MA-LDP: Regret Minimization in Decentralized Multi-Agent Locally Differentially Private Reinforcement Learning

Anonymous Author
Anonymous Institution

Abstract

This work introduces a local differential privacy (LDP) framework for fully decentralized multi-agent reinforcement learning (MARL). We propose a generic multi-agent LDP (MA-LDP) algorithm that can accommodate any noise-adding mechanism. Our MA-LDP algorithm takes an episodic form, where data from each agent undergoes anonymization through the given noise-adding mechanism in each episode. MA-LDP leverages linear function approximations of transition probabilities and reward functions. We show that the MA-LDP algorithm achieves sub-linear regret across four noise-adding mechanisms – two with bounded support and two with unbounded support. Notably, regret scales super-linearly (not quadratically), with the number of agents, an essential factor in multi-agent systems. Further, we compare these bounded and unbounded support noise-adding mechanisms about their privacy and regret performance. Notably, for well-chosen endpoints of the bounded support mechanism, the MA-LDP algorithm's regret is on par or lower than the noise mechanisms with unbounded support. Finally, we validate our theoretical results on a network MDP with a large state and action spaces. Code.

1 Introduction

The quest to design the differentially private algorithm for Reinforcement Learning (RL) is not just an academic pursuit but a practical necessity (Barrett and Stone, 2015; Li et al., 2022). For example, edge- and endpoint-AI (e.g., smartphones and vehicles) aim to reduce communication latency and preserve data privacy by moving AI closer to users. However, due to the limited capability in edges and endpoints (e.g., sensing, computing, and storage), it is still challenging for a single computing device to perform complicated tasks alone, such as autonomous driving (Yuan et al., 2022).

Despite its importance, the notion of LDP requires extensive consideration within MARL, which poses many extra challenges such as compounding of noise, balancing the trade-off between privacy and learning efficiency, and protecting the privacy of agents' rewards and actions. The LDP in the MARL setting can safeguard sensitive information in many real-life applications ranging from sensitive health data to interactions on networked platforms like social media.

To mitigate these challenges, we propose a novel and versatile fully decentralized MA-LDP algorithm that can handle various noise-adding mechanisms. In particular, we consider four noise-adding mechanisms: unbounded support (Gaussian and Laplace) and bounded support (uniform and bounded Laplace (BL)). We prove that our MA-LDP algorithm preserves data privacy and achieves sub-linear regret for each. Further, we compare MA-LDP algorithms' regret and privacy guarantees across these noise mechanisms, particularly examining "whether mechanisms with bounded support can provide the same or better privacy and regret guarantees than those with unbounded support.". We answer this question affirmatively and prove that for the suitably chosen distribution parameters and the support endpoint, the regret of the BL mechanism is on par or lower than the Laplace mechanism. Our major contributions are:

(a) Novel MA-LDP Algorithm. We present a fully decentralized MA-LDP algorithm to ensure the LDP for MARL. Our MA-LDP algorithm is versatile because it can handle any noise-adding mechanism.

(b) Privacy and Regret Guarantees. We consider four noise-adding mechanisms: two with unbounded and two with bounded support. The MA-LDP algorithm preserves the LDP property for each of these

Preliminary work. Under review by AISTATS 2025. Do not distribute.

mechanisms. The regret is also sub-linear in the number of episodes for all the noise-adding mechanisms.
(c) Comparison of noise-distribution support. We compare the unbounded support noise mechanisms with bounded support mechanisms. The regret of the bounded Laplace (BL) mechanism is on par or lower than the Laplace mechanism for a certain relationship between its end points of support and its parameters.

1.1 Related Works

This section gives a brief outline of the work closely related to ours. The idea of DP is first introduced in Dwork et al. (2006). However, DP risks data leakage and is vulnerable to membership influence attacks Shokri et al. (2017). Hence, a more vital notion of 'Locally DP' is introduced Kasiviswanathan et al. (2011); Duchi et al. (2013). Under LDP, users send privatized data to the server, and each user maintains their sensitive data. Privacy amplification in machine learning is studied in Cyffers and Bellet (2022). The notion of optimal noise-adding mechanism is available in Geng and Viswanath (2015). Apart from learning and decision-making, DP is also standard in other fields of science, including Hu and Fang (2022); Duchi et al. (2013). In particular, Hu and Fang (2022) study the K armed bandit problem with a distributional trust model of the LDP that guarantees privacy without a trustworthy server. Recently, Jia et al. (2020); Jin et al. (2020) proposed a single-agent RL scheme with linear function approximation of the transition probability function. Garcelon et al. (2021) designed the first LDP for tabular RL. However, tabular RL algorithms generally suffer from computational inefficiency. Recently, Liao et al. (2021) used the UCRL-VTR algorithm of Jia et al. (2020) and incorporated the notion of LDP into it. However, little or no work talks about LDP for multi-agent systems. The only work we found is for the LDP in multi-agent distributed optimization by Dobbe et al. (2018). To this, we propose the LDP for MARL, design the MA-LDP algorithm, and prove its sub-linear regret property. We also explore noise-adding mechanisms with bounded support and compare their regret with conventional mechanisms.

