PROVABLY EFFICIENT FEDERATED REINFORCEMENT LEARNING WITH LINEAR FUNCTION APPROXIMATION AND LOGARITHMIC COMMUNICATION COST

Anonymous authors
Paper under double-blind review

ABSTRACT

We study federated online reinforcement learning with linear function approximation. Although recent multi-agent reinforcement learning algorithms achieve strong regret guarantees, they typically require agents to share raw trajectories. This requirement leads to communication costs that scale linearly with the number of episodes and conflicts with the privacy constraints of federated settings. To address these limitations, we propose FED-LSVI, the first provably efficient federated algorithm for online reinforcement learning with linear function approximation in episodic Markov decision processes. By combining determinant-based event-triggered synchronization with a stepwise backward update mechanism, FED-LSVI enables agents to collaboratively learn an optimal policy while exchanging only compressed sufficient statistics, without sharing raw trajectories. We prove that FED-LSVI achieves a regret bound of $\widetilde{\mathcal{O}}\left(\sqrt{Md^3H^4T}\right)$, where d is the feature dimension, H is the horizon length, M is the number of agents, and T is the number of episodes per agent. This bound matches the best-known regret guarantee for multi-agent online reinforcement learning with linear function approximation. Moreover, FED-LSVI incurs a communication cost that grows only logarithmically with T while adhering to the stringent communication and privacy constraints of federated settings, substantially improving upon prior methods. Numerical experiments further corroborate our theoretical findings and demonstrate the effectiveness of the proposed algorithm.

1 INTRODUCTION

Reinforcement learning (RL) studies sequential decision-making in unknown environments and is typically formulated as learning an optimal policy within a Markov decision process (MDP) (Sutton & Barto, 2018). In modern RL, designing sample-efficient algorithms for problems with large state and action spaces remains a central challenge. A widely adopted approach is function approximation, which enables scalable learning by representing value functions within a restricted function class.

To build a rigorous foundation for this paradigm, recent work has focused on establishing provable regret guarantees for online RL with linear function approximation, where value functions are modeled as linear functions of known features. This literature is largely anchored by two prominent frameworks. The first is the linear MDP setting (Jin et al., 2020; He et al., 2023), which assumes that both transition dynamics and reward functions admit linear representations in terms of known features. The second is the linear mixture MDP setting (Ayoub et al., 2020; Zhou et al., 2021a; Zhou & Gu, 2022), which instead models transitions as a mixture over a set of base kernels.

In many real-world applications, the data collection capacity of a single agent is a primary bottleneck. This motivates cooperative multi-agent reinforcement learning (MARL), where multiple agents collaboratively explore a shared environment and jointly learn an optimal policy. By sharing exploration across agents, cooperative MARL can improve sample efficiency. Theoretical foundations for MARL have recently begun to emerge in the literature. Notably, Dubey & Pentland (2021) proposed the first provably efficient cooperative MARL algorithm with function approximation, extending the single-agent LSVI-UCB framework (Jin et al., 2020) to demonstrate that parallel exploration can be achieved with low communication overhead. Building on this work, Min et al. (2023) studied

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

cooperative linear MDPs with asynchronous communication via a central server and established both regret and communication guarantees. More recently, Hsu et al. (2024) introduced randomized exploration strategies with provable guarantees for broader function classes.

However, these cooperative approaches typically require agents to share raw trajectories for collaborative policy learning. This reliance introduces substantial communication overhead and privacy concerns, particularly in real-world applications where sequential decision-making is distributed across institutions or personal devices. In healthcare, for example, RL is increasingly used to optimize dynamic treatment regimes (Murphy, 2003), where treatment decisions are sequentially adapted to patient histories, such as sepsis management in intensive care units (Komorowski et al., 2018). In such settings, because patient-level trajectories are highly sensitive and often distributed across different hospitals or clinical sites, transmitting raw clinical data to a central learner is often rendered impractical by privacy regulations and institutional constraints. Similar bottlenecks exist in mobile edge computing and recommender systems, where local user interaction histories are generated locally and may contain sensitive behavioral patterns (Kairouz & McMahan, 2021). To circumvent these barriers, it is critical to develop collaborative RL algorithms that do not rely on raw data exchange. Federated learning (FL) (McMahan et al., 2017) provides a natural paradigm to achieve this goal: agents keep raw data locally and only transmit aggregated summary statistics or model updates to a central server. This naturally raises the following question:

Can we design federated algorithms for online RL with linear function approximation that are simultaneously sample-efficient and communication-efficient?

Designing such federated algorithms presents two closely related technical challenges. First, we must identify compressed sufficient statistics that enable the server to perform policy and value-function updates while keeping raw trajectories local. Whereas tabular methods use summaries such as state-action visitation counts (Zheng et al., 2024; 2025a), LSVI-type algorithms require covariance matrices and response vectors that summarize local data for each regression. Second, identifying these statistics is not sufficient: they must remain consistent with the backward update structure of LSVI. While covariance contributions can be aggregated directly, the response vector at step h depends on the updated value-function estimate at step $h + 1$. Consequently, response statistics computed from the same trajectories must be recomputed as the value estimates change. Under a one-shot communication scheme, agents upload response statistics for all steps before the updated next-step value estimates become available. The server, lacking access to local trajectories, must rely on stale response statistics, introducing an additional error term that complicates the regret analysis. Addressing this issue requires a synchronization mechanism that coordinates the construction of response statistics with backward value updates while keeping raw trajectories local.

In this work, we provide an affirmative answer to this important open question. To the best of our knowledge, Fed-LSVI is the first federated algorithm for episodic linear MDPs to achieve sublinear regret and logarithmic communication cost without exchanging raw trajectories. Our results show that, even when raw trajectories remain strictly local, agents can collaboratively achieve statistically efficient exploration without suffering from a communication bottleneck. Our main contributions are summarized as follows:

Algorithm Design. We propose Fed-LSVI, the first provably efficient federated algorithm for online RL with linear function approximation in the literature. Our algorithm relies on two core components: a *determinant-based event-triggered synchronization* mechanism and a *stepwise backward synchronization* protocol. The event-triggered mechanism prevents redundant communication, bounding the total communication rounds logarithmically. Meanwhile, the stepwise backward protocol introduces a dedicated solution to address the stale-response issue in a federated LSVI-type algorithm. Rather than aggregating all response statistics in a single batch, synchronization proceeds backward through the steps. Once the server updates and broadcasts the step-$(h + 1)$ model, agents use this latest global value function to construct the step-h summary statistics, which are then transmitted to the server for the step-h model update. This carefully designed stepwise backward procedure keeps each regression target consistent with the most up-to-date global value estimate, achieving statistically efficient exploration while maintaining local trajectories private.

Regret Guarantees. We establish regret upper bounds for Fed-LSVI in both homogeneous and misspecified settings, matching the best-known regret scaling in online MARL with linear function approximation (Min et al., 2023; Hsu et al., 2024). In the homogeneous setting, Fed-LSVI achieves a

regret bound of $\widetilde{\mathcal{O}}(\sqrt{Md^3H^4T})$ (Theorem 1), confirming that our communication-efficient protocol does not compromise statistical efficiency. The $\sqrt{T}$-type regret behavior is further supported by the numerical experiments detailed in Appendix B.2. Furthermore, we extend our analysis to a misspecified setting with agent-wise heterogeneity (Theorem 2), where each agent interacts with a ξ-approximate linear MDP. We show that Fed-LSVI remains robust to mild agent-wise heterogeneity. In particular, when the misspecification level satisfies $\xi = \widetilde{\mathcal{O}}(\sqrt{d/(MT)})$, our algorithm recovers the same regret bound as in the homogeneous setting. This theoretical robustness is validated by the numerical experiments in Appendix B.3.

Communication Cost. We prove that Fed-LSVI incurs a communication cost, measured by the total number of transmitted scalars, that scales only logarithmically with T in both the homogeneous and misspecified settings. This logarithmic dependence is driven by two properties: Our event-triggered mechanism ensures that the total number of communication rounds is logarithmic in T (Theorem 3), as empirically validated in Appendix B.4, and each round requires transmitting only $\mathcal{O}(Md^2H)$ scalars between the agents and the server. This represents an exponential improvement over prior cooperative RL algorithms (Dubey & Pentland, 2021; Min et al., 2023; Hsu et al., 2024), which require sharing local raw trajectories and then suffer from communication costs that scale linearly with T. Thus, Fed-LSVI significantly reduces communication overhead without transmitting raw trajectories, respecting the data-locality constraints of federated learning.

Technical Novelty. A key technical challenge in our analysis is bounding the sum of quadratic-form bonuses under delayed synchronization; see Lemma E.5. In prior cooperative MARL algorithms, agents update their local policies or covariance estimates continuously after every episode (Dubey & Pentland, 2021; Min et al., 2023; Hsu et al., 2024). In contrast, Fed-LSVI keeps each agent's policy fixed within a communication round and updates the global statistics only when synchronization triggers. As a result, the quadratic-form bonus in the regret analysis is measured with respect to a stale pre-round covariance matrix, rather than a continuously updated one. While standard elliptical-potential arguments suffice when the determinant ratio between the stale and updated matrices is small, difficulty arises when the ratio is large. To resolve this, we introduce an agent-wise pseudo covariance matrix accumulating intra-round samples collected so far from agent m, enabling a post-update elliptical-potential argument for the agent-wise bonus sum. Moreover, our determinant-based synchronization rule (see Eq. 4) controls the determinant ratio between the stale pre-round matrix and this agent-wise matrix before the final episode of each round. This allows us to tightly bound the corresponding sum of quadratic-form bonuses without losing the optimal $\sqrt{T}$ scaling. For more details, see the proof of Lemma E.5 in Appendix E.3.5.

2 PRELIMINARIES

Episodic Markov Decision Processes. In this paper, we consider a finite-horizon episodic Markov decision process $\mathcal{M} = (\mathcal{X}, \mathcal{A}, H, \{\mathbb{P}_h\}_{h=1}^H, \{r_h\}_{h=1}^H)$, where $\mathcal{X}$ is a measurable, possibly infinite state space, $\mathcal{A}$ is a finite action set, H is the horizon length, each deterministic reward function satisfies $r_h : \mathcal{X} \times \mathcal{A} \to [0, 1]$, and $\mathbb{P}_h(\cdot \mid x, a)$ denotes the transition kernel at step $h \in [H]$. Here, for any $C \in \mathbb{N}$, we use $[C]$ to denote the set $\{1, 2, \ldots, C\}$. In each episode, the agent starts from an initial state x_1. At each step $h \in [H]$, it observes a state x_h, chooses an action $a_h \in \mathcal{A}$, receives reward $r_h(x_h, a_h)$, and transitions to $x_{h+1} \sim \mathbb{P}_h(\cdot \mid x_h, a_h)$ for each $h \in [H]$.

Policies, Value Functions, and Bellman Equations. A policy is a collection $\pi = \{\pi_h\}_{h=1}^H$ with $\pi_h : \mathcal{X} \to \Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ denotes the set of probability distributions over $\mathcal{A}$.

To evaluate the performance of different policies, we define the Q-value function and V-value function under a policy π as follows: for any $(x, a) \in \mathcal{X} \times \mathcal{A}$,

$$Q_h^\pi(x, a) \coloneqq \mathbb{E}_\pi \left[\sum_{h'=h}^{H} r_{h'}(x_{h'}, a_{h'}) | x_h = x, a_h = a \right], V_h^\pi(x) = \mathbb{E}_\pi \left[\sum_{h'=h}^{H} r_{h'}(x_{h'}, a_{h'}) | x_h = x \right].$$

For any bounded function $V : \mathcal{X} \to \mathbb{R}$, define

$$[\mathbb{P}_h V](x, a) = \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot|x,a)}[V(x')].$$

Since the action space and the horizon are both finite, there exists an optimal policy π^* that achieves the optimal value $V_h^*(x) = \sup_\pi V_h^\pi(x) = V_h^{\pi^*}(x)$ for all $(x, h) \in \mathcal{X} \times [H]$ (Azar et al., 2017). For

any $(x, a, h) \in \mathcal{X} \times \mathcal{A} \times [H]$, the Bellman equation can be expressed as

$$\begin{cases} V_h^\pi(x) = \mathbb{E}_{a' \sim \pi_h(x)}[Q_h^\pi(x, a')] \\ Q_h^\pi(x, a) := r_h(x, a) + [\mathbb{P}_h V_{h+1}^\pi](x, a) \\ V_{H+1}^\pi(x) \equiv 0, \end{cases} \qquad \begin{cases} V_h^*(x) = \max_{a' \in \mathcal{A}} Q_h^*(x, a') \\ Q_h^*(x, a) := r_h(x, a) + [\mathbb{P}_h V_{h+1}^*](x, a) \\ V_{H+1}^*(x) \equiv 0. \end{cases} \tag{1}$$

Linear Markov Decision Processes. Let $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ be a known feature map satisfying $\|\phi(x, a)\|_2 \leq 1$ for all (x, a). We adopt the standard linear MDP assumption from the literature on linear function approximation (Yang & Wang, 2020; Jin et al., 2020; Wang et al., 2021).

Definition 1 (Linear MDPs)**.** For each step $h \in [H]$, there exist an unknown vector $\theta_h \in \mathbb{R}^d$ and an unknown vector of signed measures $\mu_h = (\mu_h^{(1)}, \ldots, \mu_h^{(d)})$ over $\mathcal{X}$ such that, for every $(x, a) \in \mathcal{X} \times \mathcal{A}$ and every measurable set $\mathcal{B} \subseteq \mathcal{X}$,

$$r_h(x, a) = \langle \phi(x, a), \theta_h \rangle, \qquad \mathbb{P}_h(\mathcal{B} \mid x, a) = \langle \phi(x, a), \mu_h(\mathcal{B}) \rangle.$$

Moreover, $\|\theta_h\|_2, \|\mu_h\|_{\mathrm{TV},2} \leq \sqrt{d}$ for $h \in [H]$, where $\|\mu_h\|_{\mathrm{TV},2} := \|(|\mu_h^{(1)}|(\mathcal{X}), \ldots, |\mu_h^{(d)}|(\mathcal{X}))\|_2$, and $|\mu|$ denotes the total-variation measure of the signed measure.

Federated Reinforcement Learning. We consider an FRL setting with a central server and M agents, each interacting with an independent linear MDP $\mathcal{M}$. We study a synchronous protocol: at common episode index $t \in [T]$, each agent executes its t-th episode, starting from an initial state $x_1^{m,t}$ that may be arbitrary. Thus T is the number of episodes per agent, the total number of agent-episodes is MT, and the total number of interaction steps is MTH. In episode t, agent $m \in [M]$ generates a trajectory $\{(x_h^{m,t}, a_h^{m,t}, r_h^{m,t}, x_{h+1}^{m,t})\}_{h=1}^H$. A central server coordinates occasional synchronizations, while raw trajectories always remain on the agents.

Specifically, we partition the learning process into K rounds with boundaries $1 = \tau_1 < \tau_2 < \cdots < \tau_{K+1} = T + 1$, where round k consists of episodes $\tau_k, \ldots, \tau_{k+1} - 1$. All agents execute the same policy π^k throughout round k. The cumulative regret over T synchronized episodes is defined as

$$\mathrm{Regret}(T) := \sum_{t=1}^{T} \sum_{m=1}^{M} \Big(V_{m,1}^*(x_1^{m,t}) - V_{m,1}^{\pi^{k(t)}}(x_1^{m,t}) \Big),$$

where $V_{m,1}^*$ and $V_{m,1}^\pi$ denote the optimal value function and V-value function of agent m's local MDP and $k(t)$ denotes the round containing episode t. In the homogeneous setting, where all agents share the same MDP, we omit the subscript m and write V_1^* and V_1^π for simplicity.

We also define the communication cost of an FRL algorithm as the number of scalars (integers or real numbers) communicated between the server and agents.

3 Algorithm Design

3.1 The Fed-LSVI Algorithm

We now present Fed-LSVI under the federated reinforcement learning setting described in Section 2. The complete algorithm is given in Algorithm 1 and Algorithm 2. At the beginning of learning, the central server specifies the total number of episodes per agent T, the regularization parameter λ, the trigger parameter $\gamma \geq 1$, and the exploration bonus parameter $\beta > 0$. Fed-LSVI then proceeds over communication rounds indexed by $k \in \{1, 2, \ldots, K\}$.

Global Coordination. At the beginning of round k, each agent holds the current global model $\{(w_h, \Lambda_h^{\mathrm{ser}})\}_{h=1}^H$. In the special case $k = 1$, this model is initialized and broadcast by the central server before data collection begins. For $k \geq 2$, it is the model updated and broadcast by the central server at the end of round $k - 1$. Here, $w_h \in \mathbb{R}^d$ is the regression parameter and $\Lambda_h^{\mathrm{ser}} \in \mathbb{R}^{d \times d}$ is the server-side covariance matrix at step h.

Local Exploration. Using the known feature mapping $\phi(x, a)$ together with the broadcast global model $\{(w_h, \Lambda_h^{\mathrm{ser}})\}_{h=1}^H$, for any $(x, h) \in \mathcal{X} \times [H]$, each agent can reconstruct the optimistic value estimates $Q_h(x, a)$, $V_h(x)$ according to

$$Q_h(x, a) = \Pi_{[0,H]} \left(\phi(x, a)^\top w_h + \beta \, \|\phi(x, a)\|_{(\Lambda_h^{\mathrm{ser}})^{-1}} \right), \tag{2}$$

and

$$V_h(x) = \max_{a \in \mathcal{A}} Q_h(x, a). \tag{3}$$

Here, $\Pi_{[0,H]}(z) := \min\{\max\{z, 0\}, H\}$. This projection ensures $0 \le Q_h(x, a) \le H$ and hence $0 \le V_h(x) \le H$. The policy at (x, h) is then defined by

$$\pi_h(x) = \arg\max_{a \in \mathcal{A}} Q_h(x, a).$$

During round k, all agents execute the policy π^k and repeatedly collect trajectories over multiple episodes. Each agent maintains and updates a local copy of the common episode counter t, together with a current-round episode counter Δt, which records the number of episodes collected since the beginning of round k. Under the synchronous protocol, the local copies of t remain identical across agents. We denote the trajectory collected by agent m in episode t as $\{(x_h^{m,t}, a_h^{m,t}, r_h^{m,t}, x_{h+1}^{m,t})\}_{h=1}^H$.

During data collection, each agent m maintains two types of local information for each step $h \in [H]$: a persistent buffer $\mathcal{D}_h^m$ and a round-wise local covariance matrix $\Lambda_h^{m,\mathrm{loc}}$. The buffer $\mathcal{D}_h^m$ stores all historical transition tuples $(x_h^{m,t}, a_h^{m,t}, r_h^{m,t}, x_{h+1}^{m,t})$ observed at step h throughout the learning process. In contrast, during each round k, the matrix $\Lambda_h^{m,\mathrm{loc}}$ accumulates only the step-h covariance increments from episodes collected in that round, namely $\phi(x_h^{m,t}, a_h^{m,t})\phi(x_h^{m,t}, a_h^{m,t})^\top$ for each such episode t, and is reset to zero at the beginning of the next round.

Determinant-Based Event-Triggered Synchronization. In round k, synchronization is triggered if either (i) there exist an agent $m \in [M]$ and a step $h \in [H]$ such that

$$\frac{\det\left(\Lambda_h^{\mathrm{ser}} + \Lambda_h^{m,\mathrm{loc}}\right)}{\det(\Lambda_h^{\mathrm{ser}})} \ge \frac{\gamma}{\Delta t}, \tag{4}$$

or (ii) the total episode counter reaches the overall episode budget, i.e., $t = T$. In either case, a synchronization signal together with the current episode counter t is sent to the central server, which then broadcasts the signal to all agents and terminates the exploration in round k.

Agent-Server Synchronization and Policy Update. At the end of each round, the central server updates the value-function estimates and the policy through communication with the agents. To update the step-h value-function estimate, similar to LSVI-UCB (Jin et al., 2020), the central server solves a ridge regression problem whose targets are labeled by the next-step value function V_{h+1}:

$$w_h = \arg\min_{w \in \mathbb{R}^d} \sum_{m=1}^{M} \sum_{(x,a,r,x') \in \mathcal{D}_h^m} \left(w^\top \phi(x, a) - r - V_{h+1}(x')\right)^2 + \lambda \|w\|_2^2. \tag{5}$$

The resulting estimator w_h is then used to update the value-function estimates and the policy at step h. Importantly, the value function V_{h+1} used in this regression should be the newly updated one, computed using all historical samples, including those collected in the current round. Therefore, the updates must be carried out backward over the horizon.

This motivates an interleaved, *stepwise backward* synchronization procedure. Specifically, at each step, the agents first send the necessary statistics to the central server. The server then immediately updates the corresponding value-function and policy estimates and broadcasts the updated stepwise model back to the agents before proceeding to the previous step.

Specifically, to solve the above regression problem, each agent m forms the full-history response vector b_h^m from its local buffer $\mathcal{D}_h^m$:

$$b_h^m = \sum_{(x,a,r,x') \in \mathcal{D}_h^m} \phi(x, a)\big(r + V_{h+1}(x')\big), \tag{6}$$

and sends it, together with its local covariance matrix $\Lambda_h^{m,\mathrm{loc}}$, to the central server. The server aggregates the uploaded statistics $\{(\Lambda_h^{m,\mathrm{loc}}, b_h^m)\}_{m=1}^M$ and updates the global covariance matrix Λ_h^{ser} and global response vector b_h according to

$$\Lambda_h^{\mathrm{ser}} \leftarrow \Lambda_h^{\mathrm{ser}} + \sum_{m=1}^{M} \Lambda_h^{m,\mathrm{loc}}, \quad b_h \leftarrow \sum_{m=1}^{M} b_h^m, \tag{7}$$

and then updates the regression coefficient w_h by solving the ridge regression problem in (5), which admits the closed-form solution:

$$w_h = (\Lambda_h^{\text{ser}})^{-1} b_h. \tag{8}$$

The central server broadcasts the updated step-h model $(w_h, \Lambda_h^{\text{ser}})$ to all agents before proceeding to step $h-1$. Using this newly updated step-h model, all agents can construct the same value function V_h with (2) (3), which is then used to form the response vectors b_{h-1}^m by (6) for the model update at step $h-1$. In this way, synchronization and model updating are carried out recursively from step H down to step 1. This backward, stepwise order ensures that all agents label their historical step-h samples using the same freshly updated value function V_{h+1} as in LSVI-UCB.

