specifically, let us assume the last synchronization between agents occured in episode k_t. Each agent obtains an *identical* sequence of value functions $\{Q_{\boldsymbol{\upsilon}, h}^t\}_{h \in [H]}$ by iteratively performing linear

least-squares ridge regression from the history available from the previous k_t episodes, but in contrast to the parallel setting, it now first learns a vector $Q-$function $\mathbf{Q}_{t,h}$ over $\mathbb{R}^M$, which is scalarized to obtain the $Q-$value as $Q_{\boldsymbol{\upsilon}_t,h} = \boldsymbol{\upsilon}_t^\top \mathbf{Q}_{t,h}$. Each agent m first sets $\mathbf{Q}_{t,H+1}$ to be a zero vector, and for any $h \in [H]$, solves the following sequence of regressions to obtain $Q-$values. For each $h = H, ..., 1$, for each agent computes,

$$V^t_{\boldsymbol{\upsilon}_t,h+1}(\mathbf{x}) \leftarrow \arg\max_{\mathbf{a}\in\mathcal{A}} \langle \boldsymbol{\upsilon}_t, \mathbf{Q}_{t,h+1}(\mathbf{x},\mathbf{a})\rangle\,,\ \widehat{\mathbf{Q}}_{t,h} \leftarrow \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left[\sum_{\tau\in[k_t]} \left\|\mathbf{y}_{\tau,h} - \boldsymbol{\Phi}(\mathbf{x}_\tau,\mathbf{a}_\tau)^\top \mathbf{w}\right\|_2^2 + \lambda\|\mathbf{w}\|_2^2\right].$$

$$Q_{t,h}(\mathbf{x},\mathbf{a}) \leftarrow \boldsymbol{\upsilon}_t^\top \widehat{\mathbf{Q}}_{t,h} + \beta_t \cdot \|\boldsymbol{\Phi}(\mathbf{x},\mathbf{a})^\top (\boldsymbol{\Lambda}^h_t)^{-1} \boldsymbol{\Phi}(\mathbf{x},\mathbf{a})\|_2. \tag{14}$$

Here the targets $\mathbf{y}_{\tau,h} = \mathbf{r}_h(\mathbf{x},\mathbf{a}) + \mathbf{1}_M \cdot V^t_{\boldsymbol{\upsilon}_t,h+1}$, $\mathbf{1}_M$ denotes the all-ones vector in $\mathbb{R}^M$, β_t is an appropriately chosen sequence and $\boldsymbol{\Lambda}^h_t$ is described subsequently. Once all of these quantities are computed, each agent $m = 1, .., M$ selects the action $a^t_{m,h} = [\arg\max_{\mathbf{a}\in\mathcal{A}} Q_{t,h}(\mathbf{x}^t_h,\mathbf{a})]_m$ for each $h \in [H]$. Hence, the joint action $\mathbf{a}^t_h = \{a^t_{m,h}\}_{m=1}^M = \arg\max_{\mathbf{a}\in\mathcal{A}} Q_{t,h}(\mathbf{x}^t_h,\mathbf{a})$. Observe that while the computations of the policy is decentralized, the policies coincide at all instances by the modeling assumption and the periodic synchronizations between the agents. We now present the closed form of $\widehat{\mathbf{Q}}_{t,h}$. Consider the contraction $\mathbf{z}^h_\tau = (\mathbf{x}^h_\tau, \mathbf{a}^h_\tau)$ and the map $\boldsymbol{\Phi}^h_t : \mathbb{R}^d \to \mathbb{R}^{Mt}$ given by,

$$\boldsymbol{\Phi}^h_t \boldsymbol{\theta} \triangleq \left[(\boldsymbol{\Phi}(\mathbf{z}^h_1)^\top \boldsymbol{\theta})^\top, ..., (\boldsymbol{\Phi}(\mathbf{z}^\top_t \boldsymbol{\theta})^\top\right]^\top \ \forall\ \boldsymbol{\theta} \in \mathbb{R}^d.$$

Now, consider $\boldsymbol{\Lambda}^h_t = (\boldsymbol{\Phi}^h_{k_t})^\top \boldsymbol{\Phi}^h_{k_t} + \lambda \mathbf{I}_d \in \mathbb{R}^{d\times d}$, and the matrix $\mathbf{U}^h_t = \sum_{\tau=1}^{k_t} \boldsymbol{\Phi}(\mathbf{z}^h_\tau)\mathbf{y}_{\tau,h}$. Then, we have by a multi-task concentration (see Appendix B of Chowdhury & Gopalan (2020)),

$$\widehat{\mathbf{Q}}_{t,h}(\mathbf{x},\mathbf{a}) = \boldsymbol{\Phi}(\mathbf{x},\mathbf{a})^\top \left(\boldsymbol{\Lambda}^h_t\right)^{-1} \mathbf{U}^h_t. \tag{15}$$

Now, to limit communication, the synchronization protocol is similar to that in Section 5.1 of Abbasi-Yadkori et al. (2011). Whenever $\det(\boldsymbol{\Lambda}^h_t) \geq S \cdot \det(\boldsymbol{\Lambda}^h_{k_t})$, for some constant S, the agents synchronize their rewards. The pseudocode for the algorithm is presented in Algorithm 2.

4.6 Regret Analysis

The primary challenges in bounding the regret arise from the fact that the agents are simultaneously trying to recover a class of policies (the Pareto frontier, $\boldsymbol{\Pi}^\star$) instead of any single policy within $\boldsymbol{\Pi}^\star$. This modeling choice requires us to estimate the individual rewards for each of the M agents simultaneously, and leverage vector-valued concentration such that we can bound the estimation error from the optimal policy for any $\boldsymbol{\upsilon} \in \boldsymbol{\Upsilon}$. For this, we first derive a concentration result on the ℓ_2-error on estimating vector-valued value functions ($\mathbf{V}$) which is independent of the sampled $\boldsymbol{\upsilon}_t$, by leveraging the (vector-valued) sef-normalized martingale analysis from Chowdhury & Gopalan (2020). Next, we note that the scalarization function $\mathfrak{s}$ is Lipschitz with constant 1, and therefore, for any $\boldsymbol{\upsilon} \in \boldsymbol{\Upsilon}$ and vectors $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^M$, $|\mathfrak{s}_{\boldsymbol{\upsilon}}(\mathbf{v}_1) - \mathfrak{s}_{\boldsymbol{\upsilon}}(\mathbf{v}_2)| \leq \|\mathbf{v}_1 - \mathbf{v}_2\|_2$, and hence our previous concentration result allows us to bound the estimation error *independently* of the scalarization $\boldsymbol{\upsilon}$. The structure of the rest of the bound is similar to that of the parallel MDP setting. Here we provide the primary regret bound, and defer the proof to the Appendix.

Theorem 4. *`CoopLSVI` when run on an MMDP with M agents and communication threshold S, $\beta_t = \mathcal{O}(H\sqrt{d\log(tMH)})$ and $\lambda = 1$ obtains, with probability at least $1-\alpha$, cumulative regret:*

$$\mathfrak{R}_C(T) = \widetilde{\mathcal{O}}\left(d^{\frac{3}{2}} H^2 \sqrt{ST \log\left(\frac{1}{\alpha}\right)}\right).$$

Remark 8 (Multiagent Regret Bound)**.** Theorem 4 claims in conjunction with Proposition 3 that `CoopLSVI` obtains Bayes regret of $\widetilde{\mathcal{O}}(\sqrt{T})$ even when communication is limited. Note that the dependence on T matches that of MDP algorithms (e.g., Jin et al. (2020)), and moreover, we recover the same rate (up to logarithmic factors) when $M = 1$, ensuring that the analysis is tight. Additionally, we see that this algorithm can easily be applied to an MDP by simply selecting $p_{\boldsymbol{\Upsilon}}$ to be a point mass at the appropriate $\boldsymbol{\upsilon}$, with no increase in regret. Finally, we see that `CoopLSVI` can also be emulated on a single agent with M objectives, where $S = 1$, which provides, to the best of our knowledge, the first no-regret algorithm for multi-objective reinforcement learning.

Note that as the common state $\mathbf{x}$ is visible to all agents, the agents only require communication of rewards. While the protocol in this setting essentially operates on a similar idea (communicating only when the variance of the history for any step crosses a threshold), the exact form of the threshold differs slightly in this case: instead of computing individual policy parameters using local observations, here the agents employ a *rarely-switching* strategy that delays updating global parameters until the threshold condition is met. Despite the slight difference in the communication strategy, the overall communication complexity is similar to the parallel MDP variant.

Lemma 3 (Communication Complexity)**.** *If Algorithm 2 is run with threshold $S > 1$, then the total number of episodes with communication $n \leq dH \log_S (1 + MT/d)) + H$. When $S \leq 1$, $n = T$.*

Remark 9 (Communication complexity)**.** As is with the parallel MDP setting, we can control the communication budget by adjusting the threshold parameter S. Note that when $S = 1$, we have that communication will occur each round round, as the threshold will be satisfied trivially by the rank-1 update to the covariance matrix. If the horizon T is known in advance, one can set $S = (1 + MT/d)^{1/C}$ for some independent constant $C > 1$, to ensure that the total rounds of communication is a fixed constant $(dC+1)H$, which provides us a group regret of $\widetilde{\mathcal{O}}(M^{\frac{1}{2C}} \cdot T^{\frac{1}{2}+\frac{1}{2C}})$. This dependence of the regret on the threshold S is indeed worse than the protocol for the parallel MDP, as in this case, the agents *do not* utilize local observations until they are synchronized, and merely readjust policy parameters for different scalarizations $\boldsymbol{\upsilon}$, necessitating frequent communication. A balance between communication and regret can be obtained by setting $S = C'$ for some absolute constant C', leading to a total $\mathcal{O}(\log MT)$ rounds of communication with $\widetilde{\mathcal{O}}(\sqrt{T})$ regret.

5 Conclusion

We presented `CoopLSVI`, a cooperative multi-agent reinforcement learning algorithm that attains sublinear regret while maintaining sublinear communication under linear function approximation. While most research in multi-agent RL focuses either on *fully-cooperative* settings (i.e., a multi-agent MDP with identical reward functions), or stochastic games (Shapley, 1953), the heterogeneous setting considered here allows for the agents to both observe their rewards privately, while generalizing to Bayes regret guarantees, extending several lines of prior work (Kar et al., 2013; Zhang et al., 2018b). Similarly, `CoopLSVI` is the first algorithm to provide provably sublinear regret in heterogeneous parallel MDPs. Given the rapid advancements in federated learning, we believe this to be a valuable line of inquiry, extending beyond the work that has been done in the related problem of multi-armed bandits.

There are several open questions that our work posits. A few areas include extending `CoopLSVI` to a fully-decentralized network topology; tight lower bounds on communication-regret tradeoffs; and robust estimation to avoid side information in heterogeneous settings. We believe our work will serve as a valuable stepping stone to further developments in this area.

A Omitted Algorithms

Algorithm 3 `Coop-LSVI` for Heterogeneous Rewards

Input: $T, \widetilde{\boldsymbol{\phi}}, H, S$, sequence $\beta_h = \{(\beta^t_{m,h})_{m,t}\}$.
Initialize: $\mathbf{S}^t_{m,h}, \delta\mathbf{S}^t_{m,h} = \mathbf{0}, \mathcal{U}^m_h, \mathcal{W}^m_h = \emptyset$.
for episode $t = 1, 2, ..., T$ **do**
  for agent $m \in \mathcal{M}$ **do**
    Receive initial state $x^t_{m,1}$.
    Set $V^t_{m,H+1}(\cdot) \leftarrow 0$.
    for step $h = H, ..., 1$ **do**
      Compute $\widetilde{\boldsymbol{\Lambda}}^t_{m,h} \leftarrow \mathbf{S}^t_{m,h} + \delta\mathbf{S}^t_{m,h}$.
      Compute $\widehat{Q}^t_{m,h}$ and $\sigma^t_{m,h}$ (Eqn. 8).
      Compute $Q^t_{m,h}(\cdot,\cdot,\cdot)$ (Eqn. 10)
      Set $V^t_{m,h}(\cdot) \leftarrow \max_{a\in\mathcal{A}} Q^t_{m,h}(\cdot, a)$.
    end for
    for step $h = 1, ..., H$ **do**
      Take action $a^t_{m,h} \leftarrow \arg\max_{a\in\mathcal{A}} Q^t_{m,h}(m, x^t_{m,h}, a)$.
      Observe $r^t_{m,h}, x^t_{m,h+1}$.
      Update $\delta\mathbf{S}^t_{m,h} \leftarrow \delta\mathbf{S}^t_{m,h} + \widetilde{\boldsymbol{\phi}}(m, z^t_{m,h})\widetilde{\boldsymbol{\phi}}(m, z^t_{m,h})^\top$.
      Update $\mathcal{W}^m_h \leftarrow \mathcal{W}^m_h \cup (m, x, a, x')$.
      if $\log \frac{\det(\mathbf{S}^t_{m,h}+\delta\mathbf{S}^t_{m,h}+\lambda\mathbf{I})}{\det(\mathbf{S}^t_{m,h}+\lambda\mathbf{I})} > \frac{S}{\Delta t_{m,h}}$ **then**
        SYNCHRONIZE$\leftarrow$ TRUE.
      end if
    end for
  end for
  if SYNCHRONIZE **then**
    for step $h = H, ..., 1$ **do**
      [$\forall$ AGENTS] Send $\mathcal{W}^h_m \rightarrow$SERVER.
      [SERVER] Aggregate $\mathcal{W}^h \rightarrow \cup_{m\in\mathcal{M}}\mathcal{W}^m_h$.
      [SERVER] Communicate $\mathcal{W}^h$ to each agent.
      [$\forall$ AGENTS] Set $\delta\boldsymbol{\Lambda}^t_h \leftarrow 0, \mathcal{W}^m_h \leftarrow \emptyset$.
      [$\forall$ AGENTS] Set $\boldsymbol{\Lambda}^t_h \leftarrow \boldsymbol{\Lambda}^t_h + \sum_{(n,x,a)\in\mathcal{W}^h} \widetilde{\boldsymbol{\phi}}(n, x, a)\boldsymbol{\phi}(n, x, a)^\top$.
      [$\forall$ AGENTS] Set $\mathcal{U}^m_h \leftarrow \mathcal{U}^m_h \cup \mathcal{W}^m_h$
    end for
  end if
end for

B Parallel MDP Proofs

B.1 Proof of Lemma 1

Proof. Denote an epoch as the number of episodes between two rounds of communication. Let $q = \sqrt{\frac{ST}{d\log(1+T/d)} + 1}$. There can be at most $\lceil T/q \rceil$ rounds of communication such that they occur after an epoch of length q. On the other hand, if there is any round of communication succeeding an epoch (that begins, say at time t) of length $< n'$, then for that epoch, $\log \frac{\det\left(\mathbf{S}^t_{m,h} + \delta\mathbf{S}^t_{m,h} + \lambda\mathbf{I}_d\right)}{\det\left(\mathbf{S}^t_{m,h} + \lambda\mathbf{I}_d\right)} \geq \frac{S}{q}$. Let the communication occur at a set of episodes $t'_1, ..., t'_n$. Now, since:

$$\sum_{i=1}^{n-1} \log \frac{\det\left(\mathbf{S}^{t_{i+1}}_{m,h}\right)}{\det\left(\mathbf{S}^{t_i}_{m,h}\right)} = \log \frac{\det\left(\mathbf{\Lambda}^T_h\right)}{\det\left(\mathbf{\Lambda}^0_h\right)} \leq d\log(1 + T/(d)), \tag{16}$$

We have that the total number of communication rounds succeeding epochs of length less than n' is upper bounded by $\log \frac{\det\left(\mathbf{\Lambda}^T_h\right)}{\det\left(\mathbf{\Lambda}^0_h\right)} \leq d\log(1 + T/(d)) \cdot (q/S)$. Combining both the results together, we have the total rounds of communication as:

$$\begin{aligned} n &\leq \lceil T/q \rceil + \lceil d\log(1 + T/(d)) \cdot (q/S) \rceil & (17)\\ &\leq T/q + d\log(1 + T/(d)) \cdot (q/S) + 2 & (18) \end{aligned}$$

Replacing q from earlier and summing over $h \in [H]$ (as communication may be triggered by any of the steps satisfying the condition) gives us the final result. □

B.2 Proof of Theorem 1 (Homogenous Setting)

We first present our primary concentration result to bound the error in the least-squares value iteration.

Lemma 4. *Under the setting of Therorem 1, let c_β be the constant defining β, and $\mathbf{S}^t_{m,h}$ and $\mathbf{\Lambda}^k_t$ be defined as follows.*

$$\begin{aligned} \mathbf{S}^t_{m,h} = \sum_{n=1}^{M}\sum_{\tau=1}^{k_t} \boldsymbol{\phi}(x^\tau_{n,h}, a^\tau_{n,h}) \left[V^t_{m,h+1}(x^\tau_{n,h+1}) - (\mathbb{P}_h V^t_{m,h+1})(x^\tau_{n,h}, a^\tau_{n,h})\right] \\ + \sum_{\tau=k_t+1}^{t-1} \boldsymbol{\phi}(x^\tau_{m,h}, a^\tau_{m,h}) \left[V^t_{m,h+1}(x^\tau_{m,h+1}) - (\mathbb{P}_h V^t_{m,h+1})(x^\tau_{m,h}, a^\tau_{m,h})\right], \end{aligned}$$

$$\mathbf{\Lambda}^t_{m,h} = \sum_{n=1}^{M}\sum_{\tau=1}^{k_t} \boldsymbol{\phi}(x^\tau_{n,h}, a^\tau_{n,h})\boldsymbol{\phi}(x^\tau_{n,h}, a^\tau_{n,h})^\top + \sum_{\tau=k_t+1}^{t-1} \boldsymbol{\phi}(x^\tau_{m,h}, a^\tau_{m,h})\boldsymbol{\phi}(x^\tau_{m,h}, a^\tau_{m,h})^\top + \lambda\mathbf{I}_d.$$

Where $V \in \mathcal{V}$ and $\mathcal{N}_\epsilon$ denotes the $\epsilon-$covering of the value function space $\mathcal{V}$. Then, there exists an absolute constant c_β independent of M, T, H, d, such that, with probability at least $1 - \delta'/2$ for all $m \in \mathcal{M}, t \in [T], h \in [H]$ simultaneously,

$$\left\|\mathbf{S}^t_{m,h}\right\|_{(\mathbf{\Lambda}^t_{m,h})^{-1}} \leq c_\beta \cdot dH\sqrt{2\log\left(\frac{dMTH}{\delta'}\right)}.$$

Proof. The proof is done in two steps. The first step is to bound the deviations in $\mathbf{S}$ for any fixed function V by a martingale concentration. The second step is to bound the resulting concentration over all functions V by a covering argument. Finally, we select appropriate constants to provide the form of the result required.

Step 1. Note that for any agent m, the function $V^t_{m,h+1}$ depends on the historical data from all M agents from the first k_t episodes, and the personal historical data for the first $(t-1)$ episodes, and depends on

$$\mathcal{U}^m_h(t) = \left(\cup_{n\in[M],\tau\in[k_t]}\{(x^\tau_{n,h}, a^\tau_{n,h}, x^\tau_{n,h+1})\}\right) \bigcup \left(\cup_{\tau\in[k_t+1,t-1}\{(x^\tau_{m,h}, a^\tau_{m,h}, x^\tau_{m,h+1})\}\right). \tag{19}$$

To bound the term we will construct an appropriate filtration to use a self-normalized concentration defined on elements of $\mathcal{U}^m_h(t)$. We highlight that in the multi-agent case with stochastic communication, it is not straightforward to provide a uniform martingale concentration that holds for all $t \in [T]$ simultaneously (as is done in the single-agent case), as the stochasticity in the environment dictates when communication will take place, and subsequently the quantity considered within self-normalization will depend on this communication itself. To circumvent this issue, we will first fix $k_t \leq t$ and obtain a filtration for a fixed k_t. Then, we will take a union bound over all $k_t \in [t]$ to provide the final self-normalized bound. We first fix k_t and define the following mappings where $i \in [M(t-1)]\,, l \in [t-1]$, and $n \in [M]$.

$$\mu(i) = \left\lceil \frac{i}{M} \right\rceil, \nu(i) = i(\text{mod } M), \text{ and, } \eta(l,n) = l\cdot(M+1) + n - 1.$$

Now, for a fixed k_t, consider the stochastic processes $\{\tilde{x}_\tau\}_{\tau=1}^\infty$ and $\{\tilde{\boldsymbol{\phi}}_\tau\}_{\tau=1}^\infty$, where,

$$\tilde{\boldsymbol{\phi}}_i = \boldsymbol{\phi}(x^{\nu(i)}_{\mu(i),h+1}) \otimes \mathbb{1}_d\left\{(\mu(i) = m) \vee (\nu(i) \leq k_t)\right\}$$

Here $\otimes$ denotes the Hadamard product, and $\mathbb{1}_d$ is the indicator function in $\mathbb{R}^d$. Consider now the filtration $\{\mathcal{F}_\tau\}_{\tau=0}^\infty$, where $\mathcal{F}_0$ is empty, and $\mathcal{F}_\tau = \sigma\left(\left\{\bigcup(\tilde{x}_i, \tilde{\boldsymbol{\phi}}_i)\right\}_{i\leq\tau}\right)$, where $\sigma(\cdot)$ denotes the corresponding $\sigma-$algebra formed by the set.

At any instant t for any agent m, the function $V^t_{m,h+1}$ and features $\boldsymbol{\phi}(x^t_{m,h}, a^t_{m,h})$ depend only on historical data from all other agents $[M]\setminus\{m\}$ up to the last episode of synchronization $k_t \leq t-1$ and depend on the personal data up to episode $t-1$. Hence, both are $V^t_{m,h+1}$ and $\boldsymbol{\phi}(x^t_{m,h}, a^t_{m,h})$ are measurable with respect to

$$\sigma\left(\left\{\bigcup_{l=1}^{k_t}\bigcup_{n=1}^{M}(\tilde{x}_{\eta(l,n)}, \tilde{\boldsymbol{\phi}}_{\eta(l,n)})\right\}\bigcup\left\{\bigcup_{l=k_t+1}^{t-1}(\tilde{x}_{\eta(l,m)}, \tilde{\boldsymbol{\phi}}_{\eta(l,m)})\right\}\right).$$

This is a subset of $\mathcal{F}_{\eta(t,m)}$. Therefore $V^t_{m,h+1}$ is $\mathcal{F}_{\eta(t,m)}-$measurable for fixed k_t. Now, consider $\mathcal{U}^m_h(\tau)$, the set of features available to agent m at episode $\tau \leq t$. We therefore have that, for any value function V,

$$\begin{aligned}
&\sum_{\tau=1}^{M(t-1)} \widetilde{\boldsymbol{\phi}}_{m,h}(\tau)\left\{V(\tilde{x}_\tau) - \mathbb{E}[V(\tilde{x}_\tau)|\mathcal{F}_{\tau-1}]\right\} \\
&= \sum_{\tau=1}^{M(t-1)} \left[\boldsymbol{\phi}(x^{\nu(i)}_{\mu(i),h+1}) \otimes \mathbb{1}_d\left\{(\mu(i) = m) \vee (\nu(i) \leq k_t)\right\}\right]\left\{V(\tilde{x}_\tau) - \mathbb{E}\left[V(\tilde{x}_\tau)|\mathcal{F}_{\tau-1}\right]\right\} \\
&= \sum_{(x_\tau, a_\tau, x'_\tau)\in\mathcal{U}^m_h(t)} \boldsymbol{\phi}(x_\tau, a_\tau)\left\{V(x'_\tau) - \mathbb{E}[V(x'_\tau)|\mathcal{F}_{\tau-1}]\right\}.
\end{aligned}$$

Now, when $V = V^t_{m,h+1}$, we have from the above,

$$\sum_{\tau=1}^{M(t-1)} \widetilde{\boldsymbol{\phi}}_\tau \left\{V^t_{h,m+1}(\tilde{x}_\tau) - \mathbb{E}[V^t_{h,m+1}(\tilde{x}_\tau)|\mathcal{F}_{\tau-1}]\right\}$$
$$= \sum_{(n_\tau, x_\tau, a_\tau, x'_\tau) \in \mathcal{U}^m_h(t)} \boldsymbol{\phi}(x_\tau, a_\tau) \left\{V^t_{m,h+1}(x'_\tau) - \mathbb{E}[V^t_{m,h+1}(x'_\tau)|\mathcal{F}_{\tau-1}]\right\} = \mathbf{S}^t_{m,h}.$$

Furthermore, consider $\widetilde{\boldsymbol{\Lambda}}^t_{m,h} = \lambda \mathbf{I}_d + \sum_{\tau=1}^{M(t-1)} \widetilde{\boldsymbol{\phi}}_\tau \widetilde{\boldsymbol{\phi}}_\tau^\top$. For the second term, we have,

$$\begin{aligned}
\widetilde{\boldsymbol{\Lambda}}^t_{m,h} &= \lambda \mathbf{I}_d + \sum_{\tau=1}^{M(t-1)} \widetilde{\boldsymbol{\phi}}_\tau \widetilde{\boldsymbol{\phi}}_\tau^\top \\
&= \lambda \mathbf{I}_d \\
&\quad + \sum_{\tau=1}^{M(t-1)} \left[\boldsymbol{\phi}(x^{\nu(i)}_{\mu(i),h+1}) \otimes \mathbb{1}_d \left\{\mu(i) = m \vee \nu(i) \leq k_t\right\}\right] \left[\boldsymbol{\phi}(x^{\nu(i)}_{\mu(i),h+1}) \otimes \mathbb{1}_d \left\{\mu(i) = m \vee \nu(i) \leq k_t\right\}\right]^\top \\
&= \lambda \mathbf{I}_d + \sum_{(n_\tau, x_\tau, a_\tau, x'_\tau) \in \mathcal{U}^m_h(t)} \boldsymbol{\phi}(x_\tau, a_\tau) \boldsymbol{\phi}(x_\tau, a_\tau)^\top = \boldsymbol{\Lambda}^t_{m,h}.
\end{aligned}$$

To complete the proof, we bound $\left\|\sum_{\tau=1}^{M(t-1)} \widetilde{\boldsymbol{\phi}}_{m,h}(\tau) \left\{V^t_{h,m+1}(\tilde{x}_\tau) - \mathbb{E}[V^t_{h,m+1}(\tilde{x}_\tau)|\mathcal{F}_{\tau-1}]\right\}\right\|_{(\widetilde{\boldsymbol{\Lambda}}^t_{m,h})^{-1}}$ over all $k_t \in [t]$. We proceed following a self-normalized martingale bound and a covering argument, as done in Yang et al. (2020).