2 Preliminaries

Local Differential Privacy Dwork et al. (2006): LDP is a well-known notion for the data privacy preserving method in which each user adds randomness or noise to its sensitive information before sharing it with the server. Formally, for given privacy parameters $\epsilon, \delta > 0$, the mechanism $\mathcal{M}$ is (ϵ, δ)-LDP if for any two neighboring data sets d, d', auxiliary input $\mathbf{aux}$, and an outcome $o \in \mathbb{R}$, we have that

$$\mathbb{P}(\mathcal{M}(\mathbf{aux}, d) = o) \leq e^{\epsilon}\mathbb{P}(\mathcal{M}(\mathbf{aux}, d') = o) + \delta \quad (1)$$

Here, the parameters ϵ and δ control the privacy and confidence level, respectively. $c(o; \mathcal{M}, \mathbf{aux}, d, d') := \frac{\mathbb{P}(\mathcal{M}(\mathbf{aux},d)=o)}{\mathbb{P}(\mathcal{M}(\mathbf{aux},d')=o)}$ is the *privacy loss* that quantifies the impact on the outcome o due to the substitution of data d with d'. So, from equation 1, (ϵ, δ)-LDP ensures that for the neighboring data points d and d', the absolute privacy loss is limited to ϵ with a probability of at least $1 - \delta$.

Multi-Agent MDP: Let $N = \{1, 2, \ldots, n\}$ be the set of agents. An instance of a multi-agent time-inhomogeneous episodic Markov Decision Process (MDP) is defined as $(N, \mathcal{S}, \{\mathcal{A}^i\}_{i\in N}, H, \{r_h^i\}_{i\in N, h\in H}, \{\mathbb{P}_h\}_{h\in H}, \{\mathcal{G}_t\}_{t\geq 0})$. Here $\mathcal{S}$ denotes the finite set of global states. Each agent $i \in N$ independently takes an action $\boldsymbol{a}^i(\boldsymbol{s})$ from the set of local actions $\mathcal{A}^i(\boldsymbol{s})$ available to agent $i \in N$ when the system is in global state $\boldsymbol{s} \in \mathcal{S}$. Therefore, the global action in the state $\boldsymbol{s}$ is $\boldsymbol{a}(\boldsymbol{s}) = (\boldsymbol{a}^1(\boldsymbol{s}), \boldsymbol{a}^2(\boldsymbol{s}), \ldots, \boldsymbol{a}^n(\boldsymbol{s}))$, where $\boldsymbol{a}^i(\boldsymbol{s}) \in \mathcal{A}^i(\boldsymbol{s})$. Each agent $i \in N$ realizes a deterministic local reward $r_h^i(\boldsymbol{s}, \boldsymbol{a}) \in [0, 1]$ at each intermediate stage h in the planning horizon H. Notably, the reward for agent i depends on the global action and state. This reward is a private information to agent $i \in N$ and remains unknown to other agents. $\mathbb{P}_h(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a})$ denotes the probability of transitioning to state $\boldsymbol{s}'$ once the action $\boldsymbol{a}$ in taken in the state $\boldsymbol{s}$ at stage h. Let K be the total number of episodes, each with planning horizon H, so the total number of interactions is $T = KH$. For each $t = kh$, $\mathcal{G}_t$ denotes the time-varying network that enables information sharing among agents at time t in our decentralized setup.

Value functions: In decentralized MARL, we aim to learn an optimal policy $\pi = \{\pi_h\}_{h=1}^H$ such that the following global state-action value function is maximized.

$$Q_h^{\pi}(\boldsymbol{s}, \boldsymbol{a}) = \bar{r}_h(\boldsymbol{s}, \boldsymbol{a}) + \mathbb{E}_{\pi}\Big[\sum_{\tau=h+1}^{H} \bar{r}_{\tau}(\boldsymbol{s}_{\tau}, \pi_{\tau}(\boldsymbol{s}_{\tau}))\Big], \quad (2)$$

where $\bar{r}_h = \frac{1}{n}\sum_{i\in N} r_h^i$ is the global reward function, and $\boldsymbol{s}_h = \boldsymbol{s}$, $\boldsymbol{a}_h = \boldsymbol{a}$ and $\boldsymbol{s}_{h+1} \sim \mathbb{P}_h(\cdot|\boldsymbol{s}_h, \boldsymbol{a}_h)$. Moreover, let $V_h^{\pi}(\boldsymbol{s}) = Q_h^{\pi}(\boldsymbol{s}, \pi_h(\boldsymbol{s}))$ be the global state value function at stage h. Since $\bar{r}_h^i(\cdot, \cdot)$ is a bounded function, both $V_h^{\pi}(\cdot)$ and $Q_h^{\pi}(\cdot, \cdot)$ are also bounded.

3 LDP for Decentralized MARL

We begin by defining the multi-agent LDP for decentralized MARL. This definition extends the single agent LDP introduced in Kasiviswanathan et al.

(2011); Duchi et al. (2013). It is important to note that in MA-LDP settings, the term "user" (episode) is distinct from "agents" – a distinction that prevents potential confusion in this multi-agent setup. The term 'server' refers to the entity or system that collects and processes data from multiple users while preserving their privacy. MA-LDP ensures the privacy of the sensitive information of any agent $i \in N$ with the server; thus, the server is agnostic to the sensitive data.