Algorithm 1 Central server

Input: regularizer $\lambda > 0$, trigger parameter $\gamma \geq 1$, bonus scale $\beta > 0$, hard episode budget T.
Init: for each $h \in [H]$, set $\Lambda_h^{\text{ser}} \leftarrow \lambda I_d$ and $w_h \leftarrow 0$; set the common episode counter $t \leftarrow 0$.
Broadcast parameters $\{w_h, \Lambda_h^{\text{ser}}, \gamma, \beta, T\}$ to all agents.
while $t < T$ **do**
  Wait until receiving a synchronization signal and set t to the received current episode counter.
  Broadcast the synchronization signal to all agents.
  for $h = H, H-1, \ldots, 1$ **do**
    Receive $\{(\Lambda_h^{m,\text{loc}}, b_h^m)\}_{m=1}^M$ from all agents.
    $\Lambda_h^{\text{ser}} \leftarrow \Lambda_h^{\text{ser}} + \sum_{m=1}^M \Lambda_h^{m,\text{loc}}$.
    $b_h \leftarrow \sum_{m=1}^M b_h^m$.
    Update $w_h \leftarrow (\Lambda_h^{\text{ser}})^{-1} b_h$.
    Broadcast $(w_h, \Lambda_h^{\text{ser}})$ to all agents.
  end for
end while

Algorithm 2 Agent m

Init: Set $\mathcal{D}_h^m \leftarrow \emptyset$ for all $h \in [H]$ and $t \leftarrow 0$.
Receive parameters $\{w_h, \Lambda_h^{\text{ser}}, \gamma, \beta, T\}$ from the central server and then initialize Q_h and V_h by (2),(3).
while $t < T$ **do**
  Set the round length counter $\Delta t \leftarrow 0$.
  for $h = 1, 2, \ldots, H$ **do**
    $\Lambda_h^{m,\text{loc}} \leftarrow 0$.
  end for
  while no synchronization signal has been received and $t < T$ **do**
    $t \leftarrow t+1$ and $\Delta t \leftarrow \Delta t + 1$.
    Observe the initial state $x_1^{m,t}$.
    for $h = 1, 2, \ldots, H$ **do**
      Evaluate the latest optimistic model and pick $a_h^{m,t} \leftarrow \arg\max_a Q_h(x_h^{m,t}, a)$.
      Execute $a_h^{m,t}$, observe reward $r_h^{m,t}$ and next state $x_{h+1}^{m,t}$.
      Append $(x_h^{m,t}, a_h^{m,t}, r_h^{m,t}, x_{h+1}^{m,t})$ to $\mathcal{D}_h^m$.
      $\Lambda_h^{m,\text{loc}} \leftarrow \Lambda_h^{m,\text{loc}} + \phi(x_h^{m,t}, a_h^{m,t})\phi(x_h^{m,t}, a_h^{m,t})^\top$.
    end for
    if $t < T$ and $\exists h \in [H]$ such that $\dfrac{\det(\Lambda_h^{\text{ser}} + \Lambda_h^{m,\text{loc}})}{\det(\Lambda_h^{\text{ser}})} \geq \dfrac{\gamma}{\Delta t}$ **then**
      Send synchronization signal and current episode counter t to the server.
    end if
  end while
  if $t = T$ **then**
    Send synchronization signal and current episode counter t to the server.
  end if
  Synchronization (backward, stepwise): set $V_{H+1}(\cdot) \equiv 0$.
  for $h = H, H-1, \ldots, 1$ **do**
    Compute the full-history target b_h^m as (6).
    Send $(\Lambda_h^{m,\text{loc}}, b_h^m)$ to the server.
    Receive the updated step-h model $(w_h, \Lambda_h^{\text{ser}})$ from the server.
    Update Q_h and V_h according to (2),(3).
  end for
end while

3.2 Intuition Behind the Algorithm Design

Sufficient Statistics for Trajectory-Free Communication. At step h, after the next-step value function V_{h+1} has been updated and sent to each agent, the central server updates the step-h model by solving for the linear coefficient w_h in the regression problem (5). By the closed-form solution in (8), this update only requires the aggregated covariance matrix Λ_h^{ser} and response vector b_h:

$$\Lambda_h^{\mathrm{ser}} = \lambda I_d + \sum_{m=1}^{M} \sum_{(x,a,r,x') \in \mathcal{D}_h^m} \phi(x,a)\phi(x,a)^\top, b_h = \sum_{m=1}^{M} \sum_{(x,a,r,x') \in \mathcal{D}_h^m} \phi(x,a)\big(r + V_{h+1}(x')\big).$$

Therefore, each agent m only needs to upload the local statistics

$$\Lambda_h^{m,\mathrm{loc}} = \sum_{(x,a,r,x') \in \mathcal{D}_h^{m,k}} \phi(x,a)\phi(x,a)^\top, \quad b_h^m = \sum_{(x,a,r,x') \in \mathcal{D}_h^m} \phi(x,a)\big(r + V_{h+1}(x')\big),$$

where $\mathcal{D}_h^{m,k}$ collects the transitions at step h of agent m in round k. The server can then aggregate these quantities to reproduce the centralized step-h update exactly, without collecting raw trajectories. Hence, $\Lambda_h^{m,\mathrm{loc}}$ and b_h^m serve as sufficient statistics for the step-h regression. This statistic-based communication substantially reduces the per-round communication burden while preserving privacy.

Stepwise Backward Synchronization and Update Mechanism. The need for a stepwise backward update comes from the recursive structure of LSVI-type algorithms. To update the step-h model, the server must solve the regression (5) whose response vector b_h depends on the latest next-step value function V_{h+1}. Therefore, before the server can update w_h, each agent needs to construct and upload the response vector b_h^m using the newly updated V_{h+1} from the current communication round.

This requirement naturally leads to a stepwise backward synchronization procedure. The server first updates the step-$(h+1)$ model by aggregating the uploaded summary statistics, then immediately broadcasts the updated model back to all agents. After receiving this model, all agents can reconstruct the same fresh value function V_{h+1} and upload the resulting response statistics b_h^m for the step-h update. Repeating this procedure from H down to 1 ensures that every response vector b_h is computed using the most recent globally updated value function.

In contrast, a conventional one-shot synchronization scheme would require agents to upload the response statistics for all steps simultaneously. Since the fresh V_{h+1} is not yet available when these statistics are formed, each agent would have to compute b_h^m using a stale value function from the previous round. This creates an additional stale-response drift term in b_h of the form

$$\sum_{(x,a,r,x') \in \mathcal{D}_h^m} \phi(x,a)\big(V_{h+1}^{\mathrm{stale}}(x') - V_{h+1}^{\mathrm{fresh}}(x')\big),$$

which is not a martingale term and is difficult to control by self-normalized concentration arguments.

Determinant-Based Event-Triggered Synchronization. Similar to prior multi-agent RL works (Dubey & Pentland, 2021; Min et al., 2023; Hsu et al., 2024), Fed-LSVI uses the determinant-based condition in (4) to determine the end of each communication round. The key idea is that, except for the final round, any round of length at most $\gamma/2$ must at least double the server-side covariance determinant at some step h, and hence also the product of these determinants across all steps. Because the feature vectors are bounded, this product grows at most polynomially in MT, which implies that only logarithmically many such rounds can occur. Meanwhile, the number of rounds longer than $\gamma/2$ is at most $2T/\gamma$. Therefore, with $\lambda = 1$ and $\gamma = \max\{T/(Md), 1\}$, Theorem 3 implies that the number of communication rounds is logarithmic in T. Since per-round communication is independent of T, the total communication cost is therefore logarithmic in T.

4 Theoretical Guarantees

4.1 Homogeneous Setting

We first consider the homogeneous setting, where all agents interact with identical copies of the same linear MDP. In this setting, we present the regret and communication guarantees of Fed-LSVI.

Theorem 1. *For Algorithms 1 and 2 with trigger parameter $\gamma \geq 1$ and regularization parameter $\lambda = 1$, there exists a positive constant $c_\beta > 0$ such that, for any confidence level $p \in (0,1)$, if we choose $\beta = c_\beta Hd\sqrt{\log(1 + dMHT/p)}$, then with probability at least $1 - 2p$,*

$$\mathrm{Regret}(T) \;\leq\; \widetilde{\mathcal{O}}\Big(\sqrt{Md^3H^4T} + MH^2d^2\sqrt{\gamma}\Big).$$

The complete expression is provided in (13) *of Appendix E.*

The proof of Theorem 1 is provided in Section E. Choosing $\gamma = \max\{T/(Md), 1\}$, Theorem 1 implies $\mathrm{Regret}(T) \leq \widetilde{\mathcal{O}}(Md^2H^2)$ when $T \leq Md$ and $\mathrm{Regret}(T) \leq \widetilde{\mathcal{O}}(\sqrt{Md^3H^4T})$ when $T \geq Md$. In the latter regime, Fed-LSVI matches the best-known regret scaling for multi-agent RL with linear function approximation (Min et al., 2023; Hsu et al., 2024), while being tailored to the stricter federated regime, where raw trajectories are not shared. To the best of our knowledge, this is the first regret guarantee for online federated reinforcement learning under linear function approximation.

Moreover, for $T \geq Md$, the average regret per agent satisfies $\widetilde{\mathcal{O}}(\sqrt{d^3H^4T/M})$, exhibiting a $\sqrt{M}$-type multi-agent speedup, consistent with worst-case results in federated tabular RL (Zheng et al., 2024; 2025a; Labbi et al., 2025; Zhang et al., 2025b) and in multi-agent RL with linear function approximation (Dubey & Pentland, 2021; Min et al., 2023; Hsu et al., 2024). Numerical experiments in Appendix B.2 further support the $\sqrt{T}$-type regret scaling and the multi-agent speedup.

4.2 Misspecified Setting

We next consider a heterogeneous extension under the misspecified setting, where agents interact with agent-specific MDPs, each of which is ξ-approximate to a common reference linear MDP. We show that the proposed algorithm is robust to such mild agent-wise heterogeneity.

Definition 2 (Misspecified Setting)**.** For any $0 < \xi \leq 1$, and for any agent $m \in [M]$, the corresponding MDP $(\mathcal{X}, \mathcal{A}, H, \mathbb{P}_m, r_m)$ is a ξ-approximate linear MDP with a feature map $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$. Specifically, for each step $h \in [H]$, there exist d unknown signed measures $\mu_h = \big(\mu_h^{(1)}, \ldots, \mu_h^{(d)}\big)$ over $\mathcal{X}$ and an unknown vector $\theta_h \in \mathbb{R}^d$ such that, for all $(x,a) \in \mathcal{X} \times \mathcal{A}$,

$$\big\|\mathbb{P}_{m,h}(\cdot \mid x,a) - \langle \phi(x,a), \mu_h(\cdot)\rangle\big\|_{\mathrm{TV}} \leq \xi, \quad \big|r_{m,h}(x,a) - \langle \phi(x,a), \theta_h\rangle\big| \leq \xi.$$

Here, we assume $\|\phi(x,a)\|_2 \leq 1$ for all $(x,a) \in \mathcal{X} \times \mathcal{A}$, and $\max\big\{\|\mu_h\|_{\mathrm{TV},2}, \|\theta_h\|_2\big\} \leq \sqrt{d}$. Moreover, $\{P_h\}_{h=1}^H$, with $P_h(\cdot \mid x,a) := \langle \phi(x,a), \mu_h(\cdot)\rangle$, forms a valid transition kernel of the common reference linear MDP.

Our misspecified setting Definition 2 is a bounded-variation version of the setting considered in Hsu et al. (2024), which generalizes the single-agent misspecified setting in Jin et al. (2020). Moreover, it also covers the small heterogeneity setting studied in Dubey & Pentland (2021).

We now present the regret guarantee under the misspecified setting described in Definition 2.

Theorem 2. *Under Definition 2, for Algorithms 1 and 2 with trigger parameter $\gamma \geq 1$ and regularization parameter $\lambda = 1$, there exists a constant $c'_\beta > 0$ such that, for any confidence level $\bar{p} \in (0,1)$, if we choose $\beta = c'_\beta Hd\sqrt{\log(1 + 2dMHT/\bar{p})} + 2H\xi\sqrt{dMT}$, then with probability at least $1 - 2\bar{p}$,*

$$\mathrm{Regret}(T) \;\leq\; \widetilde{\mathcal{O}}\Big(\sqrt{Md^3H^4T} + Md^2H^2\sqrt{\gamma} + \xi MdH^2T + \xi\sqrt{M^3d^3H^4\gamma T}\Big).$$

The complete expression is provided in (30) *of Appendix F.*

The proof is deferred to Appendix F. This result establishes a regret guarantee for the misspecified setting that matches the best-known multi-agent RL result in Hsu et al. (2024), and demonstrates that the proposed algorithm is robust to bounded agent-wise heterogeneity. In particular, when $\xi = \widetilde{\mathcal{O}}(\sqrt{d/(MT)})$, the regret bound matches that of the homogeneous setting in Theorem 1. This robustness is further corroborated by numerical experiments in Appendix B.3.

4.3 Communication Cost

We next present a unified result of the communication cost of Fed-LSVI in both the homogeneous and misspecified settings.

Theorem 3. *For Algorithms 1 and 2, with $\lambda = 1$ and a total of T episodes per agent, the total number of communication rounds K satisfies*

$$K \;\le\; 1 + \frac{2T}{\gamma} + \frac{dH}{\log 2} \log\left(1 + \frac{MT}{d}\right).$$

The proof is deferred to Appendix G. If we choose $\gamma = \max\{T/(Md), 1\}$, then Theorem 3 yields

$$K \;\le\; 1 + 2Md + \frac{dH}{\log 2} \log\left(1 + \frac{MT}{d}\right).$$

This logarithmic dependence on T is further supported by the experiments in Appendix B.4. Compared with the communication-round bounds of $\tilde{\mathcal{O}}(dHM)$ in Dubey & Pentland (2021) and Hsu et al. (2024), and $\tilde{\mathcal{O}}(dHM^2)$ in Min et al. (2023) for the asynchronous setting, our result improves the dependence on the number of agents M in the leading logarithmic term. Moreover, in each round, our algorithm communicates only $\mathcal{O}(MHd^2)$ scalars between the agents and the central server. Consequently, the total communication cost of Fed-LSVI is

$$\mathcal{O}\left(M^2 d^3 H + Md^3 H^2 \log\left(1 + \frac{MT}{d}\right)\right).$$

In contrast, existing multi-agent online RL methods (Dubey & Pentland, 2021; Min et al., 2023; Hsu et al., 2024) typically require sharing raw trajectories from local agents, leading to a communication cost of $\mathcal{O}(MHT)$, which scales linearly with T. Therefore, Fed-LSVI reduces the communication cost from linear to logarithmic in T, yielding a significant improvement.

Remark 1. We note an issue in the proof of the communication-round bound in Appendix C of Hsu et al. (2024). They use a trigger condition similar to (4) (see (36) in Appendix H). However, their proof treats all synchronization episodes as if they were triggered by a fixed agent-step pair (m, h), whereas the triggering pair may vary across synchronization episodes. We detail this issue in Appendix I.1. Building on the proof of Theorem 3 in Appendix G, we also provide a corrected proof in Appendix I.2, recovering the $\tilde{O}(dHM)$ communication-round upper bound in Hsu et al. (2024).

Remark 2. We also note that the statistics-only communication claim in Hsu et al. (2024) appears to be overstated. In the linear MDP setting, they suggest that trajectory sharing can be replaced by communicating agent-wise summary statistics. The key issue concerns the response statistic used in the Bellman regression. For example, in our Fed-LSVI update, the corresponding response statistic is b_h in (7), which depends on the next-step value function used to label the regression targets. Since Hsu et al. (2024) perform local value updates for each agent at each episode, a response statistic uploaded by agent n at a synchronization episode is computed using agent n's value function at that episode, whereas a later local update by agent m requires evaluating the same past next states using agent m's current value function. This creates both an agent-wise mismatch and an episode-wise mismatch in the response vector. Consequently, the correct Bellman response vector cannot in general be reconstructed from previously transmitted summary statistics alone without access to the raw next-state observations. See Appendix H for a detailed discussion.

When $M = 1$, Theorems 1 to 3 show that our method reduces to a low-switching variant of LSVI-UCB (Jin et al., 2020), retaining the $\tilde{\mathcal{O}}(\sqrt{d^3H^4T})$ regret with only logarithmic switching cost.

5 Conclusion

We introduced Fed-LSVI, the first provably efficient algorithm for federated online reinforcement learning under linear MDPs. By integrating a determinant-based event-triggered synchronization rule with a stepwise backward update mechanism, Fed-LSVI achieves statistical and communication efficiency without sharing raw local trajectories. Fed-LSVI attains $\sqrt{T}$-type regret guarantees with a multi-agent speedup, while incurring a logarithmic communication cost. We extend our analysis to the misspecified setting, showing that Fed-LSVI remains robust to mild agent-wise heterogeneity. Numerical experiments corroborate our findings, including the $\sqrt{T}$-type regret scaling, the multi-agent speedup in average regret, and the logarithmic growth of communication rounds.

AI USE STATEMENT

We used generative AI tools to assist with implementing and debugging experimental code, mathematical proofreading, and manuscript writing and language editing to improve clarity and readability. Our use of generative AI was limited to these purposes. The authors take full responsibility for the experimental work, the reported results, and the final content of the manuscript. All AI-assisted experimental work and manuscript text were reviewed by the authors.

REPRODUCIBILITY STATEMENT

The problem settings and assumptions are specified in Sections 2 and 4, and Fed-LSVI is described in Algorithms 1 and 2. Proofs of the theoretical results and supporting lemmas are provided in Appendices D–G. Appendix B describes the synthetic MDP construction, experimental settings, parameter choices, evaluation procedures, and computing resources. Experimental results are averaged over five training seeds and reported with one-standard-deviation bands. The code for method implementation is included in the supplementary materials.

ETHICS STATEMENT

This work studies the theoretical foundations of federated reinforcement learning, with experiments conducted exclusively on synthetic MDPs. The reported experiments involve no human subjects or personal data. Deployment in healthcare or other sensitive domains would require additional privacy protections and domain-specific safety validation beyond the scope of this work.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24, 2011.

Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pp. 67–83. PMLR, 2020.

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

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in Neural Information Processing Systems*, 21, 2008.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pp. 463–474. PMLR, 2020.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 263–272. PMLR, 2017. URL https://proceedings.mlr.press/v70/azar17a.html.

Yu Bai, Tengyang Xie, Nan Jiang, and Yu-Xiang Wang. Provably efficient Q-learning with low switching cost. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/473803f0f2ebd77d83ee60daaa61f381-Abstract.html.

Nicolò Dal Fabbro, Aritra Mitra, and George J. Pappas. Federated TD learning over finite-rate erasure channels: Linear speedup under markovian sampling. *IEEE Control Systems Letters*, 7:2461–2466, 2023. URL https://arxiv.org/abs/2305.08104.

Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pp. 1507–1516. PMLR, 2019.

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

Swetha Ganesh, Jiayu Chen, Gugan Thoppe, and Vaneet Aggarwal. Global convergence guarantees for federated policy gradient methods with adversaries. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=Ea0LrPORzM.

Jiafan He, Heyang Zhao, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for linear Markov Decision Processes. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12790–12822. PMLR, 2023. URL https://proceedings.mlr.press/v202/he23d.html.

Hao-Lun Hsu, Weixin Wang, Miroslav Pajic, and Pan Xu. Randomized exploration in cooperative multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, 2024. URL https://neurips.cc/virtual/2024/poster/96449.

Pihe Hu, Yu Chen, and Longbo Huang. Nearly minimax optimal reinforcement learning with linear function approximation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8971–9019. PMLR, 2022. URL https://proceedings.mlr.press/v162/hu22a.html.

Zeyu Jia, Lin Yang, Csaba Szepesvari, and Mengdi Wang. Model-based reinforcement learning with value-targeted regression. In *Learning for Dynamics and Control*, pp. 666–686. PMLR, 2020.

Chi Jin, Zeyuan Allen-Zhu, Sébastien Bubeck, and Michael I. Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/hash/d3b1fb02964aa64e257f9f26a31f72cf-Abstract.html.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Proceedings of the Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2137–2143. PMLR, 2020. URL https://proceedings.mlr.press/v125/jin20a.html.

Peter Kairouz and H Brendan McMahan. Advances and open problems in federated learning. *Foundations and trends in machine learning*, 14(1-2):1–210, 2021.

Sajad Khodadadian, Pranay Sharma, Gauri Joshi, and Siva Theja Maguluri. Federated reinforcement learning: Linear speedup under markovian sampling. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 10997–11057. PMLR, 2022. URL https://proceedings.mlr.press/v162/khodadadian22a.html.

Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

Safwan Labbi, Daniil Tiapkin, Lorenzo Mancini, Paul Mangold, and Eric Moulines. Federated UCBVI: Communication-efficient federated regret minimization with heterogeneous agents. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1315–1323. PMLR, 2025. URL https://proceedings.mlr.press/v258/labbi25a.html.

Guangchen Lan, Dong-Jun Han, Abolfazl Hashemi, Vaneet Aggarwal, and Christopher G. Brinton. Asynchronous federated reinforcement learning with policy gradient updates: Algorithm design and convergence analysis. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=5DUekOKWcS.

Gen Li, Laixi Shi, Yuxin Chen, and Yuejie Chi. Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. *Information and Inference: A Journal of the IMA*, 12 (2):969–1043, 2023.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Pierre Ménard, Omar Darwiche Domingues, Xuedong Shang, and Michal Valko. UCB momentum Q-learning: Correcting the bias without forgetting. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7609–7618. PMLR, 2021. URL https://proceedings.mlr.press/v139/menard21b.html.

Yifei Min, Jiafan He, Tianhao Wang, and Quanquan Gu. Cooperative multi-agent reinforcement learning: Asynchronous communication and linear function approximation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24785–24811. PMLR, 2023. URL https://proceedings.mlr.press/v202/min23a.html.

Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pp. 2010–2020. PMLR, 2020.

Susan A Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(2):331–355, 2003.

Dan Qiao, Ming Yin, Ming Min, and Yu-Xiang Wang. Sample-efficient reinforcement learning with loglog(T) switching cost. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18031–18061. PMLR, 2022. URL https://proceedings.mlr.press/v162/qiao22a.html.

Sudeep Salgia and Yuejie Chi. The sample-communication complexity trade-off in federated Q-learning. In *Advances in Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=6YIpvnkjUK.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018.

Grigoris Velegkas, Zhuoran Yang, and Amin Karbasi. Reinforcement learning with logarithmic regret and policy switches. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/ea318cbc405c9803925e188e5d6836c6-Abstract-Conference.html.

Andrew J Wagenmaker, Yifang Chen, Max Simchowitz, Simon Du, and Kevin Jamieson. First-order regret in reinforcement learning with linear function approximation: A robust estimation approach. In *International Conference on Machine Learning*, pp. 22384–22429. PMLR, 2022.

Han Wang, Aritra Mitra, Hamed Hassani, George J. Pappas, and James Anderson. Federated TD learning with linear function approximation under environmental heterogeneity. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=hdQspgyFrk.

Tianhao Wang, Dongruo Zhou, and Quanquan Gu. Provably efficient reinforcement learning with linear function approximation under adaptivity constraints. In *Advances in Neural Information Processing Systems*, volume 34, pp. 13524–13536. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/70a32110fff0f26d301e58ebbca9cb9f-Paper.pdf.

Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, and Rahul Jain. Learning infinite-horizon average-reward MDPs with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3007–3015. PMLR, 2021.

Jiin Woo, Gauri Joshi, and Yuejie Chi. The blessing of heterogeneity in federated q-learning: Linear speedup and beyond. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 37157–37216. PMLR, 2023. URL https://proceedings.mlr.press/v202/woo23a.html.

Kunhe Yang, Lin Yang, and Simon Du. Q-learning with logarithmic regret. In *International Conference on Artificial Intelligence and Statistics*, pp. 1576–1584. PMLR, 2021.

Lin F. Yang and Mengdi Wang. Sample-optimal parametric Q-learning using linearly additive features. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6995–7004. PMLR, 2019. URL https://proceedings.mlr.press/v97/yang19b.html.

Lin F. Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10746–10756. PMLR, 2020. URL https://proceedings.mlr.press/v119/yang20h.html.

Tong Yang, Shicong Cen, Yuting Wei, Yuxin Chen, and Yuejie Chi. Federated natural policy gradient and actor critic methods for multi-task reinforcement learning. In *Advances in Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=DUFD6vsyF8.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pp. 7304–7312. PMLR, 2019.