Applying Lemma 28 to $\left\|\sum_{\tau=1}^{M(t-1)} \widetilde{\boldsymbol{\phi}}_{m,h}(\tau) \left\{V^t_{h,m+1}(\tilde{x}_\tau) - \mathbb{E}[V^t_{h,m+1}(\tilde{x}_\tau)|\mathcal{F}_{\tau-1}]\right\}\right\|_{(\widetilde{\boldsymbol{\Lambda}}^t_{m,h})^{-1}}$ under the filtration $\{\mathcal{F}_\tau\}_{\tau=0}^\infty$ described earlier, we have that with probability at least $1-\delta'$,

$$\begin{aligned}
\left\|\mathbf{S}^t_{m,h}\right\|^2_{(\boldsymbol{\Lambda}^t_{m,h})^{-1}} &= \left\|\sum_{\tau=1}^{M(t-1)} \widetilde{\boldsymbol{\phi}}_\tau \left\{V^t_{h,m+1}(\tilde{x}_\tau) - \mathbb{E}[V^t_{h,m+1}(\tilde{x}_\tau)|\mathcal{F}_{\tau-1}]\right\}\right\|^2_{(\widetilde{\boldsymbol{\Lambda}}^t_{m,h})^{-1}} \\
&\leq \sup_{V \in \mathcal{V}} \left\|\sum_{\tau=1}^{M(t-1)} \widetilde{\boldsymbol{\phi}}_\tau \left\{V(\tilde{x}_\tau) - \mathbb{E}[V(\tilde{x}_\tau)|\mathcal{F}_{\tau-1}]\right\}\right\|^2_{(\widetilde{\boldsymbol{\Lambda}}^t_{m,h})^{-1}} \\
&\leq 4H^2 \cdot \log \frac{\det\left(\widetilde{\boldsymbol{\Lambda}}^t_{m,h}\right)}{\det\left(\lambda \mathbf{I}_d\right)} + 8H^2 \log(|\mathcal{N}_\epsilon|/\delta') + 8M^2t^2\epsilon^2/\lambda.
\end{aligned}$$

Where $\mathcal{N}_\epsilon$ is an $\epsilon-$covering of $\mathcal{V}$. Therefore, we have that, with probability at least $1-\delta'$, for any fixed $k_t \leq t$,

$$\left\|\mathbf{S}^t_{m,h}\right\|_{(\boldsymbol{\Lambda}^t_{m,h})^{-1}} \leq 2H \sqrt{\log\left(\frac{\det\left(\widetilde{\boldsymbol{\Lambda}}^t_{m,h}\right)}{\det\left(\lambda \mathbf{I}_d\right)}\right) + 2\log\left(\frac{|\mathcal{N}_\epsilon|}{\delta}\right) + \frac{2M^2t^2\epsilon^2}{H^2\lambda}}.$$

Taking a union bound over all $k_t \in [t]$, $m \in \mathcal{M}, t \in [T], h \in [H]$ and replacing $\delta' = \delta/(MHT^2)$ gives

us that with probability at least $1-\delta'/2$ for all $m \in \mathcal{M}, t \in [T], h \in [H]$ simultaneously,

$$\left\|\mathbf{S}_{m,h}^t\right\|_{(\mathbf{\Lambda}_{m,h}^t)^{-1}} \leq 2H\sqrt{\log\left(\frac{\det\left(\widetilde{\mathbf{\Lambda}}_{m,h}^t\right)}{\det\left(\lambda\mathbf{I}_d\right)}\right) + \log\left(MHT^2 \cdot \frac{|\mathcal{N}_\epsilon|}{\delta'}\right) + \frac{2M^2t^2\epsilon^2}{H^2\lambda}}. \tag{20}$$

$$\leq 2H\sqrt{\log\left(\frac{\det\left(\mathbf{\Lambda}_h^t\right)}{\det\left(\lambda\mathbf{I}_d\right)}\right) + 2\log\left(\frac{MHT^2|\mathcal{N}_\epsilon|}{\delta'}\right) + \frac{2t^2\epsilon^2}{H^2\lambda}} \tag{21}$$

$$\leq 2H\sqrt{d\log\frac{t+\lambda}{\lambda} + 4\log(MHT) + 2\log\left(\frac{|\mathcal{N}_\epsilon|}{\delta'}\right) + \frac{2t^2\epsilon^2}{H^2\lambda}}$$

(AM $\geqslant$ GM; determinant-trace inequality)

Step 2. Here $\mathcal{N}_\epsilon$ is an $\epsilon-$covering of the function class $\mathcal{V}_{\text{UCB}}$ for any $h \in [H], m \in [M]$ or $t \in [T]$ under the distance function $\text{dist}(V, V') = \sup_{x\in\mathcal{S}} |V(x) - V'(x)|$. To bound this quantity by the appropriate covering number, we first observe that for any $V \in \mathcal{V}_{\text{UCB}}$, we have that the policy weights are bounded as $2H\sqrt{dMT/\lambda}$ (Lemma 26). Therefore, by Lemma 31 we have for any constant B such that $\beta_{m,h}^t \leq B$,

$$\log|\mathcal{N}_\varepsilon| \leq d\log\left(1 + 8H\sqrt{\frac{dMT}{\lambda\epsilon^2}}\right) + d^2\log\left(1 + \frac{8d^{1/2}B^2}{\lambda\epsilon^2}\right). \tag{22}$$

Recall that we select the hyperparameters $\lambda = 1$ and $\beta = \mathcal{O}(dH\sqrt{\log(TMH)})$, and to balance the terms in $\bar{\beta}_h^t$ we select $\epsilon = \epsilon^\star = dH/T$. Finally, we obtain that for some absolute constant c_β, by replacing the above values,

$$\log|\mathcal{N}_\varepsilon| \leq d\log\left(1 + 8\sqrt{\frac{MT^3}{d}}\right) + d^2\log\left(1 + 8c_\beta d^{1/2}T^2\log(TMH)\right). \tag{23}$$

Therefore, for some absolute constant C' independent of M, T, H, d and c_β, we have,

$$\log|\mathcal{N}_\varepsilon| \leq C'd^2\log\left(CdT\log(TMH)\right). \tag{24}$$

Replacing this result in the result from Step 1, we have that with probability at least $1-\delta'/2$ for all $m \in \mathcal{M}, t \in [T], h \in [H]$ simultaneously,

$$\left\|\mathbf{S}_{m,h}^t\right\|_{(\mathbf{\Lambda}_{m,h}^t)^{-1}}$$
$$\leq 2H\sqrt{(d+2)\log\frac{t+\lambda}{\lambda} + 2\log\left(\frac{1}{\alpha}\right) + C'd^2\log\left(c_\beta dT\log(TMH)\right) + 2 + 4\log(TMH)}.$$

This implies that there exists an absolute constant C independent of M, T, H, d and c_β, such that, with probability at least $1-\delta'/2$ for all $m \in \mathcal{M}, t \in [T], h \in [H]$ simultaneously,

$$\left\|\mathbf{S}_{m,h}^t\right\|_{(\mathbf{\Lambda}_{m,h}^t)^{-1}} \leq C \cdot dH\sqrt{2\log\left(\frac{(c_\beta+2)dMTH}{\delta'}\right)}. \tag{25}$$

Now, following the procedure in Lemma B.4 of Jin et al. (2020), we can select c_β such that we have,

$$\left\|\mathbf{S}_{m,h}^t\right\|_{(\mathbf{\Lambda}_{m,h}^t)^{-1}} \leq c_\beta \cdot dH\sqrt{2\log\left(\frac{dMTH}{\delta'}\right)}. \tag{26}$$

This finishes the proof. □

Next, we present the key result for cooperative value iteration, which demonstrates that for any agent the estimated $Q-$values have bounded error for any policy π. This result is an extension of Lemma B.4 of Jin et al. (2020) on to the homogenous setting.

Lemma 5. *There exists an absolute constant c_β such that for $\beta^t_{m,h} = c_\beta \cdot dH\sqrt{\log(2dMHT/\delta')}$ for any policy π, such that for each $x \in \mathcal{S}, a \in \mathcal{A}$ we have for all $m \in \mathcal{M}, t \in [T], h \in [H]$ simultaneously, with probability at least $1 - \delta'/2$,*

$$\left|\langle \boldsymbol{\phi}(x,a), \mathbf{w}^t_{m,h} - \mathbf{w}^\pi_h \rangle\right| \leq \mathbb{P}_h(V^t_{m,h+1} - V^\pi_{m,h+1})(x,a) + c_\beta \cdot dH \cdot \|\boldsymbol{\phi}(z)\|_{(\boldsymbol{\Lambda}^t_{m,h})^{-1}} \cdot \sqrt{2\log\left(\frac{dMTH}{\delta'}\right)}.$$

Proof. By the Bellman equation and the assumption of the linear MDP (Definition 1), we have that for any policy $\boldsymbol{\pi}$, there exist weights $\mathbf{w}^\pi_h$ such that, for all $z \in \mathcal{Z}$,

$$\langle \boldsymbol{\phi}(z), \mathbf{w}^\pi_h \rangle = r_h(z) + \mathbb{P}_h V^\pi_{h+1}(z). \tag{27}$$

The set of all observations available to any agent at instant t is given by $\mathcal{U}^m_h(t)$ for step h, with the cardinality of this set being $U^h_m(t)$. For convenience, let us assume an ordering $\tau = 1, ..., U^h_m(t)$ over this set and use the shorthand $U_m = U^h_m(t)$. Therefore, we have, for any $m \in \mathcal{M}$,

$$\mathbf{w}^t_{m,h} - \mathbf{w}^\pi_h = (\boldsymbol{\Lambda}^t_{m,h})^{-1} \sum_{\tau=1}^{U_m} \left[\boldsymbol{\phi}_\tau[(r_h + V^t_{m,h+1})(x_\tau)]\right] - \mathbf{w}^\pi_h \tag{28}$$

$$= (\boldsymbol{\Lambda}^t_{m,h})^{-1} \left\{ -\lambda \mathbf{w}^\pi_h + \sum_{\tau=1}^{U_m} \left[\boldsymbol{\phi}_\tau[V^t_{m,h+1}(x'_\tau) - \mathbb{P}_h V^\pi_{m,h+1}(x_\tau, a_\tau)]\right] \right\}. \tag{29}$$

$$\implies \mathbf{w}^t_{m,h} - \mathbf{w}^\pi_h = \underbrace{-\lambda(\boldsymbol{\Lambda}^t_{m,h})^{-1}\mathbf{w}^\pi_h}_{\mathbf{v}_1} + \underbrace{(\boldsymbol{\Lambda}^t_{m,h})^{-1} \left\{ \sum_{\tau=1}^{U_m} \left[\boldsymbol{\phi}_\tau[V^t_{m,h+1}(x'_\tau) - \mathbb{P}_h V^t_{m,h+1}(z_\tau)]\right] \right\}}_{\mathbf{v}_2}$$

$$+ \underbrace{(\boldsymbol{\Lambda}^t_{m,h})^{-1} \left\{ \sum_{\tau=1}^{U_m} \left[\boldsymbol{\phi}_\tau[\mathbb{P}_h V^t_{m,h+1} - \mathbb{P}_h V^\pi_{m,h+1})(z_\tau)]\right] \right\}}_{\mathbf{v}_3}. \tag{30}$$

Now, we know that for any $z \in \mathcal{Z}$ for any policy π,

$$|\langle \boldsymbol{\phi}(z), \mathbf{v}_1 \rangle| \leq \lambda \left|\langle \boldsymbol{\phi}(z), \boldsymbol{\Lambda}^t_{m,h})^{-1}\mathbf{w}^\pi_h \rangle\right| \leq \lambda \cdot \|\mathbf{w}^\pi_h\| \|\boldsymbol{\phi}(z)\|_{(\boldsymbol{\Lambda}^t_{m,h})^{-1}} \leq 2H\lambda\sqrt{d}\|\boldsymbol{\phi}(z)\|_{(\boldsymbol{\Lambda}^t_{m,h})^{-1}} \tag{31}$$

Here the last inequality follows from Lemma 24. For the second term, we have by Lemma 4 that there exists an absolute constant C independent of M, T, H, d and c_β, such that, with probability at least $1 - \delta'/2$ for all $m \in \mathcal{M}, t \in [T], h \in [H]$ simultaneously,

$$|\langle \boldsymbol{\phi}(z), \mathbf{v}_2 \rangle| \leq \|\boldsymbol{\phi}(z)\|_{(\boldsymbol{\Lambda}^t_{m,h})^{-1}} \cdot c_\beta \cdot dH \cdot \sqrt{2\log\left(\frac{dMTH}{\delta'}\right)}. \tag{32}$$

For the last term, note that,

$$|\langle \boldsymbol{\phi}(z), \mathbf{v}_3 \rangle| \tag{33}$$

$$= \left\langle \boldsymbol{\phi}(z), (\boldsymbol{\Lambda}^t_{m,h})^{-1} \left\{ \sum_{\tau=1}^{U_m} \left[\boldsymbol{\phi}_\tau [\mathbb{P}_h V^t_{m,h+1} - \mathbb{P}_h V^\pi_{m,h+1})(z_\tau)] \right] \right\} \right\rangle \tag{34}$$

$$= \left\langle \boldsymbol{\phi}(z), (\boldsymbol{\Lambda}^t_{m,h})^{-1} \sum_{\tau=1}^{U_m} \left[\boldsymbol{\phi}_\tau \boldsymbol{\phi}_\tau^\top \int (V^t_{m,h+1} - V^\pi_{m,h+1})(x') d\boldsymbol{\mu}_h(x') \right] \right\rangle \tag{35}$$

$$= \left\langle \boldsymbol{\phi}(z), \int (V^t_{m,h+1} - V^\pi_{m,h+1})(x') d\boldsymbol{\mu}_h(x') \right\rangle - \lambda \left\langle \boldsymbol{\phi}(z), (\boldsymbol{\Lambda}^t_{m,h})^{-1} \int (V^t_{m,h+1} - V^\pi_{m,h+1})(x') d\boldsymbol{\mu}_h(x') \right\rangle \tag{36}$$

$$= \mathbb{P}_h (V^t_{m,h+1} - V^\pi_{m,h+1})(x,a) - \lambda \left\langle \boldsymbol{\phi}(z), (\boldsymbol{\Lambda}^t_{m,h})^{-1} \int (V^t_{m,h+1} - V^\pi_{m,h+1})(x') d\boldsymbol{\mu}_h(x') \right\rangle \tag{37}$$

$$= \mathbb{P}_h (V^t_{m,h+1} - V^\pi_{m,h+1})(x,a) + 2H\sqrt{d\lambda} \|\boldsymbol{\phi}(z)\|_{(\boldsymbol{\Lambda}^t_{m,h})^{-1}}. \tag{38}$$

Putting it all together, we have that since $\langle \boldsymbol{\phi}(z), \mathbf{w}^t_{m,h} - \mathbf{w}^\pi_h \rangle = \langle \boldsymbol{\phi}(z), \mathbf{v}_1 + \mathbf{v}_2 + \mathbf{v}_3 \rangle$, there exists an absolute constant C independent of M, T, H, d and c_β, such that, with probability at least $1 - \delta'/2$ for all $m \in \mathcal{M}, t \in [T], h \in [H]$ simultaneously,

$$\begin{aligned} \left| \langle \boldsymbol{\phi}(x,a), \mathbf{w}^t_{m,h} - \mathbf{w}^\pi_h \rangle \right| &\leq \mathbb{P}_h (V^t_{m,h+1} - V^\pi_{m,h+1})(x,a) \\ &\quad + \|\boldsymbol{\phi}(z)\|_{(\boldsymbol{\Lambda}^t_{m,h})^{-1}} \left(C \cdot dH \cdot \sqrt{2 \log \left((c_\beta + 2) \frac{dMTH}{\delta'} \right)} + 2H\sqrt{d\lambda} + 2H\lambda\sqrt{d} \right) \end{aligned}$$

Since $\lambda \leq 1$ and since C is independent of c_β, we can select c_β such that we have the following for any $(x,a) \in \mathcal{S} \times \mathcal{A}$ with probability $1 - \delta'/2$ simultaneously for all $h \in [H], m \in \mathcal{M}, t \in [T]$,

$$\left| \langle \boldsymbol{\phi}(x,a), \mathbf{w}^t_{m,h} - \mathbf{w}^\pi_h \rangle \right| \leq \mathbb{P}_h (V^t_{m,h+1} - V^\pi_{m,h+1})(x,a) + c_\beta \cdot dH \cdot \|\boldsymbol{\phi}(z)\|_{(\boldsymbol{\Lambda}^t_{m,h})^{-1}} \cdot \sqrt{2 \log \left(\frac{dMTH}{\delta'} \right)}. \tag{39}$$

□

Lemma 6 (UCB in the Homogenous Setting)**.** *With probability at least $1 - \delta'/2$, we have that for all $(x,a,h,t,m) \in \mathcal{S} \times \mathcal{A} \times [H] \times [T] \times \mathcal{M}$, $Q^t_{m,h}(x,a) \geq Q^\star_{m,h}(x,a)$.*

Proof. The proof is done by induction, identical to the proof in Lemma B.5 of Jin et al. (2020), and we urge the reader to refer to the aforementioned source. □

Lemma 7 (Recursive Relation in Homogenous Settings)**.** *Let $\delta^t_{m,h} = V^t_{m,h}(x^t_{m,h}) - V^{\pi_t}_{m,h}(x^t_{m,h})$, and $\xi^t_{m,h+1} = \mathbb{E}\left[\delta^t_{m,h} | x^t_{m,h}, a^t_{m,h} \right] - \delta^t_{m,h}$. Then, with probability at least $1 - \alpha$, for all $(t,m,h) \in [T] \times \mathcal{M} \times [H]$ simultaneously,*

$$\delta^t_{m,h} \leq \delta^t_{m,h+1} + \xi^t_{m,h+1} + 2 \left\| \boldsymbol{\phi}(x^t_{m,h}, a^t_{m,h}) \right\|_{(\boldsymbol{\Lambda}^t_{m,h})^{-1}} \cdot c_\beta \cdot dH \cdot \sqrt{2 \log \left(\frac{dMTH}{\alpha} \right)}. \tag{40}$$

Proof. By Lemma 5, we have that for any $(x, a, h, m, t) \in \mathcal{S} \times \mathcal{A} \times [H] \times \mathcal{M} \times [T]$ with probability at least $1 - \alpha/2$,

$$Q^t_{m,h}(x,a) - Q^{\pi_t}_{m,h}(x,a) \leq \mathbb{P}_h(V^t_{m,h+1} - V^{\pi_t}_{m,h})(x,a) + 2\left\|\boldsymbol{\phi}(x,a)\right\|_{(\boldsymbol{\Lambda}^t_{m,h})^{-1}} \cdot c_\beta \cdot dH \cdot \sqrt{2\log\left(\frac{dMTH}{\alpha}\right)}.$$

Replacing the definition of $\delta^t_{m,h}$ and $V^{\pi_t}_{m,h}$ finishes the proof. □

Lemma 8. *For $\xi^t_{m,h}$ as defined earlier and any $\delta \in (0,1)$, we have with probability at least $1 - \delta/2$,*

$$\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{h=1}^{H} \xi^t_{m,h} \leq \sqrt{2H^3MT\log\left(\frac{2}{\alpha}\right)}. \tag{41}$$

Proof. We generalize the procedure from Jin et al. (2018), by demonstrating that the overall sums can be written as bounded martingale difference sequences with respect to an appropriately chosen filtration. For any $(t, m, h) \in [T] \times [M] \times [H]$, we define the σ-algebra $\mathcal{F}_{t,m,h}$ as,

$$\mathcal{F}_{t,m,h} = \sigma\left(\left\{\left(x^\tau_{l,i}, a^\tau_{l,i}\right)\right\}_{(\tau,l,i)\in[t-1]\times[M]\times[H]} \cup \left\{\left(x^t_{l,i}, a^t_{l,i}\right)\right\}_{(i,l)\in[h]\times[m-1]} \cup \left\{\left(x^t_{m,i}, a^t_{m,i}\right)\right\}_{i\in[h]}\right) \tag{42}$$

Where we denote the $\sigma-$algebra generated by a finite set by $\sigma(\cdot)$. For any $t \in [T], m \in [M], h \in [H]$, we can define the timestamp index $\tau(t, m, h)$ as $\tau(t, m, h) = (t-1) \cdot HM + h(m-1) + (h-1)$. We see that this ordering ensures that the $\sigma-$algebras from earlier form a filtration. We can see that for any agent $m \in [M]$, $Q^t_{m,h}$ and $V^t_{m,h}$ are both obtained based on the trajectories of the first $(t-1)$ episodes, and are both measurable with respect to $\mathcal{F}_{t,1,1}$ (which is a subset of $\mathcal{F}_{t,m,h}$ for all $h \in [H]$ and $m \in [M]$). Moreover, note that $a^t_{m,h} \sim \pi_{m,t}(\cdot|x^t_{m,h})$ and $x^t_{m,h+1} \sim \mathbb{P}_{m,h}(\cdot|x^t_{m,h}, a^t_{m,h})$. Therefore,

$$\mathbb{E}_{\mathbb{P}_{m,h}}[\xi^t_{m,h}|\mathcal{F}_{t,m,h}] = 0. \tag{43}$$

where we set $\mathcal{F}_{1,0,0}$ with the empty set. We define the martingale $\{U_{t,m,h}\}_{(t,h,m)\in[T]\times[M]\times[H]}$ indexed by $\tau(t, m, h)$ defined earlier, as follows. For any $(t, m, h) \in [T] \times [M] \times [H]$, we define

$$U_{t,m,h} = \left\{\sum_{(a,b,c)} \xi^a_{b,c} : \tau(a,b,c) \leq \tau(t,m,h)\right\}, \tag{44}$$

Additionally, we have that

$$U_{T,M,H} = \sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{h=1}^{H} \xi^t_{m,h}. \tag{45}$$

Now, we have that for each $m \in \mathcal{M}$, $V^t_{m,h}, Q^t_{m,h}, V^{\pi_{m,t}}_{m,h}, Q^{\pi_{m,t}}_{m,h}$ take values in $[0, H]$. Therefore, wh have that $\xi^t_{m,h} \leq 2H$ for all $(t, m, h) \in [T] \times [M] \times [H]$. This allows us to apply the Azuma-Hoeffding inequality (Azuma, 1967) to $U_{T,M,H}$. We therefore obtain that for all $\tau > 0$,

$$\mathbb{P}\left(\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{h=1}^{H} \xi^t_{m,h} > \tau\right) \leq \exp\left(\frac{-\tau^2}{2H^3MT}\right). \tag{46}$$

Setting the RHS as $\alpha/2$, we obtain that with probability at least $1-\alpha/2$,

$$\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{h=1}^{H}\xi_{m,h}^{t} \leq \sqrt{2H^3MT\log\left(\frac{2}{\alpha}\right)}. \tag{47}$$

□

Lemma 9 (Variance control via communication in homogenous factored environments)**.** *Let Algorithm 1 be run for any $T>0$ and $M\geq 1$, with S as the communication control factor. Then, the following holds for the cumulative variance.*

$$\sum_{m=1}^{M}\sum_{t=1}^{T}\left\|\boldsymbol{\phi}(z_{m,h}^{t})\right\|_{(\boldsymbol{\Lambda}_{m,h}^{t})^{-1}} \leq 2\log\left(\frac{\det\left(\boldsymbol{\Lambda}_{h}^{T}\right)}{\det\left(\lambda\mathbf{I}_{d}\right)}\right)\left(\frac{M}{\log 2}\right)\sqrt{S}+2\sqrt{2MT\log\left(\frac{\det\left(\boldsymbol{\Lambda}_{h}^{T}\right)}{\det\left(\lambda\mathbf{I}_{d}\right)}\right)}. \tag{48}$$

Proof. Consider the following mappings $\nu_M, \nu_T : [MT] \to [M]\times[T]$.

$$\nu_M(\tau) = \tau(\text{mod } M), \text{and } \nu_T = \left\lceil \frac{\tau}{M} \right\rceil. \tag{49}$$