Definition 1 (MA-LDP). *For the given privacy parameters $\epsilon, \delta \geq 0$, a randomized mechanism $\mathcal{M}$ preserves (ϵ, δ)-MA-LDP if, for any two users u, u', and their corresponding datasets $\boldsymbol{D}_u = (D_u^1, D_u^2, \cdots D_u^n) \in \mathcal{U}$ and $\boldsymbol{D}_{u'} = (D_{u'}^1, D_{u'}^2, \ldots, D_{u'}^n) \in \mathcal{U}$, the following condition holds for all $U \in \mathcal{U}$*

$$\mathbb{P}(\mathcal{M}(\boldsymbol{D}_u) \in U) \leq e^{\epsilon}\mathbb{P}(\mathcal{M}(\boldsymbol{D}_{u'}) \in U) + \delta, \quad (3)$$

where for each agent $i \in N$, D_u^i and $D_{u'}^i$ differ in exactly one component.

Apart from the LDP, a key challenge in decentralized MARL is to preserve the agent's reward and action privacy. To address this, we propose the following:

Privacy of agent's reward. Recall the state-action value function Q as defined in equation 2 require globally averaged reward $\bar{r}_h(\cdot, \cdot)$. However, none of the agents know this averaged reward in our decentralized model. So, in our decentralized setup, each agent maintains an estimate of $\bar{r}_h(\cdot, \cdot)$ by parameterizing it.

Assumption 1. *There exists a $\boldsymbol{w}^\star \in \mathbb{R}^p$ such that $\bar{r}(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{w}^\star) = \langle \psi(\boldsymbol{s}, \boldsymbol{a}), \boldsymbol{w}^\star \rangle$, for all $(\boldsymbol{s}, \boldsymbol{a})$. Further, the feature matrix Ψ with $[\psi_m(\boldsymbol{s}, \boldsymbol{a})]^\top$ as its m-th column, has full column rank.*

To obtain the true parameters $\boldsymbol{w}^\star$ we can minimize the following objective: $\min_{\boldsymbol{w}} \mathbb{E}_{\boldsymbol{s},\boldsymbol{a}}[\bar{r}_h(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{w}) - \bar{r}_h(\boldsymbol{s}, \boldsymbol{a})]^2$. This optimization problem is equivalently characterized (same stationary points) as $\min_{\boldsymbol{w}} \sum_{i=1}^n \mathbb{E}_{\boldsymbol{s},\boldsymbol{a}}[\bar{r}_h(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{w}) - r_h^i(\boldsymbol{s}, \boldsymbol{a})]^2$. Based on this, each agent updates the reward function parameters as follows:

$$\widetilde{\boldsymbol{w}}_{k,h}^i \leftarrow \boldsymbol{w}_{k,h}^i + \gamma_{k,h} \cdot [r_h^i - \bar{r}_h(\boldsymbol{w}_{k,h}^i)] \cdot \nabla_{\boldsymbol{w}} \bar{r}_h(\boldsymbol{w}_{k,h}^i)$$
$$\boldsymbol{w}_{k+1,h}^i = \sum_{j \in N} l_{k,h}(i,j) \cdot \widetilde{\boldsymbol{w}}_{k,h}^j, \quad (4)$$

where $l_{k,h}(i,j)$ is the (i,j)-th entry of the consensus matrix $L_{k,h}$ obtained using communication network $\mathcal{G}_{k,h}$ in the stage h of the k-th episode. $\gamma_{k,h}$ is the step-size satisfying $\sum_{k,h} \gamma_{k,h} = \infty$ and $\sum_{k,h} \gamma_{k,h}^2 < \infty$. The consensus matrix satisfies the standard assumption in Appendix A (Zhang et al., 2018; Trivedi and Hemachandra, 2022, 2023).

Linear function approximations of transition probability. Apart from the reward function parameterization, we also assume that the transition probabilities $\mathbb{P}_h(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a})$ are written as the linear mixture of given basis functions. This assumption is common in many RL works (Min et al., 2022; Vial et al., 2022; Liao et al., 2021), and is an essential requirement while proving the sub-linear regret.

Assumption 2. *There exists a $\boldsymbol{\theta}_h^\star \in \mathbb{R}^{nd}$ such that $\mathbb{P}_h(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a}) = \langle \phi(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a}), \boldsymbol{\theta}_h^\star \rangle$, $\quad \forall \boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}'$.*

Using above assumption, for any function $V : \mathcal{S} \to [0, H]$, we define the cost-to-go function as

$$\mathbb{P}_h V(\boldsymbol{s}, \boldsymbol{a}) = \sum_{\boldsymbol{s}' \in \mathcal{S}} \mathbb{P}_h(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a}) V(\boldsymbol{s}') = \langle \phi_V(\boldsymbol{s}, \boldsymbol{a}), \boldsymbol{\theta}_h^\star \rangle,$$

where $\phi_V(\boldsymbol{s}, \boldsymbol{a}) = \sum_{\boldsymbol{s}' \in \mathcal{S}} \phi(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a}) V(\boldsymbol{s}')$, are the cost-to-go features such that $||\phi_V(\boldsymbol{s}, \boldsymbol{a})||_2 \leq H$.