Chenyu Zhang, Han Wang, Aritra Mitra, and James Anderson. Finite-time analysis of on-policy heterogeneous federated reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=D2eOVqPX9g.

Haochen Zhang, Zhong Zheng, and Lingzhou Xue. Gap-dependent bounds for federated q-learning. In *Forty-second International Conference on Machine Learning*, 2025a. URL https://openreview.net/forum?id=0n2nXmOxZS.

Haochen Zhang, Zhong Zheng, and Lingzhou Xue. Regret-optimal Q-learning with low cost for single-agent and federated reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=fNOCsycDG4.

Haochen Zhang, Zhong Zheng, and Lingzhou Xue. Gap-dependent bounds for nearly minimax optimal reinforcement learning with linear function approximation. *arXiv preprint arXiv:2602.20297*, 2026a.

Haochen Zhang, Zhong Zheng, and Lingzhou Xue. Q-learning with fine-grained gap-dependent regret. In *International Conference on Learning Representations*, 2026b.

Junkai Zhang, Weitong Zhang, and Quanquan Gu. Optimal horizon-free reward-free exploration for linear mixture MDPs. In *International Conference on Machine Learning*, pp. 41902–41930. PMLR, 2023.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learning via reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33: 15198–15207, 2020.

Zihan Zhang, Xiangyang Ji, and Simon Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In *Conference on Learning Theory*, pp. 4528–4531. PMLR, 2021.

Zihan Zhang, Yuxin Chen, Jason D Lee, and Simon S Du. Settling the sample complexity of online reinforcement learning. In *Conference on Learning Theory*, pp. 5213–5219. PMLR, 2024b.

Zihan Zhang, Jason D Lee, Yuxin Chen, and Simon Shaolei Du. Horizon-free regret for linear Markov Decision Processes. In *The Twelfth International Conference on Learning Representations*, 2024c.

Zhong Zheng, Fengyu Gao, Lingzhou Xue, and Jing Yang. Federated q-learning: Linear regret speedup with low communication cost. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=fe6ANBxcKM.

Zhong Zheng, Haochen Zhang, and Lingzhou Xue. Federated q-learning with reference-advantage decomposition: Almost optimal regret and logarithmic communication cost. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Zhong Zheng, Haochen Zhang, and Lingzhou Xue. Gap-dependent bounds for Q-learning using reference-advantage decomposition. In *International Conference on Learning Representations*, 2025b.

Dongruo Zhou and Quanquan Gu. Computationally efficient horizon-free reinforcement learning for linear mixture MDPs. *Advances in neural information processing systems*, 35:36337–36349, 2022.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvári. Nearly minimax optimal reinforcement learning for linear mixture Markov Decision Processes. In *Proceedings of the Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4532–4576. PMLR, 2021a. URL https://proceedings.mlr.press/v134/zhou21a.html.

Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted MDPs with feature mapping. In *International Conference on Machine Learning*, pp. 12793–12802. PMLR, 2021b.

Runlong Zhou, Zihan Zhang, and Simon Shaolei Du. Sharp variance-dependent bounds in reinforcement learning: Best of both worlds in stochastic and deterministic environments. In *International Conference on Machine Learning*, pp. 42878–42914. PMLR, 2023.

A Related Work

Near-Optimal Reinforcement Learning. In the single-agent setting, the literature most relevant to our work spans near-optimal tabular RL and RL with linear function approximation. In tabular RL, algorithms are typically categorized into model-based and model-free approaches. A large body of work has focused on model-based algorithms (Agarwal et al., 2020; Agrawal & Jia, 2017; Auer et al., 2008; Azar et al., 2017; Dann et al., 2019; Zanette & Brunskill, 2019; Zhang et al., 2024b; 2021; Zhou et al., 2023). Notably, Zhang et al. (2024b) proposed an algorithm achieving a regret bound of $\tilde{O}(\min\{\sqrt{SAH^3T},\, HT\})$, which matches the information-theoretic lower bound. There is also a large body of work on model-free methods that has progressively narrowed the theoretical gap, including Jin et al. (2018); Li et al. (2023); Ménard et al. (2021); Yang et al. (2021); Zhang et al. (2020); Zheng et al. (2025b); Zhang et al. (2025b; 2026b). Several works (Zhang et al., 2020; Ménard et al., 2021; Li et al., 2023; Zhang et al., 2025b) achieved the near-optimal regret bound $\tilde{O}(\sqrt{SAH^3T})$.

The literature on RL with linear function approximation can in turn be divided according to the structural assumptions imposed on the MDP. One line studies linear MDPs (Yang & Wang, 2019; Jin et al., 2020; Wei et al., 2021; Wagenmaker et al., 2022; He et al., 2023; Zhang et al., 2024c; 2026a). Foundational works established sample-efficient learning under generative-model or feature-space assumptions (Yang & Wang, 2019; 2020), and Jin et al. (2020) later introduced LSVI-UCB, the first provably efficient online algorithm without access to a generative model. Subsequent works substantially sharpened the regret guarantees for linear MDPs, culminating in nearly minimax-optimal results (Hu et al., 2022; He et al., 2023). Another line studies linear-mixture MDPs (Jia et al., 2020; Ayoub et al., 2020; Modi et al., 2020; Zhou et al., 2021a;b; Zhou & Gu, 2022). The near-minimax-optimal regret guarantees are also available (Zhou et al., 2021a). Furthermore, Zhou & Gu (2022) proposed a near-optimal, horizon-free algorithm for time-homogeneous linear mixture MDPs. Zhang et al. (2023) provided the near-optimal horizon-free sample complexity in the reward-free time-homogeneous setting.

Low-Switching Reinforcement Learning. A closely related direction studies learning with limited switching or adaptivity, motivated by the cost of frequent policy deployment. In the tabular setting, Bai et al. (2019) established Q-learning with low switching cost, Qiao et al. (2022) characterized near-optimal regret–switching trade-offs up to logarithmic factors, and Velegkas et al. (2022) showed that positive-gap instances permit both logarithmic regret and logarithmically many policy switches. Under linear function approximation, Wang et al. (2021) proved that rare policy updates can still be combined with statistically efficient learning.

Multi-Agent RL (MARL) with Event-Triggered Communications. In the multi-agent setting, the literature most relevant to our work can be divided into two directions: cooperative or parallel exploration with a shared objective, and decentralized equilibrium learning in Markov games under function approximation. The first direction is closest to our setting. Dubey & Pentland (2021) initiated provably efficient cooperative multi-agent reinforcement learning with function approximation and showed that parallel exploration can be combined with limited inter-agent communication. Building on this line, Min et al. (2023) studied cooperative linear MDPs with asynchronous communication through a central server and established regret and communication guarantees, while Hsu et al. (2024) introduced randomized exploration in cooperative multi-agent reinforcement learning and analyzed its performance under approximately linear transitions.

Federated Reinforcement Learning. Recent finite-time analyses of federated reinforcement learning have mainly focused on tabular control, policy evaluation, or policy-gradient methods. In tabular value-based RL, prior works established linear speedup, communication efficiency, and heterogeneity-dependent guarantees for federated Q-learning and related methods (Khodadadian et al., 2022; Woo et al., 2023; Zheng et al., 2024; 2025a; Salgia & Chi, 2024; Labbi et al., 2025; Zhang et al., 2025a;b). A complementary line studies federated policy evaluation or on-policy learning with function approximation, including TD-learning under Markovian sampling and heterogeneous environments (Dal Fabbro et al., 2023; Wang et al., 2024; Zhang et al., 2024a), as well as federated policy-gradient and actor-critic methods (Yang et al., 2024; Ganesh et al., 2024; Lan et al., 2025).

B Numerical Experiments

In this section, we present synthetic simulation[1] results to examine whether the empirical behavior of Fed-LSVI aligns with the theoretical predictions in Section 4, namely an $\widetilde{\mathcal{O}}(\sqrt{T})$ regret scaling, a multi-agent speedup in average regret per agent, and a logarithmic growth in the number of communication rounds.

B.1 Synthetic Linear MDP Environment

We construct a finite-horizon, finite-state linear MDP with S states as follows. Each feature vector $\phi(s,a)$ is sampled from a d-dimensional Dirichlet distribution, each transition basis μ_h^i is sampled from an S-dimensional Dirichlet distribution, and each reward parameter θ_h has i.i.d. entries drawn uniformly from $[0,1]$. The transition kernel and reward function are then induced by the linear structure in Definition 1. Since $\phi(s,a)$ lies on the probability simplex and each μ_h^i is a valid distribution over the state space, $\mathbb{P}_h(\cdot \mid s,a)$ is a proper transition distribution by construction; furthermore, because $\phi(s,a), \theta_h \in [0,1]^d$, the resulting rewards satisfy $r_h(s,a) \in [0,1]$.

In the homogeneous setting, all agents share the same generated linear MDP. In the misspecified setting, we first generate a common reference linear MDP. Then for each agent, we independently generate an agent-specific linear MDP and construct the agent's local reward function and transition kernel as convex combinations of the common and agent-specific components, with weights $1-\xi$ and ξ, respectively.

All runs follow Algorithms 1 and 2. We use the simplified exploration bonus $\beta = Hd$ in both the homogeneous and misspecified experiments. We set the trigger parameter to $\gamma = \max\{T/(dM), 1\}$, the regularization parameter to $\lambda = 1$, and, in the misspecified setting, the heterogeneity level to $\xi = \sqrt{1/T}$. Each agent runs a total of T episodes. Training episodes use a mixed reset rule, starting from state 0 with probability 0.3 and uniformly from the nonzero states with probability 0.7.

B.2 Numerical Results for Regret in the Homogeneous Setting

B.2.1 Regret Scaling in T

To examine whether the regret exhibits the $\widetilde{\mathcal{O}}(\sqrt{T})$ scaling predicted by Theorem 1, we fix the number of agents at $M = 20$ and vary the number of episodes per agent

$$T \in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 3072, 4096, 5120, 6144, 7168, 8192\}.$$

In each run, we record the final cumulative regret for the given T, average over five training seeds, and report the mean together with one-standard-deviation bands.

We use four representative synthetic linear-MDP families,

$$G_1 : (S, A, H, d) = (10, 5, 5, 4), \quad G_2 : (14, 7, 7, 5), \quad G_3 : (16, 8, 8, 6), \quad G_4 : (22, 10, 10, 7).$$

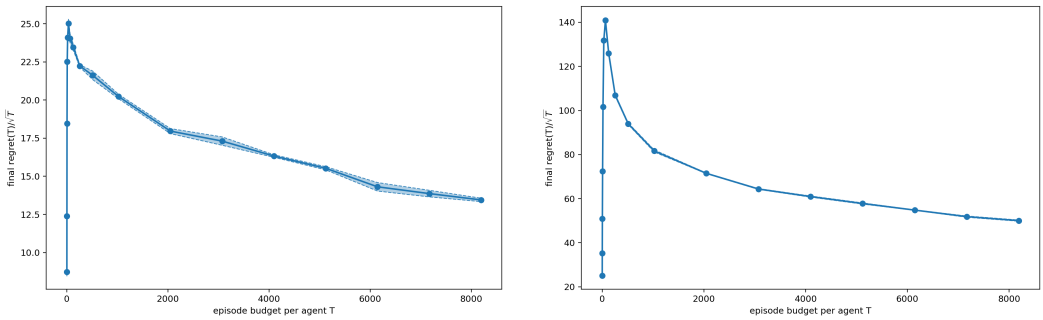
The results are reported in Figure 1, which plots the normalized final regret $\mathrm{Regret}(T)/\sqrt{T}$ as a function of T. The solid lines represent the means over five training seeds, while the dashed lines indicate the boundaries of the one-standard-deviation bands. As T becomes sufficiently large, the curves decrease and then stabilize across all groups, indicating that the regret follows the predicted $\widetilde{\mathcal{O}}(\sqrt{T})$ dependence on T, consistent with Theorem 1.

B.2.2 Speedup in Average Regret

We next examine the collaboration gain with respect to the number of agents in terms of the average regret per agent. In this study, we use the same four synthetic families as in Section B.2.1. We fix $T = 8192$ for G_1, G_2, G_3, and G_4, and vary

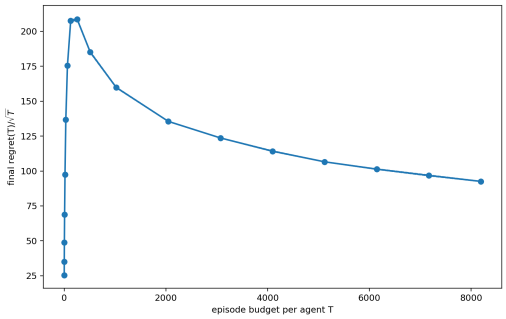
$$M \in \{1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40\}.$$

[1] All experiments were run on a local workstation with an Intel Core i7-14650HX CPU (16 physical cores and 24 logical threads) and 16GB RAM. We used at most 3 concurrent experimental runs and did not explicitly pin the experiments to a fixed number of CPU cores. In our runs, the full set of experiments was completed within approximately 12 hours on this setup.

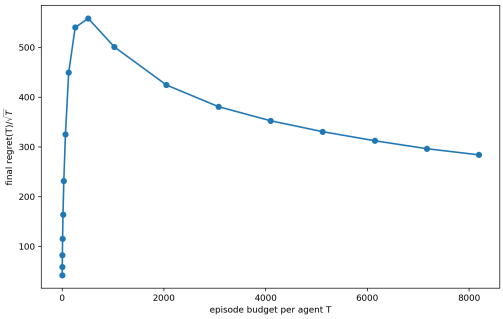


(a) $G_1 : (S, A, H, d) = (10, 5, 5, 4)$.

(b) $G_2 : (S, A, H, d) = (14, 7, 7, 5)$.



(c) $G_3 : (S, A, H, d) = (16, 8, 8, 6)$.



(d) $G_4 : (S, A, H, d) = (22, 10, 10, 7)$.

Figure 1: Regret performance of Fed-LSVI over different T in the homogeneous setting

For each value of M, we compute the average regret per agent $\mathrm{Regret}(T)/M$ for each run, and report its mean over five training seeds together with one-standard-deviation bands.

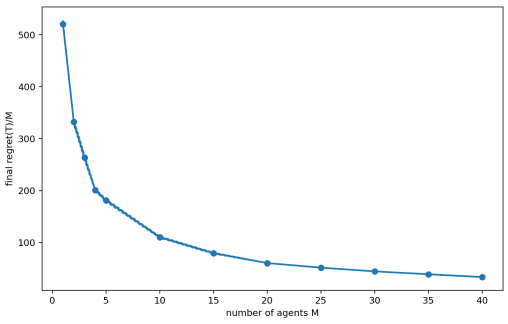
Figure 2 shows the average regret per agent $\mathrm{Regret}(T)/M$ as a function of the number of agents M. The solid lines represent the means over five training seeds, while the dashed lines indicate the boundaries of the one-standard-deviation bands. Across all families, increasing the number of agents does not cause the average regret to diverge; instead, the curves decrease as M increases. This behavior demonstrates a clear speedup pattern in the average regret per agent with respect to the number of agents M.

B.3 Numerical Results for Regret in the Misspecified Setting

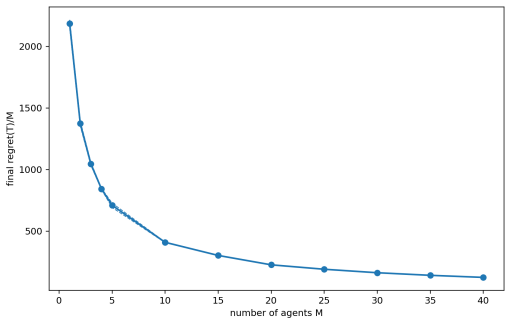
In this subsection, we present regret results of Fed-LSVI under the misspecified setting in Definition 2 to demonstrate its robustness to mild heterogeneity. The experimental setup is identical to that in Section B.2, except for the heterogeneity level ξ, which are specified in Section B.1. For each group of experiments, we report results averaged over five training seeds (solid lines) under both the misspecified setting and the corresponding homogeneous baseline, together with one-standard-deviation bands whose boundaries are indicated by dashed lines.

B.3.1 Regret Scaling in T

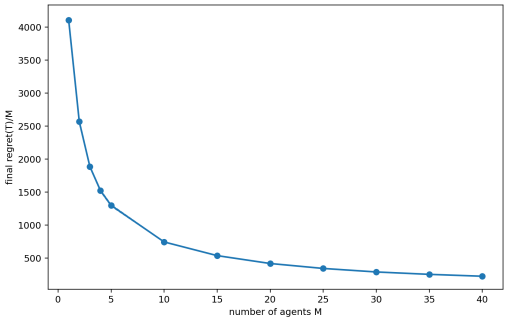
Figure 3 presents the normalized final regret $\mathrm{Regret}(T)/\sqrt{T}$ as a function of T for the same four experimental groups G_1 to G_4 with $M = 20$ as in Appendix B.2.1, evaluated under both the homogeneous and misspecified settings. The results show that, under mild heterogeneity, Fed-LSVI achieves performance comparable to that in the homogeneous setting, demonstrating its robustness.



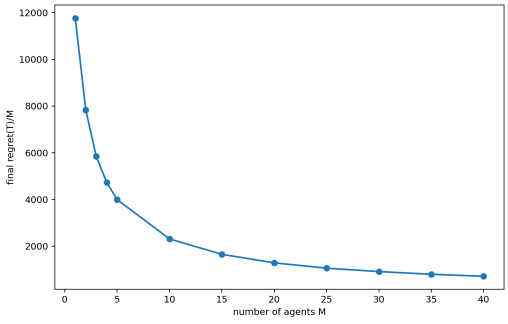
(a) $G_1 : (S, A, H, d) = (10, 5, 5, 4), T = 8192.$



(b) $G_2 : (S, A, H, d) = (14, 7, 7, 5), T = 8192.$

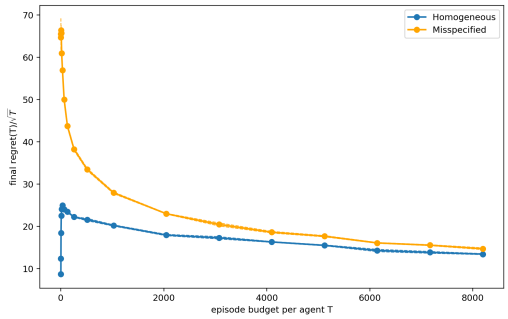


(c) $G_3 : (S, A, H, d) = (16, 8, 8, 6), T = 8192.$

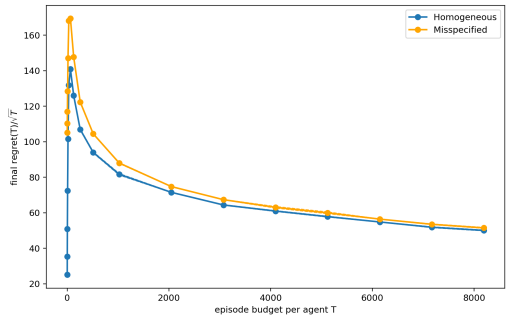


(d) $G_4 : (S, A, H, d) = (22, 10, 10, 7), T = 8192.$

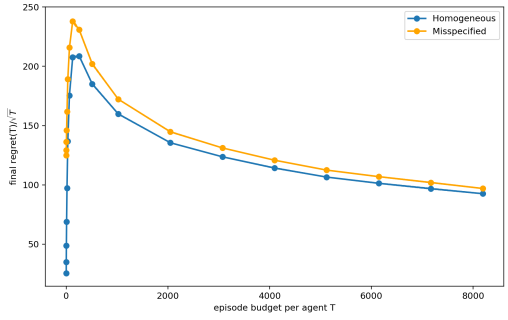
Figure 2: Regret performance of Fed-LSVI over different M in the homogeneous setting



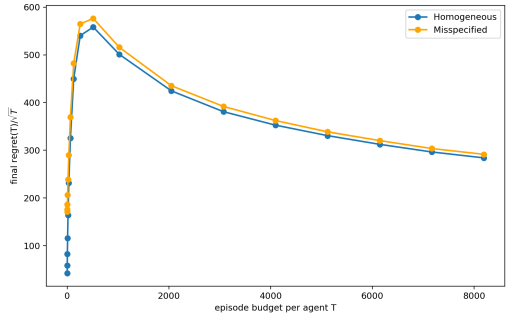
(a) $G_1 : (S, A, H, d) = (10, 5, 5, 4).$



(b) $G_2 : (S, A, H, d) = (14, 7, 7, 5).$



(c) $G_3 : (S, A, H, d) = (16, 8, 8, 6).$

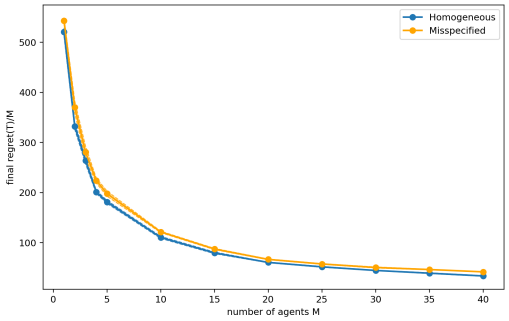


(d) $G_4 : (S, A, H, d) = (22, 10, 10, 7).$

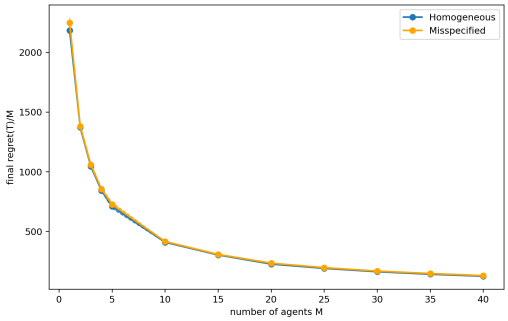
Figure 3: Regret performance of Fed-LSVI over different T in misspecified setting

B.3.2 Speedup in Average Regret

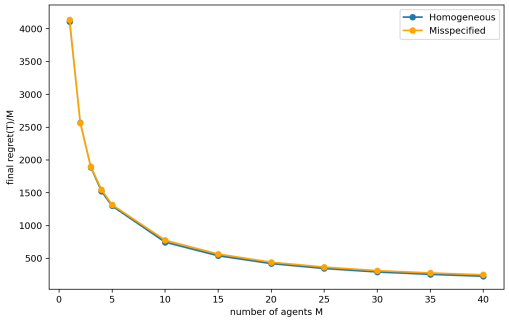
Figure 4 presents the average regret per agent, $\mathrm{Regret}(T)/M$, as a function of M for the same four experimental groups G_1–G_4 with $T = 8192$ as in Appendix B.2.2, evaluated under both the homogeneous and misspecified settings. The results show that, even under mild heterogeneity, Fed-LSVI achieves a clear speedup in the average regret per agent as M increases.