Now, consider $\bar{\boldsymbol{\Lambda}}_h^{\tau} = \lambda\mathbf{I}_d + \sum_{u=1}^{\tau}\boldsymbol{\phi}\left(z_{\nu_M(u),h}^{\nu_T(u)}\right)\boldsymbol{\phi}\left(z_{\nu_M(u),h}^{\nu_T(u)}\right)^{\top}$ for $\tau>0$ and $\bar{\boldsymbol{\Lambda}}_h^{0}=\lambda\mathbf{I}_d$. Furthermore, assume that global synchronizations occur at round $\boldsymbol{\sigma}=(\sigma_1,...,\sigma_n)$ where there are a total of $n-1$ rounds of synchronization and $\sigma_i\in[T]\forall\ i\in[N-1]$ and $\sigma_n=T$, i.e., the final round. Let $R_h = \left\lceil \log\left(\frac{\det\left(\bar{\boldsymbol{\Lambda}}_h^{T}\right)}{\det\left(\lambda\mathbf{I}_d\right)}\right)\right\rceil$. It follows that there exist at most R_h periods between synchronization (i.e., intervals σ_{k-1} to σ_k for $k\in[N]$) in which the following does not hold true:

$$1 \leq \frac{\det(\bar{\boldsymbol{\Lambda}}_h^{\sigma_k})}{\det(\bar{\boldsymbol{\Lambda}}_h^{k-1})} \leq 2. \tag{50}$$

Let us denote the event when Equation (50) does holds for an interval σ_{k-1} to σ_k as E. Now, for any $t\in[\sigma_{k-1},\sigma_k]$, we have, for any $m\in[M]$,

$$\left\|\boldsymbol{\phi}(z_{m,h}^{t})\right\|_{(\boldsymbol{\Lambda}_{m,h}^{t})^{-1}} \leq \left\|\boldsymbol{\phi}(z_{m,h}^{t})\right\|_{(\bar{\boldsymbol{\Lambda}}_{h}^{t})^{-1}}\sqrt{\frac{\det\left(\bar{\boldsymbol{\Lambda}}_{h}^{t}\right)}{\det\left(\boldsymbol{\Lambda}_{m,h}^{t}\right)}} \leq \left\|\boldsymbol{\phi}(z_{m,h}^{t})\right\|_{(\bar{\boldsymbol{\Lambda}}_{h}^{t})^{-1}}\sqrt{\frac{\det\left(\bar{\boldsymbol{\Lambda}}_{h}^{\sigma_k}\right)}{\det\left(\bar{\boldsymbol{\Lambda}}_{h}^{\sigma_{k-1}}\right)}} \tag{51}$$

$$\leq 2\left\|\boldsymbol{\phi}(z_{m,h}^{t})\right\|_{(\bar{\boldsymbol{\Lambda}}_{h}^{t})^{-1}}. \tag{52}$$

Here, the first inequality follows from the fact that $\boldsymbol{\Lambda}_{m,h}^{t} \preccurlyeq \bar{\boldsymbol{\Lambda}}_h^{t}$, the second inequality follows from the fact that $\boldsymbol{\Lambda}_{m,h}^{t} \preccurlyeq \bar{\boldsymbol{\Lambda}}_h^{\sigma_k} \implies \det(\boldsymbol{\Lambda}_{m,h}^{t}) \leqslant \det(\bar{\boldsymbol{\Lambda}}_h^{\sigma_k})$, and $\boldsymbol{\Lambda}_{m,h}^{t} \succcurlyeq \bar{\boldsymbol{\Lambda}}_h^{\sigma_{k-1}} \implies \det(\boldsymbol{\Lambda}_{m,h}^{t}) \geqslant \det(\bar{\boldsymbol{\Lambda}}_h^{\sigma_{k-1}})$; and the final inequality follows from the fact that event E holds. Now, we can consider the partial sums only in the intervals for which event E holds. For any $t\in[T]$, consider $\sigma(t) = \max_{i\in[N]}\{\sigma_i|\sigma_i\leq t\}$ denote the last round of synchronization prior to episode t. Then,

$$\sum_{t:E \text{ is true}}^{T}\sum_{m=1}^{M}\left\|\boldsymbol{\phi}(z_{m,h}^{t})\right\|_{(\boldsymbol{\Lambda}_{m,h}^{t})^{-1}} \leq \sqrt{MT\sum_{m=1}^{M}\sum_{t:E \text{ is true}}^{T}\left\|\boldsymbol{\phi}(z_{m,h}^{t})\right\|^2_{(\boldsymbol{\Lambda}_{m,h}^{t})^{-1}}} \tag{53}$$

$$\leq 2\sqrt{MT\sum_{m=1}^{M}\sum_{t:E \text{ is true}}^{T}\left\|\boldsymbol{\phi}(z_{m,h}^{t})\right\|^2_{(\bar{\boldsymbol{\Lambda}}_{h}^{t})^{-1}}} \leq 2\sqrt{MT\sum_{m=1}^{M}\sum_{t=1}^{T}\left\|\boldsymbol{\phi}(z_{m,h}^{t})\right\|^2_{(\bar{\boldsymbol{\Lambda}}_{h}^{t})^{-1}}} \tag{54}$$

$$= 2\sqrt{MT\sum_{m=1}^{M}\sum_{\tau=1}^{T}\left\|\boldsymbol{\phi}(z_{\nu_M(\tau),h}^{\nu_T(\tau)})\right\|^2_{(\bar{\boldsymbol{\Lambda}}_{h}^{t})^{-1}}} = 2\sqrt{MT\log\left(\frac{\det\left(\boldsymbol{\Lambda}_{h}^{T}\right)}{\det\left(\lambda\mathbf{I}_{d}\right)}\right)}. \tag{55}$$

Here, the first inequality follows from Cauchy-Schwarz, the second inequality follows from the fact that event E holds, and the final equality follows from Lemma 33. Now, we sum up the cumulative sum for episodes when E does not hold. Consider an interval σ_{k-1} to σ_k for $k \in [N]$ of length $\Delta_k = \sigma_k - \sigma_{k-1}$ in which E does not hold. We have that,

$$\sum_{m=1}^{M} \sum_{t=\sigma_{k-1}}^{\sigma_k} \left\|\boldsymbol{\phi}(z_{m,h}^t)\right\|_{(\boldsymbol{\Lambda}_{m,h}^t)^{-1}} \leq \sum_{m=1}^{M} \sqrt{\Delta_{k,h} \sum_{t=\sigma_{k-1}}^{\sigma_k} \left\|\boldsymbol{\phi}(z_{m,h}^t)\right\|^2_{(\boldsymbol{\Lambda}_{m,h}^t)^{-1}}} \tag{56}$$

$$\leq \sum_{m=1}^{M} \sqrt{\Delta_{k,h} \cdot \log_\omega \left(\frac{\det(\boldsymbol{\Lambda}_{m,h}^{\sigma_k})}{\det(\boldsymbol{\Lambda}_{m,h}^{\sigma_{k-1}})}\right)} \leq \sum_{m=1}^{M} \sqrt{\Delta_{k,h} \cdot \log_\omega \left(\frac{\det(\bar{\boldsymbol{\Lambda}}_{h}^{\sigma_k})}{\det(\bar{\boldsymbol{\Lambda}}_{h}^{\sigma_{k-1}})}\right)} \leq M\sqrt{S}. \tag{57}$$

The last inequality follows from the synchronization criterion. Now, note that there are at most R_h periods in which event E does not hold, and hence the total sum in this period can be bound as,

$$\sum_{(t:E \text{ is not true})}^{T} \sum_{m=1}^{M} \left\|\boldsymbol{\phi}(z_{m,h}^t)\right\|_{(\boldsymbol{\Lambda}_{m,h}^t)^{-1}} \leq R_h M\sqrt{S} \leq \left(\log\left(\frac{\det\left(\boldsymbol{\Lambda}_h^T\right)}{\det\left(\lambda \mathbf{I}_d\right)}\right) + 1\right) M\sqrt{S}. \tag{58}$$

Therefore, we can bound the total variance as,

$$\sum_{m=1}^{M} \sum_{t=1}^{T} \left\|\boldsymbol{\phi}(z_{m,h}^t)\right\|_{(\boldsymbol{\Lambda}_{m,h}^t)^{-1}} \leq \left(\log\left(\frac{\det\left(\boldsymbol{\Lambda}_h^T\right)}{\det\left(\lambda \mathbf{I}_d\right)}\right) + 1\right) M\sqrt{S} + 2\sqrt{MT \log\left(\frac{\det\left(\boldsymbol{\Lambda}_h^T\right)}{\det\left(\lambda \mathbf{I}_d\right)}\right)}$$

$$\leq 2\log\left(\frac{\det\left(\boldsymbol{\Lambda}_h^T\right)}{\det\left(\lambda \mathbf{I}_d\right)}\right)\left(\frac{M}{\log 2}\right)\sqrt{S} + 2\sqrt{2MT \log\left(\frac{\det\left(\boldsymbol{\Lambda}_h^T\right)}{\det\left(\lambda \mathbf{I}_d\right)}\right)}.$$

□

We are now ready to prove Theorem 1. We first restate the Theorem for completeness.

Theorem 1 (Homogenous Regret). Algorithm 1 when run on M agents with communication threshold S, $\beta_t = \mathcal{O}(H\sqrt{d\log(tMH)})$ and $\lambda = 1$ obtains the following cumulative regret after T episodes, with probability at least $1-\alpha$,

$$\mathfrak{R}(T) = \widetilde{\mathcal{O}}\left(d^{\frac{3}{2}} H^2 \left(M\sqrt{S} + \sqrt{MT}\right)\sqrt{\log\left(\frac{1}{\alpha}\right)}\right).$$

Proof. We have by the definition of group regret:

$$\mathfrak{R}(T) = \sum_{m=1}^{M} \sum_{t=1}^{T} V_{m,1}^{\star}(x_{m,1}^t) - V_{m,1}^{\pi_t}(x_{m,1}^t) \leq \sum_{m=1}^{M} \sum_{t=1}^{T} \delta_{m,1}^t \tag{59}$$

$$\leq \sum_{m=1}^{M} \sum_{t=1}^{T} \sum_{h=1}^{H} \xi_{m,h}^t + 2c_\beta \cdot dH \cdot \sqrt{2\log\left(\frac{dMTH}{\alpha}\right)} \left(\sum_{t=1}^{T} \sum_{m=1}^{M} \sum_{h=1}^{H} \|\boldsymbol{\phi}(x,a)\|_{(\boldsymbol{\Lambda}_{m,h}^t)^{-1}}\right). \tag{60}$$

Where the last inequality holds with probability at least $1-\alpha/2$, via Lemma 7 and Lemma 6. Next, we can bound the first term via Lemma 8. We have with probability at least $1-\alpha$, for some absolute constant c_β,

$$\mathfrak{R}(T) \leq \sqrt{2H^3 MT \log\left(\frac{2}{\alpha}\right)} + 2c_\beta \cdot dH \cdot \sqrt{2\log\left(\frac{dMTH}{\alpha}\right)} \left(\sum_{t=1}^{T} \sum_{m=1}^{M} \sum_{h=1}^{H} \|\boldsymbol{\phi}(x,a)\|_{(\boldsymbol{\Lambda}_{m,h}^t)^{-1}}\right). \tag{61}$$

Finally, to bound the summation, we use Lemma 9. We have that,

$$\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{h=1}^{H}\|\boldsymbol{\phi}(x,a)\|_{(\boldsymbol{\Lambda}_{m,h}^{t})^{-1}} \leq 2\sum_{h=1}^{H}\left(\log\left(\frac{\det\left(\boldsymbol{\Lambda}_{h}^{T}\right)}{\det\left(\lambda\mathbf{I}_{d}\right)}\right)\left(\frac{M}{\log 2}\right)\sqrt{S}+2\sqrt{2MT\log\left(\frac{\det\left(\boldsymbol{\Lambda}_{h}^{T}\right)}{\det\left(\lambda\mathbf{I}_{d}\right)}\right)}\right) \tag{62}$$

$$\leq 2H\log(dMT)M\sqrt{S}+2\sqrt{2dMT\log(MT)}. \tag{63}$$

Where the last inequality is an application of the determinant-trace inequality and using the fact that $\|\boldsymbol{\phi}(\cdot)\|_2 \leq 1$. Replacing this result, we have that with probability at least $1-\alpha$,

$$\mathfrak{R}(T) \leq \sqrt{2H^3MT\log\left(\frac{2}{\alpha}\right)}+2c_\beta dH^2\sqrt{2\log\left(\frac{dMTH}{\alpha}\right)}\left(2\log(dMT)M\sqrt{S}+2\sqrt{2dMT\log(MT)}\right) \tag{64}$$

$$=\widetilde{\mathcal{O}}\left(d^{3/2}H^2\left(M\sqrt{S}+\sqrt{MT}\right)\sqrt{\log\left(\frac{1}{\alpha}\right)}\right). \tag{65}$$

□

B.3 Proof of Theorem 2 (Small Heterogeneity)

The proof for this section is very similar to that of Theorem 1 with some modifications to handle the differences between MDPs as a case of model misspecification. First, we must bound the difference in the projected $Q-$values for any pair of MDPs under the small heterogeneity condition.

Lemma 10. *Under the small heterogeneity condition (Assumption 1), for any policy $\boldsymbol{\pi}$ over $\mathcal{S} \times \mathcal{A}$, let the corresponding weights at step h for two MDPs $m, m' \in \mathcal{M}$ be given by $\mathbf{w}^{\pi}_{m,h}, \mathbf{w}^{\pi}_{m',h}$ respectively, i.e., $Q^{\pi}_{m,h}(x,a) = \langle \boldsymbol{\phi}(x,a), \mathbf{w}^{\pi}_{m,h} \rangle$ and $Q^{\pi}_{m',h}(x,a) = \langle \boldsymbol{\phi}(x,a), \mathbf{w}^{\pi}_{m',h} \rangle$. Then, we have for any $x, a \in \mathcal{S} \times \mathcal{A}$,*

$$\left| \langle \boldsymbol{\phi}(x,a), \mathbf{w}^{\pi}_{m,h} - \mathbf{w}^{\pi}_{m',h} \rangle \right| \leq 2H\xi.$$

Proof. The proof follows from the fact that for any $h \in [H]$,

$$\begin{aligned}
&\left| \langle \boldsymbol{\phi}(x,a), \mathbf{w}^{\pi}_{m,h} - \mathbf{w}^{\pi}_{m',h} \rangle \right| && (66)\\
&= \left| Q^{\pi}_{m,h}(x,a) - Q^{\pi}_{m',h}(x,a) \right| && (67)\\
&\leq |r_{m,h}(x,a) - r_{m',h}(x,a)| + \left| \mathbb{P}_{m,h} V^{\pi}_{m,h+1}(x,a) - \mathbb{P}_{m',h} V^{\pi}_{m',h+1}(x,a) \right| && (68)\\
&\leq |r_{m,h}(x,a) - r_{m',h}(x,a)| + \sup_{x' \in \mathcal{S}} |V^{\pi}_{m,h+1}(x') - V^{\pi}_{m',h+1}(x')| \cdot \|(\mathbb{P}_{m,h} - \mathbb{P}_{m',h})(x,a)\|_{\mathrm{TV}} && (69)\\
&\leq |r_{m,h}(x,a) - r_{m',h}(x,a)| + H \cdot \|(\mathbb{P}_{m,h} - \mathbb{P}_{m',h})(x,a)\|_{\mathrm{TV}} && (70)\\
&\leq 2H\xi. && (71)
\end{aligned}$$

Here the last inequality follows from Assumption 1. □

Now, we reproduce a general result bounding bias introduced by the potentially adversarial noise due to misspecification.

Lemma 11. *Let $\{\varepsilon_\tau\}_{\tau=1}^{t}$ be a sequence such that $|\varepsilon_\tau| \leq B$. We have, for any $(h,t,m) \in [H] \times [T] \times \mathcal{M}$, and $\boldsymbol{\phi} \in \mathbb{R}^d$,*

$$|\boldsymbol{\phi}^\top (\boldsymbol{\Lambda}^t_{m,h})^{-1} \sum_{\tau=1}^{U^m_h(t)} \boldsymbol{\phi}_\tau \varepsilon_\tau| \leq B\sqrt{dMt} \|\boldsymbol{\phi}\|_{(\boldsymbol{\Lambda}^t_{m,h})^{-1}}.$$

Proof. Recall that at any instant the collective set of observations possessed by an agent is given by $\mathcal{U}^m_h(t)$ with size $U^m_h(t) \leq Mt$. We have that,

$$\begin{aligned}
|\boldsymbol{\phi}^\top (\boldsymbol{\Lambda}^t_{m,h})^{-1} \sum_{\tau=1}^{U^m_h(t)} \boldsymbol{\phi}_\tau \varepsilon_\tau| &\leq B \cdot |\boldsymbol{\phi}^\top (\boldsymbol{\Lambda}^t_{m,h})^{-1} \sum_{\tau=1}^{U^m_h(t)} \boldsymbol{\phi}_\tau| && (72)\\
&\leq B \cdot \sqrt{\left[\sum_{\tau=1}^{U^m_h(t)} \boldsymbol{\phi}^\top (\boldsymbol{\Lambda}^t_{m,h})^{-1} \boldsymbol{\phi} \right] \cdot \left[\sum_{\tau=1}^{U^m_h(t)} \boldsymbol{\phi}_\tau^\top (\boldsymbol{\Lambda}^t_{m,h})^{-1} \boldsymbol{\phi}_\tau \right]} && (73)\\
&\leq B\sqrt{dMt} \|\boldsymbol{\phi}\|_{(\boldsymbol{\Lambda}^t_{m,h})^{-1}}. && (74)
\end{aligned}$$

□

Now we present the primary concentration result for the small heterogeneity setting.

Lemma 12. *There exists an absolute constant c_β such that for $\beta^t_{m,h} = c_\beta \cdot dH(\sqrt{\log(2dMHT/\delta')} + \xi\sqrt{dMT})$ for any policy π, there exists a constant c_β such that for each $x \in \mathcal{S}, a \in \mathcal{A}$ we have for*

all $m \in \mathcal{M}, t \in [T], h \in [H]$ *simultaneously, with probability at least* $1 - \delta'/2$,

$$\left|\langle \boldsymbol{\phi}(x,a), \mathbf{w}^t_{m,h} - \mathbf{w}^\pi_{m,h} \rangle\right| \leq$$
$$\mathbb{P}_{m,h}(V^t_{m,h+1} - V^\pi_{m,h+1})(x,a) + c_\beta \cdot dH \cdot \|\boldsymbol{\phi}(z)\|_{(\boldsymbol{\Lambda}^t_{m,h})^{-1}} \left(\sqrt{2\log\left(\frac{dMTH}{\delta'}\right)} + 2\xi\sqrt{dMT}\right).$$

Proof. By the Bellman equation and the assumption of the linear MDP (Definition 1), we have that for any policy $\boldsymbol{\pi}$, there exist weights $\mathbf{w}^\pi_{m,h}$ such that, for all $z \in \mathcal{Z}$,

$$\langle \boldsymbol{\phi}(z), \mathbf{w}^\pi_{m,h} \rangle = r_{m,h}(z) + \mathbb{P}_{m,h} V^\pi_{h+1}(z). \tag{75}$$

Recall that the set of all observations available to any agent at instant t is given by $\mathcal{U}^m_h(t)$ for step h, with the cardinality of this set being $U^h_m(t)$. For convenience, let us assume an ordering $\tau = 1, ..., U^h_m(t)$ over this set and use the shorthand $U_m = U^h_m(t)$. Therefore, we have, for any $m \in \mathcal{M}$,

$$\mathbf{w}^t_{m,h} - \mathbf{w}^\pi_{m,h} \tag{76}$$
$$= (\boldsymbol{\Lambda}^t_{m,h})^{-1} \sum_{\tau=1}^{U_m} \left[\boldsymbol{\phi}_\tau[(r_h + V^t_{m,h+1})(x_\tau)]\right] - \mathbf{w}^\pi_{m,h} \tag{77}$$
$$= (\boldsymbol{\Lambda}^t_{m,h})^{-1} \left\{-\lambda \mathbf{w}^\pi_h + \sum_{\tau=1}^{U_m} \left[\boldsymbol{\phi}_\tau[V^t_{m,h+1}(x'_\tau) - \mathbb{P}_{m_\tau,h} V^\pi_{m,h+1}(x_\tau, a_\tau)]\right]\right\}. \tag{78}$$

$$\implies \mathbf{w}^t_{m,h} - \mathbf{w}^\pi_{m,h} = \underbrace{-\lambda(\boldsymbol{\Lambda}^t_{m,h})^{-1}\mathbf{w}^\pi_{m,h}}_{\mathbf{v}_1} + \underbrace{(\boldsymbol{\Lambda}^t_{m,h})^{-1}\left\{\sum_{\tau=1}^{U_m}\left[\boldsymbol{\phi}_\tau[V^t_{m,h+1}(x'_\tau) - \mathbb{P}_{m,h}V^t_{m,h+1}(z_\tau)]\right]\right\}}_{\mathbf{v}_2}$$
$$+ \underbrace{(\boldsymbol{\Lambda}^t_{m,h})^{-1}\left\{\sum_{\tau=1}^{U_m}\left[\boldsymbol{\phi}_\tau[\mathbb{P}_{m,h}V^t_{m,h+1} - \mathbb{P}_{m,h}V^\pi_{m,h+1})(z_\tau)]\right]\right\}}_{\mathbf{v}_3}$$
$$+ \underbrace{(\boldsymbol{\Lambda}^t_{m,h})^{-1}\left\{\sum_{\tau=1}^{U_m}\left[\boldsymbol{\phi}_\tau[\mathbb{P}_{m,h}V^t_{m,h+1} - \mathbb{P}_{m_\tau,h}V^t_{m,h+1})(z_\tau)]\right]\right\}}_{\mathbf{v}_4}$$
$$+ \underbrace{(\boldsymbol{\Lambda}^t_{m,h})^{-1}\left\{\sum_{\tau=1}^{U_m}\left[\boldsymbol{\phi}_\tau[\mathbb{P}_{m,h}V^\pi_{m,h+1} - \mathbb{P}_{m_\tau,h}V^\pi_{m,h+1})(z_\tau)]\right]\right\}}_{\mathbf{v}_5}. \tag{79}$$

Now, we know that for any $z \in \mathcal{Z}$ for any policy π,

$$|\langle \boldsymbol{\phi}(z), \mathbf{v}_1\rangle| \leq \lambda \left|\langle \boldsymbol{\phi}(z), \boldsymbol{\Lambda}^t_{m,h})^{-1}\mathbf{w}^\pi_h\rangle\right| \leq \lambda \cdot \|\mathbf{w}^\pi_h\| \|\boldsymbol{\phi}(z)\|_{(\boldsymbol{\Lambda}^t_{m,h})^{-1}} \leq 2H\lambda\sqrt{d}\|\boldsymbol{\phi}(z)\|_{(\boldsymbol{\Lambda}^t_{m,h})^{-1}} \tag{80}$$

Here the last inequality follows from Lemma 24. For the second term, we have by Lemma 4 that there exists an absolute constant C independent of M, T, H, d and c_β, such that, with probability at least $1 - \delta'/2$ for all $m \in \mathcal{M}, t \in [T], h \in [H]$ simultaneously,

$$|\langle \boldsymbol{\phi}(z), \mathbf{v}_2\rangle| \leq \|\boldsymbol{\phi}(z)\|_{(\boldsymbol{\Lambda}^t_{m,h})^{-1}} \cdot c_\beta \cdot dH \cdot \sqrt{2\log\left(\frac{dMTH}{\delta'}\right)}. \tag{81}$$

For the third term, note that,

$$|\langle \boldsymbol{\phi}(z), \mathbf{v}_3 \rangle| \tag{82}$$

$$= \left\langle \boldsymbol{\phi}(z), (\boldsymbol{\Lambda}^t_{m,h})^{-1} \left\{ \sum_{\tau=1}^{U_m} \left[\boldsymbol{\phi}_\tau [\mathbb{P}_h V^t_{m,h+1} - \mathbb{P}_h V^\pi_{m,h+1})(z_\tau)] \right] \right\} \right\rangle \tag{83}$$

$$= \left\langle \boldsymbol{\phi}(z), (\boldsymbol{\Lambda}^t_{m,h})^{-1} \sum_{\tau=1}^{U_m} \left[\boldsymbol{\phi}_\tau \boldsymbol{\phi}_\tau^\top \int (V^t_{m,h+1} - V^\pi_{m,h+1})(x') d\boldsymbol{\mu}_h(x') \right] \right\rangle \tag{84}$$

$$= \left\langle \boldsymbol{\phi}(z), \int (V^t_{m,h+1} - V^\pi_{m,h+1})(x') d\boldsymbol{\mu}_h(x') \right\rangle - \lambda \left\langle \boldsymbol{\phi}(z), (\boldsymbol{\Lambda}^t_{m,h})^{-1} \int (V^t_{m,h+1} - V^\pi_{m,h+1})(x') d\boldsymbol{\mu}_h(x') \right\rangle \tag{85}$$

$$= \mathbb{P}_h (V^t_{m,h+1} - V^\pi_{m,h+1})(x,a) - \lambda \left\langle \boldsymbol{\phi}(z), (\boldsymbol{\Lambda}^t_{m,h})^{-1} \int (V^t_{m,h+1} - V^\pi_{m,h+1})(x') d\boldsymbol{\mu}_h(x') \right\rangle \tag{86}$$

$$= \mathbb{P}_h (V^t_{m,h+1} - V^\pi_{m,h+1})(x,a) + 2H\sqrt{d\lambda} \|\boldsymbol{\phi}(z)\|_{(\boldsymbol{\Lambda}^t_{m,h})^{-1}} \tag{87}$$

$$\tag{88}$$

For the fourth and fifth terms, we have that both $[\mathbb{P}_{m,h} V^\pi_{m,h+1} - \mathbb{P}_{m_\tau,h} V^\pi_{m,h+1})(z_\tau)]$ and $[\mathbb{P}_{m,h} V^t_{m,h+1} - \mathbb{P}_{m_\tau,h} V^t_{m,h+1})(z_\tau)]$ are bounded by $H\xi$ (from Assumption 1 and the fact that the value functions are always smaller than H). This gives us, by Lemma 11,

$$|\langle \boldsymbol{\phi}(z), \mathbf{v}_4 + \mathbf{v}_5 \rangle| \leq 2H\xi\sqrt{dMt} \|\boldsymbol{\phi}(z)\|_{(\boldsymbol{\Lambda}^t_{m,h})^{-1}} \tag{89}$$

Putting it all together, we have that since $\langle \boldsymbol{\phi}(z), \mathbf{w}^t_{m,h} - \mathbf{w}^\pi_{m,h} \rangle = \langle \boldsymbol{\phi}(z), \mathbf{v}_1 + \mathbf{v}_2 + \mathbf{v}_3 + \mathbf{v}_4 + \mathbf{v}_5 \rangle$, there exists an absolute constant C independent of M, T, H, d and c_β, such that, with probability at least $1 - \delta'/2$ for all $m \in \mathcal{M}, t \in [T], h \in [H]$ simultaneously,

$$\left| \langle \boldsymbol{\phi}(x,a), \mathbf{w}^t_{m,h} - \mathbf{w}^\pi_{m,h} \rangle \right| \leq \mathbb{P}_{m,h} (V^t_{m,h+1} - V^\pi_{m,h+1})(x,a) + \\ \|\boldsymbol{\phi}(z)\|_{(\boldsymbol{\Lambda}^t_{m,h})^{-1}} \left(C \cdot dH \cdot \sqrt{2 \log \left((c_\beta + 2) \frac{dMTH}{\delta'} \right)} + 2H\sqrt{d\lambda} + 2H\lambda\sqrt{d} + 2H\xi\sqrt{dMT} \right) \tag{90}$$

Since $\lambda \leq 1$ and since C is independent of c_β, we can select c_β such that we have the following for any $(x,a) \in \mathcal{S} \times \mathcal{A}$ with probability $1 - \delta'/2$ simultaneously for all $h \in [H], m \in \mathcal{M}, t \in [T]$,

$$\left| \langle \boldsymbol{\phi}(x,a), \mathbf{w}^t_{m,h} - \mathbf{w}^\pi_{m,h} \rangle \right| \leq \\ \mathbb{P}_{m,h} (V^t_{m,h+1} - V^\pi_{m,h+1})(x,a) + c_\beta \cdot dH \cdot \|\boldsymbol{\phi}(z)\|_{(\boldsymbol{\Lambda}^t_{m,h})^{-1}} \left(\sqrt{2 \log \left(\frac{dMTH}{\delta'} \right)} + 2\xi\sqrt{dMT} \right). \tag{91}$$

□

We now present an analogous recursive relationship in the small heterogeneity setting.