Modified Bellman Optimality Equation. Let $V_h^i(\cdot)$ and $Q_h^i(\cdot, \cdot)$ be an estimate of the global $V_h(\cdot)$ and $Q_h(\cdot, \cdot)$, respectively made by an agent $i \in N$. Further, let $Q_h^{\star,i}$ and $V_h^{\star,i}$ are the optimistic estimators of these value functions, respectively. Using this, the modified Bellman optimality equation for any $(\boldsymbol{s}, \boldsymbol{a})$ and for all agent $i \in N$ can be written as

$$Q_h^{\star,i}(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{w}_{k,h}^i) = \bar{r}_h(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{w}_{k,h}^i) + \mathbb{P}_h V_{h+1}^{\star,i}(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{w}_{k,h}^i);$$
$$V_{h+1}^{\star,i}(\boldsymbol{s}; \boldsymbol{w}_{k,h}^i) = \max_{\boldsymbol{a} \in \mathcal{A}} Q_h^{\star,i}(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{w}_{k,h}^i); V_{H+1}^{\star,i}(\boldsymbol{s}; \boldsymbol{w}_{k,h}^i) = 0.$$

Since $\bar{r}_h(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{w}_{k,h}^i)$, $Q_h^{\star,i}(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{w}_{k,h}^i)$ and $V_h^{\star,i}(\boldsymbol{s}; \boldsymbol{w}_{k,h}^i)$ are continuous functions of $\boldsymbol{w}_{k,h}^i$, and as $k \to \infty$, $\boldsymbol{w}_{k,h}^i \to \boldsymbol{w}^\star$ *a.s* for all $i \in N$ (see lemma 1), we have

$$Q_h^{\star,i}(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{w}_{k,h}^i) \to Q_h^\star(\boldsymbol{s}, \boldsymbol{a}); V_h^{\star,i}(\boldsymbol{s}; \boldsymbol{w}_{k,h}^i) \to V_h^\star(\boldsymbol{s}) \quad (5)$$

as $k \to \infty$, where $Q_h^\star(\boldsymbol{s}, \boldsymbol{a}) = \bar{r}_h(\boldsymbol{s}, \boldsymbol{a}) + \mathbb{P}_h V_h^\star(\boldsymbol{s}, \boldsymbol{a})$; and $V_h^\star(\boldsymbol{s}) = \max_{\boldsymbol{a} \in \mathcal{A}} Q_h^\star(\boldsymbol{s}, \boldsymbol{a})$.

Performance metric: We take regret as the performance metric. It is defined as the difference between a fully privatized optimal value (baseline) and the value of our MA-LDP algorithm. Note that this difference is more appropriate than using a non-private mechanism as the baseline; hence, it is a natural choice in our setup. Formally, it is defined as:

Definition 2 (Regret). *The total expected regret in K episodes is defined as*

$$R_K = \sum_{k=1}^{K} \left(\frac{1}{n} \sum_{i \in N} \{V_1^{\star,i}(\boldsymbol{s}_1^k) - V_1^i(\boldsymbol{s}_1^k)\} \right). \quad (6)$$

We use $V_1^{\star,i}(\boldsymbol{s}_1^k)$ instead of $V_1^{\star}(\boldsymbol{s}_1^k)$ in the above regret definition. This is because of equation 5; the converged $V_1^{\star}(\cdot)$ being independent of agents is a good sign of a decentralized algorithm. We now provide our decentralized MA-LDP algorithm that preserves user data privacy and achieves sub-linear regret.

4 MA-LDP Algorithm

The MA-LDP algorithm we provide draws inspiration from the UCRL-VTR algorithm (Jia et al., 2020) and incorporates some aspects from the UCRL-VTR-LDP algorithm (Liao et al., 2021). Notably, our algorithm is flexible in terms of the noise-adding mechanism. In sections 6 and 7, we analyze the effect of various noise-adding mechanisms on regret. We now present a brief outline of the MA-LDP algorithm.

Algorithm 1 MA-LDP

Require: Parameters ϵ, δ; parameter η, $\boldsymbol{w}_{0,0}^i = 0$, $\forall\, i \in N$; consensus matrices $L_{k,h}$.
for user $k = 1, \ldots, K$ **do**
  From server receive $I_{s,k}^i = \{\Sigma_{k,1}^i, \ldots, \Sigma_{k,H}^i, \hat{\boldsymbol{\theta}}_{k,1}^i, \ldots, \hat{\boldsymbol{\theta}}_{k,H}^i\}$ for each agent $i \in N$
  Call **User sub-routine** with information $I_{s,k}^i$
  Set $D_k^i = \{\Delta\Lambda_{k,1}^i, \ldots, \Delta\Lambda_{k,H}^i, \Delta u_{k,1}^i, \ldots, \Delta u_{k,H}^i\}$ for each agent $i \in N$
  Send $\mathbf{D}_k = (D_k^1, D_k^2, \ldots, D_k^n)$ to the server
  Call **Server sub-routine** with $\mathbf{D}_k$ and obtain $I_{s,k+1}^i$ for each agent $i \in N$
  Update $\boldsymbol{w}_{k+1,h}^i = \sum_{j \in N} l_{k,h}(i,j)\widetilde{\boldsymbol{w}}_{k,h}^j, \forall h \in [H]$
end for