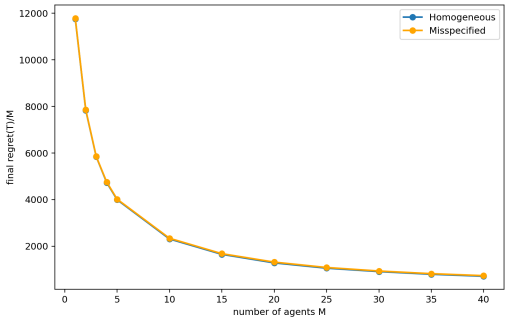
(a) $G_1 : (S, A, H, d) = (10, 5, 5, 4), T = 8192.$



(b) $G_2 : (S, A, H, d) = (14, 7, 7, 5), T = 8192.$



(c) $G_3 : (S, A, H, d) = (16, 8, 8, 6), T = 8192.$



(d) $G_4 : (S, A, H, d) = (22, 10, 10, 7), T = 8192.$

Figure 4: Regret performance of Fed-LSVI over different M in the misspecified setting

B.4 Numerical Results for Communication Complexity

We next study the communication burden under the event-triggered rule in both the homogeneous and misspecified settings. We consider four representative groups:

$$C_1 : (S, A, H, d) = (10, 5, 5, 4), \quad C_2 : (14, 7, 7, 5), \quad C_3 : (16, 8, 8, 6), \quad C_4 : (22, 10, 10, 7).$$

For each group, we fix $M = 20$ and vary the total number of episodes per agent

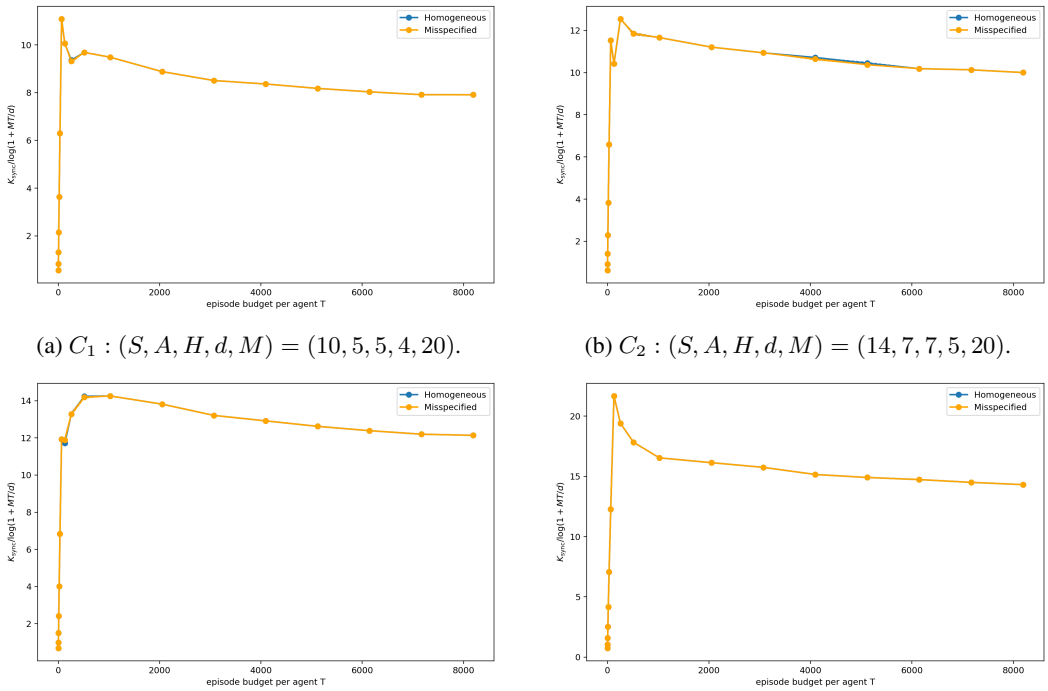
$$T \in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 3072, 4096, 5120, 6144, 7168, 8192\}.$$

Let K denote the number of communication rounds. We plot the average value of

$$\frac{K}{\log(1 + MT/d)}$$

over five training seeds as a function of T together with one-standard-deviation bands, whose boundaries are indicated by dashed lines.

As shown in Figure 5, the results under the misspecified and homogeneous settings are nearly identical, indicating that the communication complexity is uniformly bounded across both settings, consistent with Theorem 3. Across all four groups, the curves exhibit a similar pattern: following an initial burn-in phase, the normalized quantity remains nearly flat over a wide range of episode budgets. This latter behavior further aligns with the logarithmic growth in the number of communication rounds predicted by Theorem 3.



(a) $C_1 : (S, A, H, d, M) = (10, 5, 5, 4, 20)$.

(b) $C_2 : (S, A, H, d, M) = (14, 7, 7, 5, 20)$.

(c) $C_3 : (S, A, H, d, M) = (16, 8, 8, 6, 20)$.

(d) $C_4 : (S, A, H, d, M) = (22, 10, 10, 7, 20)$.

Figure 5: The number of communication rounds of Fed-LSVI

Summary. The empirical results support three key theoretical predictions. First, the regret exhibits the predicted $\widetilde{\mathcal{O}}(\sqrt{T})$ scaling. Second, the average regret per agent decreases as the number of agents M increases, demonstrating a clear multi-agent speedup. Third, the number of communication rounds grows logarithmically with T. Importantly, these behaviors remain essentially unchanged under the misspecified setting, indicating that the proposed method is robust to mild heterogeneity.

C Auxiliary Notation

We collect the notation used throughout the proofs. The index $m \in [M]$ denotes an agent, $k \in [K]$ denotes a communication round, and $t \in [T]$ denotes the *common* episode index shared by all agents. The function $k(t)$ returns the index of the round containing episode t. For a vector $u, v \in \mathbb{R}^d$ and a positive semi-definite matrix $A \in \mathbb{R}^{d\times d}$, we define $\langle u, v\rangle_A = u^\top Av$ and $\|u\|_A := \sqrt{u^\top Au}$.

To define a single filtration over all agents and episodes, we use the linearization

$$\ell(m, t) := m + M(t - 1), \qquad \nu_T(\zeta) := \left\lceil \frac{\zeta}{M} \right\rceil, \qquad \nu_M(\zeta) := ((\zeta - 1) \bmod M) + 1.$$

Then $(m, t) \mapsto \ell(m, t)$ and $\zeta \mapsto (\nu_M(\zeta), \nu_T(\zeta))$ form a pair of inverse mappings. Accordingly, we define the ordering $(m, t) \le (m', t')$ to mean $\ell(m, t) \le \ell(m', t')$.

For each step $h \in [H]$, let Λ_h^k denote the server-side Gram matrix Λ_h^{ser} at the beginning of round k, and let $\Lambda_h^{m,k,\mathrm{loc}}$ denote the information increment collected by agent m during round k. By the design of the algorithm, these quantities satisfy

$$\Lambda_h^k := \lambda I_d + \sum_{m=1}^{M} \sum_{t=1}^{\tau_k - 1} \phi(x_h^{m,t}, a_h^{m,t})\phi(x_h^{m,t}, a_h^{m,t})^\top,$$

$$\Lambda_h^{m,k,\mathrm{loc}} := \sum_{t=\tau_k}^{\tau_{k+1}-1} \phi(x_h^{m,t}, a_h^{m,t})\phi(x_h^{m,t}, a_h^{m,t})^\top.$$

By the definitions above, for any $k \in [K]$, the quantities Λ_h^{k+1}, Λ_h^k, and $\Lambda_h^{m,k,\mathrm{loc}}$ further satisfy

$$\Lambda_h^{k+1} = \Lambda_h^k + \sum_{m=1}^{M} \Lambda_h^{m,k,\mathrm{loc}},$$

where

$$\Lambda_h^{K+1} \coloneqq \lambda I_d + \sum_{m=1}^{M} \sum_{t=1}^{T} \phi(x_h^{m,t}, a_h^{m,t}) \phi(x_h^{m,t}, a_h^{m,t})^\top. \tag{9}$$

In addition, we define the following three auxiliary statistics, which will be used in the subsequent analysis:

$$\bar{\Lambda}_h^{m,t} \coloneqq \Lambda_h^{k(t)} + \sum_{j=\tau_{k(t)}}^{t} \phi(x_h^{m,j}, a_h^{m,j}) \phi(x_h^{m,j}, a_h^{m,j})^\top, \ \forall t \in [T]$$

$$\tilde{\Lambda}_h^{m,t} \coloneqq \lambda I_d + \sum_{(n,j) \le (m,t)} \phi(x_h^{n,j}, a_h^{n,j}) \phi(x_h^{n,j}, a_h^{n,j})^\top, \ \forall t \in [T]$$

$$\check{\Lambda}_h^{t} \coloneqq \lambda I_d + \sum_{m=1}^{M} \sum_{j=1}^{t-1} \phi(x_h^{m,j}, a_h^{m,j}) \phi(x_h^{m,j}, a_h^{m,j})^\top, \ \forall t \in [T+1].$$

The value of w_h at the beginning of round k, denoted by w_h^k, is given by

$$w_h^k = (\Lambda_h^k)^{-1} \sum_{m=1}^{M} \sum_{t=1}^{\tau_k - 1} \phi(x_h^{m,t}, a_h^{m,t}) \big[r_h(x_h^{m,t}, a_h^{m,t}) + V_{h+1}^k(x_{h+1}^{m,t}) \big].$$

In the heterogeneous setting, it can be written as

$$w_h^k = (\Lambda_h^k)^{-1} \sum_{m=1}^{M} \sum_{t=1}^{\tau_k - 1} \phi(x_h^{m,t}, a_h^{m,t}) \big[r_{m,h}(x_h^{m,t}, a_h^{m,t}) + V_{h+1}^k(x_{h+1}^{m,t}) \big].$$

D Auxiliary Lemmas

Lemma D.1 (Lemma 12 in Abbasi-Yadkori et al. (2011))**.** *Let B be a positive definite matrix and C be a positive semi-definite matrix and let $A = B + C$. Then, we have that*

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

Lemma D.2 (Proposition 2.3 and Lemma B.1 in Jin et al. (2020))**.** *Assume the linear MDP parameters satisfy $\|\theta_h\|_2 \le \sqrt{d}$ and $\|\mu_h\|_{\mathrm{TV},2} \le \sqrt{d}$ for all $h \in [H]$. For any policy π, define*

$$w_h^\pi \coloneqq \theta_h + \int_{\mathcal{X}} V_{h+1}^\pi(x') \, \mu_h(\mathrm{d}x').$$

Then $Q_h^\pi(x,a) = \langle \phi(x,a), w_h^\pi \rangle$ for any $(x,a,h) \in \mathcal{X} \times \mathcal{A} \times [H]$ and $\|w_h^\pi\|_2 \le 2H\sqrt{d}$ for any $h \in [H]$.

Lemma D.3 (Norm control for the step-wise regression vector)**.** *For every step $h \in [H]$ and round $k \in [K]$, it holds that*

$$\|w_h^k\|_2 \le 2H \sqrt{\frac{dM\tau_k}{\lambda}}.$$

Proof of Lemma D.3. For any vector $v \in \mathbb{R}^d$, since

$$w_h^k = (\Lambda_h^k)^{-1} \sum_{m=1}^{M} \sum_{j=1}^{\tau_k - 1} \phi(x_h^{m,j}, a_h^{m,j}) \big[r_h(x_h^{m,j}, a_h^{m,j}) + V_{h+1}^k(x_{h+1}^{m,j}) \big],$$

we have

$$\left|\langle v, w_h^k\rangle\right| = \left|\left\langle v, \sum_{m=1}^{M}\sum_{j=1}^{\tau_k-1} \phi(x_h^{m,j}, a_h^{m,j})\left[r_h(x_h^{m,j}, a_h^{m,j}) + V_{h+1}^k(x_{h+1}^{m,j})\right]\right\rangle_{(\Lambda_h^k)^{-1}}\right|.$$

Since rewards are bounded in $[0,1]$ and $0 \le V_{h+1}^k \le H$, each regression target satisfies

$$\left|r_h(x_h^{m,j}, a_h^{m,j}) + V_{h+1}^k(x_{h+1}^{m,j})\right| \le 2H.$$

Applying this bound gives

$$\begin{aligned}
\left|\langle v, w_h^k\rangle\right| &\le 2H \sum_{m=1}^{M}\sum_{j=1}^{\tau_k-1} \left|\langle v, \phi(x_h^{m,j}, a_h^{m,j})\rangle_{(\Lambda_h^k)^{-1}}\right| \\
&\le 2H \sum_{m=1}^{M}\sum_{j=1}^{\tau_k-1} \sqrt{\langle v, v\rangle_{(\Lambda_h^k)^{-1}} \langle \phi(x_h^{m,j}, a_h^{m,j}), \phi(x_h^{m,j}, a_h^{m,j})\rangle_{(\Lambda_h^k)^{-1}}} \\
&\le 2H \left\|v\right\|_2 \lambda^{-1/2} \sqrt{M\tau_k \sum_{m=1}^{M}\sum_{j=1}^{\tau_k-1} \phi(x_h^{m,j}, a_h^{m,j})^\top (\Lambda_h^k)^{-1} \phi(x_h^{m,j}, a_h^{m,j})}. \qquad (10)
\end{aligned}$$

The second inequality follows from Cauchy–Schwarz inequality. The last inequality follows from $\langle v, v\rangle_{(\Lambda_h^k)^{-1}} \le \|v\|_2^2/\lambda$ since $\Lambda_h^k \succeq \lambda I_d$. Since Λ_h^k is a positive-definite matrix, we further note that,

$$\begin{aligned}
&\sum_{m=1}^{M}\sum_{j=1}^{\tau_k-1} \phi(x_h^{m,j}, a_h^{m,j})^\top (\Lambda_h^k)^{-1} \phi(x_h^{m,j}, a_h^{m,j}) \\
&= \mathrm{tr}\Big((\Lambda_h^k)^{-1} \sum_{m=1}^{M}\sum_{j=1}^{\tau_k-1} \phi(x_h^{m,j}, a_h^{m,j}) \phi(x_h^{m,j}, a_h^{m,j})^\top\Big) \\
&= \mathrm{tr}\big((\Lambda_h^k)^{-1}(\Lambda_h^k - \lambda I_d)\big) \\
&= \mathrm{tr}\big(I_d - \lambda(\Lambda_h^k)^{-1}\big) \le d. \qquad (11)
\end{aligned}$$

Combining (10) and (11) yields

$$\left|\langle v, w_h^k\rangle\right| \le 2H \left\|v\right\|_2 \sqrt{\frac{dM\tau_k}{\lambda}}.$$

Taking the supremum over all $\|v\|_2 = 1$ concludes the proof. □

Lemma D.4 (Self-Normalized Concentration Inequality, Theorem 1 in Abbasi-Yadkori et al. (2011))**.** *Let $\{\eta_t\}_{t\ge1}$ be a real-valued martingale difference sequence adapted to a filtration $\{\mathcal{F}_t\}$, such that η_t is conditionally R-sub-Gaussian, i.e.,*

$$\mathbb{E}[\exp(\lambda\eta_t) \mid \mathcal{F}_{t-1}] \le \exp\left(\frac{\lambda^2 R^2}{2}\right) \quad \textit{for all } \lambda \in \mathbb{R}.$$

Let $\{x_t\}_{t\ge1}$ be an $\mathbb{R}^d$-valued predictable process (i.e., x_t is $\mathcal{F}_{t-1}$-measurable), and define

$$V_t = \lambda I + \sum_{s=1}^{t} x_s x_s^\top, \quad S_t = \sum_{s=1}^{t} x_s \eta_s,$$

for some $\lambda > 0$.

Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, for all $t \ge 0$ simultaneously,

$$\|S_t\|_{V_t^{-1}}^2 \;\le\; 2R^2 \log\left(\frac{\det(V_t)^{1/2}}{\det(\lambda I)^{1/2}\,\delta}\right).$$

Lemma D.5 (Self-normalized concentration for a fixed target function)**.** *Fix a step $h \in [H]$ and a function $V : \mathcal{X} \to [0, H]$. Then, for any confidence level $p \in (0, 1)$, with probability at least $1 - p$, simultaneously for all $t \in \{1, \dots, T + 1\}$,*

$$\left\| \sum_{m=1}^{M} \sum_{j=1}^{t-1} \phi(x_h^{m,j}, a_h^{m,j}) \Big[V(x_{h+1}^{m,j}) - \mathbb{P}_h V(x_h^{m,j}, a_h^{m,j}) \Big] \right\|_{(\check{\Lambda}_h^t)^{-1}} \leq 2H \sqrt{\frac{d}{2} \log\Big(1 + \frac{MT}{d\lambda}\Big) + 2 \log\Big(\frac{1}{p}\Big)}.$$

Proof of Lemma D.5. Let

$$\mathcal{G}_{m,t,h} := \sigma\left(\Big\{ (x_\ell^{n,j}, a_\ell^{n,j}, x_{\ell+1}^{n,j}) \Big\}_{(n,j)<(m,t),\, \ell \in [H]} \cup \Big\{ (x_\ell^{m,t}, a_\ell^{m,t}) \Big\}_{\ell \leq h} \right),$$

$\eta_{m,j} \coloneqq V(x_{h+1}^{m,j}) - \mathbb{P}_h V(x_h^{m,j}, a_h^{m,j})$ is a bounded martingale difference with respect to $\mathcal{G}_{m,j,h}$. Applying Lemma D.4 gives that with probability at least $1 - p$, simultaneously for all t,

$$\left\| \sum_{m=1}^{M} \sum_{j=1}^{t-1} \phi(x_h^{m,j}, a_h^{m,j}) \eta_{m,j} \right\|_{(\check{\Lambda}_h^t)^{-1}}^2 \leq 2H^2 \log\left(\frac{\det(\check{\Lambda}_h^{T+1})^{1/2}}{p \, \det(\check{\Lambda}_h^1)^{1/2}} \right) \leq 4H^2 \log\left(\frac{\det(\check{\Lambda}_h^{T+1})^{1/2}}{p^2 \det(\check{\Lambda}_h^1)^{1/2}} \right).$$

Finally, we upper bound the log-determinant term by using $\det(\lambda I + \sum_{i=1}^{MT} \phi_i \phi_i^\top) \leq \lambda^d \big(1 + MT/(\lambda d)\big)^d$. This yields

$$\log\left(\frac{\det(\check{\Lambda}_h^{T+1})^{1/2}}{p^2 \det(\check{\Lambda}_h^1)^{1/2}} \right) \leq \frac{d}{2} \log\Big(1 + \frac{MT}{d\lambda}\Big) + 2 \log\Big(\frac{1}{p}\Big),$$

which concludes the proof. □

Lemma D.6 (Post-update elliptical potential lemma)**.** *Let $V_0 \in \mathbb{R}^{d \times d}$ be symmetric positive definite, and let $x_1, \dots, x_N \in \mathbb{R}^d$. Define*

$$V_t = V_0 + \sum_{s=1}^{t} x_s x_s^\top, \qquad t = 1, \dots, N.$$

Then

$$\sum_{t=1}^{N} \|x_t\|_{V_t^{-1}}^2 \leq \log \frac{\det(V_N)}{\det(V_0)}.$$

In particular, if $V_0 = \lambda I_d$ for some $\lambda > 0$ and $\|x_t\|_2 \leq 1$ for all t, then

$$\sum_{t=1}^{N} \|x_t\|_{V_t^{-1}}^2 \leq \log \frac{\det(V_N)}{\det(\lambda I_d)} \leq d \log\left(1 + \frac{N}{d\lambda}\right).$$

Proof. For each t, let $u_t = x_t^\top V_{t-1}^{-1} x_t \geq 0$. Since $V_t = V_{t-1} + x_t x_t^\top$, the Sherman–Morrison formula gives

$$x_t^\top V_t^{-1} x_t = u_t - \frac{u_t^2}{1 + u_t} = \frac{u_t}{1 + u_t}.$$

For any $u \geq 0$,

$$\log(1 + u) = \int_0^u \frac{1}{1 + s} \, ds \geq \frac{u}{1 + u}.$$

Together with the matrix determinant lemma, this yields

$$\|x_t\|_{V_t^{-1}}^2 \leq \log(1 + u_t) = \log \frac{\det(V_t)}{\det(V_{t-1})}.$$

Summing over $t = 1, \dots, N$ and telescoping proves the first claim.

For the special case $V_0 = \lambda I_d$ and $\|x_t\|_2 \le 1$,

$$\operatorname{tr}(V_N) = d\lambda + \sum_{t=1}^{N} \|x_t\|_2^2 \le d\lambda + N.$$

Applying the arithmetic–geometric mean inequality to the eigenvalues of V_N, we obtain

$$\det(V_N) \le \left(\frac{\operatorname{tr}(V_N)}{d}\right)^d \le \left(\lambda + \frac{N}{d}\right)^d.$$

Since $\det(\lambda I_d) = \lambda^d$, it follows that

$$\log \frac{\det(V_N)}{\det(\lambda I_d)} \le d \log\left(1 + \frac{N}{d\lambda}\right),$$

which completes the proof. □

Lemma D.7 (Lemma D.6 in Jin et al. (2020))**.** *Let $\mathcal{V}$ denote a class of functions of the form*

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

where the parameters satisfy $\|w\|_2 \le L$, $\beta \in [0, B]$, and $\lambda_{\min}(\Lambda) \ge \lambda$. Assume $\|\phi(x, a)\|_2 \le 1$ for all (x, a). Let $\mathcal{N}_\epsilon$ be the ϵ-covering number of $\mathcal{V}$ under $\operatorname{dist}(V, V') = \sup_x |V(x) - V'(x)|$. Then

$$\log \mathcal{N}_\epsilon \le d \log\Big(1 + \frac{4L}{\epsilon}\Big) + d^2 \log\Big(1 + \frac{8\sqrt{d}\, B^2}{\lambda\epsilon^2}\Big).$$

Lemma D.8 (Covering-number bound for optimistic value functions)**.** *Let $\mathcal{V}$ denote a class of functions of the form*

$$V(\cdot) = \Pi_{[0,H]}\Big(\max_{a\in\mathcal{A}}\big[w^\top \phi(\cdot, a) + \beta\sqrt{\phi(\cdot, a)^\top \Lambda^{-1}\phi(\cdot, a)}\big]\Big),$$

where the parameters satisfy $\|w\|_2 \le L$, $\beta \in [0, B]$, and $\lambda_{\min}(\Lambda) \ge \lambda$. Assume $\|\phi(x, a)\|_2 \le 1$ for all (x, a). Let $\mathcal{N}_\epsilon$ be the ϵ-covering number of $\mathcal{V}$ under $\operatorname{dist}(V, V') = \sup_x |V(x) - V'(x)|$. Then

$$\log \mathcal{N}_\epsilon \le d \log\Big(1 + \frac{4L}{\epsilon}\Big) + d^2 \log\Big(1 + \frac{8\sqrt{d}\, B^2}{\lambda\epsilon^2}\Big).$$

Proof. Let $\bar{\mathcal{V}}$ denote the value-function class before the additional lower clipping, namely

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

After adding the projection onto $[0, H]$, the new value-function class can be written as

$$\mathcal{V} = \left\{V(\cdot) = [\bar{V}(\cdot)]_+ : \bar{V} \in \bar{\mathcal{V}}\right\},$$

where $[z]_+ = \max\{z, 0\}$. Since the map $z \mapsto [z]_+$ is 1-Lipschitz, for any $\bar{V}, \bar{V}' \in \bar{\mathcal{V}}$,

$$\|[\bar{V}]_+ - [\bar{V}']_+\|_\infty \le \|\bar{V} - \bar{V}'\|_\infty.$$

Therefore, if $\{\bar{V}_i\}_{i=1}^{N_\epsilon}$ is an ϵ-net of $\bar{\mathcal{V}}$ under the sup norm, then $\{[\bar{V}_i]_+\}_{i=1}^{N_\epsilon}$ is an ϵ-net of $\mathcal{V}$. Hence

$$N_\epsilon(\mathcal{V}) \le N_\epsilon(\bar{\mathcal{V}}).$$

Applying Lemma D.7 gives the same bound for the projected class $\mathcal{V}$. □

Lemma D.9 (Uniform concentration for the data-dependent target V_{h+1}^k)**.** *Fix any confidence level $p \in (0, 1)$ and $\lambda = 1$, and set $\beta = c_\beta H d\sqrt{\log(1 + dMHT/p)}$, where c_β is a sufficiently large positive constant. Let $\mathcal{C}$ be the event that simultaneously for all rounds $k \in [K]$ and steps $h \in [H]$,*

$$\left\|\sum_{m=1}^{M}\sum_{j=1}^{\tau_k - 1} \phi(x_h^{m,j}, a_h^{m,j})\Big[V_{h+1}^k(x_{h+1}^{m,j}) - \mathbb{P}_h V_{h+1}^k(x_h^{m,j}, a_h^{m,j})\Big]\right\|^2_{(\check{\Lambda}_h^{\tau_k})^{-1}} \le CH^2 d^2 \chi.$$

where $\chi = \log(1 + (c_\beta + 1)dMHT/p)$. Then, $\mathbb{P}(\mathcal{C}) \ge 1 - p$.