Lemma 13 (Recursive Relation in Small Heterogeneous Settings)**.** *Let* $\delta^t_{m,h} = V^t_{m,h}(x^t_{m,h}) - V^{\pi_t}_{m,h}(x^t_{m,h})$, *and* $\xi^t_{m,h+1} = \mathbb{E}\left[\delta^t_{m,h} | x^t_{m,h}, a^t_{m,h}\right] - \delta^t_{m,h}$. *Then, with probability at least* $1 - \alpha$, *for all* $(t,m,h) \in [T] \times \mathcal{M} \times [H]$ *simultaneously,*

$$\delta^t_{m,h} \leq \delta^t_{m,h+1} + \xi^t_{m,h+1} + c_\beta \cdot dH \cdot \|\boldsymbol{\phi}(x,a)\|_{(\boldsymbol{\Lambda}^t_{m,h})^{-1}} \left(\sqrt{2 \log \left(\frac{dMTH}{\delta'} \right)} + 2\xi\sqrt{dMT} \right).$$

Proof. By Lemma 12, we have that for any $(x, a, h, m, t) \in \mathcal{S} \times \mathcal{A} \times [H] \times \mathcal{M} \times [T]$ with probability at least $1 - \alpha/2$,

$$Q^t_{m,h}(x,a) - Q^{\pi_t}_{m,h}(x,a) = \left|\langle \boldsymbol{\phi}(x,a), \mathbf{w}^t_{m,h} - \mathbf{w}^{\pi}_{m,h}\rangle\right| \leq$$
$$\mathbb{P}_{m,h}(V^t_{m,h+1} - V^{\pi}_{m,h+1})(x,a) + c_\beta \cdot dH \cdot \|\boldsymbol{\phi}(x,a)\|_{(\boldsymbol{\Lambda}^t_{m,h})^{-1}} \left(\sqrt{2\log\left(\frac{dMTH}{\delta'}\right)} + 2\xi\sqrt{dMT}\right). \tag{92}$$

Replacing the definition of $\delta^t_{m,h}$ and $V^{\pi_t}_{m,h}$ finishes the proof. □

We are now ready to prove Theorem 2. We first restate the Theorem for completeness.

Theorem 2. Algorithm 1 when run on M agents with parameter S in the small deviation setting (Assumption 1), with $\beta_t = \mathcal{O}(H\sqrt{d\log(tMH)} + \xi\sqrt{dMT})$ and $\lambda = 1$ obtains the following cumulative regret after T episodes, with probability at least $1 - \alpha$,

$$\mathfrak{R}(T) = \widetilde{\mathcal{O}}\left(d^{3/2}H^2\left(M\sqrt{S} + \sqrt{MT}\right)\left(\sqrt{\log\left(\frac{1}{\alpha}\right)} + 2\xi\sqrt{dMT}\right)\right).$$

Proof. We have by the definition of group regret:

$$\mathfrak{R}(T) \tag{93}$$
$$= \sum_{m=1}^{M}\sum_{t=1}^{T} V^{\star}_{m,1}(x^t_{m,1}) - V^{\pi_t}_{m,1}(x^t_{m,1}) \leq \sum_{m=1}^{M}\sum_{t=1}^{T}\delta^t_{m,1} \tag{94}$$
$$\leq \sum_{m=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H}\xi^t_{m,h} + 4c_\beta \cdot dH \cdot \left(\sqrt{\log\left(\frac{dMTH}{\alpha}\right)} + \xi\sqrt{dMT}\right)\left(\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{h=1}^{H}\|\boldsymbol{\phi}(x,a)\|_{(\boldsymbol{\Lambda}^t_{m,h})^{-1}}\right). \tag{95}$$

Where the last inequality holds with probability at least $1 - \alpha/2$, via Lemma 13 and Lemma 6. Next, we can bound the first term via Lemma 8. We have with probability at least $1 - \alpha$, for some absolute constant c_β,

$$\mathfrak{R}(T) \leq \sqrt{2H^3MT\log\left(\frac{2}{\alpha}\right)}$$
$$+ 4c_\beta \cdot dH \cdot \left(\sqrt{\log\left(\frac{dMTH}{\alpha}\right)} + \xi\sqrt{dMT}\right)\left(\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{h=1}^{H}\|\boldsymbol{\phi}(x,a)\|_{(\boldsymbol{\Lambda}^t_{m,h})^{-1}}\right).$$

Finally, to bound the summation, we use Lemma 9. We have that,

$$\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{h=1}^{H}\|\boldsymbol{\phi}(x,a)\|_{(\boldsymbol{\Lambda}^t_{m,h})^{-1}} \leq 2\sum_{h=1}^{H}\left(\log\left(\frac{\det\left(\boldsymbol{\Lambda}^T_h\right)}{\det\left(\lambda\mathbf{I}_d\right)}\right)\left(\frac{M}{\log 2}\right)\sqrt{S} + 2\sqrt{2MT\log\left(\frac{\det\left(\boldsymbol{\Lambda}^T_h\right)}{\det\left(\lambda\mathbf{I}_d\right)}\right)}\right) \tag{96}$$
$$\leq 2H\log(dMT)M\sqrt{S} + 2\sqrt{2dMT\log(MT)}. \tag{97}$$

Where the last inequality is an application of the determinant-trace inequality and using the fact that $\|\boldsymbol{\phi}(\cdot)\|_2 \leq 1$. Replacing this result, we have that with probability at least $1-\alpha$,

$$\begin{aligned}\mathfrak{R}(T) &\leq \sqrt{2H^3 MT \log\left(\frac{2}{\alpha}\right)} \\ &\quad + 4c_\beta \cdot dH^2 \cdot \left(\sqrt{\log\left(\frac{dMTH}{\alpha}\right)} + \xi\sqrt{dMT}\right)\left(2\log(dMT)M\sqrt{S} + 2\sqrt{2dMT\log(MT)}\right) \\ &\implies \mathfrak{R}(T) = \widetilde{\mathcal{O}}\left(d^{3/2}H^2\left(M\sqrt{S}+\sqrt{MT}\right)\left(\sqrt{\log\left(\frac{1}{\alpha}\right)} + 2\xi\sqrt{dMT}\right)\right).\end{aligned}$$

□

B.4 Proof for Theorem 3 (Large Heterogeneity)

The proof for this section is largely similar to that of Theorem 1, however since we use the modified feature, the analysis differs in several key places. First we introduce the basic result which relates the variance with the coefficient of heterogeneity.

Lemma 14 (Variance Decomposition)**.** *Under the heterogeneous parallel MDP assumption (Definition 2) and coefficient of heterogeneity defined in Definition 3, we have that,*

$$\max_{h\in[H]} \log\det\left(\widetilde{\mathbf{\Lambda}}_h^T\right) \leq (d+\lambda+\chi)\log(MT).$$

Proof. We know, from the form of $\widetilde{\mathbf{\Lambda}}_h^T$ that,

$$\log\det\left(\widetilde{\mathbf{\Lambda}}_h^T\right) = \log\det\left((\widetilde{\mathbf{\Phi}}_h^T)^\top(\widetilde{\mathbf{\Phi}}_h^T\right) + \lambda\mathbf{I}_{d+k}) = \log\det\left((\widetilde{\mathbf{\Phi}}_h^T)(\widetilde{\mathbf{\Phi}}_h^T)^\top + \lambda\mathbf{I}_{MT}\right). \tag{98}$$

Here, $\widetilde{\mathbf{\Phi}}_h^T \in \mathbb{R}^{MT\times(d+k)}$ is the matrix of all features $\widetilde{\phi}(x,a,m)$ for step h until episode T. Now, observe that the matrix $(\widetilde{\mathbf{\Phi}}_h^T)(\widetilde{\mathbf{\Phi}}_h^T)^\top$ can be rewritten as the sum of two matrices $(\widetilde{\mathbf{\Phi}}_h^T)(\widetilde{\mathbf{\Phi}}_h^T)^\top = (\mathbf{\Phi}_h^T)(\mathbf{\Phi}_h^T)^\top + \widetilde{\mathbf{K}}_h^T$, where $[\widetilde{\mathbf{K}}_h^T]_{i,j} = \boldsymbol{\nu}(m_i)^\top\boldsymbol{\nu}(m_j), \widetilde{\mathbf{K}}_h^T \in \mathbb{R}^{MT\times MT}$, i.e., the corresponding dot-product contribution from the agent-specific features between any pair of transitions, and $(\mathbf{\Phi}_h^T)(\mathbf{\Phi}_h^T)^\top$ refers to the regular (agent-agnostic) features, i.e., $[(\mathbf{\Phi}_h^T)(\mathbf{\Phi}_h^T)^\top]_{i,j} = \boldsymbol{\phi}_i^\top\boldsymbol{\phi}_j$. Now, from Theorem IV of Madiman (2008), we have that,

$$\begin{aligned}
\log\det\left((\widetilde{\mathbf{\Phi}}_h^T)(\widetilde{\mathbf{\Phi}}_h^T)^\top + \lambda\mathbf{I}_{MT}\right) &\leq \log\det\left((\mathbf{\Phi}_h^T)(\mathbf{\Phi}_h^T)^\top + \lambda\mathbf{I}_{MT}\right) + \log\det\left(\widetilde{\mathbf{K}}_h^T + \lambda\mathbf{I}_{MT}\right) && (99)\\
&= \log\det\left((\mathbf{\Phi}_h^T)^\top(\mathbf{\Phi}_h^T) + \lambda\mathbf{I}_d\right) + \log\det\left(\widetilde{\mathbf{K}}_h^T + \lambda\mathbf{I}_{MT}\right) && (100)\\
&\leq d\log(MT) + \log\det\left(\widetilde{\mathbf{K}}_h^T + \lambda\mathbf{I}_{MT}\right) && (101)\\
&\leq d\log(MT) + \lambda\log(MT) + \operatorname{rank}(\widetilde{\mathbf{K}}_h^T)\cdot\log(MT) && (102)\\
&= (d+\lambda)\log(MT) + \operatorname{rank}(\mathbf{K}_h^\kappa)\log(MT). && (103)
\end{aligned}$$

The second inequality follows from $\|\phi(\cdot)\| \leq 1$ and then applying an AM-GM inequality followed by the determinant-trace inequality (as is common in bandit analyses). The final equality follows by the fact that since $\widetilde{\mathbf{K}}_h^T$ is $T\times T$ tiles of $\mathbf{K}_h^\kappa$ followed by permutations, which implies that $\operatorname{rank}(\widetilde{\mathbf{K}}_h^T) = \operatorname{rank}(\mathbf{K}_h^\kappa)$. Taking the maximum over all $h\in[H]$ and gives us the result. □

We first present a variant of the previous concentration result to bound the least-squares value iteration error (analog of Lemma 4).

Lemma 15. *Under the setting of Theorem 3, let c'_β be the constant defining β, and $\widetilde{\mathbf{S}}_{m,h}^t$ and $\widetilde{\mathbf{\Lambda}}_t^k$ be defined as follows.*

$$\begin{aligned}
\widetilde{\mathbf{S}}_{m,h}^t = &\sum_{n=1}^{M}\sum_{\tau=1}^{k_t}\widetilde{\phi}(n,x_{n,h}^\tau,a_{n,h}^\tau)\left[V_{m,h+1}^t(n,x_{n,h+1}^\tau) - (\mathbb{P}_{m,h}V_{m,h+1}^t)(n,x_{n,h}^\tau,a_{n,h}^\tau)\right]\\
&+\sum_{\tau=k_t+1}^{t-1}\widetilde{\phi}(n,x_{m,h}^\tau,a_{m,h}^\tau)\left[V_{m,h+1}^t(m,x_{m,h+1}^\tau) - (\mathbb{P}_{m,h}V_{m,h+1}^t)(m,x_{m,h}^\tau,a_{m,h}^\tau)\right],
\end{aligned}$$

$$\widetilde{\mathbf{\Lambda}}_{m,h}^t = \sum_{n=1}^{M}\sum_{\tau=1}^{k_t}\widetilde{\phi}(n,x_{n,h}^\tau,a_{n,h}^\tau)\widetilde{\phi}(n,x_{n,h}^\tau,a_{n,h}^\tau)^\top + \sum_{\tau=k_t+1}^{t-1}\widetilde{\phi}(n,x_{m,h}^\tau,a_{m,h}^\tau)\widetilde{\phi}(n,x_{m,h}^\tau,a_{m,h}^\tau)^\top + \lambda\mathbf{I}_{d+k}.$$

Where $V \in \mathcal{V}$ and $\mathcal{N}_\epsilon$ denotes the $\epsilon-$covering of the value function space $\mathcal{V}$. Then, there exists an absolute constant C independent of M, T, H, d and c'_β, such that, with probability at least $1-\delta'/2$ for all $m \in \mathcal{M}, t \in [T], h \in [H]$ simultaneously,

$$\left\|\widetilde{\mathbf{S}}^t_{m,h}\right\|_{(\widetilde{\mathbf{\Lambda}}^t_{m,h})^{-1}} \leq C \cdot (d+k)H\sqrt{2\log\left(\frac{(c'_\beta+2)(d+k)MTH}{\delta'}\right)}.$$

Proof. The proof is identical to that of Lemma 4, except that we utilize the combined features of dimensionality $(d+k)$, which requires us to select an alternative constant c'_β in the bound. □

Lemma 16. *There exists an absolute constant c'_β such that for $\beta^t_{m,h} = c'_\beta \cdot dH\sqrt{\log(2(d+k)MHT/\delta')}$ for any policy π, there exists a constant c_β such that for each $x \in \mathcal{S}, a \in \mathcal{A}$ we have for all $m \in \mathcal{M}, t \in [T], h \in [H]$ simultaneously, with probability at least $1-\delta'/2$,*

$$\left|\langle\widetilde{\boldsymbol{\phi}}(n,x,a), \mathbf{w}^t_{m,h} - \mathbf{w}^\pi_{m,h}\rangle\right| \leq \mathbb{P}_{m,h}(V^t_{m,h+1} - V^\pi_{m,h+1})(n,x,a) + c_\beta \cdot (d+k)H \cdot \|\widetilde{\boldsymbol{\phi}}(n,z)\|_{(\widetilde{\mathbf{\Lambda}}^t_{m,h})^{-1}} \cdot \sqrt{2\log\left(\frac{(d+k)MTH}{\delta'}\right)}.$$

Proof. The proof for this is identical to Lemma 5, however we modify the application of Lemma 4 with Lemma 15 instead. □

Lemma 17 (UCB in the Heterogeneous Setting)**.** *With probability at least $1-\delta'/2$, we have that for all $(x,a,h,t,m) \in \mathcal{S} \times \mathcal{A} \times [H] \times [T] \times \mathcal{M}$, $Q^t_{m,h}(x,a) \geq Q^\star_{m,h}(x,a)$.*

Proof. The proof is done by induction, identical to the proof in Lemma B.5 of Jin et al. (2020), and we urge the reader to refer to the aforementioned source. □

Lemma 18 (Recursive Relation in Heterogeneous Settings)**.** *Let $\delta^t_{m,h} = V^t_{m,h}(x^t_{m,h}) - V^{\pi_t}_{m,h}(x^t_{m,h})$, and $\xi^t_{m,h+1} = \mathbb{E}\left[\delta^t_{m,h} | x^t_{m,h}, a^t_{m,h}\right] - \delta^t_{m,h}$. Then, with probability at least $1-\alpha$, for all $(t,m,h) \in [T] \times \mathcal{M} \times [H]$ simultaneously,*

$$\delta^t_{m,h} \leq \delta^t_{m,h+1} + \xi^t_{m,h+1} + 2\left\|\widetilde{\boldsymbol{\phi}}(m, x^t_{m,h}, a^t_{m,h})\right\|_{(\widetilde{\mathbf{\Lambda}}^t_{m,h})^{-1}} \cdot c'_\beta \cdot (d+k)H \cdot \sqrt{2\log\left(\frac{(d+k)MTH}{\alpha}\right)}. \tag{104}$$

Proof. By Lemma 16, we have that for any $(x,a,h,m,t) \in \mathcal{S} \times \mathcal{A} \times [H] \times \mathcal{M} \times [T]$ with probability at least $1-\alpha/2$,

$$Q^t_{m,h}(x,a) - Q^{\pi_t}_{m,h}(x,a) \leq \mathbb{P}_{m,h}(V^t_{m,h+1} - V^{\pi_t}_{m,h})(x,a) + 2\left\|\widetilde{\boldsymbol{\phi}}(x,a)\right\|_{(\widetilde{\mathbf{\Lambda}}^t_{m,h})^{-1}} \cdot c'_\beta \cdot (d+k)H \cdot \sqrt{2\log\left(\frac{(d+k)MTH}{\alpha}\right)}. \tag{105}$$

Replacing the definition of $\delta^t_{m,h}$ and $V^{\pi_t}_{m,h}$ finishes the proof. □

We are now ready to prove Theorem 3. We first restate the Theorem for completeness.

Theorem 3. Algorithm 3 when run on M agents with parameter S in the heterogeneous setting (Definition 2), with $\beta_t = \mathcal{O}(H\sqrt{(d+k)\log(tMH)})$ and $\lambda = 1$ obtains the following cumulative regret after T episodes, with probability at least $1-\alpha$,

$$\mathfrak{R}(T) = \widetilde{\mathcal{O}}\left((d+k)H^2\left(M(d+\chi)\sqrt{S} + \sqrt{(d+\chi)MT}\right)\sqrt{\log\left(\frac{1}{\alpha}\right)}\right).$$

Proof. We have by the definition of group regret:

$$\mathfrak{R}(T) \tag{106}$$

$$= \sum_{m=1}^{M}\sum_{t=1}^{T} V^{\star}_{m,1}(x^t_{m,1}) - V^{\pi_t}_{m,1}(x^t_{m,1}) \leq \sum_{m=1}^{M}\sum_{t=1}^{T}\delta^t_{m,1} \tag{107}$$

$$\leq \sum_{m=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H}\xi^t_{m,h} + 2c'_\beta\cdot(d+k)H\cdot\sqrt{2\log\left(\frac{(d+k)MTH}{\alpha}\right)}\left(\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{h=1}^{H}\left\|\widetilde{\boldsymbol{\phi}}(m,x,a)\right\|_{(\widetilde{\boldsymbol{\Lambda}}^t_{m,h})^{-1}}\right). \tag{108}$$

Where the last inequality holds with probability at least $1-\alpha/2$, via Lemma 18 and Lemma 17. Next, we can bound the first term via Lemma 8. We have with probability at least $1-\alpha$, for some absolute constant c'_β,

$$\mathfrak{R}(T) \leq \sqrt{2H^3MT\log\left(\frac{2}{\alpha}\right)} + 2c'_\beta\cdot(d+k)H\cdot\sqrt{2\log\left(\frac{(d+k)MTH}{\alpha}\right)}\left(\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{h=1}^{H}\left\|\widetilde{\boldsymbol{\phi}}(m,x,a)\right\|_{(\widetilde{\boldsymbol{\Lambda}}^t_{m,h})^{-1}}\right). \tag{109}$$

Finally, to bound the summation, we use Lemma 9. We have that,

$$\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{h=1}^{H}\left\|\widetilde{\boldsymbol{\phi}}(x,a)\right\|_{(\widetilde{\boldsymbol{\Lambda}}^t_{m,h})^{-1}} \tag{110}$$

$$\leq 2\sum_{h=1}^{H}\left(\log\left(\frac{\det\left(\widetilde{\boldsymbol{\Lambda}}^T_h\right)}{\det\left(\lambda\mathbf{I}_d\right)}\right)\left(\frac{M}{\log 2}\right)\sqrt{S} + 2\sqrt{2MT\log\left(\frac{\det\left(\widetilde{\boldsymbol{\Lambda}}^T_h\right)}{\det\left(\lambda\mathbf{I}_d\right)}\right)}\right) \tag{111}$$

$$\leq 2H(d+\chi)\log(MT)M\sqrt{S} + 2H\sqrt{2(d+\chi)MT\log(MT)}. \tag{112}$$

Where the last inequality is an application of the variance decomposition (Lemma 14) and using the fact that $\|\boldsymbol{\phi}(\cdot)\|_2 \leq 1$. Replacing this result, we have that with probability at least $1-\alpha$,

$$\mathfrak{R}(T) \leq \sqrt{2H^3MT\log\left(\frac{2}{\alpha}\right)} + 2c'_\beta\cdot(d+k)H^2\cdot\sqrt{2\log\left(\frac{dMTH}{\alpha}\right)}\left(2\log(MT)M(d+\chi)\sqrt{\cdot S} + 2\sqrt{2(d+\chi)MT\log(MT)}\right). \tag{113}$$

$$\implies \mathfrak{R}(T) = \widetilde{\mathcal{O}}\left((d+k)H^2\left(M(d+\chi)\sqrt{S} + \sqrt{(d+\chi)MT}\right)\sqrt{\log\left(\frac{1}{\alpha}\right)}\right). \tag{114}$$

□

C Multiagent MDP Proofs

C.1 Proof of Proposition 1

We restate the Proposition for clarity.