(Lines 1-6, Algo 2). For every stage $h \in [H]$ and for every agent $i \in N$, the local user $k = 1$ receives $\Lambda_{1,h}^i = \Sigma_{1,h}^i = \lambda\mathbf{I}$; $\widehat{\boldsymbol{\theta}}_{1,h}^i = \mathbf{0}_{nd}$. Using this information $\Lambda_{k,h}^i$ and $u_{k,h}^i$ for each agent $i \in N$, the local user k employs a backward induction algorithm with an extra UCB bonus term to obtain the following optimistic estimate of the optimal state-action value function

$$Q_{k,h}^i(\cdot,\cdot) \leftarrow \min\{\bar{r}_h(\boldsymbol{w}_{k,h}^i) + \beta_{k,h}\|\Sigma_{k,h}^{i^{-1/2}}\phi_{V_{k,h+1}^i}(\cdot,\cdot)\|_2 + \langle\hat{\boldsymbol{\theta}}_{k,h}^i, \phi_{V_{k,h+1}^i}(\cdot,\cdot)\rangle, H+1-h\}. \quad (7)$$

(Lines 7-12, Algo 2). Each agent $i \in N$ observes the global state $\boldsymbol{s}_{k,h}$ and takes an action via current optimistic estimate $Q_{k,h}^i$. Agents use the max min strategy to select the action, optimizing their actions while accounting for potential worst-case actions by others. Further, each agent maintains intermediate reward function parameters, $\widetilde{\boldsymbol{w}}_{k,h}^i$ for each stage $h \in [H]$.

(Lines 13-18, Algo 2). Once an action is taken, a new state is generated according to the unknown distribution $\mathbb{P}_h(\cdot|\boldsymbol{s}_h, \boldsymbol{a}_h)$. To estimate the true transition

Algorithm 2 User/Episode sub-routine

for $i = 1, \ldots, n$ **do**
  for $h = H, \ldots, 1$ **do**
    Update $Q_{k,h}^i(\cdot,\cdot)$ as in equation 7
    $V_{k,h}^i \leftarrow \max_{a^i \in \mathcal{A}^i} Q_{k,h}^i(\cdot, a^i, \boldsymbol{a}_{k,h}^{-i})$
  end for
end for
Receive the initial state $s_{k,1}$
for $h = 1, \ldots, H$ **do**
  for $i = 1, \ldots, n$ **do**
    Take action $\boldsymbol{a}_{k,h}^i \leftarrow \arg\max_{\boldsymbol{a} \in \mathcal{A}^i} \min_{\boldsymbol{a}^{-i} \in \mathcal{A}^{-i}} Q_{k,h}^i(\boldsymbol{s}_{k,h}, \boldsymbol{a}, \boldsymbol{a}^{-i})$
    Update $\widetilde{\boldsymbol{w}}_{k,h}^i$ according to equation 4
  end for
  Set $\boldsymbol{a}_{k,h} = (\boldsymbol{a}_{k,h}^1, \boldsymbol{a}_{k,h}^2, \ldots, \boldsymbol{a}_{k,h}^n)$ and get $s_{k,h+1}$
  for $i = 1, \ldots, n$ **do**
    $\Delta\widetilde{\Lambda}_{k,h}^i = \phi_{V_{k,h+1}^i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})\phi_{V_{k,h+1}^i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})^\top$
    $\Delta\widetilde{u}_{k,h}^i = \phi_{V_{k,h+1}^i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) \quad V_{k,h+1}^i(s_{k,h+1})$
    $\Delta\Lambda_{k,h}^i \leftarrow \Delta\widetilde{\Lambda}_{k,h}^i + \mathbf{W}_{k,h}^i$
    $\Delta u_{k,h}^i \leftarrow \Delta\widetilde{u}_{k,h}^i + \boldsymbol{\xi}_{k,h}^i$
    $\{\mathbf{W}_{k,h}^i$ and $\boldsymbol{\xi}_{k,h}^i$ are chosen suitably$\}$
  end for
end for

probability parameters, each agent must send some information to the server. In the MA-LDP algorithm, data shared with the server from each agent is obtained via the ridge regression-based minimization of the transition probability parameters. Thus, for each agent $i \in N$, and the stage h of the episode k, the server requires the following sensitive information

$$\Delta\widetilde{\Lambda}_{k,h}^i = \phi_{V_{k,h+1}^i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})\phi_{V_{k,h+1}^i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})^\top, \quad (8)$$

$$\Delta\widetilde{u}_{k,h}^i = \phi_{V_{k,h+1}^i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})V_{k,h+1}^i(s_{k,h+1}). \quad (9)$$

However, we privatize the above sensitive information using noise-adding mechanisms before sending it to the server. In particular, to the true information $\Delta\widetilde{\Lambda}_{k,h}^i$, we add the noise matrix $\mathbf{W}_{k,h}^i$, where each entry of the matrix $\mathbf{W}_{k,h}^i$ is drawn according to the noise adding mechanism's distribution. Similarly, to the true information $\Delta\widetilde{u}_{k,h}^i$, we add a noise vector $\boldsymbol{\xi}_{k,h}^i$ with each entry of $\boldsymbol{\xi}_{k,h}^i$ drawn from the corresponding noise distribution. Let $\Delta\Lambda_{k,h}^i$ and $\Delta u_{k,h}^i$ be the anonymized information shared to the server, i.e.,

$$\Delta\Lambda_{k,h}^i \leftarrow \Delta\widetilde{\Lambda}_{k,h}^i + \mathbf{W}_{k,h}^i; \; \Delta u_{k,h}^i \leftarrow \Delta\widetilde{u}_{k,h}^i + \boldsymbol{\xi}_{k,h}^i. \quad (10)$$