Proof of Lemma D.9. The challenge is that V_{h+1}^k is data dependent. We therefore discretize the relevant function class and then transfer the bound from the net to the whole class.

By construction, every V_{h+1}^k belongs to the class $\mathcal{V}$ in Lemma D.8. Moreover, Lemma D.3 implies that the regression parameter can be taken in a ball of radius $L = 2H\sqrt{dMT/\lambda}$, because $\tau_k \le T$ for all k.

Let $\{V^{(1)}, \dots, V^{(\mathcal{N}_\epsilon)}\}$ be an ϵ-net of $\mathcal{V}$ under the sup norm. Applying Lemma D.5 to each $V^{(i)}$ and taking a union bound over i and h shows that, with probability at least $1-p$, for every $h \in [H]$, every $k \in [K]$, and every net point $V^{(i)}$,

$$\|S_{h,k}^{(i)}\|^2_{(\check{\Lambda}_h^{\tau_k})^{-1}} \le 2H^2 d \log\Big(1 + \frac{MT}{d\lambda}\Big) + 8H^2 \log\Big(\frac{H\mathcal{N}_\epsilon}{p}\Big).$$

Here,

$$S_{h,k}^{(i)} := \sum_{m=1}^{M} \sum_{j=1}^{\tau_k - 1} \phi(x_h^{m,j}, a_h^{m,j}) \Big[V^{(i)}(x_{h+1}^{m,j}) - \mathbb{P}_h V^{(i)}(x_h^{m,j}, a_h^{m,j})\Big].$$

Fix any $V \in \mathcal{V}$ and choose $V^{(i)}$ in the net such that $\|V - V^{(i)}\|_\infty \le \epsilon$. For every sample,

$$\Big|\big(V - V^{(i)}\big)(x_{h+1}^{m,j}) - \mathbb{P}_h\big(V - V^{(i)}\big)(x_h^{m,j}, a_h^{m,j})\Big| \le 2\epsilon.$$

Therefore,

$$\Big\| \sum_{m=1}^{M} \sum_{j=1}^{\tau_k - 1} \phi(x_h^{m,j}, a_h^{m,j}) \Big[(V - V^{(i)})(x_{h+1}^{m,j}) - \mathbb{P}_h(V - V^{(i)})(x_h^{m,j}, a_h^{m,j})\Big] \Big\|^2_{(\check{\Lambda}_h^{\tau_k})^{-1}} \le \frac{4M^2T^2\epsilon^2}{\lambda},$$

where we used $\check{\Lambda}_h^{\tau_k} \succeq \lambda I_d$ and $\|\phi(x_h^{m,j}, a_h^{m,j})\|_2 \le 1$. Combining the net bound with this approximation error, we have

$$\Big\| \sum_{m=1}^{M} \sum_{j=1}^{\tau_k - 1} \phi(x_h^{m,j}, a_h^{m,j}) \Big[V_{h+1}^k(x_{h+1}^{m,j}) - \mathbb{P}_h V_{h+1}^k(x_h^{m,j}, a_h^{m,j})\Big] \Big\|^2_{(\check{\Lambda}_h^{\tau_k})^{-1}} \le \Gamma_p.$$

where

$$\Gamma_p := 8H^2\Big(\frac{d}{2}\log\Big(1 + \frac{MT}{d\lambda}\Big) + 2\log\Big(\frac{H}{p}\Big) + 2\log\mathcal{N}_\epsilon\Big) + \frac{8M^2T^2\epsilon^2}{\lambda},$$

and $\mathcal{N}_\epsilon$ is the covering number of the value-function class from Lemma D.8. Let $\lambda = 1$, $\epsilon = Hd/MT$, $\beta = c_\beta Hd\sqrt{\log(1 + dMHT/p)}$. We obtain

$$\begin{aligned}\Gamma_p \le 8H^2 \Bigg(&\frac{d}{2}\log\Big(1 + \frac{MT}{d}\Big) + 2\log\Big(\frac{H}{p}\Big) + 2d\log\Big(1 + \frac{8(MT)^{3/2}}{\sqrt{d}}\Big) \\ &+ 2d^2\log\Big(1 + 8c_\beta^2\sqrt{d}M^2T^2\log(1 + dMHT/p)\Big)\Bigg) + 8H^2d^2.\end{aligned} \tag{12}$$

Thus, we have:

$$\Gamma_p \le CH^2d^2\chi,$$

where $\chi = \log(1 + (c_\beta + 1)dMHT/p)$ and C is an absolute constant independent of c_β. □

E Proof of Theorem 1

E.1 Supporting Lemmas

For the trajectory generated by agent m in episode t, define

$$\begin{aligned}\delta_h^{m,t} &:= V_h^{k(t)}(x_h^{m,t}) - V_h^{\pi^{k(t)}}(x_h^{m,t}),\\ \mathcal{G}_{m,t,h} &:= \sigma\Bigg(\Big\{(x_\ell^{n,j}, a_\ell^{n,j}, x_{\ell+1}^{n,j})\Big\}_{(n,j)<(m,t),\, \ell\in[H]} \cup \big\{(x_\ell^{m,t}, a_\ell^{m,t})\big\}_{\ell \le h}\Bigg),\\ \zeta_{h+1}^{m,t} &:= \mathbb{E}\big[\delta_{h+1}^{m,t} \mid \mathcal{G}_{m,t,h}\big] - \delta_{h+1}^{m,t}.\end{aligned}$$

Equivalently,

$$\zeta_{h+1}^{m,t} = \mathbb{P}_h\big(V_{h+1}^{k(t)} - V_{h+1}^{\pi^{k(t)}}\big)(x_h^{m,t}, a_h^{m,t}) - \big(V_{h+1}^{k(t)} - V_{h+1}^{\pi^{k(t)}}\big)(x_{h+1}^{m,t}).$$

Lemma E.1 (Uniform confidence decomposition for the regression error)**.** *Set $\lambda = 1$. There exists an absolute constant $c_\beta > 0$ such that for any confidence level $p \in (0,1)$, if we take $\beta = c_\beta H d\sqrt{\log(1 + dMHT/p)}$, then conditioned on the event $\mathcal{C}$ defined in Lemma D.9, the following statement holds simultaneously for all rounds $k \in [K]$, steps $h \in [H]$, policies π, and state-action pairs $(x,a) \in \mathcal{X} \times \mathcal{A}$.*

$$\big|\phi(x,a)^\top w_h^k - Q_h^\pi(x,a) - \mathbb{P}_h\big(V_{h+1}^k - V_{h+1}^\pi\big)(x,a)\big| \le \beta \,\|\phi(x,a)\|_{(\Lambda_h^k)^{-1}}.$$

Lemma E.2. *Conditioned on the event $\mathcal{C}$ defined in Lemma D.9, for every round $k \in [K]$, step $h \in [H]$ and state-action pair $(x,a) \in \mathcal{X}\times\mathcal{A}$, the optimistic estimates upper bound the corresponding value functions:*

$$Q_h^k(x,a) \ge Q_h^*(x,a), \quad V_h^k(x) \ge V_h^*(x).$$

Lemma E.3. *Conditioned on the event $\mathcal{C}$ defined in Lemma D.9, for every agent $m \in [M]$, episode $t \in [T]$, and step $h \in [H]$,*

$$\delta_h^{m,t} \le \delta_{h+1}^{m,t} + \zeta_{h+1}^{m,t} + 2\beta \,\|\phi(x_h^{m,t}, a_h^{m,t})\|_{(\Lambda_h^{k(t)})^{-1}}.$$

Lemma E.4 (Concentration of the martingale correction term)**.** *For any confidence level $p \in (0,1)$, let $\mathcal{E}$ be the event that,*

$$\sum_{m=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H} \zeta_{h+1}^{m,t} \le \sqrt{8H^3 MT \log\Big(\frac{2}{p}\Big)},$$

Then $\mathbb{P}(\mathcal{E}) \ge 1 - p$.

Lemma E.5 (Elliptical potential bound)**.** *For any $\lambda \ge 1$ and $\gamma \ge 1$,*

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{t=\tau_k}^{\tau_{k+1}-1}\sum_{m=1}^{M} \|\phi(x_h^{m,t}, a_h^{m,t})\|_{(\Lambda_h^k)^{-1}} \le$$
$$H\left(\sqrt{6dMT\log\Big(1 + \frac{MT}{d\lambda}\Big)} + \frac{Md}{\log 3}\sqrt{\gamma\log\gamma}\,\log\Big(1 + \frac{MT}{d\lambda}\Big) + \frac{Md}{\sqrt{\lambda}\,\log 3}\,\log\Big(1 + \frac{MT}{d\lambda}\Big)\right).$$

E.2 Proof of Theorem 1

Proof of Theorem 1. The event $\mathcal{C}\cap\mathcal{E}$ holds with probability at least $1 - 2p$. We now prove Theorem 1 under the event $\mathcal{C}\cap\mathcal{E}$, under which all supporting lemmas in Section E.1 hold.

We first expand the regret and reduce it to bounding $\sum_{m,t}\delta_1^{m,t}$. For each episode t and each agent m, the executed policy is exactly $\pi^{k(t)}$. Lemma E.2 implies $V_1^*(x) \le V_1^{k(t)}(x)$ for every state x, hence

$$\begin{aligned}
\mathrm{Regret}(T) &= \sum_{m=1}^{M}\sum_{t=1}^{T}\Big(V_1^*(x_1^{m,t}) - V_1^{\pi^{k(t)}}(x_1^{m,t})\Big) \\
&\le \sum_{m=1}^{M}\sum_{t=1}^{T}\Big(V_1^{k(t)}(x_1^{m,t}) - V_1^{\pi^{k(t)}}(x_1^{m,t})\Big) \\
&= \sum_{m=1}^{M}\sum_{t=1}^{T}\delta_1^{m,t}.
\end{aligned}$$

Summing Lemma E.3 over $h = 1,\dots,H$ and using $\delta_{H+1}^{m,t} = 0$ gives

$$\delta_1^{m,t} \le \sum_{h=1}^{H}\zeta_{h+1}^{m,t} + 2\beta\sum_{h=1}^{H}\|\phi(x_h^{m,t}, a_h^{m,t})\|_{(\Lambda_h^{k(t)})^{-1}}.$$

Summing over all (m,t) and regrouping by rounds yields

$$\sum_{m=1}^{M}\sum_{t=1}^{T}\delta_1^{m,t} \le \underbrace{\sum_{m=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H}\zeta_{h+1}^{m,t}}_{\text{martingale term}} + 2\beta \underbrace{\sum_{k=1}^{K}\sum_{t=\tau_k}^{\tau_{k+1}-1}\sum_{m=1}^{M}\sum_{h=1}^{H}\|\phi(x_h^{m,t},a_h^{m,t})\|_{(\Lambda_h^k)^{-1}}}_{\text{elliptical-potential term}}.$$

On the event $\mathcal{C}\cap\mathcal{E}$, we have

$$\begin{aligned}\mathrm{Regret}(T) \le\ & \sqrt{8H^3MT\log\Big(\frac{2}{p}\Big)} + 2H\beta\Bigg(\sqrt{6dMT\log\Big(1+\frac{MT}{d\lambda}\Big)} \\ & + \frac{Md}{\log 3}\sqrt{\gamma\log\gamma}\,\log\Big(1+\frac{MT}{d\lambda}\Big) + \frac{Md}{\sqrt{\lambda}\log 3}\,\log\Big(1+\frac{MT}{d\lambda}\Big)\Bigg).\end{aligned}$$

where $\beta = c_\beta Hd\sqrt{\log(1+dMHT/p)}$. By setting $\lambda = 1$ and plugging in β, the above bound directly recovers the regret bound stated in Theorem 1:

$$\begin{aligned}\mathrm{Regret}(T) \le\ & (2c_\beta\sqrt{6}+4\sqrt{2})\sqrt{Md^3H^4T\log\Big(1+\frac{dMHT}{p}\Big)\log\Big(1+\frac{MT}{d}\Big)} \\ & + \frac{2c_\beta Md^2H^2}{\log 3}(\sqrt{\gamma\log\gamma}+1)\sqrt{\log(1+dMHT/p)}\log\Big(1+\frac{MT}{d}\Big). \qquad (13)\end{aligned}$$

□

E.3 Proof of Supporting Lemmas

E.3.1 Proof of Lemma E.1

Proof of Lemma E.1. Fix a round k, a step h, a policy π, and a state-action pair (x,a). We begin from an exact algebraic decomposition of $w_h^k - w_h^\pi$. By (1) and Lemma D.2, we have

$$Q_h^\pi(x,a) = r_h(x,a) + \mathbb{P}_h V_{h+1}^\pi(x,a) = \phi(x,a)^\top w_h^\pi.$$

We obtain

$$\begin{aligned}w_h^k - w_h^\pi &= (\Lambda_h^k)^{-1}\sum_{m=1}^{M}\sum_{j=1}^{\tau_k-1}\phi(x_h^{m,j},a_h^{m,j})\big[r_h(x_h^{m,j},a_h^{m,j}) + V_{h+1}^k(x_{h+1}^{m,j})\big] - w_h^\pi \\ &= (\Lambda_h^k)^{-1}\sum_{m=1}^{M}\sum_{j=1}^{\tau_k-1}\phi(x_h^{m,j},a_h^{m,j})\Big[V_{h+1}^k(x_{h+1}^{m,j}) - \mathbb{P}_h V_{h+1}^\pi(x_h^{m,j},a_h^{m,j})\Big] - \lambda(\Lambda_h^k)^{-1}w_h^\pi \\ &= (\Lambda_h^k)^{-1}\sum_{m=1}^{M}\sum_{j=1}^{\tau_k-1}\phi(x_h^{m,j},a_h^{m,j})\Big[V_{h+1}^k(x_{h+1}^{m,j}) - \mathbb{P}_h V_{h+1}^k(x_h^{m,j},a_h^{m,j})\Big] \\ &\quad + (\Lambda_h^k)^{-1}\sum_{m=1}^{M}\sum_{j=1}^{\tau_k-1}\phi(x_h^{m,j},a_h^{m,j})\,\mathbb{P}_h\big(V_{h+1}^k - V_{h+1}^\pi\big)(x_h^{m,j},a_h^{m,j}) - \lambda(\Lambda_h^k)^{-1}w_h^\pi. \qquad (14)\end{aligned}$$

Now define

$$\nu_h^{k,\pi} := \int_{\mathcal{X}}\big(V_{h+1}^k(s') - V_{h+1}^\pi(s')\big)\,\mu_h(\mathrm{d}s').$$

By linearity of the transition probability,

$$\mathbb{P}_h\big(V_{h+1}^k - V_{h+1}^\pi\big)(x,a) = \phi(x,a)^\top\nu_h^{k,\pi}. \qquad (15)$$

Therefore,

$$\begin{aligned}(\Lambda_h^k)^{-1}\sum_{m=1}^{M}\sum_{j=1}^{\tau_k-1}\phi(x_h^{m,j},a_h^{m,j})\,\mathbb{P}_h\big(V_{h+1}^k - V_{h+1}^\pi\big)(x_h^{m,j},a_h^{m,j}) &= (\Lambda_h^k)^{-1}(\Lambda_h^k - \lambda I_d)\nu_h^{k,\pi} \\ &= \nu_h^{k,\pi} - \lambda(\Lambda_h^k)^{-1}\nu_h^{k,\pi}.\end{aligned}$$

Combined with the (14) and (15), we arrive at

$$\begin{aligned}
&\phi(x,a)^\top w_h^k - Q_h^\pi(x,a) - \mathbb{P}_h\big(V_{h+1}^k - V_{h+1}^\pi\big)(x,a) \\
&= \langle \phi(x,a), w_h^k - w_h^\pi - \nu_h^{k,\pi} \rangle \\
&= \left\langle \phi(x,a), \sum_{m=1}^{M}\sum_{j=1}^{\tau_k-1} \phi(x_h^{m,j}, a_h^{m,j}) \Big[V_{h+1}^k(x_{h+1}^{m,j}) - \mathbb{P}_h V_{h+1}^k(x_h^{m,j}, a_h^{m,j}) \Big] \right\rangle_{(\Lambda_h^k)^{-1}} \\
&\quad - \lambda \left\langle \phi(x,a), w_h^\pi + \nu_h^{k,\pi} \right\rangle_{(\Lambda_h^k)^{-1}}.
\end{aligned} \tag{16}$$

For the first term of (16), Cauchy-Schwarz and conditioned on the event $\mathcal{C}$ give

$$\begin{aligned}
&\left| \left\langle \phi(x,a), \sum_{m=1}^{M}\sum_{j=1}^{\tau_k-1} \phi(x_h^{m,j}, a_h^{m,j}) \Big[V_{h+1}^k(x_{h+1}^{m,j}) - \mathbb{P}_h V_{h+1}^k(x_h^{m,j}, a_h^{m,j}) \Big] \right\rangle_{(\Lambda_h^k)^{-1}} \right| \\
&\le \|\phi(x,a)\|_{(\Lambda_h^k)^{-1}} \left\| \sum_{m=1}^{M}\sum_{j=1}^{\tau_k-1} \phi(x_h^{m,j}, a_h^{m,j}) \Big[V_{h+1}^k(x_{h+1}^{m,j}) - \mathbb{P}_h V_{h+1}^k(x_h^{m,j}, a_h^{m,j}) \Big] \right\|_{(\Lambda_h^k)^{-1}} \\
&\le CHd\sqrt{\chi}\, \|\phi(x,a)\|_{(\Lambda_h^k)^{-1}}.
\end{aligned} \tag{17}$$

For the second term of (16), note that

$$w_h^\pi + \nu_h^{k,\pi} = \theta_h + \int_{\mathcal{X}} V_{h+1}^k(s')\, \mu_h(\mathrm{d}s').$$

Similar to the proof of Lemma D.2, we have $\|w_h^\pi + \nu_h^{k,\pi}\|_2 \le 2H\sqrt{d}$. Using again $\Lambda_h^k \succeq I_d$,

$$\lambda \left| \langle \phi(x,a), w_h^\pi + \nu_h^{k,\pi} \rangle_{(\Lambda_h^k)^{-1}} \right| \le \|\phi(x,a)\|_{(\Lambda_h^k)^{-1}} \|w_h^\pi + \nu_h^{k,\pi}\|_{(\Lambda_h^k)^{-1}} \le 2H\sqrt{d}\, \|\phi(x,a)\|_{(\Lambda_h^k)^{-1}}. \tag{18}$$

Combining the two bounds (17) and (18), we have

$$\left|\phi(x,a)^\top w_h^k - Q_h^\pi(x,a) - \mathbb{P}_h\big(V_{h+1}^k - V_{h+1}^\pi\big)(x,a)\right| \le (2H\sqrt{d} + CHd\sqrt{\chi})\, \|\phi(x,a)\|_{(\Lambda_h^k)^{-1}}. \tag{19}$$

There exists an absolute constant $c_\beta > 0$ such that for all $d, H, M, T \ge 1$,

$$2H\sqrt{d} + CHd\sqrt{\chi} \;\le\; c_\beta Hd\sqrt{\log\big(1 + dMHT/p\big)}.$$

Indeed, it suffices to verify that there exists $c_\beta > 0$ such that for all $x \ge 1$,

$$2 + C\sqrt{\log\big(1 + (c_\beta + 1)x\big)} \;\le\; c_\beta \sqrt{\log(1+x)}.$$

This condition can be satisfied by choosing $c_\beta = 8(C+1)^2$. □

E.3.2 Proof of Lemma E.2

Proof of Lemma E.2. We prove optimism by backward induction on h.

Base case $h = H$. At the last step, $V_{H+1}^k \equiv V_{H+1}^* \equiv 0$. Applying Lemma E.1 with $\pi = \pi^*$ gives

$$\phi(x,a)^\top w_H^k \ge Q_H^*(x,a) - \beta\, \|\phi(x,a)\|_{(\Lambda_H^k)^{-1}}.$$

By the definition of the optimistic estimator, for any $(x,a,k) \in \mathcal{X} \times \mathcal{A} \times [K]$, it holds that

$$Q_H^k(x,a) = \Pi_{[0,H]}\Big(\phi(x,a)^\top w_H^k + \beta\, \|\phi(x,a)\|_{(\Lambda_H^k)^{-1}}\Big) \ge Q_H^*(x,a).$$

Therefore $V_H^k(x) = \max_a Q_H^k(x,a) \ge \max_a Q_H^*(x,a) = V_H^*(x)$.

Induction step. Assume that $Q_{h+1}^k(x,a) \geq Q_{h+1}^*(x,a)$ and $V_{h+1}^k(x) \geq V_{h+1}^*(x)$ for any $(x,a,k) \in \mathcal{X} \times \mathcal{A} \times [K]$. Then $\mathbb{P}_h(V_{h+1}^k - V_{h+1}^*)(x,a) \geq 0$ for any (x,a,k).