Proposition 1. For the scalarized value function given in Equation 11, the Bellman optimality conditions are given as, for all $h \in [H], \mathbf{x} \in \mathcal{S}, \mathbf{a} \in \mathcal{A}$ for any fixed $\boldsymbol{\upsilon} \in \boldsymbol{\Upsilon}$,

$$Q^{\star}_{\boldsymbol{\upsilon},h}(\mathbf{x},\mathbf{a}) = \mathfrak{s}_{\boldsymbol{\upsilon}}\mathbf{r}_h(\mathbf{x},\mathbf{a}) + \mathbb{P}_h V^{\star}_{\boldsymbol{\upsilon},h}(\mathbf{x},\mathbf{a}), V^{\star}_{\boldsymbol{\upsilon},h}(\mathbf{x}) = \max_{\mathbf{a}\in\mathcal{A}} Q^{\star}_{\boldsymbol{\upsilon},h}(\mathbf{x},\mathbf{a}), \text{ and } V^{\star}_{\boldsymbol{\upsilon},H+1}(\mathbf{x}) = 0.$$

Proof. We prove the above result by reducing the scalarized MMDP to an equivalent MDP. Observe that for any fixed $\boldsymbol{\upsilon} \in \boldsymbol{\Upsilon}$, the (vector-valued) rewards can be scalarized to a scalar reward. For any step $h \in [H]$, for any *fixed* $\boldsymbol{\upsilon} \in \boldsymbol{\Upsilon}$, consider the MDP with state space $\mathcal{S} = \mathcal{S}_1 \times ... \times \mathcal{S}_M$, action space $\mathcal{A} = \mathcal{A}_1 \times ... \times \mathcal{A}_M$ and reward function r'_h such that for all $(\mathbf{x},\mathbf{a}) \in \mathcal{S}\times\mathcal{A}, r'_h(\mathbf{x},\mathbf{a}) = \boldsymbol{\upsilon}^{\top}\mathbf{r}_h(\mathbf{x},\mathbf{a})$. Therefore $r'_h(\mathbf{x},\mathbf{a}) \in [0,1]$ (since $\mathbf{r}_h$ lies on the $M-$dimensional simplex). Therefore, if the group of agents cooperate to optimize the scalarized reward (for any *fixed* scalarization parameter), the optimal (joint) policy coincides with the optimal policy for the aforementioned MDP defined over the *joint* state and action spaces. The optimal policy for the scalarized MDP is given by the greedy policy with respect to the following parameters:

$$Q^{\star}_{\boldsymbol{\upsilon},h}(\mathbf{x},\mathbf{a}) = r'_h(\mathbf{x},\mathbf{a}) + \mathbb{P}_h V^{\star}_{\boldsymbol{\upsilon},h}(\mathbf{x},\mathbf{a}), V^{\star}_{\boldsymbol{\upsilon},h}(\mathbf{x}) = \max_{\mathbf{a}\in\mathcal{A}} Q^{\star}_{\boldsymbol{\upsilon},h}(\mathbf{x},\mathbf{a}), \text{ and } V^{\star}_{\boldsymbol{\upsilon},H+1}(\mathbf{x}) = 0. \tag{115}$$

Replacing the reward function with the vector-valued reward in terms of $\boldsymbol{\upsilon}$ provides us the result. □

C.2 Proof of Proposition 2

We first restate the Proposition for clarity.

Proposition 2. For any parameter $\boldsymbol{\upsilon} \in \boldsymbol{\Upsilon}$, the optimal greedy policy $\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}$ with respect to the scalarized Q-value that satisfies Proposition 1 lies in the Pareto frontier $\boldsymbol{\Pi}^{\star}$.

Proof. The proof proceeds by contradiction. Assume that $\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}$ does not lie in the Pareto frontier, then there exists a policy $\boldsymbol{\pi}' \in \boldsymbol{\Pi}$ such that $\mathbf{V}_1^{\boldsymbol{\pi}'}(\mathbf{x}) \succeq \mathbf{V}_1^{\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{S}$ and $\boldsymbol{\pi} \neq \boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}$. Consider the final step H. Then, for any state $\mathbf{x} \in \mathcal{S}$, we have that if $\mathbf{V}_H^{\boldsymbol{\pi}'}(\mathbf{x}) \succeq \mathbf{V}_H^{\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}}(\mathbf{x})$, then,

$$\mathbf{r}_H(\mathbf{x},\boldsymbol{\pi}'(\mathbf{x})) \succeq \mathbf{r}_H(\mathbf{x},\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}(\mathbf{x})) \implies \mathfrak{s}_{\boldsymbol{\upsilon}}\mathbf{r}_H(\mathbf{x},\boldsymbol{\pi}'(\mathbf{x})) \geq \mathfrak{s}_{\boldsymbol{\upsilon}}\mathbf{r}_H(\mathbf{x},\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}(\mathbf{x})). \tag{116}$$

However, this is only true with equality if $\boldsymbol{\pi}'(\mathbf{x}) = \boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{S}$, as for any $\mathbf{x} \in \mathcal{S}$, $\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon},H}(\mathbf{x}) = \arg\max[\mathfrak{s}_{\boldsymbol{\upsilon}}\mathbf{r}_H(\mathbf{x},\mathbf{a})] \geq \mathfrak{s}_{\boldsymbol{\upsilon}}\mathbf{r}_H(\mathbf{x},\mathbf{a}')$ for any other $\mathbf{a}' \in \mathcal{A}$. Therefore, we have that $\boldsymbol{\pi}'_H(\mathbf{x}) = \boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon},H}(\mathbf{x})$ for each $\mathbf{x} \in \mathcal{S}$, and that $\mathbf{V}_H^{\boldsymbol{\pi}'}(\mathbf{x}) = \mathbf{V}_H^{\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}}(\mathbf{x})$. This implies that $\mathbb{P}_H\mathbf{V}_H^{\boldsymbol{\pi}'}(\mathbf{x},\mathbf{a}) = \mathbb{P}_H\mathbf{V}_H^{\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}}(\mathbf{x},\mathbf{a})$ for all $\mathbf{x} \in \mathcal{S}$ and $\mathbf{a} \in \mathcal{A}$. Now, if $\mathbf{V}_{H-1}^{\boldsymbol{\pi}'}(\mathbf{x}) \succeq \mathbf{V}_{H-1}^{\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}}(\mathbf{x})$, then we have that,

$$\mathbf{r}_{H-1}(\mathbf{x},\boldsymbol{\pi}'_{H-1}(\mathbf{x})) + \mathbb{E}_{\mathbf{x}'\sim\mathbb{P}_H(\cdot|\mathbf{x},\boldsymbol{\pi}'_{H-1}(\mathbf{x}))}\left[\mathbf{V}_H^{\boldsymbol{\pi}'}(\mathbf{x}')\right] \succeq \mathbf{r}_{H-1}(\mathbf{x},\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}(\mathbf{x})) + \mathbb{P}_H\mathbf{V}_H^{\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}}(\mathbf{x},\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}(\mathbf{x})) \tag{117}$$

$$\implies \mathbf{r}_{H-1}(\mathbf{x},\boldsymbol{\pi}'_{H-1}(\mathbf{x})) + \mathbb{E}_{\mathbf{x}'\sim\mathbb{P}_H(\cdot|\mathbf{x},\boldsymbol{\pi}'_{H-1}(\mathbf{x}))}\left[\mathbf{V}_H^{\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}}(\mathbf{x}')\right] \succeq \mathbf{r}_{H-1}(\mathbf{x},\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}(\mathbf{x})) + \mathbb{P}_H\mathbf{V}_H^{\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}}(\mathbf{x},\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}(\mathbf{x})) \tag{118}$$

$$\implies \mathfrak{s}_{\boldsymbol{\upsilon}}\left(\mathbf{r}_{H-1}(\mathbf{x},\boldsymbol{\pi}'_{H-1}(\mathbf{x})) + \mathbb{E}_{\mathbf{x}'\sim\mathbb{P}_H(\cdot|\mathbf{x},\boldsymbol{\pi}'_{H-1}(\mathbf{x}))}\left[\mathbf{V}_H^{\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}}(\mathbf{x}')\right]\right) \geq \mathfrak{s}_{\boldsymbol{\upsilon}}\left(\mathbf{r}_{H-1}(\mathbf{x},\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}(\mathbf{x})) + \mathbb{P}_H\mathbf{V}_H^{\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}}(\mathbf{x},\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}(\mathbf{x}))\right) \tag{119}$$

$$\implies \mathfrak{s}_{\boldsymbol{\upsilon}}\mathbf{r}_{H-1}(\mathbf{x},\boldsymbol{\pi}'_{H-1}(\mathbf{x})) + \mathbb{E}_{\mathbf{x}'\sim\mathbb{P}_H(\cdot|\mathbf{x},\boldsymbol{\pi}'_{H-1}(\mathbf{x}))}\left[\mathbf{V}_H^{\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}}(\mathbf{x}')\right] \geq \mathfrak{s}_{\boldsymbol{\upsilon}}\mathbf{r}_{H-1}(\mathbf{x},\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}(\mathbf{x})) + \mathbb{P}_H\mathbf{V}_H^{\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}}(\mathbf{x},\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}(\mathbf{x})). \tag{120}$$

This is true only if $\boldsymbol{\pi}'_{H-1}(\mathbf{x}) = \boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon},H}(\mathbf{x})$ for each $\mathbf{x} \in \mathcal{S}$, as $\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon},H}$ is the greedy policy with respect to $\mathfrak{s}_{\boldsymbol{\upsilon}}\mathbf{r}_{H-1}(\mathbf{x},\mathbf{a}) + \mathbb{P}_H\mathbf{V}_H^{\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon},H}}(\mathbf{x},\mathbf{a})$. Continuing this argument inductively for $h = H-2, H-3, ..., 1$ we obtain that $\mathbf{V}_1^{\boldsymbol{\pi}'}(\mathbf{x}) \succeq \mathbf{V}_1^{\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}}(\mathbf{x})$ for each $\mathbf{x} \in \mathcal{S}$ only if $\boldsymbol{\pi}' = \boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}$. This is a contradiction as we assumed that $\boldsymbol{\pi}' \neq \boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}$, and hence $\boldsymbol{\pi}^{\star}_{\boldsymbol{\upsilon}}$ lies in $\boldsymbol{\Pi}^{\star}$. □

C.3 Proof of Proposition 3

We first restate Proposition 3 for clarity.

Proposition 3. For any scalarization $\mathfrak{s}$ that is Lipschitz and bounded $\boldsymbol{\Upsilon}$, we have that $\mathfrak{R}_B(T) \leq \frac{1}{T}\mathfrak{R}_C(T) + o(1)$.

Proof. We will follow the approach in Paria et al. (2020) (Appendix B.3). Recall that $\boldsymbol{\Upsilon}$ is a bounded subset of $\mathbb{R}^M$. Now, we have that since $\mathfrak{s}_{\boldsymbol{\upsilon}}(\cdot) = \boldsymbol{\upsilon}^\top(\cdot)$, we have that $\mathfrak{s}_{\boldsymbol{\upsilon}}$ is Lipschitz with constant M with respect to the ℓ_1-norm, i.e., for any $\mathbf{y} \in \mathbb{R}^M$,

$$|\mathfrak{s}_{\boldsymbol{\upsilon}}(\mathbf{y}) - \mathfrak{s}_{\boldsymbol{\upsilon}'}(\mathbf{y})| \leq M\|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}'\|_1. \tag{121}$$

Now, consider the Wasserstein distance conditioned on the history $\mathcal{H}$ between the sampling distribution $p_{\boldsymbol{\Upsilon}}$ on $\boldsymbol{\Upsilon}$ and the empirical distribution $\widehat{p}_{\boldsymbol{\Upsilon}}$ corresponding to $\{\boldsymbol{\upsilon}_t\}_{t=1}^T$,

$$W_1(p_{\boldsymbol{\Upsilon}}, \widehat{p}_{\boldsymbol{\Upsilon}}) = \inf_q \left\{\mathbb{E}_q\|X - Y\|_1, q(X) = p_{\boldsymbol{\Upsilon}}, q(Y) = \widehat{p}_{\boldsymbol{\Upsilon}}\right\}, \tag{122}$$

where q is a joint distribution on the RVs X, Y with marginal distributions equal to $p_{\boldsymbol{\Upsilon}}$ and $\widehat{p}_{\boldsymbol{\Upsilon}}$. We therefore have for some randomly drawn samples $\boldsymbol{\upsilon}_1, \boldsymbol{\upsilon}_2, ..., \boldsymbol{\upsilon}_T$ and for any arbitrary sequence of (joint) policies $\widehat{\boldsymbol{\Pi}}_T = \{\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_T\}$, for any state $\mathbf{x} \in \mathcal{S}$,

$$\frac{1}{T}\sum_{t=1}^T \max_{\mathbf{x}\in\mathcal{S}} \left[V^{\boldsymbol{\pi}_t}_{\boldsymbol{\upsilon}_t,1}(\mathbf{x}) - \mathbb{E}_{\boldsymbol{\upsilon}\in\boldsymbol{\Upsilon}}\left[\max_{\boldsymbol{\pi}\in\widehat{\boldsymbol{\Pi}}_T} V^{\boldsymbol{\pi}}_{\boldsymbol{\upsilon},1}(\mathbf{x})\right]\right] \leq \frac{1}{T}\sum_{t=1}^T \max_{\mathbf{x}\in\mathcal{S}} \left[V^{\boldsymbol{\pi}_t}_{\boldsymbol{\upsilon}_t,1}(\mathbf{x}) - \mathbb{E}_{\boldsymbol{\upsilon}\in\boldsymbol{\Upsilon}}\left[\max_{\boldsymbol{\pi}\in\widehat{\boldsymbol{\Pi}}_T} V^{\boldsymbol{\pi}}_{\boldsymbol{\upsilon},1}(\mathbf{x})\right]\right] \tag{123}$$

$$\leq \mathbb{E}_{q(X,Y)}\left[\max_{\mathbf{x}\in\mathcal{S}}\left[\max_{\boldsymbol{\pi}\in\widehat{\boldsymbol{\Pi}}_T} V^{\boldsymbol{\pi}}_{X,1}(\mathbf{x}) - \max_{\boldsymbol{\pi}\in\widehat{\boldsymbol{\Pi}}_T} V^{\boldsymbol{\pi}}_{Y,1}(\mathbf{x})\right]\right] \tag{124}$$

$$\leq \mathbb{E}_{q(X,Y)}\left[M\|X - Y\|_1\right]. \tag{125}$$

Taking an expectation with respect to $\mathcal{H} = \{\boldsymbol{\upsilon}_1, .., \boldsymbol{\upsilon}_T\}$, we have,

$$\mathfrak{R}_B(T) - \frac{1}{T}\mathfrak{R}_C(T) \tag{126}$$

$$= \mathbb{E}_{\boldsymbol{\upsilon}\in\boldsymbol{\Upsilon}}\left[\max_{\mathbf{x}\in\mathcal{S}}\left[V^\star_{\boldsymbol{\upsilon},1}(\mathbf{x}) - \max_{\boldsymbol{\pi}\in\widehat{\boldsymbol{\Pi}}_T} V^{\boldsymbol{\pi}}_{\boldsymbol{\upsilon},1}(\mathbf{x})\right]\right] - \mathbb{E}_{\mathcal{H}}\left[\frac{1}{T}\sum_{t=1}^T \max_{\mathbf{x}\in\mathcal{S}}\left[V^\star_{\boldsymbol{\upsilon}_t,1}(\mathbf{x}) - V^{\boldsymbol{\pi}_t}_{\boldsymbol{\upsilon}_t,1}(\mathbf{x})\right]\right] \tag{127}$$

$$= \mathbb{E}_{\mathcal{H}}\left[\frac{1}{T}\sum_{t=1}^T \max_{\mathbf{x}\in\mathcal{S}}\left[V^\star_{\boldsymbol{\upsilon}_t,1}(\mathbf{x}) - \max_{\boldsymbol{\pi}\in\widehat{\boldsymbol{\Pi}}_T} V^{\boldsymbol{\pi}}_{\boldsymbol{\upsilon}_t,1}(\mathbf{x})\right]\right] - \mathbb{E}_{\mathcal{H}}\left[\frac{1}{T}\sum_{t=1}^T \max_{\mathbf{x}\in\mathcal{S}}\left[V^\star_{\boldsymbol{\upsilon}_t,1}(\mathbf{x}) - V^{\boldsymbol{\pi}_t}_{\boldsymbol{\upsilon}_t,1}(\mathbf{x})\right]\right] \tag{128}$$

$$\leq \mathbb{E}_{\mathcal{H}}\left[\frac{1}{T}\sum_{t=1}^T \max_{\mathbf{x}\in\mathcal{S}}\left[V^{\boldsymbol{\pi}_t}_{\boldsymbol{\upsilon}_t,1}(\mathbf{x}) - \mathbb{E}_{\boldsymbol{\upsilon}\in\boldsymbol{\Upsilon}}\left[\max_{\boldsymbol{\pi}\in\widehat{\boldsymbol{\Pi}}_T} V^{\boldsymbol{\pi}}_{\boldsymbol{\upsilon},1}(\mathbf{x})\right]\right]\right] \tag{129}$$

$$\leq M\mathbb{E}_{q(X,Y)}\left[\|X - Y\|_1\right]. \tag{130}$$

The penultimate inequality follows from max being a contraction mapping in bounded domains, and the final inequality follows from the previous analysis. To complete the proof, we first take an infimum over q and observe that the subsequent RHS converges at a rate of $\widetilde{\mathcal{O}}(T^{-1/M})$ under mild regulatory conditions, as shown by Canas & Rosasco (2012). □

C.4 Proof of Theorem 4

The proof in the multiagent MDP setting is similar to that of the parallel MDP setting. There are several differences: first, in each episode, since we sample a scalarization parameter $\boldsymbol{\upsilon}_t$ from $\boldsymbol{\Upsilon}_T$,

we would like to derive concentration results *independent* of the scalarization parameter. We do this by utilizing *vector-valued* concentration results and utilizing the monotonicity of the scalarized $Q-$value. We first present a vector-valued concentration result.

Lemma 19. *For any $m \in [M], h \in [H]$ and $t \in [T]$, let k_t denote the episode after which the last global synchronization has taken place, and $\mathbf{S}_t^h$ and $\mathbf{\Lambda}_t^h$ be defined as follows.*

$$\mathbf{S}_{\boldsymbol{\upsilon},t}^h = \sum_{\tau=1}^{k_t} \mathbf{\Phi}(\mathbf{x}_h^\tau, \mathbf{a}_h^\tau) \left[\mathbf{v}_{\boldsymbol{\upsilon},h+1}^t(\mathbf{x}_{h+1}^\tau) - (\mathbb{P}_h \mathbf{v}_{\boldsymbol{\upsilon},h+1}^t)(\mathbf{x}_h^\tau, \mathbf{a}_h^\tau)\right], \ \mathbf{\Lambda}_t^h = \lambda \mathbf{I}_d + (\mathbf{\Phi}_h^{k_t})^\top (\mathbf{\Phi}_h^{k_t}).$$

Where $\mathbf{v}_{\boldsymbol{\upsilon},h+1}^t(\mathbf{x}) = \mathbf{1}_M \cdot V_{\boldsymbol{\upsilon},h+1}^t(\mathbf{x}) \ \forall \ \mathbf{x} \in \mathcal{S}$, $\mathbf{1}_M$ denotes the all-ones vector in $\mathbb{R}^M$, and C_β is the constant such that $\beta_h^t = C_\beta \cdot dH\sqrt{\log(TMH)}$. Then, there exists a constant C such that with probability at least $1-\delta$,

$$\sup_{\boldsymbol{\upsilon} \in \mathbf{\Upsilon}} \left\|\mathbf{S}_{\boldsymbol{\upsilon},h}^t\right\|_{(\mathbf{\Lambda}_h^t)^{-1}} \leq C \cdot dH \sqrt{2\log\left(\frac{(C_\beta+2)dMTH}{\delta'}\right)}.$$

Proof. The proof is done in two steps. The first step is to bound the deviations in $\mathbf{S}$ for any fixed function V by a martingale concentration. The second step is to bound the resulting concentration over all functions V by a covering argument. Finally, we select appropriate constants to provide the form of the result required.

Step 1. Recall that $\mathbf{S}_{\boldsymbol{\upsilon},h}^t = \sum_{\tau=1}^{k_t} \mathbf{\Phi}(\mathbf{z}_h^\tau)[V_{\boldsymbol{\upsilon},h+1}^t(\mathbf{x}_{h+1}^\tau) - (\mathbb{P}_h V_{\boldsymbol{\upsilon},h+1}^t)(\mathbf{z}_h^\tau)]$, where $\mathbf{v}_{\boldsymbol{\upsilon},h+1}^t$ is the vector with each entry being $V_{\boldsymbol{\upsilon},h+1}^t$. We have that $V_{\boldsymbol{\upsilon},h+1}^t(\mathbf{x}_{h+1}^\tau) - (\mathbb{P}_h V_{\boldsymbol{\upsilon},h+1}^t)(\mathbf{z}_h^\tau) = \mathbf{v}_{\boldsymbol{\upsilon},h+1}^t - \mathbb{P}_h \mathbf{v}_{\boldsymbol{\upsilon},h+1}^t$. Consider the following distance metric $\text{dist}_{\mathbf{\Upsilon}}$,

$$\text{dist}_{\mathbf{\Upsilon}}(\mathbf{v}, \mathbf{v}') = \sup_{\mathbf{x} \in \mathcal{S}, \boldsymbol{\upsilon} \in \mathbf{\Upsilon}} \left\|\mathbf{v}(\mathbf{x}) - \mathbf{v}(\mathbf{x}')\right\|_1. \tag{131}$$

Let $\mathcal{V}_{\mathbf{\Upsilon}}$ be the family of all vector-valued UCB value functions that can be produced by the algorithm, and now let $\mathcal{N}_\epsilon$ be an $\epsilon-$covering of $\mathcal{V}_{\mathbf{\Upsilon}}$ under $\text{dist}_{\mathbf{\Upsilon}}$, i.e., for every $\mathbf{v} \in \mathcal{V}_{\mathbf{\Upsilon}}$, there exists $\mathbf{v}' \in \mathcal{N}_\epsilon$ such that $\text{dist}_{\mathbf{\Upsilon}}(\mathbf{v}, \mathbf{v}') \leq \epsilon$. Now, here again, we adopt a similar strategy as the independent case. To bound the RHS, we decompose $\mathbf{S}_{\boldsymbol{\upsilon},h}^t$ in terms of the covering described earlier. We know that since $\mathcal{N}_\epsilon$ is an $\epsilon-$covering of $\mathcal{V}_{\mathbf{\Upsilon}}$, there exists a $\mathbf{v}' \in \mathcal{N}_\epsilon$ and $\mathbf{\Delta} = \mathbf{v}_{\boldsymbol{\upsilon},h+1}^t - \mathbf{v}'$ such that,

$$\mathbf{S}_{\boldsymbol{\upsilon},h}^t = \sum_{\tau=1}^{k_t} \mathbf{\Phi}(\mathbf{z}_h^\tau) \left[\mathbf{v}'(\mathbf{x}_{h+1}^\tau) - \mathbb{P}_h \mathbf{v}'(\mathbf{z}_h^\tau)\right] + \sum_{\tau=1}^{k_t} \mathbf{\Phi}(\mathbf{z}_h^\tau) \left[\mathbf{\Delta}(\mathbf{x}_{h+1}^\tau) - \mathbb{P}_h \mathbf{\Delta}(\mathbf{z}_h^\tau)\right]. \tag{132}$$

Now, observe that by the definition of the covering, we have that $\|\mathbf{\Delta}\|_1 \leq \epsilon$. Therefore, we have that $\|\mathbf{\Delta}(\mathbf{x})\|_{(\mathbf{\Lambda}_h^t)^{-1}} \leq \epsilon/\sqrt{\lambda}$, and $\|\mathbb{P}_h \mathbf{\Delta}(\mathbf{z})\|_{(\mathbf{\Lambda}_h^t)^{-1}} \leq \epsilon/\sqrt{\lambda}$ for all $\mathbf{z} \in \mathcal{Z}, \mathbf{x} \in \mathcal{S}, h \in [H]$. Therefore, since $\|\mathbf{\Phi}(\mathbf{z})\|_2 \leq \sqrt{M}$,

$$\left\|\mathbf{S}_{\boldsymbol{\upsilon},h}^t\right\|_{(\mathbf{\Lambda}_h^t)^{-1}}^2 \leq 2 \left\|\sum_{\tau=1}^{k_t} \mathbf{\Phi}(\mathbf{z}_h^\tau) \left[\mathbf{v}'(\mathbf{x}_{h+1}^\tau) - \mathbb{P}_h \mathbf{v}'(\mathbf{z}_h^\tau)\right]\right\|_{(\mathbf{\Lambda}_h^t)^{-1}} + \frac{8Mt^2\epsilon^2}{\lambda}. \tag{133}$$

Consider the substitution $\boldsymbol{\varepsilon}_{\tau,h}^t = \mathbf{v}'(\mathbf{x}_{h+1}^\tau) - \mathbb{P}_h \mathbf{v}'(\mathbf{z}_h^\tau)$. To bound the first term on the RHS, we consider the filtration $\{\mathcal{F}_\tau\}_{\tau=0}^\infty$ where $\mathcal{F}_0$ is empty, and $\mathcal{F}_\tau = \sigma\left(\left\{\bigcup\left(\mathbf{x}_{h+1}^i, \boldsymbol{\phi}(\mathbf{z}_h^i)\right)\right\}_{i \leq \tau}\right)$, and σ denotes the $\sigma-$algebra generated by a finite set. Then, we have that,

$$\begin{aligned}
&\left\|\sum_{\tau=1}^{k_t} \mathbf{\Phi}(\mathbf{z}_h^\tau) \left[\mathbf{v}'(\mathbf{x}_{h+1}^\tau) - \mathbb{P}_h \mathbf{v}'(\mathbf{z}_h^\tau)\right]\right\|_{(\mathbf{\Lambda}_h^t)^{-1}} \\
&= \left\|\sum_{\tau=1}^{k_t} \mathbf{\Phi}(\mathbf{z}_h^\tau) \left[\mathbf{v}'(\mathbf{x}_{h+1}^\tau) - \mathbb{E}\left[\mathbf{v}'(\mathbf{x}_{h+1}^\tau) | \mathcal{F}_{\tau-1}\right]\right]\right\|_{(\mathbf{\Lambda}_h^t)^{-1}} = \left\|\sum_{\tau=1}^{k_t} \mathbf{\Phi}(\mathbf{z}_h^\tau) \boldsymbol{\varepsilon}_{\tau,h}^t\right\|_{(\mathbf{\Lambda}_h^t)^{-1}}.
\end{aligned}$$

Note that for each $\boldsymbol{\varepsilon}^t_{\tau,h}$, each entry is bounded by $2H$, and therefore we have that the vector $\boldsymbol{\varepsilon}^t_{\tau,h}$ is $H-$sub-Gaussian. Then, applying Lemma 29, we have that,

$$\left\|\sum_{\tau=1}^{k_t}\boldsymbol{\Phi}(\mathbf{z}^\tau_h)\boldsymbol{\varepsilon}^t_{\tau,h}\right\|_{(\boldsymbol{\Lambda}^t_h)^{-1}} \leq H^2\log\left(\frac{\det\left(\boldsymbol{\Lambda}^t_h\right)}{\det\left(\lambda\mathbf{I}_d\right)\delta^2}\right) \leq H^2\log\left(\frac{\det\left(\bar{\boldsymbol{\Lambda}}^t_h\right)}{\det\left(\lambda\mathbf{I}_d\right)\delta^2}\right). \tag{134}$$

Replacing this result for each $\mathbf{v}\in\mathcal{N}_\epsilon$, we have by a union bound over each $t\in[T], h\in[H]$, we have with probability at least $1-\delta$, simultaneously for each $t\in[T], h\in[H]$,

$$\sup_{\boldsymbol{\upsilon}_t\in\boldsymbol{\Upsilon},\mathbf{v}\in\mathcal{V}_{\boldsymbol{\Upsilon}}}\left\|\mathbf{S}^t_{\boldsymbol{\upsilon},h}\right\|_{(\boldsymbol{\Lambda}^t_h)^{-1}} \leq 2H\sqrt{\log\left(\frac{\det\left(\bar{\boldsymbol{\Lambda}}^t_h\right)}{\det\left(\lambda\mathbf{I}_d\right)}\right)+\log\left(\frac{HT|\mathcal{N}_\epsilon|}{\delta}\right)+\frac{2Mt^2\epsilon^2}{H^2\lambda}} \tag{135}$$

$$\leq 2H\sqrt{d\log\left(\frac{Mt+\lambda}{\lambda}\right)+\log\left(\frac{|\mathcal{N}_\epsilon|}{\delta}\right)+\log(HT)+\frac{2Mt^2\epsilon^2}{H^2\lambda}}. \tag{136}$$

The last step follows once again by first noticing that $\|\boldsymbol{\Phi}(\cdot)\|\leq\sqrt{M}$ and then applying an AM-GM inequality, and then using the determinant-trace inequality.