(Algo 3). The server collects this anonymized information and uses it to update $\Sigma_{k,h}^i$ and $u_{k,h}^i$, thereby providing the next estimate of the transition probability parameters. In an ideal situation, the server should use only the anonymized information and estimate

the true transition probability parameters. However, adding noise might not preserve the positive semidefinite (PSD) property. Therefore, we introduce a matrix shift by adding $\eta\mathbf{I}$. Apart from executing a policy that uses an optimistic estimator, each agent $i \in N$ also updates the $\Sigma^i_{k,h}$ and $u^i_{k,h}$. These updates are used to estimate the actual model parameters $\hat{\boldsymbol{\theta}}^i_{k+1,h}$, inspired by the minimizer of a regularized linear regression problem similar to Zhou et al. (2021).

Algorithm 3 Server sub-routine

1: **for** $h = 1, \ldots, H$ **do**
2: **for** $i = 1, \ldots, n$ **do**
3: $\Lambda^i_{k+1,h} \leftarrow \Lambda^i_{k,h} + \Delta\Lambda^i_{k,h}$
$u^i_{k+1,h} \leftarrow u^i_{k,h} + \Delta u^i_{k,h}$
4: $\Sigma^i_{k+1,h} \leftarrow \Lambda^i_{k+1,h} + \eta I$
$\hat{\boldsymbol{\theta}}^i_{k+1,h} \leftarrow (\Sigma^i_{k+1,h})^{-1} u^i_{k+1,h}$
5: **end for**
6: **end for**

(Line 8, Algo 1). Finally, to preserve the privacy of the reward of the i^{th} agent, we update the parameters $\boldsymbol{w}^i$ associated to the reward functions. We employ a stochastic approximation-based method that prioritizes using the most recent parameters of the global reward function. These updates are directed through the consensus matrix.

5 Main Results

Let $d(\boldsymbol{s})$ be the stationary distribution of the Markov chain $\{\boldsymbol{s}_t\}_{t\geq 0}$ under the policy π and $D^{\boldsymbol{s},\boldsymbol{a}}$ be the diagonal matrix with $d(\boldsymbol{s}) \cdot \pi(\boldsymbol{s}, \boldsymbol{a})$ as diagonal entries. We have the following lemma for the reward function parameter convergence (Zhang et al., 2018; Trivedi and Hemachandra, 2022, 2023).

Lemma 1. *Under assumption 1 for the sequence* $\{\boldsymbol{w}^i_{k,h}\}$, *we have* $lim_k\ \boldsymbol{w}^i_{k,h} = \boldsymbol{w}^\star$ *a.s. for each agent* $i \in N$, *and for all* $h \in [H]$, *where* $\boldsymbol{w}^\star$ *is unique solution to* $\Psi^\top D^{s,a}(\Psi \boldsymbol{w}^\star - \bar{r}) = 0$.

The detailed proof is in Appendix A. Next, we show that in our MA-LDP algorithm, every agent uses an optimistic estimator of the state-action value function to identify the best action for each step.

Lemma 2. *Let* $Q^i_{k,h}$ *and* $V^i_{k,h}$ *be the estimate of the global state-action value and global state value functions by agent* $i \in N$. *Then, for any pair* $(\boldsymbol{s}, \boldsymbol{a}, k, h) \in \mathcal{S} \times \mathcal{A} \times [K] \times [H]$, *we have* $Q^{\star,i}_h(\boldsymbol{s}, \boldsymbol{a}) \leq Q^i_{k,h}(\boldsymbol{s}, \boldsymbol{a})$ *and* $V^{\star,i}_h(\boldsymbol{s}) \leq V^i_{k,h}(\boldsymbol{s})$.

The proof of this lemma is available in Appendix B of the SM and relies on the principle of Mathematical induction over h. Moreover, we also require the following bound on the model parameter estimates.

Lemma 3. *If* $\eta = 1$, *then for any fixed policy* π *and all pairs* (s, a, h, k), *with probability at least* $1 - \alpha/2$, *we have* $||(\Sigma^i_{k,h})^{1/2}(\hat{\boldsymbol{\theta}}^i_{k,h} - \boldsymbol{\theta}^\star_h)|| \leq \beta_k,\ \forall\ i \in N$.

The proof of this theorem is available in Appendix C. The above lemma ensures that the model parameters obtained from the MA-LDP algorithm converge to the true model parameters $\boldsymbol{\theta}^\star$ within the confidence radius β_k. We want to emphasize that while updating $Q^i_{k,h}$, we use the most recent available reward function parameters $w^i_{k,h}$ that converges to the same $w^\star$ for all agents $i \in N$ almost surely (Lemma 1). Nevertheless, as $k \to \infty$, $Q^{\star,i}_{k,1}$ converges to the same $Q^\star$ for all agents $i \in N$, which is the true optimal value. We now present the regret bound of the MA-LDP algorithm.