Applying Lemma E.1 with $\pi = \pi^*$ yields

$$\begin{aligned}\phi(x,a)^\top w_h^k &\geq Q_h^*(x,a) + \mathbb{P}_h\big(V_{h+1}^k - V_{h+1}^*\big)(x,a) - \beta \left\|\phi(x,a)\right\|_{(\Lambda_h^k)^{-1}} \\ &\geq Q_h^*(x,a) - \beta \left\|\phi(x,a)\right\|_{(\Lambda_h^k)^{-1}}.\end{aligned}$$

Therefore, for any $(x,a,k) \in \mathcal{X} \times \mathcal{A} \times [K]$, we have

$$Q_h^k(x,a) = \Pi_{[0,H]}\Big(\phi(x,a)^\top w_h^k + \beta \left\|\phi(x,a)\right\|_{(\Lambda_h^k)^{-1}}\Big) \geq Q_h^*(x,a)$$

and thus

$$V_h^k(x) = \max_a Q_h^k(x,a) \geq \max_a Q_h^*(x,a) = V_h^*(x).$$

This completes the proof for step h, and hence the conclusion follows by backward induction. □

E.3.3 Proof of Lemma E.3

Proof of Lemma E.3. Let $k = k(t)$ be the round index that the t-th episode is in. Fix an agent m, an episode t. Since $\pi^{k(t)}$ is greedy with respect to Q_h^k,

$$V_h^{k(t)}(x_h^{m,t}) = Q_h^{k(t)}(x_h^{m,t}, a_h^{m,t}), \qquad V_h^{\pi^{k(t)}}(x_h^{m,t}) = Q_h^{\pi^{k(t)}}(x_h^{m,t}, a_h^{m,t}).$$

Therefore,

$$\delta_h^{m,t} = Q_h^{k(t)}(x_h^{m,t}, a_h^{m,t}) - Q_h^{\pi^{k(t)}}(x_h^{m,t}, a_h^{m,t}).$$

Applying Lemma E.1 with $\pi = \pi^{k(t)}$ and using the definition of $Q_h^{k(t)}$ gives

$$\delta_h^{m,t} \leq \mathbb{P}_h\big(V_{h+1}^{k(t)} - V_{h+1}^{\pi^{k(t)}}\big)(x_h^{m,t}, a_h^{m,t}) + 2\beta \left\|\phi(x_h^{m,t}, a_h^{m,t})\right\|_{(\Lambda_h^k)^{-1}}.$$

The preceding argument applies unless the lower clipping is active, namely when the unprojected optimistic score is negative. In this case $Q_h^k = 0$, and thus

$$\begin{aligned}\delta_h^{m,t} = -Q_h^{\pi^{k(t)}}(x_h^{m,t}, a_h^{m,t}) &= -r_h(x_h^{m,t}, a_h^{m,t}) - \mathbb{P}_h V_{h+1}^{\pi^{k(t)}}(x_h^{m,t}, a_h^{m,t}) \\ &\leq \mathbb{P}_h\big(V_{h+1}^{k(t)} - V_{h+1}^{\pi^{k(t)}}\big)(x_h^{m,t}, a_h^{m,t}),\end{aligned}$$

because $r_h \geq 0$ and $V_{h+1}^{k(t)} \geq 0$. Combining the two cases and noting that

$$\mathbb{P}_h\big(V_{h+1}^{k(t)} - V_{h+1}^{\pi^{k(t)}}\big)(x_h^{m,t}, a_h^{m,t}) = \mathbb{E}\big[\delta_{h+1}^{m,t} \mid \mathcal{G}_{m,t,h}\big] = \delta_{h+1}^{m,t} + \zeta_{h+1}^{m,t},$$

we finish the proof of the conclusion. □

E.3.4 Proof of Lemma E.4

Proof of Lemma E.4. By construction, $\mathbb{E}[\zeta_{h+1}^{m,t} \mid \mathcal{G}_{m,t,h}] = 0$, so the sequence $\{\zeta_{h+1}^{m,t}\}_{m,t,h}$ is a martingale-difference array under the linearized ordering. Moreover, both $V_{h+1}^{k(t)}$ and $V_{h+1}^{\pi^{k(t)}}$ lie in $[0,H]$, hence $|\zeta_{h+1}^{m,t}| \leq 2H$. Applying Azuma–Hoeffding to the sum over all MTH terms gives

$$\mathbb{P}\left(\Big|\sum_{m=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H}\zeta_{h+1}^{m,t}\Big| > u\right) \leq 2\exp\Big(-\frac{u^2}{8H^3MT}\Big).$$

Setting $u = \sqrt{8H^3MT\log(2/p)}$ yields the claim. □

E.3.5 Proof of Lemma E.5

Proof of Lemma E.5. Fix a step $h \in [H]$. We bound

$$\sum_{k=1}^{K}\sum_{t=\tau_k}^{\tau_{k+1}-1}\sum_{m=1}^{M}\left\|\phi(x_h^{m,t}, a_h^{m,t})\right\|_{(\Lambda_h^k)^{-1}},$$

and then sum the result over $h = 1, \ldots, H$.

Step 1: Split rounds by global determinant growth. Define

$$\mathcal{K}_1(h) \coloneqq \Big\{k : \frac{\det(\Lambda_h^{k+1})}{\det(\Lambda_h^k)} \le 3\Big\}, \qquad \mathcal{K}_2(h) \coloneqq \Big\{k : \frac{\det(\Lambda_h^{k+1})}{\det(\Lambda_h^k)} > 3\Big\}.$$

Step 2: Analysis for rounds with small determinant growth. Fix $k \in \mathcal{K}_1(h)$ and any (m,t) with $t \in [\tau_k, \tau_{k+1}-1]$. Since $\Lambda_h^k \preceq \Lambda_h^{k+1}$, for any $x \in \mathbb{R}^d$, due to Lemma D.1

$$\|x\|^2_{(\Lambda_h^k)^{-1}} \le \frac{\det(\Lambda_h^{k+1})}{\det(\Lambda_h^k)} \|x\|^2_{(\Lambda_h^{k+1})^{-1}}.$$

Hence

$$\|x\|_{(\Lambda_h^k)^{-1}} \le \sqrt{3}\, \|x\|_{(\Lambda_h^{k+1})^{-1}}. \tag{20}$$

Recall that

$$\begin{aligned}\tilde{\Lambda}_h^{m,t} &= \lambda I_d + \sum_{(n,j)\le(m,t)} \phi(x_h^{n,j}, a_h^{n,j})\phi(x_h^{n,j}, a_h^{n,j})^\top \\ &= \Lambda_h^k + \sum_{\substack{(n,j)\le(m,t)\\ k(j)=k}} \phi(x_h^{n,j}, a_h^{n,j})\phi(x_h^{n,j}, a_h^{n,j})^\top.\end{aligned}$$

Here the last summation is over those pairs $(n,j) \le (m,t)$ such that episode j lies in round k. Hence, it holds that

$$\Lambda_h^{k+1} \succeq \tilde{\Lambda}_h^{m,t} \succeq \Lambda_h^k.$$

and thus together with (20), we have

$$\|\phi(x_h^{m,t}, a_h^{m,t})\|_{(\Lambda_h^k)^{-1}} \le \sqrt{3}\, \|\phi(x_h^{m,t}, a_h^{m,t})\|_{(\Lambda_h^{k+1})^{-1}} \le \sqrt{3}\, \|\phi(x_h^{m,t}, a_h^{m,t})\|_{(\tilde{\Lambda}_h^{m,t})^{-1}}.$$

Summing the inequality above and applying Cauchy–Schwarz inequality yields

$$\begin{aligned}\sum_{k\in\mathcal{K}_1(h)} \sum_{t=\tau_k}^{\tau_{k+1}-1} \sum_{m=1}^{M} \|\phi(x_h^{m,t}, a_h^{m,t})\|_{(\Lambda_h^k)^{-1}} &\le \sqrt{3} \sum_{t=1}^{T} \sum_{m=1}^{M} \|\phi(x_h^{m,t}, a_h^{m,t})\|_{(\tilde{\Lambda}_h^{m,t})^{-1}} \\ &\le \sqrt{3MT \sum_{t=1}^{T} \sum_{m=1}^{M} \|\phi(x_h^{m,t}, a_h^{m,t})\|^2_{(\tilde{\Lambda}_h^{m,t})^{-1}}}.\end{aligned}$$

Applying Lemma D.6 in the linearized order of (m,t), with $V_0 = \lambda I_d$, gives

$$\sum_{t=1}^{T} \sum_{m=1}^{M} \|\phi(x_h^{m,t}, a_h^{m,t})\|^2_{(\tilde{\Lambda}_h^{m,t})^{-1}} \le 2 \log \frac{\det(\tilde{\Lambda}_h^{M,T})}{\det(\lambda I_d)} \le 2d \log\Big(1 + \frac{MT}{d\lambda}\Big).$$

Therefore,

$$\sum_{k\in\mathcal{K}_1(h)} \sum_{t=\tau_k}^{\tau_{k+1}-1} \sum_{m=1}^{M} \|\phi(x_h^{m,t}, a_h^{m,t})\|_{(\Lambda_h^k)^{-1}} \le \sqrt{6dMT \log\Big(1 + \frac{MT}{d\lambda}\Big)}. \tag{21}$$

Step 3: Analysis for rounds with large determinant growth. Fix $k \in \mathcal{K}_2(h)$ and suppose first that $\tau_{k+1} - \tau_k \ge 2$. If $\tau_{k+1} - \tau_k = 1$, the round contains only its final episode, which is covered directly by Step 4 below. For one agent m, define the **agent-wise pseudo covariance matrix**

$$\bar{\Lambda}_h^{m,t} = \Lambda_h^k + \sum_{j=\tau_k}^{t} \phi(x_h^{m,j}, a_h^{m,j})\phi(x_h^{m,j}, a_h^{m,j})^\top, \qquad t \in [\tau_k, \tau_{k+1}-1].$$

Recall that

$$\Lambda_h^{m,k,\mathrm{loc}} = \sum_{t=\tau_k}^{\tau_{k+1}-1} \phi(x_h^{m,t}, a_h^{m,t})\phi(x_h^{m,t}, a_h^{m,t})^\top$$

and

$$\Lambda_h^{k+1} = \Lambda_h^k + \sum_{m=1}^{M} \Lambda_h^{m,k,\mathrm{loc}},$$

we have

$$\Lambda_h^k \preceq \bar{\Lambda}_h^{m,t} \preceq \Lambda_h^k + \Lambda_h^{m,k,\mathrm{loc}} = \bar{\Lambda}_h^{m,\tau_{k+1}-1}.$$

For any $t \in [\tau_k, \tau_{k+1} - 2]$, by Lemma D.1, we have

$$\|\phi(x_h^{m,t}, a_h^{m,t})\|_{(\Lambda_h^k)^{-1}} \le \sqrt{\frac{\det(\bar{\Lambda}_h^{m,\tau_{k+1}-2})}{\det(\Lambda_h^k)}}\, \|\phi(x_h^{m,t}, a_h^{m,t})\|_{(\bar{\Lambda}_h^{m,t})^{-1}}.$$

Summing this inequality over $t = \tau_k, \ldots, \tau_{k+1} - 2$ and applying Cauchy–Schwarz inequality, we reach

$$\begin{aligned}
&\sum_{t=\tau_k}^{\tau_{k+1}-2} \|\phi(x_h^{m,t}, a_h^{m,t})\|_{(\Lambda_h^k)^{-1}} \\
&\le \left[(\tau_{k+1} - \tau_k - 1)\, \frac{\det(\bar{\Lambda}_h^{m,\tau_{k+1}-2})}{\det(\Lambda_h^k)} \cdot \sum_{t=\tau_k}^{\tau_{k+1}-2} \|\phi(x_h^{m,t}, a_h^{m,t})\|^2_{(\bar{\Lambda}_h^{m,t})^{-1}} \right]^{1/2} \\
&\le \left[(\tau_{k+1} - \tau_k - 1)\, \frac{\det(\bar{\Lambda}_h^{m,\tau_{k+1}-2})}{\det(\Lambda_h^k)} \cdot \log \frac{\det(\bar{\Lambda}_h^{m,\tau_{k+1}-2})}{\det(\Lambda_h^k)} \right]^{1/2}. \qquad (22)
\end{aligned}$$

The last inequality follows from Lemma D.6 applied to the episode-wise rank-one updates defining $\bar{\Lambda}_h^{m,t}$, initialized at Λ_h^k and ending at $\bar{\Lambda}_h^{m,\tau_{k+1}-2}$. Now use the stopping rule. At time $t = \tau_{k+1} - 2$ the round has not yet stopped, so the determinant trigger has not fired. Hence

$$\frac{\det(\bar{\Lambda}_h^{m,\tau_{k+1}-2})}{\det(\Lambda_h^k)} \le \frac{\gamma}{\tau_{k+1} - \tau_k - 1}.$$

Combining this bound and (22) yields

$$\sum_{t=\tau_k}^{\tau_{k+1}-2} \|\phi(x_h^{m,t}, a_h^{m,t})\|_{(\Lambda_h^k)^{-1}} \le \sqrt{\gamma \log\Big(\frac{\gamma}{\tau_{k+1} - \tau_k - 1}\Big)} \le \sqrt{\gamma \log \gamma}.$$

Summing over m and then over $k \in \mathcal{K}_2(h)$ gives

$$\sum_{k \in \mathcal{K}_2(h)} \sum_{t=\tau_k}^{\tau_{k+1}-2} \sum_{m=1}^{M} \|\phi(x_h^{m,t}, a_h^{m,t})\|_{(\Lambda_h^k)^{-1}} \le M\, |\mathcal{K}_2(h)|\, \sqrt{\gamma \log \gamma}.$$

To bound $|\mathcal{K}_2(h)|$, note that

$$|\mathcal{K}_2(h)| \log 3 \le \sum_{k \in \mathcal{K}_2(h)} \log \frac{\det(\Lambda_h^{k+1})}{\det(\Lambda_h^k)} \le \log \frac{\det(\Lambda_h^{K+1})}{\det(\Lambda_h^1)} \le d \log\Big(1 + \frac{MT}{d\lambda}\Big).$$

Therefore,

$$\sum_{k \in \mathcal{K}_2(h)} \sum_{t=\tau_k}^{\tau_{k+1}-2} \sum_{m=1}^{M} \|\phi(x_h^{m,t}, a_h^{m,t})\|_{(\Lambda_h^k)^{-1}} \le \frac{Md}{\log 3} \sqrt{\gamma \log \gamma}\, \log\Big(1 + \frac{MT}{d\lambda}\Big). \qquad (23)$$

Step 4: Analysis for residual last-episode terms. For the remaining terms at $t = \tau_{k+1} - 1$ with $k \in \mathcal{K}_2(h)$, we simply use

$$\|\phi(x, a)\|_{(\Lambda_h^k)^{-1}} \le \frac{1}{\sqrt{\lambda}},$$

which follows from $\Lambda_h^k \succeq \lambda I_d$ and $\|\phi(x, a)\|_2 \le 1$. Hence

$$\sum_{k \in \mathcal{K}_2(h)} \sum_{m=1}^{M} \|\phi(x_h^{m,\tau_{k+1}-1}, a_h^{m,\tau_{k+1}-1})\|_{(\Lambda_h^k)^{-1}} \le \frac{M}{\sqrt{\lambda}}\, |\mathcal{K}_2(h)| \le \frac{Md}{\sqrt{\lambda} \log 3} \log\Big(1 + \frac{MT}{d\lambda}\Big). \quad (24)$$

Step 5: combine and sum over steps. Adding the bounds (21), (23), (24) and then summing over $h = 1, \dots, H$ proves the lemma. □

F Proof of Theorem 2

Throughout this section, we consider the misspecified setting in Definition 2. Define $\{(r_h, \mathbb{P}_h)\}_{h=1}^H$ to be the common reference linear reward and transition kernel in Definition 2, where

$$r_h(x,a) = \langle \phi(x,a), \theta_h \rangle, \qquad \mathbb{P}_h(B \mid x,a) = \langle \phi(x,a), \mu_h(B) \rangle.$$

For each agent $m \in [M]$, we write $Q^\pi_{m,h}$ and $V^\pi_{m,h}$ for the value functions induced by the local MDP of agent m under policy π, and $Q^*_{m,h}, V^*_{m,h}$ for the corresponding optimal functions under the optimal policy π^*_m.

F.1 Heterogeneity-Specific Lemmas

Lemma F.1. *For any agent $m \in [M]$, policy π, and step $h \in [H]$, define*

$$w^\pi_{m,h} := \theta_h + \int_{\mathcal{X}} V^\pi_{m,h+1}(x')\, \mu_h(\mathrm{d}x').$$

Then, for every $(x,a) \in \mathcal{X} \times \mathcal{A}$,

$$\left| Q^\pi_{m,h}(x,a) - \phi(x,a)^\top w^\pi_{m,h} \right| \le H\xi.$$

Moreover, $\|w^\pi_{m,h}\|_2 \le 2H\sqrt{d}$.

Proof of Lemma F.1. By the linearity of the reference model,

$$\phi(x,a)^\top w^\pi_{m,h} = r_h(x,a) + \mathbb{P}_h V^\pi_{m,h+1}(x,a).$$

Therefore,

$$\begin{aligned} \left| Q^\pi_{m,h}(x,a) - \phi(x,a)^\top w^\pi_{m,h} \right| &\le \left| r_{m,h}(x,a) - r_h(x,a) \right| + \left| (\mathbb{P}_{m,h} - \mathbb{P}_h) V^\pi_{m,h+1}(x,a) \right| \\ &\le \xi + (H-h)\xi \le H\xi, \end{aligned}$$

where we used $0 \le V^\pi_{m,h+1} \le H - h$. The norm bound follows from

$$\|w^\pi_{m,h}\|_2 \le \|\theta_h\|_2 + \left\| \int_{\mathcal{X}} V^\pi_{m,h+1}(x')\, \mu_h(\mathrm{d}x') \right\|_2 \le \sqrt{d} + H\sqrt{d} \le 2H\sqrt{d}.$$

□

Lemma F.2. *Let $\{\varepsilon^{m,t}\}_{m\in[M],\, t\in[T]}$ be any collection satisfying $|\varepsilon^{m,t}| \le B$ for all (m,t). Then, for every step $h \in [H]$, round $k \in [K]$, and vector $u \in \mathbb{R}^d$,*

$$\left| u^\top (\Lambda^k_h)^{-1} \sum_{m=1}^{M} \sum_{t=1}^{\tau_k - 1} \phi(x^{m,t}_h, a^{m,t}_h)\, \varepsilon^{m,t} \right| \le B\sqrt{dM\tau_k}\, \|u\|_{(\Lambda^k_h)^{-1}}.$$

Proof of Lemma F.2. Using the triangle inequality,

$$\begin{aligned} &\left| u^\top (\Lambda^k_h)^{-1} \sum_{m=1}^{M} \sum_{t=1}^{\tau_k - 1} \phi(x^{m,t}_h, a^{m,t}_h)\, \varepsilon^{m,t} \right| \\ &\le B \sum_{m=1}^{M} \sum_{t=1}^{\tau_k - 1} \left| u^\top (\Lambda^k_h)^{-1} \phi(x^{m,t}_h, a^{m,t}_h) \right| \\ &\le B \sum_{m=1}^{M} \sum_{t=1}^{\tau_k - 1} \left(\|u\|_{(\Lambda^k_h)^{-1}} \sqrt{\phi(x^{m,t}_h, a^{m,t}_h)^\top (\Lambda^k_h)^{-1} \phi(x^{m,t}_h, a^{m,t}_h)} \right) \\ &\le B\sqrt{M(\tau_k - 1)}\, \|u\|_{(\Lambda^k_h)^{-1}} \left[\sum_{m=1}^{M} \sum_{t=1}^{\tau_k - 1} \phi(x^{m,t}_h, a^{m,t}_h)^\top (\Lambda^k_h)^{-1} \phi(x^{m,t}_h, a^{m,t}_h) \right]^{1/2}. \end{aligned}$$

The last two inequalities hold due to the Cauchy–Schwarz inequality. Moreover, note that

$$\sum_{m=1}^{M}\sum_{t=1}^{\tau_k-1}\phi(x_h^{m,t},a_h^{m,t})^\top(\Lambda_h^k)^{-1}\phi(x_h^{m,t},a_h^{m,t}) = \mathrm{tr}\Big((\Lambda_h^k)^{-1}(\Lambda_h^k-\lambda I_d)\Big) \le d.$$

Combining the two displays and using $M(\tau_k-1)\le M\tau_k$ proves the claim. □

Lemma F.3 (Uniform concentration with agent-dependent transitions)**.** *Fix $\bar{p}\in(0,1)$ and set $\lambda=1$. Suppose the optimistic estimates are formed with bonus parameter*

$$\beta = c'_\beta Hd\sqrt{\log\Big(1+\frac{2dMHT}{\bar{p}}\Big)} + 2H\xi\sqrt{dMT},$$

where $c'_\beta>0$ is a sufficiently large positive constant. Then there exists an absolute constant $C>0$ such that, if we define an event $\mathcal{C}_{\mathrm{het}}$ as, simultaneously for all steps $h\in[H]$ and rounds $k\in[K]$,

$$\Bigg\|\sum_{m=1}^{M}\sum_{t=1}^{\tau_k-1}\phi(x_h^{m,t},a_h^{m,t})\Big[V_{h+1}^k(x_{h+1}^{m,t})-\mathbb{P}_{m,h}V_{h+1}^k(x_h^{m,t},a_h^{m,t})\Big]\Bigg\|_{(\Lambda_h^k)^{-1}} \le CHd\sqrt{\log\Big(1+\frac{(c'_\beta+1)dMHT}{\bar{p}}\Big)}.$$

Then $\mathbb{P}(\mathcal{C}_{\mathrm{het}})\ge 1-\bar{p}$.

Proof of Lemma F.3. The proof follows the same argument as in Lemma D.9. The only differences are the choice of β and the replacement of the common transition kernel $\mathbb{P}_h$ by the agent-dependent transition kernel $\mathbb{P}_{m,h}$. Neither difference affects the proof, since the argument only relies on boundedness and the martingale property.