Step 2. Here $\mathcal{N}_\epsilon$ is an $\epsilon-$covering of the function class $\mathcal{V}_{\boldsymbol{\Upsilon}}$ for any $h\in[H], m\in[M]$ or $t\in[T]$ under the distance function $\mathrm{dist}_{\boldsymbol{\Upsilon}}(\mathbf{v},\mathbf{v}')=\sup_{\mathbf{x}\in\mathcal{S},\boldsymbol{\upsilon}\in\boldsymbol{\Upsilon}}\|\mathbf{v}(\mathbf{x})-\mathbf{v}(\mathbf{x}')\|_1$. To bound this quantity by the appropriate covering number, we first observe that for any $V\in\mathcal{V}_{\boldsymbol{\Upsilon}}$, we have that the policy weights are bounded as $2HM\sqrt{dT/\lambda}$ (Lemma 27). Therefore, by Lemma 32 we have for any constant B such that $\beta^t_h\leq B$,

$$\log\left(\mathcal{N}_\varepsilon\right)\leq d\cdot\log\left(1+\frac{8HM^3}{\varepsilon}\sqrt{\frac{dT}{\lambda}}\right)+d^2\log\left(1+\frac{8Md^{1/2}B^2}{\lambda\varepsilon^2}\right). \tag{137}$$

Recall that we select the hyperparameters $\lambda=1$ and $\beta=\mathcal{O}(dH\sqrt{\log(TMH)}$, and to balance the terms in $\bar{\beta}^t_h$ we select $\epsilon=\epsilon^\star=dH/\sqrt{MT^2}$. Finally, we obtain that for some absolute constant C_β, by replacing the above values,

$$\log\left(\mathcal{N}_\varepsilon\right)\leq d\cdot\log\left(1+\frac{8M^{7/2}T^{3/2}}{d^{1/2}}\right)+d^2\log\left(1+8C_\beta d^{1/2}MT^2\log(TMH)\right). \tag{138}$$

Therefore, for some absolute constant C' independent of M,T,H,d and C_β, we have,

$$\log|\mathcal{N}_\varepsilon|\leq C'd^2\log\left(C_\beta\cdot dMT\log(TMH)\right). \tag{139}$$

Replacing this result in the result from Step 1, we have that with probability at least $1-\delta'/2$ for all $t\in[T], h\in[H]$ simultaneously,

$$\left\|\mathbf{S}^t_{\boldsymbol{\upsilon},h}\right\|_{(\boldsymbol{\Lambda}^t_h)^{-1}}$$
$$\leq 2H\sqrt{(d+2)\log\frac{MT+\lambda}{\lambda}+2\log\left(\frac{1}{\delta'}\right)+C'd^2\log\left(C_\beta\cdot dMT\log(TMH)\right)+2+4\log(TH)}.$$

This implies that there exists an absolute constant C independent of M,T,H,d and C_β, such that, with probability at least $1-\delta'/2$ for all $t\in[T], h\in[H], \boldsymbol{\upsilon}\in\boldsymbol{\Upsilon}$ simultaneously,

$$\left\|\mathbf{S}^t_{\boldsymbol{\upsilon},h}\right\|_{(\boldsymbol{\Lambda}^t_h)^{-1}}\leq C\cdot dH\sqrt{2\log\left(\frac{(C_\beta+2)dMTH}{\delta'}\right)}. \tag{140}$$

□

Next, we present the key result for cooperative value iteration, which demonstrates that for any agent the estimated $Q-$values have bounded error for any policy π. This result is an extension of Lemma B.4 of Jin et al. (2020) on to the multiagent MDP setting.

Lemma 20. *There exists an absolute constant c_β such that for $\beta^t_{m,h} = c_\beta \cdot dH\sqrt{\log(2dMHT/\delta')}$ for any policy π, there exists a constant C'_β such that for each $x \in \mathcal{S}, a \in \mathcal{A}$ we have for all $m \in \mathcal{M}, t \in [T], h \in [H]$ simultaneously, with probability at least $1 - \delta'/2$,*

$$\left|\langle \boldsymbol{\phi}(x,a), \mathbf{w}^t_{m,h} - \mathbf{w}^\pi_h \rangle\right| \leq \mathbb{P}_h(V^t_{m,h+1} - V^\pi_{m,h+1})(x,a) + C'_\beta \cdot dH \cdot \|\boldsymbol{\phi}(z)\|_{(\boldsymbol{\Lambda}^t_{m,h})^{-1}} \cdot \sqrt{2\log\left(\frac{dMTH}{\delta'}\right)}.$$

Proof. By the Bellman equation and the assumption of the linear MMDP (Definition 5), we have that for any policy $\boldsymbol{\pi}$, and $\boldsymbol{\upsilon} \in \boldsymbol{\Upsilon}$, there exist weights $\mathbf{w}^\pi_{\boldsymbol{\upsilon},h}$ such that, for all $\mathbf{z} \in \mathcal{Z} = \mathcal{S} \times \mathcal{A}$,

$$\boldsymbol{\upsilon}^\top \boldsymbol{\Phi}(\mathbf{z})^\top \mathbf{w}^\pi_{\boldsymbol{\upsilon},h} = \boldsymbol{\upsilon}^\top \mathbf{r}_h(\mathbf{z}) + \mathbb{P}_h V^\pi_{\boldsymbol{\upsilon},h+1}(\mathbf{z}) = \boldsymbol{\upsilon}^\top \left(\mathbf{r}_h(\mathbf{z}) + \mathbf{1}_M \cdot \mathbb{P}_h V^\pi_{\boldsymbol{\upsilon},h+1}(\mathbf{z})\right). \tag{141}$$

We have,

$$\mathbf{w}^t_{\boldsymbol{\upsilon},h} - \mathbf{w}^\pi_{\boldsymbol{\upsilon},h} = (\boldsymbol{\Lambda}^t_h)^{-1} \sum_{\tau=1}^{k_t} \left[\boldsymbol{\Phi}_\tau(\mathbf{x}_\tau, \mathbf{a}_\tau)[\mathbf{r}_h(\mathbf{x}_\tau, \mathbf{a}_\tau) + \mathbf{1}_M \cdot V^t_{\boldsymbol{\upsilon},h+1}(\mathbf{x}_\tau)]\right] - \mathbf{w}^\pi_{\boldsymbol{\upsilon},h} \tag{142}$$

$$= (\boldsymbol{\Lambda}^t_h)^{-1} \left\{ -\lambda \mathbf{w}^\pi_{\boldsymbol{\upsilon},h} + \sum_{\tau=1}^{k_t} \left[\boldsymbol{\Phi}_\tau(\mathbf{x}_\tau, \mathbf{a}_\tau)[\mathbf{1}_M \cdot (V^t_{\boldsymbol{\upsilon},h+1}(\mathbf{x}'_\tau) - \mathbb{P}_h V^\pi_{\boldsymbol{\upsilon},h+1}(\mathbf{x}_\tau, \mathbf{a}_\tau))]\right] \right\}. \tag{143}$$

$$\begin{aligned} \mathbf{w}^t_{\boldsymbol{\upsilon},h} - \mathbf{w}^\pi_{\boldsymbol{\upsilon},h} = {} & \underbrace{-\lambda(\boldsymbol{\Lambda}^t_h)^{-1}\mathbf{w}^\pi_{\boldsymbol{\upsilon},h}}_{\mathbf{v}_1} \\ & + \underbrace{(\boldsymbol{\Lambda}^t_h)^{-1} \left\{ \sum_{\tau=1}^{k_t} \left[\boldsymbol{\Phi}_\tau(\mathbf{x}_\tau, \mathbf{a}_\tau)[\mathbf{1}_M \cdot (V^t_{\boldsymbol{\upsilon},h+1}(\mathbf{x}'_\tau) - \mathbb{P}_h V^t_{\boldsymbol{\upsilon},h+1}(\mathbf{x}_\tau, \mathbf{a}_\tau))]\right] \right\}}_{\mathbf{v}_2} \\ & + \underbrace{(\boldsymbol{\Lambda}^t_{m,h})^{-1} \left\{ \sum_{\tau=1}^{k_t} \left[\boldsymbol{\Phi}_\tau(\mathbf{x}_\tau, \mathbf{a}_\tau)[\mathbf{1}_M \cdot (\mathbb{P}_h V^t_{\boldsymbol{\upsilon},h+1} - \mathbb{P}_h V^\pi_{\boldsymbol{\upsilon},h+1})(\mathbf{x}_\tau, \mathbf{a}_\tau)]\right] \right\}}_{\mathbf{v}_3} \end{aligned} \tag{144}$$

Now, we know that for any $\mathbf{z} \in \mathcal{Z}$ for any policy π,

$$\begin{aligned} & \|\langle \boldsymbol{\Phi}(\mathbf{z}), \mathbf{v}_1 \rangle\|_2 \\ & \leq \lambda \|\langle \boldsymbol{\Phi}(\mathbf{z}), (\boldsymbol{\Lambda}^t_h)^{-1}\mathbf{w}^\pi_{\boldsymbol{\upsilon},h} \rangle\|_2 \leq \lambda \cdot \|\mathbf{w}^\pi_{\boldsymbol{\upsilon},h}\| \|\boldsymbol{\Phi}(\mathbf{z})\|_{(\boldsymbol{\Lambda}^t_h)^{-1}} \leq 2HM\lambda\sqrt{d} \cdot \|\boldsymbol{\Phi}(\mathbf{z})\|_{(\boldsymbol{\Lambda}^t_h)^{-1}} \end{aligned}$$

Here the last inequality follows from Lemma 25. For the second term, we have by Lemma 19 that there exists an absolute constant C_β, independent of M, T, H, d such that, with probability at least $1 - \delta'/2$ for all $t \in [T], h \in [H], \boldsymbol{\upsilon} \in \boldsymbol{\Upsilon}$ simultaneously,

$$\|\langle \boldsymbol{\Phi}(\mathbf{z}), \mathbf{v}_2 \rangle\|_2 \leq \|\boldsymbol{\Phi}(\mathbf{z})\|_{(\boldsymbol{\Lambda}^t_h)^{-1}} \cdot C_\beta \cdot dH \cdot \sqrt{2\log\left(\frac{dMTH}{\delta'}\right)}. \tag{145}$$

For the last term, note that,

$$\langle \mathbf{\Phi}(\mathbf{x},\mathbf{a}), \mathbf{v}_3 \rangle \tag{146}$$

$$= \left\langle \mathbf{\Phi}(\mathbf{z}), (\mathbf{\Lambda}_h^t)^{-1} \left\{ \sum_{\tau=1}^{k_t} \mathbf{\Phi}(\mathbf{x}_\tau, \mathbf{a}_\tau)[\mathbf{1}_M \cdot (\mathbb{P}_h V_{\boldsymbol{\upsilon},h+1}^t - \mathbb{P}_h V_{\boldsymbol{\upsilon},h+1}^\pi)(\mathbf{x}_\tau, \mathbf{a}_\tau)] \right\} \right\rangle \tag{147}$$

$$= \left\langle \mathbf{\Phi}(\mathbf{z}), (\mathbf{\Lambda}_h^t)^{-1} \left\{ \sum_{\tau=1}^{k_t} \mathbf{\Phi}(\mathbf{x}_\tau, \mathbf{a}_\tau)\mathbf{\Phi}(\mathbf{x}_\tau, \mathbf{a}_\tau)^\top \int (V_{\boldsymbol{\upsilon},h+1}^t - V_{\boldsymbol{\upsilon},h+1}^\pi)(\mathbf{x}') d\boldsymbol{\mu}_h(\mathbf{x}') \right\} \right\rangle \tag{148}$$

$$= \left\langle \mathbf{\Phi}(\mathbf{z}), (\mathbf{\Lambda}_h^t)^{-1} \left\{ \sum_{\tau=1}^{k_t} \mathbf{\Phi}(\mathbf{x}_\tau, \mathbf{a}_\tau)\mathbf{\Phi}(\mathbf{x}_\tau, \mathbf{a}_\tau)^\top \int (V_{\boldsymbol{\upsilon},h+1}^t - V_{\boldsymbol{\upsilon},h+1}^\pi)(\mathbf{x}') d\boldsymbol{\mu}_h(\mathbf{x}') \right\} \right\rangle \tag{149}$$

$$= \left\langle \mathbf{\Phi}(\mathbf{z}), \int (V_{\boldsymbol{\upsilon},h+1}^t - V_{\boldsymbol{\upsilon},h+1}^\pi)(\mathbf{x}') d\boldsymbol{\mu}_h(\mathbf{x}') \right\rangle - \lambda \left\langle \mathbf{\Phi}(\mathbf{z}), (\mathbf{\Lambda}_h^t)^{-1} \int (V_{\boldsymbol{\upsilon},h+1}^t - V_{\boldsymbol{\upsilon},h+1}^\pi)(\mathbf{x}') d\boldsymbol{\mu}_h(\mathbf{x}') \right\rangle \tag{150}$$

$$= \int (V_{\boldsymbol{\upsilon},h+1}^t - V_{\boldsymbol{\upsilon},h+1}^\pi)(\mathbf{x}') \langle \mathbf{\Phi}(\mathbf{z}), \boldsymbol{\mu}_h(\mathbf{x}') \rangle - \lambda \left\langle \mathbf{\Phi}(\mathbf{z}), (\mathbf{\Lambda}_h^t)^{-1} \int (V_{\boldsymbol{\upsilon},h+1}^t - V_{\boldsymbol{\upsilon},h+1}^\pi)(\mathbf{x}') d\boldsymbol{\mu}_h(\mathbf{x}') \right\rangle \tag{151}$$

$$= \mathbf{1}_M \cdot \left(\mathbb{P}_h (V_{\boldsymbol{\upsilon},h+1}^t - V_{\boldsymbol{\upsilon},h+1}^\pi)(\mathbf{x},\mathbf{a}) \right) - \lambda \left\langle \mathbf{\Phi}(\mathbf{z}), (\mathbf{\Lambda}_h^t)^{-1} \int (V_{\boldsymbol{\upsilon},h+1}^t - V_{\boldsymbol{\upsilon},h+1}^\pi)(\mathbf{x}') d\boldsymbol{\mu}_h(\mathbf{x}') \right\rangle \tag{152}$$

$$\leq \mathbf{1}_M \cdot \left(\mathbb{P}_h (V_{\boldsymbol{\upsilon},h+1}^t - V_{\boldsymbol{\upsilon},h+1}^\pi)(\mathbf{x},\mathbf{a}) + 2H\sqrt{d\lambda} \|\mathbf{\Phi}(\mathbf{x},\mathbf{a})\|_{(\mathbf{\Lambda}_h^t)^{-1}} \right). \tag{153}$$

Putting it all together, we have that since $\langle \mathbf{\Phi}(\mathbf{x},\mathbf{a}), \mathbf{w}_{\boldsymbol{\upsilon},h}^t - \mathbf{w}_{\boldsymbol{\upsilon},h}^\pi \rangle = \langle \mathbf{\Phi}(\mathbf{x},\mathbf{a}), \mathbf{v}_1 + \mathbf{v}_2 + \mathbf{v}_3 \rangle$, there exists an absolute constant C_β independent of M, T, H, d, such that, with probability at least $1 - \delta'/2$ for all $t \in [T], h \in [H], \boldsymbol{\upsilon} \in \mathbf{\Upsilon}$ simultaneously,

$$\begin{aligned} \left| \langle \boldsymbol{\upsilon}^\top \mathbf{\Phi}(\mathbf{x},\mathbf{a}), \mathbf{w}_{\boldsymbol{\upsilon},h}^t - \mathbf{w}_{\boldsymbol{\upsilon},h}^\pi \rangle \right| &\leq \boldsymbol{\upsilon}^\top \mathbf{1}_M \cdot \left(\mathbb{P}_h (V_{\boldsymbol{\upsilon},h+1}^t - V_{\boldsymbol{\upsilon},h+1}^\pi)(\mathbf{x},\mathbf{a}) \right) \\ &+ \|\mathbf{\Phi}(\mathbf{x},\mathbf{a})\|_{(\mathbf{\Lambda}_h^t)^{-1}} \|\boldsymbol{\upsilon}\|_2 \left(2H\sqrt{d\lambda} + C_\beta \cdot dH \cdot \sqrt{2\log\left(\frac{dMTH}{\delta'}\right)} + 2HM\lambda\sqrt{d} \right) \end{aligned} \tag{154}$$

Since $\lambda \leq 1$ and $\|\boldsymbol{\upsilon}\|_2 \leq 1$, there exists a constant C'_β that we have the following for any $(x,a) \in \mathcal{S} \times \mathcal{A}$ with probability $1 - \delta'/2$ simultaneously for all $h \in [H], \boldsymbol{\upsilon} \in \mathbf{\Upsilon}, t \in [T]$,

$$\begin{aligned} &\left| \langle \boldsymbol{\upsilon}^\top \mathbf{\Phi}(\mathbf{x},\mathbf{a}), \mathbf{w}_{\boldsymbol{\upsilon},h}^t - \mathbf{w}_{\boldsymbol{\upsilon},h}^\pi \rangle \right| \\ &\leq \mathbb{P}_h (V_{\boldsymbol{\upsilon},h+1}^t - V_{\boldsymbol{\upsilon},h+1}^\pi)(\mathbf{x},\mathbf{a}) + C'_\beta \cdot dMH \cdot \|\Phi(\mathbf{z})\|_{(\mathbf{\Lambda}_h^t)^{-1}} \cdot \sqrt{2\log\left(\frac{dMTH}{\delta'}\right)}. \end{aligned}$$

□

Lemma 21 (UCB in the Multiagent Setting)**.** *With probability at least $1 - \delta'/2$, we have that for all $(x,a,h,t,\boldsymbol{\upsilon}) \in \mathcal{S} \times \mathcal{A} \times [H] \times [T] \times \mathbf{\Upsilon}$, $Q_{\upsilon,h}^t(x,a) \geq Q_{\upsilon,h}^\star(x,a)$.*

Proof. We prove this result by induction. First, for the last step H, note that the statement holds as $Q_{\boldsymbol{\upsilon},H}^t(\mathbf{x},\mathbf{a}) \geq Q_{\boldsymbol{\upsilon},H}^\star(\mathbf{x},\mathbf{a})$ for all $\boldsymbol{\upsilon}$. Recall that the value function at step $H+1$ is zero. Therefore, by Lemma 20, we have that, for any $\boldsymbol{\upsilon} \in \mathbf{\Upsilon}$,

$$\left| \langle \boldsymbol{\upsilon}^\top \mathbf{\Phi}(\mathbf{x},\mathbf{a}), \mathbf{w}_{\boldsymbol{\upsilon},H}^t \rangle - Q_{\boldsymbol{\upsilon},H}^\star(\mathbf{x},\mathbf{a}) \right| \leq C'_\beta \cdot dH \cdot \|\Phi(\mathbf{z})\|_{(\mathbf{\Lambda}_h^t)^{-1}} \cdot \sqrt{2\log\left(\frac{dMTH}{\delta'}\right)}. \tag{155}$$

We have $Q^{\star}_{\boldsymbol{\upsilon},H}(\mathbf{x},\mathbf{a}) \leq \langle \boldsymbol{\upsilon}^{\top}\boldsymbol{\Phi}(\mathbf{x},\mathbf{a}), \mathbf{w}^{t}_{\boldsymbol{\upsilon},H}\rangle + C'_{\beta}\cdot dH\cdot \|\Phi(\mathbf{z})\|_{(\boldsymbol{\Lambda}^{t}_{h})^{-1}}\cdot\sqrt{2\log\left(\frac{dMTH}{\delta'}\right)} = Q^{t}_{\boldsymbol{\upsilon},H}$. Now, for the inductive case, we have by Lemma 20 for any $h\in[H], \boldsymbol{\upsilon}\in\boldsymbol{\Upsilon}$,

$$\begin{aligned}&\left|\langle \boldsymbol{\upsilon}^{\top}\boldsymbol{\Phi}(\mathbf{x},\mathbf{a}), \mathbf{w}^{t}_{\boldsymbol{\upsilon},h} - \mathbf{w}^{\star}_{\boldsymbol{\upsilon},h}\rangle - \left(\mathbb{P}_h V^{\star}_{\boldsymbol{\upsilon},h+1}(\mathbf{x},\mathbf{a}) - \mathbb{P}_h V^{t}_{\boldsymbol{\upsilon},h+1}(\mathbf{x},\mathbf{a})\right)\right| \\ &\leq C'_{\beta}\cdot dH\cdot \|\Phi(\mathbf{z})\|_{(\boldsymbol{\Lambda}^{t}_{h})^{-1}}\cdot\sqrt{2\log\left(\frac{dMTH}{\delta'}\right)}.\end{aligned}$$

By the inductive assumption we have $Q^{t}_{\boldsymbol{\upsilon},h+1}(\mathbf{x},\mathbf{a}) \geq Q^{\star}_{\boldsymbol{\upsilon},h+1}(\mathbf{x},\mathbf{a})$ implying $\mathbb{P}_h V^{\star}_{\boldsymbol{\upsilon},h+1}(\mathbf{x},\mathbf{a}) - \mathbb{P}_h V^{t}_{\boldsymbol{\upsilon},h+1}(\mathbf{x},\mathbf{a}) \geq 0$. Substituting the appropriate Q value formulations we have,

$$Q^{\star}_{\boldsymbol{\upsilon},h} \leq \langle \boldsymbol{\upsilon}^{\top}\boldsymbol{\Phi}(\mathbf{x},\mathbf{a}), \mathbf{w}^{t}_{\boldsymbol{\upsilon},h}\rangle + C'_{\beta}\cdot dH\cdot \|\Phi(\mathbf{z})\|_{(\boldsymbol{\Lambda}^{t}_{h})^{-1}}\cdot\sqrt{2\log\left(\frac{dMTH}{\delta'}\right)} = Q^{t}_{\boldsymbol{\upsilon},h}(\mathbf{x},\mathbf{a}). \tag{156}$$

□

Lemma 22 (Recursive Relation in Multiagent MDP Settings)**.** *For any $\boldsymbol{\upsilon}\in\boldsymbol{\Upsilon}$, let $\delta^{t}_{\boldsymbol{\upsilon},h} = V^{t}_{\boldsymbol{\upsilon},h}(\mathbf{x}^{t}_{h}) - V^{\pi_t}_{\boldsymbol{\upsilon},h}(\mathbf{x}^{t}_{h})$, and $\xi^{t}_{\boldsymbol{\upsilon},h+1} = \mathbb{E}\left[\delta^{t}_{\boldsymbol{\upsilon},h}|\mathbf{x}^{t}_{h},\mathbf{a}^{t}_{h}\right] - \delta^{t}_{\boldsymbol{\upsilon},h}$. Then, with probability at least $1-\alpha$, for all $(t,h)\in[T]\times[H]$ simultaneously,*

$$\delta^{t}_{\boldsymbol{\upsilon},h} \leq \delta^{t}_{\boldsymbol{\upsilon},h+1} + \xi^{t}_{\boldsymbol{\upsilon},h+1} + 2\left\|\boldsymbol{\Phi}(\mathbf{x}^{t}_{h},\mathbf{a}^{t}_{h})\right\|_{(\boldsymbol{\Lambda}^{t}_{h})^{-1}}\cdot C'_{\beta}\cdot dH\cdot\sqrt{2\log\left(\frac{dMTH}{\alpha}\right)}. \tag{157}$$

Proof. By Lemma 20, we have that for any $(x,a,h,\boldsymbol{\upsilon},t)\in\mathcal{S}\times\mathcal{A}\times[H]\times\boldsymbol{\Upsilon}\times[T]$ with probability at least $1-\alpha/2$,

$$\begin{aligned}&Q^{t}_{\boldsymbol{\upsilon},h}(\mathbf{x},\mathbf{a}) - Q^{\pi_t}_{\boldsymbol{\upsilon},h}(\mathbf{x},\mathbf{a}) \\ &\leq \mathbb{P}_h(V^{t}_{\boldsymbol{\upsilon},h+1} - V^{\pi_t}_{\boldsymbol{\upsilon},h})(\mathbf{x},\mathbf{a}) + 2\left\|\boldsymbol{\phi}(\mathbf{x},\mathbf{a})\right\|_{(\boldsymbol{\Lambda}^{t}_{h})^{-1}}\cdot C_{\beta}\cdot dH\cdot\sqrt{2\log\left(\frac{dMTH}{\alpha}\right)}.\end{aligned}$$

Replacing the definition of $\delta^{t}_{\boldsymbol{\upsilon},h}$ and $V^{\pi_t}_{\boldsymbol{\upsilon},h}$ finishes the proof. □

Lemma 23. *Let $\boldsymbol{\upsilon}_1,\boldsymbol{\upsilon}_2,...,\boldsymbol{\upsilon}_T$ be T i.i.d. samples from $\boldsymbol{\Upsilon}$. For $\xi^{t}_{\boldsymbol{\upsilon}_t,h}$ as defined earlier and any $\delta\in(0,1)$, we have with probability at least $1-\delta/2$,*

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\xi^{t}_{\boldsymbol{\upsilon}_t,h} \leq \sqrt{2H^3T\log\left(\frac{2}{\alpha}\right)}. \tag{158}$$

Proof. The proof is identical to Lemma 8, since $|\xi^{t}_{\boldsymbol{\upsilon}_t,h}| \leq H$ regardless of $\boldsymbol{\upsilon}_t$, which allows us to apply the Martingale concentration with the same analysis. □

We are now ready to prove Theorem 4. We first restate the Theorem for completeness.