Theorem 1 (Informal). *The regret of the MA-LDP algorithm in K episodes is given by*

$$R_K \leq H\beta_K\sqrt{2ndK\log(1 + K/\lambda)} + 4H\sqrt{2KH\log(2H/\alpha)}. \quad (11)$$

The proof follows by utilizing lemmas 1, 2, 3, and using various concentration inequalities. The precise statement and the proof details are given in Appendix E. Since the MA-LDP algorithm's regret depends on the noise-adding mechanism via β_k (Lemma 3), we now explore different noise-adding mechanisms and address the following two questions for each.
Q1. What privacy guarantees does our MA-LDP algorithm provide for each noise-adding mechanism?
Q2. What are the corresponding regret bounds associated with these noise-adding mechanisms?

6 Gaussian and Laplace Noise Adding Mechanisms (Unbounded Support)

First, we consider two well-known unbounded support noise-adding mechanisms - Gaussian and Laplace.

6.1 Gaussian Noise Adding Mechanism

For the Gaussian mechanism each entry (l, m) of $\mathbf{W}^i_{k,h}$ is sampled from the Gaussian distribution $\mathcal{N}(0, \sigma^2)$ such that $\mathbf{W}^i_{k,h}$ remains symmetric. Moreover, to $u^i_{k,h}$ we add $\boldsymbol{\xi}^i_{k,h}$ sampled from $\mathcal{N}(\mathbf{0}_{nd}, \sigma^2\mathbf{I}_{nd\times nd})$ distribution. The following theorem guarantees the (ϵ, δ)-MA-LDP privacy of the MA-LDP algorithm for the Gaussian mechanism.

Theorem 2. *If we chose the parameter σ of the Gaussian distribution $\mathcal{N}(0, \sigma^2)$ such that $\sigma = 4H^3\sqrt{2\log(2.5H/\delta)}/\epsilon$, then the MA-LDP algorithm with Gaussian mechanism preserves (ϵ, δ)-MA-LDP privacy.*

The proof of this theorem is given in Appendix D.1. Next, we show that the MA-LDP algorithm with the Gaussian mechanism achieves a sub-linear regret.

Theorem 3. *Consider the Gaussian noise-adding mechanism with parameter σ as in Theorem 2. For any user k, with probability at least $1-\alpha$, the total regret of MA-LDP algorithm in the first T steps is at most $\widetilde{\mathcal{O}}((nd)^{5/4}H^{7/4}T^{3/4}\log(ndT/\alpha)(\log(H/\delta))^{1/4}\sqrt{1/\epsilon})$.*

The proof follows by identifying suitable β_k and substituting it in Theorem 1. For proof, see Appendix E.1.

6.2 Laplace Noise Adding Mechanism

For the Laplace noise mechanism, each entry of $\mathbf{W}^i_{k,h}$ and $\boldsymbol{\xi}^i_{k,h}$ are sampled from the Laplace distribution $\mathcal{L}(b)$ with distribution parameter b.

Theorem 4. *If the parameter of the Laplace distribution is set to $b = 4H^3\sqrt{nd}/\epsilon$, then the MA-LDP algorithm with Laplace mechanism preserves $(\epsilon, 0)$-MA-LDP privacy.*

The proof is deferred to Appendix D.2. The regret bound of the MA-LDP algorithm with the Laplace mechanism is given below.

Theorem 5. *Consider the Laplace noise-adding mechanism with parameter b as in Theorem 4. For any user k, with probability at least $1-\alpha$, the total regret of MA-LDP algorithm in the first T steps is at most $\widetilde{\mathcal{O}}((nd)^{5/4}H^{7/4}T^{3/4}\log(ndT/\alpha)\sqrt{1/\epsilon})$.*

The proof follows by identifying β_k for the Laplace noise mechanism and substituting it in the regret bound given in Theorem 1. The details are given in Appendix E.2.

7 Uniform and Bounded Laplace Noise Adding Mechanisms

We now delve into noise mechanisms with bounded support. In particular, we consider two noise-adding mechanisms – uniform and bounded Laplace.

The bounded support noise-adding mechanisms are motivated by the fact that unbounded mechanisms may produce noise values of arbitrary magnitude, raising questions about their effects on privacy and regret. We aim to investigate whether restricting the noise to bounded support impacts privacy and regret, and if so, how. If both privacy and regret remain unaffected, injecting excessive noise may be unnecessary. We show that this is indeed the case. For specific conditions on distribution parameters and bounded mechanism support, our MA-LDP algorithm achieves equivalent regret results, which differ only from constants. This stands out as a significant contribution.

7.1 Uniform Noise Adding Mechanism

In the uniform noise adding mechanism each entry of $\mathbf{W}^i_{k,h}$ and $\boldsymbol{\xi}^i_{k,h}$ are sampled from uniform distribution $\mathcal{U}[-a, a]$ distribution.

Theorem 6. *If we choose $a = 4H^3\sqrt{\log(2H/\delta)}$ for the uniform distribution $\mathcal{U}[-a, a]$, then MA-LDP algorithm with uniform mechanism preserves $(0, \delta)$-MA-LDP privacy.*

The proof of this theorem is deferred to Appendix D.3. The regret of the MA-LDP algorithm with a uniform mechanism is given in the following theorem. For proof, see Appendix E.3.