□

Lemma F.4. *Assume $\lambda=1$. There exists an absolute constant $c'_\beta>0$ such that, for every confidence level $\bar{p}\in(0,1)$, if*

$$\beta = c'_\beta Hd\sqrt{\log\Big(1+\frac{2dMHT}{\bar{p}}\Big)} + 2H\xi\sqrt{dMT},$$

then conditioned on the event $\mathcal{C}_{\mathrm{het}}$ defined in Lemma F.3, simultaneously for all agents $m\in[M]$, rounds $k\in[K]$, steps $h\in[H]$, policies π, and state-action pairs (x,a),

$$\Big|\phi(x,a)^\top w_h^k - Q_{m,h}^\pi(x,a) - \mathbb{P}_h\big(V_{h+1}^k - V_{m,h+1}^\pi\big)(x,a)\Big| \le \beta\,\|\phi(x,a)\|_{(\Lambda_h^k)^{-1}} + H\xi.$$

Proof of Lemma F.4. Fix m,k,h,π, and (x,a). Define

$$\nu_{m,h}^{k,\pi} := \int_{\mathcal{X}}\big(V_{h+1}^k(x') - V_{m,h+1}^\pi(x')\big)\,\mu_h(\mathrm{d}x').$$

Then

$$\mathbb{P}_h\big(V_{h+1}^k - V_{m,h+1}^\pi\big)(x,a) = \phi(x,a)^\top\nu_{m,h}^{k,\pi}. \tag{25}$$

In the misspecified setting, we have

$$\begin{aligned} w_h^k &= (\Lambda_h^k)^{-1}\sum_{n=1}^{M}\sum_{t=1}^{\tau_k-1}\phi(x_h^{n,t},a_h^{n,t})\big[r_{n,h}(x_h^{n,t},a_h^{n,t}) + V_{h+1}^k(x_{h+1}^{n,t})\big] \\ &= (\Lambda_h^k)^{-1}\sum_{n=1}^{M}\sum_{t=1}^{\tau_k-1}\phi(x_h^{n,t},a_h^{n,t})\Big[\Delta_{n,t,h}^k + \eta_{n,t,h}^k + r_h(x_h^{n,t},a_h^{n,t}) + \mathbb{P}_h V_{h+1}^k(x_h^{n,t},a_h^{n,t})\Big], \end{aligned}$$

where

$$\Delta_{n,t,h}^k := \big[r_{n,h}-r_h\big](x_h^{n,t},a_h^{n,t}) + \big(\mathbb{P}_{n,h}-\mathbb{P}_h\big)V_{h+1}^k(x_h^{n,t},a_h^{n,t}),$$

$$\eta^k_{n,t,h} := V^k_{h+1}(x^{n,t}_{h+1}) - \mathbb{P}_{n,h}V^k_{h+1}(x^{n,t}_h, a^{n,t}_h).$$

According to the definition of $w^\pi_{m,h}$ and $\nu^{k,\pi}_{m,h}$, we have

$$r_h(x,a) + \mathbb{P}_h V^k_{h+1}(x,a) = \phi(x,a)^\top \big(w^\pi_{m,h} + \nu^{k,\pi}_{m,h}\big),$$

and then together with (25),

$$\begin{aligned}
&\phi(x,a)^\top w^k_h - \phi(x,a)^\top w^\pi_{m,h} - \mathbb{P}_h\big(V^k_{h+1} - V^\pi_{m,h+1}\big)(x,a) \\
&= \phi(x,a)^\top(\Lambda^k_h)^{-1}\sum_{n=1}^{M}\sum_{t=1}^{\tau_k-1}\phi(x^{n,t}_h, a^{n,t}_h)\,\Delta^k_{n,t,h} + \phi(x,a)^\top(\Lambda^k_h)^{-1}\sum_{n=1}^{M}\sum_{t=1}^{\tau_k-1}\phi(x^{n,t}_h, a^{n,t}_h)\,\eta^k_{n,t,h} \\
&\quad + \phi(x,a)^\top(\Lambda^k_h)^{-1}\sum_{n=1}^{M}\sum_{t=1}^{\tau_k-1}\phi(x^{n,t}_h, a^{n,t}_h)\phi(x^{n,t}_h, a^{n,t}_h)^\top\big(w^\pi_{m,h} + \nu^{k,\pi}_{m,h}\big) \\
&\quad - \phi(x,a)^\top\big(w^\pi_{m,h} + \nu^{k,\pi}_{m,h}\big) \\
&= \phi(x,a)^\top(\Lambda^k_h)^{-1}\sum_{n=1}^{M}\sum_{t=1}^{\tau_k-1}\phi(x^{n,t}_h, a^{n,t}_h)\,\Delta^k_{n,t,h} + \phi(x,a)^\top(\Lambda^k_h)^{-1}\sum_{n=1}^{M}\sum_{t=1}^{\tau_k-1}\phi(x^{n,t}_h, a^{n,t}_h)\,\eta^k_{n,t,h} \\
&\quad - \lambda\,\phi(x,a)^\top(\Lambda^k_h)^{-1}\big(w^\pi_{m,h} + \nu^{k,\pi}_{m,h}\big). \qquad (26)
\end{aligned}$$

In the last equation, we use

$$\sum_{n=1}^{M}\sum_{t=1}^{\tau_k-1}\phi(x^{n,t}_h, a^{n,t}_h)\phi(x^{n,t}_h, a^{n,t}_h)^\top = \Lambda^k_h - \lambda I_d.$$

For the first term in (26), Definition 2 and $0 \le V^k_{h+1} \le H$ imply $|\Delta^k_{n,t,h}| \le 2H\xi$, so Lemma F.2 gives

$$\left|\phi(x,a)^\top(\Lambda^k_h)^{-1}\sum_{n=1}^{M}\sum_{t=1}^{\tau_k-1}\phi(x^{n,t}_h, a^{n,t}_h)\,\Delta^k_{n,t,h}\right| \le 2H\xi\sqrt{dMT}\,\|\phi(x,a)\|_{(\Lambda^k_h)^{-1}}.$$

For the second martingale term in (26), Cauchy-Schwarz inequality and Lemma F.3 yield

$$\begin{aligned}
&\left|\phi(x,a)^\top(\Lambda^k_h)^{-1}\sum_{n=1}^{M}\sum_{t=1}^{\tau_k-1}\phi(x^{n,t}_h, a^{n,t}_h)\,\eta^k_{n,t,h}\right| \\
&\le \|\phi(x,a)\|_{(\Lambda^k_h)^{-1}}\left\|\sum_{n=1}^{M}\sum_{t=1}^{\tau_k-1}\phi(x^{n,t}_h, a^{n,t}_h)\,\eta^k_{n,t,h}\right\|_{(\Lambda^k_h)^{-1}} \\
&\le CHd\sqrt{\log\Big(1 + \frac{(c'_\beta+1)dMHT}{\bar{p}}\Big)}\,\|\phi(x,a)\|_{(\Lambda^k_h)^{-1}}.
\end{aligned}$$

Moreover, for the last term in (26)

$$\big\|w^\pi_{m,h} + \nu^{k,\pi}_{m,h}\big\|_2 = \left\|\theta_h + \int_{\mathcal{X}} V^k_{h+1}(x')\,\mu_h(\mathrm{d}x')\right\|_2 \le \sqrt{d} + H\sqrt{d} \le 2H\sqrt{d},$$

so by Cauchy-Schwarz inequality and $\Lambda^k_h \succeq \lambda I$,

$$\begin{aligned}
\lambda\left|\phi(x,a)^\top(\Lambda^k_h)^{-1}(w^\pi_{m,h} + \nu^{k,\pi}_{m,h})\right| &\le \lambda\,\|\phi(x,a)\|_{(\Lambda^k_h)^{-1}}\|w^\pi_{m,h} + \nu^{k,\pi}_{m,h}\|_{(\Lambda^k_h)^{-1}} \\
&\le 2H\sqrt{d\lambda}\,\|\phi(x,a)\|_{(\Lambda^k_h)^{-1}}.
\end{aligned}$$

Finally, Lemma F.1 gives

$$\left|\phi(x,a)^\top w^\pi_{m,h} - Q^\pi_{m,h}(x,a)\right| \le H\xi.$$

Combining this with the preceding upper bounds for the three terms in (26), and using $\tau_k \le T$ and $\lambda = 1$, we obtain

$$\begin{aligned}&\left|\phi(x,a)^\top w_h^k - Q_{m,h}^\pi(x,a) - \mathbb{P}_h\big(V_{h+1}^k - V_{m,h+1}^\pi\big)(x,a)\right| \\ &\le \left(CHd\sqrt{\log\Big(1 + \frac{(c'_\beta+1)dMHT}{\bar{p}}\Big)} + 2H\xi\sqrt{dMT} + 2H\sqrt{d}\right)\|\phi(x,a)\|_{(\Lambda_h^k)^{-1}} + H\xi.\end{aligned}$$

Choosing c'_β sufficiently large absorbs the first and third terms into the displayed value of β, which proves the claim. □

Lemma F.5. *Under the event $\mathcal{C}_{\mathrm{het}}$, for every agent $m \in [M]$, round $k \in [K]$, step $h \in [H]$, and state-action pair (x,a),*

$$Q_h^k(x,a) \ge Q_{m,h}^*(x,a) - H(H+1-h)\xi, \quad V_h^k(x) \ge V_{m,h}^*(x) - H(H+1-h)\xi.$$

Proof of Lemma F.5. We argue by backward induction on h.

For $h = H$, we have $V_{H+1}^k \equiv V_{m,H+1}^* \equiv 0$. Applying Lemma F.4 with $\pi = \pi_m^*$ gives

$$\phi(x,a)^\top w_H^k \ge Q_{m,H}^*(x,a) - \beta\,\|\phi(x,a)\|_{(\Lambda_H^k)^{-1}} - H\xi.$$

Hence, by the definition of Q_H^k,

$$Q_H^k(x,a) = \Pi_{[0,H]}\Big(\phi(x,a)^\top w_H^k + \beta\,\|\phi(x,a)\|_{(\Lambda_H^k)^{-1}}\Big) \ge Q_{m,H}^*(x,a) - H\xi.$$

and thus

$$V_H^k(x) = \max_a Q_H^k(x,a) \ge \max_a Q_{m,H}^*(x,a) - H\xi = V_{m,H}^*(x) - H\xi.$$

Now assume the claim holds at step $h+1$. Then, by induction, for all $(x,k) \in \mathcal{X} \times [K]$, we have

$$V_{h+1}^k(x) \ge V_{m,h+1}^*(x) - H(H-h)\xi.$$

Therefore,

$$\mathbb{P}_h\big(V_{h+1}^k - V_{m,h+1}^*\big)(x,a) \ge -H(H-h)\xi.$$

Applying Lemma F.4 with $\pi = \pi_m^*$ yields

$$\phi(x,a)^\top w_h^k \ge Q_{m,h}^*(x,a) - H(H-h)\xi - H\xi - \beta\,\|\phi(x,a)\|_{(\Lambda_h^k)^{-1}}.$$

Thus,

$$Q_h^k(x,a) = \Pi_{[0,H]}\Big(\phi(x,a)^\top w_h^k + \beta\,\|\phi(x,a)\|_{(\Lambda_h^k)^{-1}}\Big) \ge Q_{m,h}^*(x,a) - H(H+1-h)\xi.$$

and thus

$$V_h^k(x) = \max_a Q_h^k(x,a) \ge \max_a Q_{m,h}^*(x,a) - H(H+1-h)\xi = V_{m,h}^*(x) - H(H+1-h)\xi.$$

This completes the proof for step h, and hence the conclusion follows by backward induction. □

For the trajectory generated by agent m in episode t, define

$$\begin{aligned}\delta_h^{m,t} &:= V_h^{k(t)}(x_h^{m,t}) - V_{m,h}^{\pi^{k(t)}}(x_h^{m,t}), \\ \mathcal{G}_{m,t,h} &:= \sigma\left(\Big\{(x_\ell^{n,j}, a_\ell^{n,j}, x_{\ell+1}^{n,j})\Big\}_{(n,j)<(m,t),\,\ell\in[H]} \cup \Big\{(x_\ell^{m,t}, a_\ell^{m,t})\Big\}_{\ell\le h}\right), \\ \zeta_{h+1}^{m,t} &:= \mathbb{E}\big[\delta_{h+1}^{m,t} \mid \mathcal{G}_{m,t,h}\big] - \delta_{h+1}^{m,t}.\end{aligned}$$

Equivalently,

$$\zeta_{h+1}^{m,t} = \mathbb{P}_{m,h}\big(V_{h+1}^{k(t)} - V_{m,h+1}^{\pi^{k(t)}}\big)(x_h^{m,t}, a_h^{m,t}) - \big(V_{h+1}^{k(t)} - V_{m,h+1}^{\pi^{k(t)}}\big)(x_{h+1}^{m,t}).$$

Lemma F.6. *Conditioned on $\mathcal{C}_{\mathrm{het}}$, for every agent $m \in [M]$, episode $t \in [T]$, and step $h \in [H]$,*

$$\delta_h^{m,t} \le \delta_{h+1}^{m,t} + \zeta_{h+1}^{m,t} + 2\beta \left\|\phi(x_h^{m,t}, a_h^{m,t})\right\|_{(\Lambda_h^{k(t)})^{-1}} + 3H\xi.$$

Proof of Lemma F.6. Fix m, t. Let $k = k(t)$ be the round index that the t-th episode is in. Since $\pi^{k(t)}$ is greedy with respect to $Q_h^{k(t)}$,

$$V_h^{k(t)}(x_h^{m,t}) = Q_h^{k(t)}(x_h^{m,t}, a_h^{m,t}), \qquad V_{m,h}^{\pi^{k(t)}}(x_h^{m,t}) = Q_{m,h}^{\pi^{k(t)}}(x_h^{m,t}, a_h^{m,t}).$$

Hence

$$\delta_h^{m,t} = Q_h^{k(t)}(x_h^{m,t}, a_h^{m,t}) - Q_{m,h}^{\pi^{k(t)}}(x_h^{m,t}, a_h^{m,t}).$$

Under the event $\mathcal{C}_{\mathrm{het}}$, Lemma F.4 implies

$$\delta_h^{m,t} \le \mathbb{P}_h\big(V_{h+1}^{k(t)} - V_{m,h+1}^{\pi^{k(t)}}\big)(x_h^{m,t}, a_h^{m,t}) + 2\beta \left\|\phi(x_h^{m,t}, a_h^{m,t})\right\|_{(\Lambda_h^k)^{-1}} + H\xi. \tag{27}$$

Now compare the reference and local transition operators. For any (x, a),

$$\begin{aligned}
&\mathbb{P}_h\big(V_{h+1}^{k(t)} - V_{m,h+1}^{\pi^{k(t)}}\big)(x,a) - \mathbb{P}_{m,h}\big(V_{h+1}^{k(t)} - V_{m,h+1}^{\pi^{k(t)}}\big)(x,a) \\
&= \big(\mathbb{P}_h - \mathbb{P}_{m,h}\big)V_{h+1}^{k(t)}(x,a) + \big(\mathbb{P}_{m,h} - \mathbb{P}_h\big)V_{m,h+1}^{\pi^{k(t)}}(x,a).
\end{aligned}$$

Since both value functions lie in $[0, H]$, the right-hand side has upper bound at most $2H\xi$. Therefore,

$$\mathbb{P}_h\big(V_{h+1}^{k(t)} - V_{m,h+1}^{\pi^{k(t)}}\big)(x_h^{m,t}, a_h^{m,t}) \le \mathbb{P}_{m,h}\big(V_{h+1}^{k(t)} - V_{m,h+1}^{\pi^{k(t)}}\big)(x_h^{m,t}, a_h^{m,t}) + 2H\xi.$$

Substituting this inequality into (27) gives

$$\delta_h^{m,t} \le \mathbb{P}_{m,h}\big(V_{h+1}^{k(t)} - V_{m,h+1}^{\pi^{k(t)}}\big)(x_h^{m,t}, a_h^{m,t}) + 2\beta \left\|\phi(x_h^{m,t}, a_h^{m,t})\right\|_{(\Lambda_h^k)^{-1}} + 3H\xi.$$

The only additional point caused by the lower clipping is the case where the unprojected optimistic score is negative. In this case $Q_h^{k(t)} = 0$, so

$$\begin{aligned}
\delta_h^{m,t} &= -Q_{m,h}^{\pi^{k(t)}}(x_h^{m,t}, a_h^{m,t}) \\
&= -r_{m,h}(x_h^{m,t}, a_h^{m,t}) - P_{m,h}V_{m,h+1}^{\pi^{k(t)}}(x_h^{m,t}, a_h^{m,t}) \\
&\le P_{m,h}\big(V_{h+1}^{k(t)} - V_{m,h+1}^{\pi^{k(t)}}\big)(x_h^{m,t}, a_h^{m,t}).
\end{aligned}$$

because $r_{m,h} \ge 0$ and $V_{h+1}^{k(t)} \ge 0$. Hence the same recursion follows, since the bonus term and the heterogeneity error term are nonnegative.

Finally, by noting that

$$\mathbb{P}_{m,h}\big(V_{h+1}^{k(t)} - V_{m,h+1}^{\pi^{k(t)}}\big)(x_h^{m,t}, a_h^{m,t}) = \mathbb{E}\big[\delta_{h+1}^{m,t} \mid \mathcal{G}_{m,t,h}\big] = \delta_{h+1}^{m,t} + \zeta_{h+1}^{m,t},$$

we finish the proof of the recursion. □

F.2 Proof of Theorem 2

Proof of Theorem 2. We first note that similar to Lemma E.4, the following event $\mathcal{E}_{\mathrm{het}}$ holds with probability at least $1 - \bar{p}$:

$$\sum_{m=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H} \zeta_{h+1}^{m,t} \le \sqrt{8H^3MT\log\Big(\frac{2}{\bar{p}}\Big)}. \tag{28}$$

Then the event $\mathcal{C}_{\mathrm{het}} \cap \mathcal{E}_{\mathrm{het}}$ holds with probability at least $1 - 2\bar{p}$. We now prove Theorem 2 under the event $\mathcal{C}_{\mathrm{het}} \cap \mathcal{E}_{\mathrm{het}}$, under which all lemmas in Section F.1 hold.

Lemma F.5 with $h = 1$ gives

$$V_{m,1}^*(x_1^{m,t}) \le V_1^{k(t)}(x_1^{m,t}) + H^2\xi.$$

Hence

$$\begin{aligned}\mathrm{Regret}(T) &= \sum_{m=1}^{M}\sum_{t=1}^{T}\Big(V_{m,1}^{*}(x_1^{m,t}) - V_{m,1}^{\pi^{k(t)}}(x_1^{m,t})\Big)\\ &\le \sum_{m=1}^{M}\sum_{t=1}^{T}\Big(V_1^{k(t)}(x_1^{m,t}) - V_{m,1}^{\pi^{k(t)}}(x_1^{m,t})\Big) + H^2MT\xi\\ &= \sum_{m=1}^{M}\sum_{t=1}^{T}\delta_1^{m,t} + H^2MT\xi. \end{aligned} \tag{29}$$

Next, summing Lemma F.6 over $h = 1, \dots, H$ and using $\delta_{H+1}^{m,t} = 0$ yields

$$\delta_1^{m,t} \le \sum_{h=1}^{H}\zeta_{h+1}^{m,t} + 2\beta\sum_{h=1}^{H}\|\phi(x_h^{m,t}, a_h^{m,t})\|_{(\Lambda_h^{k(t)})^{-1}} + 3H^2\xi.$$

Summing over all (m, t) and regrouping by rounds,

$$\sum_{m=1}^{M}\sum_{t=1}^{T}\delta_1^{m,t} \le \sum_{m=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H}\zeta_{h+1}^{m,t} + 2\beta\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{t=\tau_k}^{\tau_{k+1}-1}\sum_{m=1}^{M}\|\phi(x_h^{m,t}, a_h^{m,t})\|_{(\Lambda_h^{k})^{-1}} + 3H^2MT\xi.$$

Therefore, together with (29), we have

$$\mathrm{Regret}(T) \le \sum_{m=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H}\zeta_{h+1}^{m,t} + 2\beta\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{t=\tau_k}^{\tau_{k+1}-1}\sum_{m=1}^{M}\|\phi(x_h^{m,t}, a_h^{m,t})\|_{(\Lambda_h^{k})^{-1}} + 4H^2MT\xi.$$

Applying the martingale bound in (28) and Lemma E.5 (which is unchanged because it depends only on the Gram matrices and the synchronization rule) gives

$$\begin{aligned}\mathrm{Regret}(T) \le &\sqrt{8H^3MT\log\Big(\frac{2}{\bar{p}}\Big)} + 4H^2MT\xi\\ &+ 2\beta H\left(\sqrt{6dMT\log\Big(1+\frac{MT}{d}\Big)} + \frac{Md}{\log 3}(\sqrt{\gamma\log\gamma}+1)\,\log\Big(1+\frac{MT}{d}\Big)\right).\end{aligned}$$

Substituting

$$\beta = c'_\beta Hd\sqrt{\log\Big(1+\frac{2dMHT}{\bar{p}}\Big)} + 2H\xi\sqrt{dMT}$$

and collecting terms yield

$$\begin{aligned}\mathrm{Regret}(T) \le\ & \big(2c'_\beta\sqrt{6}+4\sqrt{2}\big)\sqrt{Md^3H^4T\,\log\Big(1+\frac{2dMHT}{\bar{p}}\Big)\,\log\Big(1+\frac{MT}{d}\Big)}\\ &+ \frac{2c'_\beta Md^2H^2}{\log 3}\big(\sqrt{\gamma\log\gamma}+1\big)\sqrt{\log\Big(1+\frac{2dMHT}{\bar{p}}\Big)}\,\log\Big(1+\frac{MT}{d}\Big)\\ &+ 4H^2MT\xi + 4\sqrt{6}\,\xi MdH^2T\sqrt{\log\Big(1+\frac{MT}{d}\Big)}\\ &+ \frac{4\xi\sqrt{M^3d^3H^4T}}{\log 3}\,\big(\sqrt{\gamma\log\gamma}+1\big)\,\sqrt{\log(1+2dMHT/\bar{p})}\log\Big(1+\frac{MT}{d}\Big).\end{aligned} \tag{30}$$

which is exactly the claimed bound in Theorem 2. □

Extending our communication-efficient federated RL guarantees in the misspecified setting to broader heterogeneity and participation models, such as stronger distributional shifts across agents or asynchronous and partial client participation, remains an interesting direction for future work.

G PROOF OF THEOREM 3

Proof of Theorem 3. Fix any $\alpha \in (0, \gamma)$ and divide the rounds into three classes:

$$I_{\text{long}} \coloneqq \{k : \tau_{k+1} - \tau_k > \alpha\},$$

$$I_{\text{short,tr}} \coloneqq \{k : \tau_{k+1} - \tau_k \le \alpha \text{ and round } k \text{ ends because the determinant trigger fires}\},$$

$$I_{\text{short,term}} \coloneqq \{k : \tau_{k+1} - \tau_k \le \alpha \text{ and round } k \text{ ends only because the hard cap } T \text{ is reached}\}.$$

By definition, these three sets form a partition of $[K]$. Moreover, $|I_{\text{short,term}}| \le 1$, because only the final round can terminate solely due to the hard stop.

Long rounds. For each $k \in I_{\text{long}}$ we have $\tau_{k+1} - \tau_k > \alpha$. Summing the round lengths and using that the total number of episodes per agent is T gives

$$\alpha |I_{\text{long}}| \le \sum_{k \in I_{\text{long}}} (\tau_{k+1} - \tau_k) \le \sum_{k=1}^{K} (\tau_{k+1} - \tau_k) = T,$$

so

$$|I_{\text{long}}| \le \frac{T}{\alpha}. \tag{31}$$

Short rounds that end by a determinant trigger. Fix any $k \in I_{\text{short,tr}}$. Then some agent m and some step h satisfy the trigger condition at the end of round k:

$$\frac{\det(\Lambda_h^k + \Lambda_h^{m,k,\text{loc}})}{\det(\Lambda_h^k)} \ge \frac{\gamma}{\tau_{k+1} - \tau_k} \ge \frac{\gamma}{\alpha}.$$

Since the global update dominates the single-agent update in the Loewner order,

$$\Lambda_h^{k+1} = \Lambda_h^k + \sum_{m=1}^{M} \Lambda_h^{m,k,\text{loc}} \succeq \Lambda_h^k + \Lambda_h^{m,k,\text{loc}},$$

and the determinant is monotone over positive semidefinite matrices. Therefore,

$$\frac{\det(\Lambda_h^{k+1})}{\det(\Lambda_h^k)} \ge \frac{\det(\Lambda_h^k + \Lambda_h^{m,k,\text{loc}})}{\det(\Lambda_h^k)} \ge \frac{\gamma}{\alpha}.$$

For every other step $\ell \ne h$, we still have $\det(\Lambda_\ell^{k+1}) / \det(\Lambda_\ell^k) \ge 1$. Hence

$$\prod_{\ell=1}^{H} \frac{\det(\Lambda_\ell^{k+1})}{\det(\Lambda_\ell^k)} \ge \frac{\gamma}{\alpha}.$$

Multiplying over all $k \in I_{\text{short,tr}}$ yields

$$\Big(\frac{\gamma}{\alpha}\Big)^{|I_{\text{short,tr}}|} \le \prod_{k \in I_{\text{short,tr}}} \prod_{\ell=1}^{H} \frac{\det(\Lambda_\ell^{k+1})}{\det(\Lambda_\ell^k)} \le \prod_{\ell=1}^{H} \frac{\det(\Lambda_\ell^{K+1})}{\det(\Lambda_\ell^1)}. \tag{32}$$

By the arithmetic-geometric mean inequality applied to the eigenvalues of $\lambda^{-1}\Lambda_\ell^{K+1}$,

$$\det(\Lambda_\ell^{K+1}) \le \lambda^d \Big(\frac{\operatorname{tr}(\Lambda_\ell^{K+1})}{\lambda d}\Big)^d \le \lambda^d \Big(1 + \frac{MT}{\lambda d}\Big)^d. \tag{33}$$

The last inequality holds due to the definition in (9) with $\|\phi(x_h^{m,t}, a_h^{m,t})\|_2 \le 1$. Since $\Lambda_\ell^1 = \lambda I_d$, we further obtain

$$\frac{\det(\Lambda_\ell^{K+1})}{\det(\Lambda_\ell^1)} \le \Big(1 + \frac{MT}{\lambda d}\Big)^d.$$

Consequently, back to (32), we have

$$|I_{\text{short,tr}}| \le \frac{dH}{\log(\gamma/\alpha)} \log\Big(1 + \frac{MT}{\lambda d}\Big). \tag{34}$$

Put the three classes together. Combining the previous bounds (31) and (34) gives

$$K = |I_{\text{long}}| + |I_{\text{short,tr}}| + |I_{\text{short,term}}| \le \frac{T}{\alpha} + \frac{dH}{\log(\gamma/\alpha)} \log\Big(1 + \frac{MT}{\lambda d}\Big) + 1.$$

Choosing $\alpha = \gamma/2$ yields the claimed bound

$$K \le 1 + \frac{2T}{\gamma} + \frac{dH}{\log 2} \log\Big(1 + \frac{MT}{d\lambda}\Big).$$

□

H On the Statistics-Only Communication Claim of Hsu et al. (2024)

This appendix revisits the logarithmic communication claim in Lemma 4.6 and Appendix B–C of Hsu et al. (2024). In particular, they claim that, in the linear MDP setting, agents can communicate only summary statistics instead of local raw trajectories, resulting in logarithmic dependence on the number of episodes in the communication cost. We explain why this claim is not supported by the update rule analyzed in their work and therefore overstates what their analysis actually establishes.