Theorem 4. `CoopLSVI` when run on a multiagent MDP with M agents and communication threshold S, $\beta_t = \mathcal{O}(dH\sqrt{\log(tMH)})$ and $\lambda = 1 - \frac{1}{MTH}$ obtains the following regret after T episodes, with probability at least $1-\alpha$,

$$\mathfrak{R}_C(T) = \widetilde{\mathcal{O}}\left(d^{\frac{3}{2}}H^2\sqrt{ST\log\left(\frac{1}{\alpha}\right)}\right).$$

Proof. We have by the definition of cumulative regret:

$$\mathfrak{R}_C(T) = \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{\upsilon}_t \sim \boldsymbol{\Upsilon}} \left[\max_{\mathbf{x}_1^t \in \mathcal{S}} \left[V^{\star}_{\boldsymbol{\upsilon}_t,1}(\mathbf{x}_1^t) - V^{\pi_t}_{\boldsymbol{\upsilon}_t,1}(\mathbf{x}_1^t) \right] \right] = \mathbb{E}_{\boldsymbol{\upsilon}_t \sim \boldsymbol{\Upsilon}} \left[\sum_{t=1}^{T} \max_{\mathbf{x}_1^t \in \mathcal{S}} \left[V^{\star}_{\boldsymbol{\upsilon}_t,1}(\mathbf{x}_1^t) - V^{\pi_t}_{\boldsymbol{\upsilon}_t,1}(\mathbf{x}_1^t) \right] \right]. \tag{159}$$

Our analysis focuses only on the term inside the expectation, which we will bound via terms that are independent of $\boldsymbol{\upsilon}_1, ..., \boldsymbol{\upsilon}_T$, bounding $\mathfrak{R}_C$. We have,

$$\begin{aligned}
&\sum_{t=1}^{T} \max_{\mathbf{x}_1^t \in \mathcal{S}} \left[V^{\star}_{\boldsymbol{\upsilon}_t,1}(\mathbf{x}_1^t) - V^{\pi_t}_{\boldsymbol{\upsilon}_t,1}(\mathbf{x}_1^t) \right] \\
&\leq \sum_{t=1}^{T} \max_{\mathbf{x}_1^t \in \mathcal{S}} \delta^t_{\boldsymbol{\upsilon}_t,1} \\
&\leq \sum_{t=1}^{T} \left[\max_{\mathbf{x}_1^t \in \mathcal{S}} \sum_{h=1}^{H} \xi^t_{\boldsymbol{\upsilon}_t,h} + 2C'_{\beta} \cdot dH \cdot \sqrt{2 \log\left(\frac{dMTH}{\alpha}\right)} \left(\sum_{t=1}^{T} \sum_{h=1}^{H} \left\| \boldsymbol{\Phi}(\mathbf{x}_h^t, \mathbf{a}_h^t) \right\|_{(\boldsymbol{\Lambda}_h^t)^{-1}} \right) \right].
\end{aligned}$$

Where the last inequality holds with probability at least $1 - \alpha/2$, via Lemma 22 and Lemma 21. Next, we can bound the first term via Lemma 23. We have with probability at least $1 - \alpha$, for some absolute constant C'_{β},

$$\begin{aligned}
&\sum_{t=1}^{T} \max_{\mathbf{x}_1^t \in \mathcal{S}} \left[V^{\star}_{\boldsymbol{\upsilon}_t,1}(\mathbf{x}_1^t) - V^{\pi_t}_{\boldsymbol{\upsilon}_t,1}(\mathbf{x}_1^t) \right] \\
&\leq \sqrt{2H^3 T \log\left(\frac{2}{\alpha}\right)} + 2C'_{\beta} \cdot dH \cdot \sqrt{2 \log\left(\frac{dMTH}{\alpha}\right)} \left(\sum_{t=1}^{T} \sum_{h=1}^{H} \left\| \boldsymbol{\Phi}(\mathbf{x}_h^t, \mathbf{a}_h^t) \right\|_{(\boldsymbol{\Lambda}_h^t)^{-1}} \right).
\end{aligned}$$

Finally, to bound the summation, we can use the technique in Theorem 4 of Abbasi-Yadkori et al. (2011). Assume that the last time synchronization of rewards occured was at instant k_T. We therefore have, by Lemma 12 of Abbasi-Yadkori et al. (2011), for any $h \in [H]$

$$\sum_{t=1}^{T} \left\| \boldsymbol{\Phi}(\mathbf{x}_h^t, \mathbf{a}_h^t) \right\|_{(\boldsymbol{\Lambda}_h^t)^{-1}} \leq \frac{\det(\bar{\boldsymbol{\Lambda}}_h^T)}{\det(\boldsymbol{\Lambda}_h^T)} \sum_{t=1}^{T} \left\| \boldsymbol{\Phi}(\mathbf{x}_h^t, \mathbf{a}_h^t) \right\|_{(\bar{\boldsymbol{\Lambda}}_h^t)^{-1}} \leq \sqrt{S} \sum_{t=1}^{T} \left\| \boldsymbol{\Phi}(\mathbf{x}_h^t, \mathbf{a}_h^t) \right\|_{(\bar{\boldsymbol{\Lambda}}_h^T)^{-1}} \tag{160}$$

Here $\bar{\boldsymbol{\Lambda}}_h^T = \sum_{t=1}^{T} \boldsymbol{\Phi}(\mathbf{x}_h^t, \mathbf{a}_h^t) \boldsymbol{\Phi}(\mathbf{x}_h^t, \mathbf{a}_h^t)^{\top}$ and the last inequality follows from the algorithms' synchronization condition. Replacing this result, we have that,

$$\sum_{t=1}^{T} \sum_{h=1}^{H} \left\| \boldsymbol{\Phi}(\mathbf{x}_h^t, \mathbf{a}_h^t) \right\|_{(\boldsymbol{\Lambda}_h^t)^{-1}} \leq 2 \sum_{h=1}^{H} \left(\sqrt{S} \sum_{t=1}^{T} \left\| \boldsymbol{\Phi}(\mathbf{x}_h^t, \mathbf{a}_h^t) \right\|_{(\bar{\boldsymbol{\Lambda}}_h^T)^{-1}} \right) \quad \leq 2H \sqrt{ST \cdot d \log \frac{MT + \lambda}{\lambda}}. \tag{161}$$

Where the last inequality is an application of Lemma 33 and using the fact that $\|\boldsymbol{\Phi}(\cdot)\|_2 \leq \sqrt{M}$. Replacing this result, we have that with probability at least $1 - \alpha$,

$$\begin{aligned}
&\sum_{t=1}^{T} \max_{\mathbf{x}_1^t \in \mathcal{S}} \left[V^{\star}_{\boldsymbol{\upsilon}_t,1}(\mathbf{x}_1^t) - V^{\pi_t}_{\boldsymbol{\upsilon}_t,1}(\mathbf{x}_1^t) \right] \\
&\leq \sqrt{2H^3 T \log\left(\frac{2}{\alpha}\right)} + 2C'_{\beta} \cdot dH^2 \cdot \sqrt{2ST \log\left(\frac{dMTH}{\alpha}\right) \cdot d \log \frac{MT + \lambda}{\lambda}}.
\end{aligned}$$

Taking expectation of the RHS over $\boldsymbol{\upsilon}_1, ..., \boldsymbol{\upsilon}_T$ gives us the final result (the $\widetilde{\mathcal{O}}$ notation hides poly-logarithmic factors). □

C.5 Proof of Lemma 3

Proof. Let the total rounds of communication triggered by the threshold condition in any step $h \in [H]$ be given by n_h. Then, we have, by the communication criterion,

$$S^{n_h} < \frac{\det\left(\mathbf{\Lambda}_h^T\right)}{\det\left(\mathbf{\Lambda}_h^0\right)} \leq (1 + MT/d)^d. \tag{162}$$

Where the last inequality follows from Lemma 33 and the fact that $\|\mathbf{\Phi}\| \leq \sqrt{M}$. This gives us that $n_h \leq d \log_S (1 + MT/d)) + 1$. Furthermore, by noticing that $n \leq \sum_{h=1}^{H} n_h$, we have the final result. □

D Weight Norm Bounds

Lemma 24 (Bound on Weights of Homogenous Value Functions, Lemma B.1 of Jin et al. (2020))**.** *Under the linear MDP Assumption (Definition 1), for any fixed policy π, let $\{\mathbf{w}_h^\pi\}_{h\in[H]}$ be the weights such that $Q_h^\pi(x,a) = \langle \boldsymbol{\phi}(x,a), \mathbf{w}_h^\pi\rangle$ for all $(x,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$ and $m \in \mathcal{M}$. Then, we have,*

$$\|\mathbf{w}_h^\pi\|_2 \le 2H\sqrt{d}.$$

Lemma 25 (Linearity of weights in MMDP)**.** *Under the linear MMDP Assumption (Definition 5), for any policy π and $\boldsymbol{\upsilon} \in \boldsymbol{\Upsilon}$, there exists weights $\{\mathbf{w}_{\boldsymbol{\upsilon},h}^\pi\}_{h\in[H]}$ such that $Q_{\boldsymbol{\upsilon},h}^\pi(\mathbf{x},\mathbf{a}) = \boldsymbol{\upsilon}^\top \boldsymbol{\Phi}(\mathbf{x},\mathbf{a})^\top \mathbf{w}_{\boldsymbol{\upsilon},h}^\pi$ for all $(\mathbf{x},\mathbf{a},h) \in \mathcal{S}\times\mathcal{A}\times[H]$, where $\|\mathbf{w}_{\boldsymbol{\upsilon},h}^\pi\|_2 \le 2H\sqrt{d}$.*

Proof. By the Bellman equation and Proposition 1, we have that for any MDP corresponding to the scalarization parameter $\boldsymbol{\upsilon} \in \boldsymbol{\Upsilon}$ and any policy $\boldsymbol{\pi}$, state $\mathbf{x} \in \mathcal{S}$, joint action $\mathbf{a} \in \mathcal{A}$,

$$\begin{aligned}
Q_{\boldsymbol{\upsilon},h}^\pi(\mathbf{x},\mathbf{a}) &= \boldsymbol{\upsilon}^\top \mathbf{r}_h(\mathbf{x},\mathbf{a}) + \mathbb{P}_h V_{\boldsymbol{\upsilon},h+1}^\pi(\mathbf{x},\mathbf{a}) && (163)\\
&= \boldsymbol{\upsilon}^\top \left(\mathbf{r}_h(\mathbf{x},\mathbf{a}) + \mathbf{1}_M \cdot \mathbb{P}_h V_{\boldsymbol{\upsilon},h+1}^\pi(\mathbf{x},\mathbf{a})\right) && (164)\\
&= \boldsymbol{\upsilon}^\top \left(\boldsymbol{\Phi}(\mathbf{x},\mathbf{a})^\top \begin{bmatrix}\boldsymbol{\theta}_h \\ \mathbf{0}_{d_2}\end{bmatrix} + \int V_{\boldsymbol{\upsilon},h+1}^\pi(\mathbf{x}')\boldsymbol{\Phi}(\mathbf{x},\mathbf{a})^\top \begin{bmatrix}\mathbf{0}_{d_1} \\ d\boldsymbol{\mu}_h(\mathbf{x}')\end{bmatrix} d\mathbf{x}'\right) && (165)\\
&= \boldsymbol{\upsilon}^\top \boldsymbol{\Phi}(\mathbf{x},\mathbf{a})^\top \mathbf{w}_{\boldsymbol{\upsilon},h}^\pi. && (166)
\end{aligned}$$

Where $\mathbf{w}_{\boldsymbol{\upsilon},h}^\pi = \begin{bmatrix}\boldsymbol{\theta}_h \\ \int V_{\boldsymbol{\upsilon},h+1}^\pi(\mathbf{x}')d\boldsymbol{\mu}(\mathbf{x}')d\mathbf{x}'\end{bmatrix}$. Therefore, since $\|\boldsymbol{\theta}_h\| \le \sqrt{d}$ and $\|\int V_{\boldsymbol{\upsilon},h+1}^\pi(\mathbf{x}')d\boldsymbol{\mu}(\mathbf{x}')\| \le H\sqrt{d}$, the result follows. □

Lemma 26 (Bound on Weights of `CoopLSVI` Policy for MDPs)**.** *At any $t \in [T]$ for any $m \in \mathcal{M}$ and all $h \in [H]$, we have that the weights $\mathbf{w}_{m,h}^t$ of Algorithm 1 satisfy,*

$$\|\mathbf{w}_{m,h}^t\|_2 \le 2H\sqrt{dMt/\lambda}.$$

Proof. For any vector $\mathbf{v} \in \mathbb{R}^d | \|\mathbf{v}\| = 1$,

$$\begin{aligned}
\left|\mathbf{v}^\top \widehat{\boldsymbol{\theta}}_{m,h}^t\right| &= \left|\mathbf{v}^\top \left(\boldsymbol{\Lambda}_{m,h}^t\right)^{-1} \left(\sum_{\tau=1}^{U_h^m(t)} \left[\boldsymbol{\phi}(x_\tau,a_\tau)\left[r(x_\tau,a_\tau) + \max_a Q_{m,h+1}(x'_\tau,a)\right]\right]\right)\right| && (167)\\
&\le 2H \cdot \left|\mathbf{v}^\top \left(\boldsymbol{\Lambda}_{m,h}^t\right)^{-1} \left(\sum_{\tau=1}^{U_h^m(t)} \boldsymbol{\phi}(x_\tau,a_\tau)\right)\right| && (168)\\
&\le 2H \cdot \sqrt{\left|\left(\sum_{\tau=1}^{U_h^m(t)} \|\mathbf{v}\|^2_{\left(\boldsymbol{\Lambda}_{m,h}^t\right)^{-1}} \|\boldsymbol{\phi}(x_\tau,a_\tau)\|^2_{\left(\boldsymbol{\Lambda}_{m,h}^t\right)^{-1}}\right)\right|} && (169)\\
&\le 2H\|\mathbf{v}\|\sqrt{dU_h^m(t)/\lambda} \le 2H\sqrt{dMt/\lambda}. && (170)
\end{aligned}$$

The penultimate inequality follows from Lemma 33 and the final inequality follows from the fact that $U_h^m(t) \le Mt$. The remainder of the proof follows from the fact that for any vector $\mathbf{w}$, $\|\mathbf{w}\| = \max_{\mathbf{v}:\|\mathbf{v}\|=1} |\mathbf{v}^\top \mathbf{w}|$. □

Lemma 27 (Bound on Weights in MMDP `CoopLSVI`)**.** *For any $t \in [T], h \in [H], \boldsymbol{\upsilon} \in \boldsymbol{\Upsilon}$, the weight $\mathbf{w}_{\boldsymbol{\upsilon},h}^t$ in `CoopLSVI` in the multiagent MDP satisfies,*

$$\|\mathbf{w}_{\boldsymbol{\upsilon},h}^t\|_2 \le 2HM\sqrt{dt/\lambda}.$$

Proof. For any vector $\mathbf{v} \in \mathbb{R}^d | \|\mathbf{v}\| = 1$,

$$\left|\mathbf{v}^\top \mathbf{w}^t_{\boldsymbol{v},h}\right| = \left|\mathbf{v}^\top \left(\mathbf{\Lambda}^t_h\right)^{-1} \left(\sum_{\tau=1}^{k_t} \left[\mathbf{\Phi}(\mathbf{x}_\tau, \mathbf{a}_\tau) \left[\mathbf{r}_h(\mathbf{x}_\tau, \mathbf{a}_\tau) + \max_{\mathbf{a}\in\mathcal{A}} Q_{\boldsymbol{v},h+1}(\mathbf{x}'_\tau, \mathbf{a})\right]\right]\right)\right| \tag{171}$$

$$\leq \sqrt{k_t \cdot \sum_{\tau=1}^{k_t} \left(\mathbf{v}^\top \left(\mathbf{\Lambda}^t_h\right)^{-1} \left[\mathbf{\Phi}(\mathbf{x}_\tau, \mathbf{a}_\tau) \left[\mathbf{r}_h(\mathbf{x}_\tau, \mathbf{a}_\tau) + \max_{\mathbf{a}\in\mathcal{A}} Q_{\boldsymbol{v},h+1}(\mathbf{x}'_\tau, \mathbf{a})\right]\right]\right)^2} \tag{172}$$

$$\leq HM\sqrt{k_t \cdot \sum_{\tau=1}^{k_t} \left\|\mathbf{v}^\top \left(\mathbf{\Lambda}^t_h\right)^{-1} \mathbf{\Phi}(\mathbf{x}_\tau, \mathbf{a}_\tau)\right\|_2^2} \tag{173}$$

$$\leq 2HM\sqrt{k_t \cdot \sum_{\tau=1}^{k_t} \|\mathbf{v}\|^2_{\left(\mathbf{\Lambda}^t_h\right)^{-1}} \|\mathbf{\Phi}(\mathbf{x}_\tau, \mathbf{a}_\tau)\|^2_{\left(\mathbf{\Lambda}^t_h\right)^{-1}}} \tag{174}$$

$$\leq 2HM\|\mathbf{v}\|\sqrt{dk_t/\lambda} \leq 2HM\sqrt{dt/\lambda}. \tag{175}$$

The penultimate inequality follows from Lemma 33 and the final inequality follows from the fact that $k_t \leq t$. The remainder of the proof follows from the fact that for any vector $\mathbf{w}, \|\mathbf{w}\| = \max_{\mathbf{v}:\|\mathbf{v}\|=1} |\mathbf{v}^\top \mathbf{w}|$. □

E Martingale Concentration Bounds

Lemma 28 (Lemma E.2 of Yang et al. (2020), Lemma D.4 of Jin et al. (2018))**.** *Let $\{x_\tau\}_{\tau=1}^\infty$ and $\{\boldsymbol{\phi}_\tau\}_{\tau=1}^\infty$ be an $\mathcal{S}$-valued and an $\mathcal{H}$-valued stochastic process adapted to filtration $\{\mathcal{F}_\tau\}_{\tau=0}^\infty$ respectively, where we assume that $\|\boldsymbol{\phi}_\tau\|_2 \leq 1$ for all $\tau \geq 1$. Besides, for any $t \geq 1$, define $\mathbf{\Lambda}_t : \mathcal{H} \to \mathcal{H}$ as $\mathbf{\Lambda}_t = \lambda \mathbf{I}_d + \sum_{\tau=1}^t \boldsymbol{\phi}_\tau \boldsymbol{\phi}_\tau^\top$ with $\lambda > 1$. Then, for any $\delta > 0$ with probability at least $1-\delta$, we have,*

$$\sup_{V\in\mathcal{V}} \left\|\sum_{\tau=1}^t \boldsymbol{\phi}_\tau \left\{V(x_\tau) - \mathbb{E}[V(x_\tau)|\mathcal{F}_{\tau-1}]\right\}\right\|^2_{\mathbf{\Lambda}_t^{-1}}$$
$$\leq 4H^2 \cdot \log \frac{\det\left(\mathbf{\Lambda}_t\right)}{\det\left(\lambda \mathbf{I}_d\right)} + 4H^2 t(\lambda - 1) + 8H^2 \log\left(\frac{|\mathcal{N}_\epsilon|)}{\delta}\right) + \frac{8t^2\epsilon^2}{\lambda}.$$

E.1 Multi-task concentration bound (Chowdhury & Gopalan, 2020)

We assume the multi-agent kernel $\mathbf{\Gamma}$ to be continuous relative to the operator norm on $\mathcal{L}(\mathbb{R}^n)$, the space of bounded linear operators from $\mathbb{R}^n$ to itself (for some $n > 0$). Then the RKHS $\mathcal{H}_{\mathbf{\Gamma}}(\mathcal{X}^n)$ associated with the kernel $\mathbf{\Gamma}$ is a subspace of the space of continuous functions from $\mathcal{X}^n$ to $\mathbb{R}^n$, and hence, $\mathbf{\Gamma}$ is a Mercer kernel. Let μ be a measure on the (compact) set $\mathcal{X}^n$. Since $\mathbf{\Gamma}$ is a Mercer kernel on $\mathcal{X}$ and $\sup_{\mathbf{X}\in\mathcal{X}^n} \|\mathbf{\Gamma}(\mathbf{X}, \mathbf{X})\| < \infty$, the RKHS $\mathcal{H}_{\mathbf{\Gamma}}(\mathcal{X}^n)$ is a subspace of $L^2(\mathcal{X}^n, \mu; \mathbb{R}^n)$, the Banach space of measurable functions $g : \mathcal{X}^n \to \mathbb{R}^n$ such that $\int_{\mathcal{X}^n} \|g(\mathbf{X})\|^2 d\mu(\mathbf{X}) < \infty$, with norm $\|g\|_{L^2} = \left(\int_{\mathcal{X}^n} \|g(\mathbf{X})\|^2 d\mu(\mathbf{X}).\right)^{1/2}$. Since $\mathbf{\Gamma}(\mathbf{X}, \mathbf{X}) \in \mathcal{L}(\mathbb{R}^n)$ is a compact operator, by the Mercer theorem

We can therefore define a feature map $\Phi : \mathcal{X}^M \to \mathcal{L}(\mathbb{R}^n, \ell^2)$ of the multi-agent kernel Γ by

$$\Phi(\mathbf{X})^\top \mathbf{y} = \left(\sqrt{\nu_1}\psi_1(\mathbf{x}_1)^\top \mathbf{y}, \sqrt{\nu_2}\psi_2(\mathbf{x}_2)^\top \mathbf{y}, ..., \sqrt{\nu_M}\psi_M(\mathbf{x}_M)^\top \mathbf{y}\right), \ \forall \mathbf{X} \in \mathcal{X}^M, \mathbf{y} \in \mathbb{R}^m. \tag{176}$$

We then obtain $F(\mathbf{X}) = \Phi(\mathbf{X})^\top \boldsymbol{\theta}^\star$ and $\Gamma(\mathbf{X}, \mathbf{X}') = \Phi(\mathbf{X})^\top \Phi(\mathbf{X}')\ \forall\ \mathbf{X}, \mathbf{X}' \in \mathcal{X}^M$.