H.1 Review of the Algorithm in Hsu et al. (2024)

Algorithm 1 of Hsu et al. (2024) provides the common outer framework for parallel data collection, event-triggered synchronization, and dataset aggregation. Combining this framework with the PHE subroutine in their Algorithm 2 yields CoopTS-PHE, whereas combining it with the LMC subroutine in their Algorithm 3 yields CoopTS-LMC. In their original formulations, both algorithms communicate raw transition sets. In Appendix B of the same work, CoopTS is introduced as a closed-form Gaussian representation of the randomized regression updates in the linear setting. Their Proposition B.1 establishes exact equality in distribution between the linear CoopTS-PHE update and the CoopTS update, while Proposition B.2 establishes asymptotic convergence of the linear CoopTS-LMC update to the CoopTS update. These relationships concern the updates when the required data are available; they do not establish that those updates can be implemented using only the proposed summary statistics. The statistics-only communication claim is made through this CoopTS representation. Accordingly, Remark 2 and Section H.2 specifically examine the statistics-only implementation proposed through CoopTS, rather than the original trajectory-sharing implementations of CoopTS-PHE and CoopTS-LMC. We first restate the relevant part of the algorithm (Algorithm 3), using notation aligned with ours.

Algorithm 3 Unified randomized-exploration framework of Hsu et al. (2024)

Initialize $U_h^{\mathrm{ser}} \leftarrow \varnothing$ for all $h \in [H]$ and $U_h^{m,\mathrm{loc}} \leftarrow \varnothing$ for all $m \in [M]$ and $h \in [H]$.
for $t = 1, 2, \ldots, T$ **do**
  for all agents $m \in [M]$ in parallel **do**
    Receive the initial state $x_1^{m,t}$.
    Set $V_{H+1}^{m,t} \leftarrow 0$.
    Update $\{Q_h^{m,t}, V_h^{m,t}\}_{h=1}^H$ via a randomized exploration algorithm (PHE or LMC) with the datasets $\{U_h^m\}_{h=1}^H$.
    for $h = 1, 2, \ldots, H$ **do**
      Select $a_h^{m,t} \in \arg\max_{a \in \mathcal{A}} Q_h^{m,t}(x_h^{m,t}, a)$.
      Observe $(r_h^{m,t}, x_{h+1}^{m,t})$.
      Update $U_h^{m,\mathrm{loc}} \leftarrow U_h^{m,\mathrm{loc}} \cup \{(x_h^{m,t}, a_h^{m,t}, x_{h+1}^{m,t})\}$.
    end for
    if the synchronization condition (36) is triggered **then**
      Set a synchronization flag.
    end if
  end for
  if the synchronization flag is set **then**
    for $h = H, H-1, \ldots, 1$ **do**
      for all agents $m \in [M]$ **do**
        Upload the raw local transition set $U_h^{m,\mathrm{loc}}$ to the server.
      end for
      Update the synchronized dataset $U_h^{\mathrm{ser}} \leftarrow U_h^{\mathrm{ser}} \cup U_h^{\mathrm{loc}}$ with $U_h^{\mathrm{loc}} \leftarrow \bigcup_{m=1}^M U_h^{m,\mathrm{loc}}$.
      Broadcast U_h^{ser} to every agent.
      for all agents $m \in [M]$ **do**
        Reset $U_h^{m,\mathrm{loc}} \leftarrow \varnothing$.
      end for
    end for
  end if
end for

Let $t_s(t)$ be the most recent synchronization episode before episode t. If no synchronization has occurred before episode t, we set $t_s(t) = 0$. For each step $h \in [H]$, define

$$\begin{aligned} U_h^{\mathrm{ser}}(t) &:= \{(x_h^{n,\tau}, a_h^{n,\tau}, x_{h+1}^{n,\tau}) : n \in [M],\ \tau \in [t_s(t)]\}, \\ U_h^{m,\mathrm{loc}}(t) &:= \{(x_h^{m,\tau}, a_h^{m,\tau}, x_{h+1}^{m,\tau}) : \tau = t_s(t)+1, \ldots, t-1\}, \\ U_h^{m,\mathrm{loc},+}(t) &:= U_h^{m,\mathrm{loc}}(t) \cup \{(x_h^{m,t}, a_h^{m,t}, x_{h+1}^{m,t})\}, \\ U_h^{m}(t) &:= U_h^{\mathrm{ser}}(t) \cup U_h^{m,\mathrm{loc}}(t). \end{aligned} \tag{35}$$

At each episode t, agent m performs local updates of the value estimates $Q_h^{m,t}$ and $V_h^{m,t}$ using a randomized exploration method, namely Perturbed-History Exploration (PHE) or Langevin Monte Carlo Exploration (LMC) in Hsu et al. (2024). These updates are based on $U_h^m(t)$, which consists of the most recently synchronized data $U_h^{\mathrm{ser}}(t)$ together with the agent's own unsynchronized local data $U_h^{m,\mathrm{loc}}(t)$. We focus on the case of linear function approximation and state the specific update rule in Section H.2.

The synchronization condition is triggered in episode t if there exist $m \in [M]$ and $h \in [H]$ such that:

$$\log \frac{\det\left(\Lambda_h^{t,\mathrm{ser}} + \Lambda_h^{m,t,\mathrm{loc}}\right)}{\det\left(\Lambda_h^{t,\mathrm{ser}}\right)} \geq \frac{\gamma}{t - t_s(t)}, \tag{36}$$

where

$$\begin{aligned} \Lambda_h^{t,\mathrm{ser}} &:= \sum_{(x_l, a_l, x_l') \in U_h^{\mathrm{ser}}(t)} \phi(x_l, a_l)\phi(x_l, a_l)^\top + \lambda I, \\ \Lambda_h^{m,t,\mathrm{loc}} &:= \sum_{(x_l, a_l, x_l') \in U_h^{m,\mathrm{loc},+}(t)} \phi(x_l, a_l)\phi(x_l, a_l)^\top. \end{aligned} \tag{37}$$

When the synchronization condition is triggered in episode t, each agent uploads its unsynchronized local transition set $U_h^{m,\mathrm{loc}} = U_h^{m,\mathrm{loc},+}(t)$ to the server for each step $h \in [H]$. The server aggregates these local sets as

$$U_h^{\mathrm{loc}} \leftarrow \bigcup_{m=1}^{M} U_h^{m,\mathrm{loc}},$$

updates the synchronized dataset by

$$U_h^{\mathrm{ser}} \leftarrow U_h^{\mathrm{ser}} \cup U_h^{\mathrm{loc}},$$

and broadcasts the updated synchronized dataset U_h^{ser} back to all agents. After synchronization, each agent resets its local dataset $U_h^{m,\mathrm{loc}}$ to the empty set.

H.2 Why the Statistics-Only Communication Claim Is Overstated in the Linear MDP Setting

We now examine whether the statistics-only implementation proposed through CoopTS reproduces its prescribed update. The issue is whether the communicated summaries suffice to reconstruct the particular statistics required by the current update. The covariance matrix can indeed be reconstructed by aggregating local covariance statistics, since its sample contributions do not depend on the value function. In contrast, the response vector requires evaluating historical next states under the updating agent's current value function, whereas the transmitted response statistics were computed using earlier, agent-specific value functions. As shown below, these agent-wise and episode-wise mismatches prevent the proposed statistics-only implementation from reproducing the exact CoopTS update in general.

We now specialize the above randomized-exploration framework to CoopTS under the homogeneous linear-MDP setting considered by Hsu et al. (2024). In this setting, the Q-function at each step h is approximated by a linear function of the feature vector $\phi(x, a)$ with parameter w_h:

$$Q_h(x, a) = w_h^\top \phi(x, a).$$

To estimate the weight vector w_h, Hsu et al. (2024) use a squared loss for Bellman regression. For each agent m, episode t, and step h, the step-h loss is

$$L_h^{m,t}(w) = \sum_{(x_l, a_l, x_l') \in U_h^m(t)} \Big(r_h(x_l, a_l) + V_{h+1}^{m,t}(x_l') - w^\top \phi(x_l, a_l) \Big)^2 + \lambda \|w\|_2^2.$$

The corresponding ridge-regression solution is

$$\widehat{w}_h^{m,t} = (\Lambda_h^{m,t})^{-1} b_h^{m,t}, \tag{38}$$

where

$$\Lambda_h^{m,t} = \sum_{(x_l, a_l, x_l') \in U_h^m(t)} \phi(x_l, a_l) \phi(x_l, a_l)^\top + \lambda I,$$
$$b_h^{m,t} := \sum_{(x_l, a_l, x_l') \in U_h^m(t)} \Big(r_h(x_l, a_l) + V_{h+1}^{m,t}(x_l') \Big) \phi(x_l, a_l).$$

They then add Gaussian noise $\mathcal{N}(\mathbf{0}, \sigma^2 (\Lambda_h^{m,t})^{-1})$ to $\widehat{w}_h^{m,t}$ and obtain the perturbed parameter $\bar{w}_h^{m,t}$. The value estimate is then updated as

$$Q_h^{m,t}(x, a) = \min \left\{ \phi(x, a)^\top \bar{w}_h^{m,t},\, H - h + 1 \right\}^+.$$

The communication-cost discussion in Appendix C of Hsu et al. (2024) suggests that, in the linear setting, for synchronization in episode τ, local trajectory sharing can be replaced by sending only agent-wise summary statistics $\Lambda_h^{m,\tau,\mathrm{loc}}$ in (37) and

$$b_h^{m,\tau,\mathrm{loc}} := \sum_{(x_l, a_l, x_l') \in U_h^{m,\mathrm{loc},+}(\tau)} \Big(r_h(x_l, a_l) + V_{h+1}^{m,\tau}(x_l') \Big) \phi(x_l, a_l)$$

for each agent m. Then the central server calculates

$$\Lambda_h^{\tau,\mathrm{loc}} := \sum_{n=1}^{M} \Lambda_h^{n,\tau,\mathrm{loc}} = \sum_{n=1}^{M} \sum_{(x_l,a_l,x_l')\in U_h^{n,\mathrm{loc},+}(\tau)} \phi(x_l,a_l)\phi(x_l,a_l)^\top$$

and

$$b_h^{\tau,\mathrm{loc}} := \sum_{n=1}^{M} b_h^{n,\tau,\mathrm{loc}} = \sum_{n=1}^{M} \sum_{(x_l,a_l,x_l')\in U_h^{n,\mathrm{loc},+}(\tau)} \Big(r_h(x_l,a_l) + V_{h+1}^{n,\tau}(x_l')\Big)\phi(x_l,a_l)$$

and then sends $\Lambda_h^{\tau,\mathrm{loc}}$ and $b_h^{\tau,\mathrm{loc}}$ back to each agent for local training. However, this statistics-only implementation is not valid as stated. We explain the reason below.

For each episode t, let $\tau_1 < \tau_2 < ... < \tau_j < t$ be all synchronization episodes before episode t. Then, with the statistics-only communication specified above, before updating $\{Q_h^{m,t}, V_h^{m,t}\}_{h=1}^H$ via randomized exploration, each agent m only has $\{(\Lambda_h^{\tau_i,\mathrm{loc}}, b_h^{\tau_i,\mathrm{loc}})\}_{i=1}^j$ and its unsynchronized local transition set $U_h^{m,\mathrm{loc}}(t)$, i.e., the raw local transitions collected after the latest synchronization in episode τ_j and before episode t (defined in (35)).

For each agent m, to update $\{Q_h^{m,t}, V_h^{m,t}\}_{h=1}^H$ in episode t, it needs to compute $\widehat{w}_h^{m,t}$ by (38). Although the statistic $\Lambda_h^{m,t}$ can be calculated by the received summary statistics $\Lambda_h^{\tau_i,\mathrm{loc}}$ and the local unsynchronized transition set $U_h^{m,\mathrm{loc}}(t)$ with

$$\begin{aligned}
\Lambda_h^{m,t} &= \sum_{(x_l,a_l,x_l')\in U_h^m(t)} \phi(x_l,a_l)\phi(x_l,a_l)^\top + \lambda I \\
&= \sum_{(x_l,a_l,x_l')\in U_h^{\mathrm{ser}}(t)} \phi(x_l,a_l)\phi(x_l,a_l)^\top + \sum_{(x_l,a_l,x_l')\in U_h^{m,\mathrm{loc}}(t)} \phi(x_l,a_l)\phi(x_l,a_l)^\top + \lambda I \\
&= \sum_{i=1}^{j} \Lambda_h^{\tau_i,\mathrm{loc}} + \sum_{(x_l,a_l,x_l')\in U_h^{m,\mathrm{loc}}(t)} \phi(x_l,a_l)\phi(x_l,a_l)^\top + \lambda I
\end{aligned}$$

since $U_h^{\mathrm{ser}}(t) = \bigcup_{i=1}^j \bigcup_{n=1}^M U_h^{n,\mathrm{loc},+}(\tau_i)$, the key problem is that $b_h^{m,t}$ depends on the agent-specific current value function $V_{h+1}^{m,t}$ evaluated on all previously collected next states and cannot be calculated by $\{b_h^{\tau_i,\mathrm{loc}}\}_{i=1}^j$ and its local unsynchronized transition set $U_h^{m,\mathrm{loc}}(t)$. To see this, we note that

$$\begin{aligned}
b_h^{m,t} &= \sum_{(x_l,a_l,x_l')\in U_h^m(t)} \Big(r_h(x_l,a_l) + V_{h+1}^{m,t}(x_l')\Big)\phi(x_l,a_l). \\
&= \left(\sum_{(x_l,a_l,x_l')\in U_h^{\mathrm{ser}}(t)} + \sum_{(x_l,a_l,x_l')\in U_h^{m,\mathrm{loc}}(t)}\right)\Big(r_h(x_l,a_l) + V_{h+1}^{m,t}(x_l')\Big)\phi(x_l,a_l) \\
&= \sum_{i=1}^{j}\sum_{n=1}^{M} \sum_{(x_l,a_l,x_l')\in U_h^{n,\mathrm{loc},+}(\tau_i)} \Big(r_h(x_l,a_l) + V_{h+1}^{m,t}(x_l')\Big)\phi(x_l,a_l) \\
&\quad + \sum_{(x_l,a_l,x_l')\in U_h^{m,\mathrm{loc}}(t)} \Big(r_h(x_l,a_l) + V_{h+1}^{m,t}(x_l')\Big)\phi(x_l,a_l).
\end{aligned}$$

Although the last term can be calculated by each agent m locally with its local unsynchronized transition set $U_h^{m,\mathrm{loc}}(t)$, there is a two-fold mismatch between the first term

$$\sum_{n=1}^{M} \sum_{(x_l,a_l,x_l')\in U_h^{n,\mathrm{loc},+}(\tau_i)} \Big(r_h(x_l,a_l) + V_{h+1}^{m,t}(x_l')\Big)\phi(x_l,a_l)$$

and the summary statistic received from the central server:

$$b_h^{\tau_i,\mathrm{loc}} = \sum_{n=1}^{M} \sum_{(x_l,a_l,x_l')\in U_h^{n,\mathrm{loc},+}(\tau_i)} \Big(r_h(x_l,a_l) + V_{h+1}^{n,\tau_i}(x_l')\Big)\phi(x_l,a_l).$$

First, there is an agent-wise value-function mismatch: the uploaded statistic $b_h^{\tau_i,\mathrm{loc}}$ was computed using agent n's value estimate V_{h+1}^{n,τ_i} evaluated at the next states in the corresponding transition set $U_h^{n,\mathrm{loc},+}(\tau_i)$, whereas agent m's update at episode t requires its locally updated value function $V_{h+1}^{m,t}$ evaluated at the next states in the same transition set. Second, there is an episode-wise value-function mismatch: the uploaded statistic $b_h^{\tau_i,\mathrm{loc}}$ was computed at the synchronization episode $\tau_i < t$ using V_{h+1}^{n,τ_i}, whereas the current update uses the value function $V_{h+1}^{m,t}$ at episode t. Due to these two mismatches, agent m cannot compute $b_h^{m,t}$ from the compressed response statistics $b_h^{\tau_i,\mathrm{loc}}$ received from the central server. Therefore, reproducing the correct Bellman response vector $b_h^{m,t}$ in their algorithm still requires access to the uploaded local transition sets from the other agents.

Thus, the claim that trajectory sharing can be replaced by statistics sharing in the algorithms of Hsu et al. (2024) is overstated and is not justified by the existing analysis. The obstruction comes from the local-update structure of their algorithms: the response vector required by agent m at episode t depends on its current local value estimate $V_{h+1}^{m,t}$, whereas the synchronized response statistic was computed using an earlier, agent-specific value estimate V_{h+1}^{n,τ_i}. In contrast, our Fed-LSVI algorithm avoids this issue because value updates are performed globally at synchronization rounds in a stepwise backward manner.

I A Correction to the Communication-Round Analysis of Hsu et al. (2024)

This appendix revisits the proof of the communication-round bound in Lemma 4.6 of Hsu et al. (2024). We identify an issue in the proof given in Appendix C of Hsu et al. (2024). Building on our communication-bound analysis in Appendix G, we also provide a corrected proof in Appendix I.2.

I.1 The Issue in the Original Analysis of Hsu et al. (2024)

In the proof of the communication-round bound in Appendix C of Hsu et al. (2024), the authors let $\sigma_1, \ldots, \sigma_n$ denote the synchronization episodes with $\sigma_0 = 0$, and define $\Delta_i = \sigma_i - \sigma_{i-1}$. They further introduce an auxiliary threshold $\alpha > 0$ and define

$$I_1 = \{i \in [n] : \Delta_i \le \alpha\}, \qquad I_2 = \{i \in [n] : \Delta_i > \alpha\}.$$

The issue arises from the inequality immediately following equation (C.1) in Appendix C of their paper:

$$\sum_{i=1}^{n} \log \frac{\det(\Lambda_{m,h}^{\sigma_i})}{\det(\Lambda_{m,h}^{\sigma_{i-1}})} \ge \sum_{i \in I_1} \frac{\gamma}{\Delta_i}. \tag{39}$$

This step is not valid because it implicitly treats all synchronization episodes as being triggered by the same agent m at the same step h. However, according to the synchronization rule in Hsu et al. (2024) (shown in (36)), the i-th synchronization at the end of episode σ_i only guarantees the existence of a triggering pair (m_i, h_i) for which the determinant ratio is larger than the threshold. This triggering pair may vary across synchronization episodes. Consequently, the synchronization triggers cannot all be assigned to a single fixed (m, h) determinant sum as in (39).

I.2 Corrected Proof of the Communication Complexity

Our analysis of the number of communication rounds in Section G generalizes to the setting of Hsu et al. (2024) and addresses the issue identified above. We present the corrected argument below with the notation defined in Section H.

Set $\lambda = 1$ as in Hsu et al. (2024). Let K denote the total number of communication rounds. We then define $\sigma_1, \ldots, \sigma_K$ as the synchronization episodes, with $\sigma_0 = 0$, and define $\Delta_i = \sigma_i - \sigma_{i-1}$. For an auxiliary threshold $\alpha > 0$, we further define

$$I_1 = \{i \in [K] : \Delta_i \le \alpha\}, \qquad I_2 = \{i \in [K] : \Delta_i > \alpha\}.$$

If the i-th synchronization is triggered by $(m_i, h_i) \in [M] \times [H]$, then the synchronization condition (36) and the monotonicity of determinant over positive semidefinite matrices give

$$\log \frac{\det\left(\Lambda_{h_i}^{\sigma_i+1,\mathrm{ser}}\right)}{\det\left(\Lambda_{h_i}^{\sigma_{i-1}+1,\mathrm{ser}}\right)} \geq \log \frac{\det\left(\Lambda_{h_i}^{\sigma_{i-1}+1,\mathrm{ser}} + \Lambda_{h_i}^{m_i,\sigma_i,\mathrm{loc}}\right)}{\det\left(\Lambda_{h_i}^{\sigma_{i-1}+1,\mathrm{ser}}\right)} \geq \frac{\gamma}{\Delta_i}.$$

Therefore,

$$\begin{aligned}
|I_1|\frac{\gamma}{\alpha} &\leq \sum_{i \in I_1} \frac{\gamma}{\Delta_i} \\
&\leq \sum_{i \in I_1} \log \frac{\det\left(\Lambda_{h_i}^{\sigma_{i-1}+1,\mathrm{ser}} + \Lambda_{h_i}^{m_i,\sigma_i,\mathrm{loc}}\right)}{\det\left(\Lambda_{h_i}^{\sigma_{i-1}+1,\mathrm{ser}}\right)} \\
&\leq \sum_{h=1}^{H} \sum_{i=1}^{K} \log \frac{\det\left(\Lambda_h^{\sigma_i+1,\mathrm{ser}}\right)}{\det\left(\Lambda_h^{\sigma_{i-1}+1,\mathrm{ser}}\right)} \\
&= \sum_{h=1}^{H} \log \frac{\det\left(\Lambda_h^{\sigma_K+1,\mathrm{ser}}\right)}{\det(I_d)} \\
&\leq dH \log\left(1 + \frac{MT}{d}\right) = dHL_T. \qquad (40)
\end{aligned}$$

The last inequality follows from the same determinant upper bound as in (33). Here we set

$$L_T \coloneqq \log\left(1 + \frac{MT}{d}\right).$$

Thus by (40), we have

$$|I_1| \leq \frac{dHL_T}{\gamma}\alpha. \qquad (41)$$

Moreover, for each $i \in I_2$ we have $\Delta_i > \alpha$. Summing the round lengths and using that the total number of episodes per agent is T gives

$$\alpha|I_2| \leq \sum_{i \in I_2} \Delta_i \leq \sum_{i=1}^{K} (\sigma_i - \sigma_{i-1}) \leq T,$$

so

$$|I_2| \leq \frac{T}{\alpha}. \qquad (42)$$

Combining the upper bounds on $|I_1|$ in (41) and $|I_2|$ in (42), we reach

$$K \leq \frac{T}{\alpha} + \frac{dHL_T}{\gamma}\alpha. \qquad (43)$$

Following the parameter choices in Hsu et al. (2024), choose $\gamma = \Theta(T/(dM))$ and $\alpha = T/d$. Then (43) yields the final upper bound for the number of communication rounds:

$$K \leq \mathcal{O}(d + dHML_T).$$