Define $\mathbf{S}_t = \sum_{\tau=1}^{t} \Phi(\mathbf{X}_\tau)^\top \varepsilon_\tau$, where $\varepsilon_1, ..., \varepsilon_t$ are the noise vectors in $\mathbb{R}^M$. Now consider $\mathcal{F}_{t-1}$, the σ-algebra generated by the random variables $\{\mathbf{X}_\tau, \varepsilon_\tau\}_{\tau=1}^{t-1}$ and $\mathbf{X}_t$. We can see that $\mathbf{S}_t$ is $\mathcal{F}_t$-measurable, and additionally, $\mathbb{E}[\mathbf{S}_t|\mathcal{F}_{t-1}] = \mathbf{S}_{t-1}$. Therefore, $\{\mathbf{S}_t\}_{t\geqslant 1}$ is a martingale with outputs in ℓ^2 space. Following Chowdhury & Gopalan (2020), consider now the map $\Phi_{\mathcal{X}_t} : \ell^2 \to \mathbb{R}^{Mt}$:

$$\Phi_{\mathcal{X}_t}\boldsymbol{\theta} = \left[\left(\Phi(\mathbf{X}_1)^\top\boldsymbol{\theta}\right)^\top, \left(\Phi(\mathbf{X}_1)^\top\boldsymbol{\theta}\right)^\top, ..., \left(\Phi(\mathbf{X}_t)^\top\boldsymbol{\theta}\right)^\top\right]^\top, \ \forall\ \boldsymbol{\theta} \in \ell^2. \tag{177}$$

Additionally, denote $\mathbf{V}_t := \Phi_{\mathcal{X}_t}^\top \Phi_{\mathcal{X}_t}$ be a map from ℓ^2 to itself, with $\mathbf{I}$ being the identity operator in ℓ^2. We have the following result from Chowdhury & Gopalan (2020) that provides us with a self-normalized martingale bound.

Lemma 29 (Lemma 3 of Chowdhury & Gopalan (2020)). *Let the noise vectors $\{\boldsymbol{\varepsilon}_t\}_{t\geqslant 1}$ be σ-sub-Gaussian. Then, for any $\eta > 0$ and $\delta \in (0, 1]$, with probability at least $1 - \delta$, the following holds uniformly over all $t \geqslant 1$:*

$$\|\mathbf{S}_t\|_{(\mathbf{V}_t+\eta\mathbf{I})^{-1}} \leqslant \sigma\sqrt{2\log(1/\delta) + \log\det(\mathbf{I} + \eta^{-1}\mathbf{V}_t)}.$$

Alternatively stated, we have again that with probability at least $1 - \delta$, the following holds uniformly over all $t \geqslant 1$:

$$\|\boldsymbol{\mathcal{E}}_t\|^2_{\left((\mathbf{K}_t+\eta\mathbf{I})^{-1}+\mathbf{I}\right)^{-1}} \leqslant 2\sigma^2 \log\left[\frac{\sqrt{\det(\mathbf{I}(1+\eta) + \mathbf{K}_t)}}{\delta}\right].$$

F Covering Number Bounds

Lemma 30 (Covering Number of the Euclidean Ball). *For any $\varepsilon > 0$, the $\varepsilon-$covering number of the Euclidean ball in $\mathbb{R}^d$ with radius $R > 0$ is less than $(1 + 2R/\varepsilon)^d$.*

Lemma 31 (Covering Number for UCB-style value functions, Lemma D.6 of Jin et al. (2020)). *Let $\mathcal{V}$ denote a class of functions mapping from $\mathcal{S}$ to $\mathbb{R}$ with the following parameteric form*

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

where the parameters $(\mathbf{w}, \beta, \boldsymbol{\Lambda})$ are such that $\|\mathbf{w}\| \leq L$, $\beta \in (0, B]$, $\|\boldsymbol{\phi}(x, a)\| \leq 1\ \forall (x, a) \in \mathcal{S} \times \mathcal{A}$, and the minimum eigenvalue of $\boldsymbol{\Lambda}$ satisfies $\lambda_{\min}(\boldsymbol{\Lambda}) \geq \lambda$. Let $\mathcal{N}_\varepsilon$ be the $\varepsilon-$covering number of $\mathcal{V}$ with respect to the distance dist$(V, V') = \sup_{x\in\mathcal{S}} |V(x) - V'(x)|$. Then,

$$\log\mathcal{N}_\varepsilon \leq d\log\left(1 + 4L/\varepsilon\right) + d^2\log\left(1 + 8d^{1/2}B^2/(\lambda\epsilon^2)\right).$$

Lemma 32 (Covering number for multiagent MDP UCB-style functions). *Let $\mathcal{V}$ denote a class of functions mapping from $\mathcal{S}$ to $\mathbb{R}$ with the following parameteric form*

$$\mathbf{v}_{\boldsymbol{\upsilon}}(\cdot) = \mathbf{1}_M \cdot \min\left\{\max_{\mathbf{a}\in\mathcal{A}}\left[\langle\boldsymbol{\upsilon}, \mathbf{v}(\cdot, \mathbf{a})\rangle + \beta\left\|\boldsymbol{\Phi}(\cdot, \mathbf{a})^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{\Phi}(\cdot, \mathbf{a})\right\|\right], H\right\}, \mathbf{v}(\cdot, \mathbf{a}) = \mathbf{w}^\top\boldsymbol{\Phi}(\cdot, \mathbf{a})$$

where the parameters $(\mathbf{w}, \beta, \boldsymbol{\Lambda})$ are such that $\mathbf{w} \in \mathbb{R}^d$, $\|\mathbf{w}\|_2 \leq L$, $\beta \in (0, B]$, $\|\boldsymbol{\Phi}(\mathbf{x}, \mathbf{a})\| \leq \sqrt{M}\ \forall (\mathbf{x}, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$, and the minimum eigenvalue of $\boldsymbol{\Lambda}$ satisfies $\lambda_{\min}(\boldsymbol{\Lambda}) \geq \lambda$. Let $\mathcal{N}_\varepsilon$ be the $\varepsilon-$covering number of $\mathcal{V}$ with respect to the distance dist$(\mathbf{v}, \mathbf{v}') = \sup_{\mathbf{x}\in\mathcal{S}, \boldsymbol{\upsilon}\in\boldsymbol{\Upsilon}} |\mathbf{v}_{\boldsymbol{\upsilon}}(\mathbf{x}) - \mathbf{v}'_{\boldsymbol{\upsilon}}(\mathbf{x})|$. Then,

$$\log\left(\mathcal{N}_\varepsilon\right) \leq d \cdot \log\left(1 + \frac{4LM^2}{\varepsilon}\right) + d^2\log\left(1 + \frac{8Md^{1/2}B^2}{\lambda\varepsilon^2}\right).$$

Proof. We have that for two matrices $\mathbf{A}_1 = \beta^2 \mathbf{\Lambda}_1^{-1}, \mathbf{A}_2 = \beta^2 \mathbf{\Lambda}_2^{-1}$ and weight matrices $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d$, by a strategy similar to that of Lemma 31,

$$\sup_{\boldsymbol{\upsilon} \in \boldsymbol{\Upsilon}, \mathbf{x} \in \mathcal{S}} |\mathbf{v}_{\boldsymbol{\upsilon}}(\mathbf{x}) - \mathbf{v}'_{\boldsymbol{\upsilon}}(\mathbf{x})|_1 \tag{178}$$

$$= M \cdot \sup_{\mathbf{x} \in \mathcal{S}, \boldsymbol{\upsilon} \in \boldsymbol{\Upsilon}} \left| \boldsymbol{\upsilon}^\top \mathbf{v}(\mathbf{x}) - \boldsymbol{\upsilon}^\top \mathbf{v}'(\mathbf{x}) \right| \tag{179}$$

$$\leq M \cdot \sup_{\mathbf{x} \in \mathcal{S}} |\mathbf{v}(\mathbf{x}) - \mathbf{v}'(\mathbf{x})|_1 \tag{180}$$

$$\leq M \cdot \sup_{\mathbf{x} \in \mathcal{S}, \mathbf{a} \in \mathcal{A}} \left| \left[\mathbf{w}_1^\top \mathbf{\Phi}(\cdot, \mathbf{a}) + \left\| \mathbf{\Phi}(\cdot, \mathbf{a})^\top \mathbf{A}_1 \mathbf{\Phi}(\cdot, \mathbf{a}) \right\|_2 \right] - \left[\mathbf{w}_2^\top \mathbf{\Phi}(\cdot, \mathbf{a}) + \left\| \mathbf{\Phi}(\cdot, \mathbf{a})^\top \mathbf{A}_2 \mathbf{\Phi}(\cdot, \mathbf{a}) \right\|_2 \right] \right|_1 \tag{181}$$

$$\leq M \cdot \sup_{\mathbf{x} \in \mathcal{S}, \mathbf{a} \in \mathcal{A}} \left| \left(\mathbf{w}_1 - \mathbf{w}_2 \right)^\top \mathbf{\Phi}(\cdot, \mathbf{a}) + \left\| \mathbf{\Phi}(\cdot, \mathbf{a})^\top \mathbf{A}_1 \mathbf{\Phi}(\cdot, \mathbf{a}) \right\|_2 - \left\| \mathbf{\Phi}(\cdot, \mathbf{a})^\top \mathbf{A}_2 \mathbf{\Phi}(\cdot, \mathbf{a}) \right\|_2 \right|_1 \tag{182}$$

$$\leq M \cdot \sup_{\mathbf{x} \in \mathcal{S}, \mathbf{a} \in \mathcal{A}} \left| \left(\mathbf{w}_1 - \mathbf{w}_2 \right)^\top \mathbf{\Phi}(\cdot, \mathbf{a}) \right|_1 + M \cdot \sup_{\mathbf{x} \in \mathcal{S}, \mathbf{a} \in \mathcal{A}} \left| \left\| \mathbf{\Phi}(\cdot, \mathbf{a})^\top \mathbf{A}_1 \mathbf{\Phi}(\cdot, \mathbf{a}) \right\|_2 - \left\| \mathbf{\Phi}(\cdot, \mathbf{a})^\top \mathbf{A}_2 \mathbf{\Phi}(\cdot, \mathbf{a}) \right\|_2 \right| \tag{183}$$

$$\leq M \cdot \sup_{\mathbf{x} \in \mathcal{S}, \mathbf{a} \in \mathcal{A}} \left| \left(\mathbf{w}_1 - \mathbf{w}_2 \right)^\top \mathbf{\Phi}(\cdot, \mathbf{a}) \right|_1 + M \cdot \sup_{\mathbf{x} \in \mathcal{S}, \mathbf{a} \in \mathcal{A}} \left\| \mathbf{\Phi}(\cdot, \mathbf{a})^\top \left(\mathbf{A}_1 - \mathbf{A}_2 \right) \mathbf{\Phi}(\cdot, \mathbf{a}) \right\|_2 \tag{184}$$

$$\leq M^{3/2} \cdot \sup_{\mathbf{\Phi} : \|\mathbf{\Phi}\| \leq \sqrt{M}} \left[\left\| \left(\mathbf{w}_1 - \mathbf{w}_2 \right)^\top \mathbf{\Phi} \right\|_2 \right] + M \cdot \sup_{\mathbf{\Phi} : \|\mathbf{\Phi}\| \leq \sqrt{M}} \left\| \mathbf{\Phi}^\top \left(\mathbf{A}_1 - \mathbf{A}_2 \right) \mathbf{\Phi} \right\|_2 \tag{185}$$

$$\leq M^2 \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|_2 + M^2 \|\mathbf{A}_1 - \mathbf{A}_2\|_2 \tag{186}$$

$$\leq M^2 \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|_2 + M^2 \|\mathbf{A}_1 - \mathbf{A}_2\|_F \tag{187}$$

$$\tag{188}$$

Now, let $\mathcal{C}_{\mathbf{w}}$ be an $\varepsilon/(2M^2)$ cover of $\left\{ \mathbf{w} \in \mathbb{R}^d \;\middle|\; \|\mathbf{w}\|_2 \leq L \right\}$ with respect to the Frobenius-norm, and $\mathcal{C}_{\mathbf{A}}$ be an $\varepsilon^2/4$ cover of $\left\{ \mathbf{A} \in \mathbb{R}^{d \times d} | \|\mathbf{A}\|_F \leq (M^2 d)^{1/2} B^2 \lambda^{-1} \right\}$ with respect to the Frobenius norm. By Lemma 30 we have,

$$|\mathcal{C}_{\mathbf{w}}| \leq (1 + 4LM^2/\varepsilon)^d, |\mathcal{C}_{\mathbf{A}}| \leq (1 + 8(M^2 d)^{1/2} B^2/(\lambda \varepsilon^2))^{d^2}. \tag{189}$$

Therefore, we can select, for any $\mathbf{v}_{\boldsymbol{\upsilon}}(\cdot)$, corresponding weight $\mathbf{w} \in \mathcal{C}_{\mathbf{w}}$, and matrix $\mathbf{A} \in \mathcal{C}_{\mathbf{A}}$. Therefore, $\mathcal{N}_\varepsilon \leq |\mathcal{C}_{\mathbf{A}}| \cdot |\mathcal{C}_{\mathbf{w}}|$. This gives us,

$$\log\left(\mathcal{N}_\varepsilon\right) \leq d \cdot \log\left(1 + \frac{4LM^2}{\varepsilon}\right) + d^2 \log\left(1 + \frac{8Md^{1/2}B^2}{\lambda \varepsilon^2}\right). \tag{190}$$

□

G Auxiliary Results

Lemma 33 (Elliptical Potential, Lemma 3 of Abbasi-Yadkori et al. (2011))**.** *Let $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n \in \mathbb{R}^d$ be vectors such that $\|\boldsymbol{x}\|_2 \leq L$. Then, for any positive definite matrix $\boldsymbol{U}_0 \in \mathbb{R}^{d \times d}$, define $\boldsymbol{U}_t := \boldsymbol{U}_{t-1} + \boldsymbol{x}_t \boldsymbol{x}_t^\top$ for all t. Then, for any $\nu > 1$,*

$$\sum_{t=1}^{n} \|\boldsymbol{x}_t\|^2_{\boldsymbol{U}_{t-1}^{-1}} \leq 2d \log_\nu \left(\frac{tr(\boldsymbol{U}_0) + nL^2}{d \det^{1/d}(\boldsymbol{U}_0)} \right).$$

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

Åström, K. J. Optimal control of markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174–205, 1965.

Azuma, K. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

Bazzan, A. L. Opportunities for multiagent systems and multiagent reinforcement learning in traffic control. *Autonomous Agents and Multi-Agent Systems*, 18(3):342, 2009.

Bistritz, I. and Leshem, A. Distributed multi-player bandits-a game of thrones approach. In *Advances in Neural Information Processing Systems*, pp. 7222–7232, 2018.

Boutilier, C. Planning, learning and coordination in multiagent decision processes. Citeseer, 1996.

Bradtke, S. J. and Barto, A. G. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1):33–57, 1996.

Canas, G. D. and Rosasco, L. Learning probability measures with respect to optimal transport metrics. *arXiv preprint arXiv:1209.1077*, 2012.

Chowdhury, S. R. and Gopalan, A. On kernelized multi-armed bandits. *arXiv preprint arXiv:1704.00445*, 2017.

Chowdhury, S. R. and Gopalan, A. No-regret algorithms for multi-task bayesian optimization. *arXiv preprint arXiv:2008.08885*, 2020.

Clemente, A. V., Castejón, H. N., and Chandra, A. Efficient parallel methods for deep reinforcement learning. *arXiv preprint arXiv:1705.04862*, 2017.

Dann, C., Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. On oracle-efficient pac rl with rich observations. *arXiv preprint arXiv:1803.00606*, 2018.

Desai, N., Critch, A., and Russell, S. J. Negotiable reinforcement learning for pareto optimal sequential decision-making. *Advances in Neural Information Processing Systems*, 31:4712–4720, 2018.

Deshmukh, A. A., Dogan, U., and Scott, C. Multi-task learning for contextual bandits. *arXiv preprint arXiv:1705.08618*, 2017.

Ding, G., Koh, J. J., Merckaert, K., Vanderborght, B., Nicotra, M. M., Heckman, C., Roncone, A., and Chen, L. Distributed reinforcement learning for cooperative multi-robot object manipulation. *arXiv preprint arXiv:2003.09540*, 2020.

Dong, K., Peng, J., Wang, Y., and Zhou, Y. Root-n-regret for learning in markov decision processes with function approximation and low bellman rank. In *Conference on Learning Theory*, pp. 1554–1557. PMLR, 2020.

Dubey, A. and Pentland, A. Kernel methods for cooperative multi-agent contextual bandits. In *International Conference on Machine Learning*, pp. 2740–2750. PMLR, 2020a.

Dubey, A. and Pentland, A. Private and byzantine-proof cooperative decision-making. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 357–365, 2020b.

Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International Conference on Machine Learning*, pp. 1407–1416. PMLR, 2018.

Grounds, M. and Kudenko, D. Parallel reinforcement learning with linear function approximation. In *Adaptive Agents and Multi-Agent Systems III. Adaptation and Multi-Agent Learning*, pp. 60–74. Springer, 2005.

Gupta, J. K., Egorov, M., and Kochenderfer, M. Cooperative multi-agent control using deep reinforcement learning. In *International Conference on Autonomous Agents and Multiagent Systems*, pp. 66–83. Springer, 2017.

Hillel, E., Karnin, Z., Koren, T., Lempel, R., and Somekh, O. Distributed exploration in multi-armed bandits. *arXiv preprint arXiv:1311.0800*, 2013.

Horgan, D., Quan, J., Budden, D., Barth-Maron, G., Hessel, M., Van Hasselt, H., and Silver, D. Distributed prioritized experience replay. *arXiv preprint arXiv:1803.00933*, 2018.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? In *Advances in neural information processing systems*, pp. 4863–4873, 2018.

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

Kar, S., Moura, J. M., and Poor, H. V. Qd-learning: A collaborative distributed strategy for multi-agent reinforcement learning through consensus innovations. *IEEE Transactions on Signal Processing*, 61(7):1848–1862, 2013.

Knowles, J. Parego: A hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. *IEEE Transactions on Evolutionary Computation*, 10(1):50–66, 2006.

Krause, A. and Ong, C. S. Contextual gaussian process bandit optimization. In *Advances in neural information processing systems*, pp. 2447–2455, 2011.

Kretchmar, R. M. Parallel reinforcement learning. Citeseer, 2002.

Krishnamurthy, A., Agarwal, A., and Langford, J. Pac reinforcement learning with rich observations. *arXiv preprint arXiv:1602.02722*, 2016.

Landgren, P., Srivastava, V., and Leonard, N. E. On distributed cooperative decision-making in multiarmed bandits. In *2016 European Control Conference (ECC)*, pp. 243–248. IEEE, 2016a.

Landgren, P., Srivastava, V., and Leonard, N. E. Distributed cooperative decision-making in multiarmed bandits: Frequentist and bayesian algorithms. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pp. 167–172. IEEE, 2016b.

Landgren, P., Srivastava, V., and Leonard, N. E. Social imitation in cooperative multiarmed bandits: Partition-based algorithms with strictly local information. In *2018 IEEE Conference on Decision and Control (CDC)*, pp. 5239–5244. IEEE, 2018.

Lauer, M. and Riedmiller, M. An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *In Proceedings of the Seventeenth International Conference on Machine Learning*. Citeseer, 2000.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 2018.

Littman, M. L. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pp. 157–163. Elsevier, 1994.

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

Liu, K. and Zhao, Q. Distributed learning in multi-armed bandit with multiple players. *IEEE Transactions on Signal Processing*, 58(11):5667–5681, 2010a.

Liu, K. and Zhao, Q. Distributed learning in cognitive radio networks: Multi-armed bandit with distributed multiple players. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3010–3013. IEEE, 2010b.

Madiman, M. On the entropy of sums. In *2008 IEEE Information Theory Workshop*, pp. 303–307. IEEE, 2008.

Martínez-Rubio, D., Kanade, V., and Rebeschini, P. Decentralized cooperative stochastic multi-armed bandits. In *Advances in Neural Information Processing Systems*, 2019.

Melo, F. S. and Ribeiro, M. I. Q-learning with linear function approximation. In *International Conference on Computational Learning Theory*, pp. 308–322. Springer, 2007.

Mossalam, H., Assael, Y. M., Roijers, D. M., and Whiteson, S. Multi-objective deep reinforcement learning. *arXiv preprint arXiv:1610.02707*, 2016.

Myerson, R. B. Optimal coordination mechanisms in generalized principal–agent problems. *Journal of mathematical economics*, 10(1):67–81, 1982.

Nair, A., Srinivasan, P., Blackwell, S., Alcicek, C., Fearon, R., De Maria, A., Panneershelvam, V., Suleyman, M., Beattie, C., Petersen, S., et al. Massively parallel methods for deep reinforcement learning. *arXiv preprint arXiv:1507.04296*, 2015.

Paria, B., Kandasamy, K., and Póczos, B. A flexible framework for multi-objective bayesian optimization using random scalarizations. In *Uncertainty in Artificial Intelligence*, pp. 766–776. PMLR, 2020.

Peteiro-Barral, D. and Guijarro-Berdiñas, B. A survey of methods for distributed machine learning. *Progress in Artificial Intelligence*, 2(1):1–11, 2013.

Polydoros, A. S. and Nalpantidis, L. Survey of model-based reinforcement learning: Applications on robotics. *Journal of Intelligent & Robotic Systems*, 86(2):153–173, 2017.

Schaerf, A., Shoham, Y., and Tennenholtz, M. Adaptive load balancing: A study in multi-agent learning. *Journal of artificial intelligence research*, 2:475–500, 1994.

Shapley, L. S. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

Sucar, L. E. Parallel markov decision processes. In *Advances in Probabilistic Graphical Models*, pp. 295–309. Springer, 2007.

Szepesvári, C. and Littman, M. L. A unified analysis of value-function-based reinforcement-learning algorithms. *Neural computation*, 11(8):2017–2060, 1999.

Tan, M. Multi-agent reinforcement learning: Independent vs. cooperative agents. 1993.

Taylor, M. E. and Stone, P. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(7), 2009.

Tuyls, K. and Nowé, A. Evolutionary game theory and multi-agent reinforcement learning. 2005.

Vamplew, P., Yearwood, J., Dazeley, R., and Berry, A. On the limitations of scalarisation for multi-objective reinforcement learning of pareto fronts. In *Australasian joint conference on artificial intelligence*, pp. 372–378. Springer, 2008.

Van Moffaert, K. and Nowé, A. Multi-objective reinforcement learning using sets of pareto dominating policies. *The Journal of Machine Learning Research*, 15(1):3483–3512, 2014.

Wai, H.-T., Yang, Z., Wang, Z., and Hong, M. Multi-agent reinforcement learning via double averaging primal-dual optimization. In *Advances in Neural Information Processing Systems*, pp. 9649–9660, 2018.

Wang, R., Salakhutdinov, R., and Yang, L. F. Provably efficient reinforcement learning with general value function approximation. *arXiv preprint arXiv:2005.10804*, 2020.

Wang, X. F. and Sandholm, T. Learning near-pareto-optimal conventions in polynomial time. 2003.

Wang, Y., Hu, J., Chen, X., and Wang, L. Distributed bandit learning: Near-optimal regret with efficient communication. *arXiv preprint arXiv:1904.06309*, 2019.

Wen, Z. and Van Roy, B. Efficient reinforcement learning in deterministic systems with value function generalization. *Mathematics of Operations Research*, 42(3):762–782, 2017.

Xie, Q., Chen, Y., Wang, Z., and Yang, Z. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on Learning Theory*, pp. 3674–3682. PMLR, 2020.

Yang, E. and Gu, D. Multiagent reinforcement learning for multi-robot systems: A survey. Technical report, tech. rep, 2004.

Yang, L. and Wang, M. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pp. 10746–10756. PMLR, 2020.

Yang, Q., Liu, Y., Chen, T., and Tong, Y. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

Yang, Z., Jin, C., Wang, Z., Wang, M., and Jordan, M. I. Provably efficient reinforcement learning with kernel and neural function approximations. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/9fa04f87c9138de23e92582b4ce549ec-Abstract.html.

Yao, Y. and Doretto, G. Boosting for transfer learning with multiple sources. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 1855–1862. IEEE, 2010.

Yoshikawa, T. Decomposition of dynamic team decision problems. *IEEE Transactions on Automatic Control*, 23(4):627–632, 1978.

Yu, S., Chen, X., Zhou, Z., Gong, X., and Wu, D. When deep reinforcement learning meets federated learning: Intelligent multi-timescale resource management for multi-access edge computing in 5g ultra dense network. *IEEE Internet of Things Journal*, 2020.

Yu, T., Wang, H., Zhou, B., Chan, K., and Tang, J. Multi-agent correlated equilibrium q (λ) learning for coordinated smart generation control of interconnected power grids. *IEEE transactions on power systems*, 30(4):1669–1679, 2014.

Zhang, K., Yang, Z., and Basar, T. Networked multi-agent reinforcement learning in continuous spaces. In *2018 IEEE Conference on Decision and Control (CDC)*, pp. 2771–2776. IEEE, 2018a.

Zhang, K., Yang, Z., Liu, H., Zhang, T., and Basar, T. Fully decentralized multi-agent reinforcement learning with networked agents. In *International Conference on Machine Learning*, pp. 5872–5881. PMLR, 2018b.

Zhang, K., Yang, Z., and Başar, T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv preprint arXiv:1911.10635*, 2019.

Zhang, Q. and Li, H. Moea/d: A multiobjective evolutionary algorithm based on decomposition. *IEEE Transactions on evolutionary computation*, 11(6):712–731, 2007.

Zhang, Q., Liu, W., Tsang, E., and Virginas, B. Expensive multiobjective optimization by moea/d with gaussian process model. *IEEE Transactions on Evolutionary Computation*, 14(3):456–474, 2009.

Zhao, Y., Borovikov, I., Rupert, J., Somers, C., and Beirami, A. On multi-agent learning in team sports games. *arXiv preprint arXiv:1906.10124*, 2019.

Zhuo, H. H., Feng, W., Xu, Q., Yang, Q., and Lin, Y. Federated reinforcement learning. *arXiv preprint arXiv:1901.08277*, 2019.