Theorem 7. *Consider the uniform noise-adding mechanism with parameter a as in Theorem 6. For any user k, with probability at least $1-\alpha$, the total regret of MA-LDP algorithm in the first T steps is at most $\widetilde{\mathcal{O}}(n^{5/4}d^{5/4}H^{7/4}\ T^{3/4}\log(ndT/\alpha)(\log(H/\delta))^{1/4})$.*

7.2 Bounded Laplace Noise Adding Mechanism

We now introduce a new noise-adding mechanism called bounded Laplace (BL), featuring bounded support. The BL mechanism is obtained by confining the support of Laplace distribution to interval $[-B, B]$, where B is a specified positive constant. The probability density of BL distribution with parameter b is

$$f_{\mathcal{BL}}(x; b) = \begin{cases} \frac{\exp(-|x|/b)}{2b(1-\exp(-B/b))}, & \forall\ x \in [-B, B] \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

The variance of the bounded Laplace distribution is given by $\zeta = 2b^2(1-\exp(\frac{-B}{b}))^{-1}$ - κ, where $\kappa = ((B+b)^2+b^2)\exp(\frac{-B}{b})(1-\exp(\frac{-B}{b}))^{-1}$. For BL mechanism, each entry of $\mathbf{W}^i_{k,h}$ and $\boldsymbol{\xi}^i_{k,h}$ is sampled from $\mathcal{BL}(b; B)$

Mechanism	Support	Privacy	Order of Regret
Gaussian	Unbounded	(ϵ, δ)	$\widetilde{\mathcal{O}}((nd)^{5/4}H^{7/4}T^{3/4}\log(ndT/\alpha)(\log(H/\delta))^{1/4}\sqrt{1/\epsilon})$
Laplace	Unbounded	$(\epsilon, 0)$	$\widetilde{\mathcal{O}}((nd)^{5/4}H^{7/4}T^{3/4}\log(ndT/\alpha)\sqrt{1/\epsilon})$
Uniform	Bounded	$(0, \delta)$	$\widetilde{\mathcal{O}}((nd)^{5/4}H^{7/4}T^{3/4}\log(ndT/\alpha)(\log(H/\delta))^{1/4}$
Bounded Laplace	Bounded	$(\epsilon, 0)$	$\widetilde{\mathcal{O}}((nd)^{5/4}\zeta^{1/4}H^{1/4}T^{3/4}\log(ndT/\alpha))$

Table 1: Summary of privacy guarantee and the order of regret for different noise adding mechanisms.

distribution. The details on sampling from the BL distribution are available in Appendix H.

Theorem 8. *If the parameters of the BL distribution are set to* $b = \frac{4H^3\sqrt{nd}}{\epsilon}$*, then the MA-LDP algorithm with BL mechanism preserves* $(\epsilon, 0)$*-MA-LDP privacy.*

The proof of the above theorem is given in Appendix D.4. Similar to the Laplace mechanism, the BL mechanism preserves $(\epsilon, 0)$-LDP. The following theorem bounds the regret of the MA-LDP algorithm with the BL mechanism. The proof is deferred to Appendix E.4.

Theorem 9. *Consider the BL noise mechanism with parameter* b *as in Theorem 8. For any user* k*, with probability at least* $1-\alpha$*, the total regret of MA-LDP algorithm in the first* T *steps is at most* $\widetilde{\mathcal{O}}(n^{5/4}d^{5/4}\zeta^{1/4}H^{1/4}T^{3/4}\log(ndT/\alpha))$.

Table 1 summarizes the regret and privacy guarantees for all the noise-adding mechanisms we consider. This answers the questions **Q1** and **Q2** for all the four noise-adding mechanisms we consider.

8 Comparison of Regret and Privacy for Different Noise Mechanisms

This section compares the regrets and privacy guarantees associated with the various noise-adding mechanisms under consideration. We have the following result for the Gaussian and Laplace mechanisms.

Theorem 10. *If privacy parameters* ϵ_1 *and* ϵ_2 *are such that* $\epsilon_1 > \epsilon_2$*. Then, for both Gaussian and Laplace mechanisms, we have that* $R_K(\epsilon_1) < R_K(\epsilon_2)$.

The above theorem holds because for both Gaussian and Laplace mechanisms we have $R_K(\epsilon_1)/R_K(\epsilon_2) = \sqrt{\epsilon_2/\epsilon_1}$ for any two privacy parameters ϵ_1 and ϵ_2 as outlined in Theorems 3 and 5.

Comparison between Gaussian and Laplace: For the privacy parameter ϵ, let $R_K^G(\epsilon), R_K^L(\epsilon)$ be the cumulative regret of the Gaussian and Laplace mechanism, respectively, then we have the following:

Theorem 11. *If* $H > 2$ *then,* $R_K^G(\epsilon) > R_K^L(\epsilon)$.

The proof follows from the fact that $R_K^G(\epsilon)/R_K^L(\epsilon) = \log(H/\delta)$, and for the Gaussian mechanism $\log(H/\delta) > 1$ for all $\delta \in (0, 1)$, and $H > 2$.

Comparison of Regret between Laplace and bounded Laplace: Recall the variance ζ of the bounded Laplace distribution, which is contingent on the values of B and b. Depending on whether B shares the same order as b, we obtain distinct expressions for ζ, leading to varying regret orders (Theorem 9). Therefore, we compare the regret of the BL mechanism with that of the Laplace mechanism, considering scenarios where $B = O(b^\gamma)$ for different potential values of γ.

$B = O(b^\gamma)$	R_K^{BL}
$0 \le \gamma \le 1$	$\widetilde{\mathcal{O}}((nd)^{5/4}H^{7/4}T^{3/4}\sqrt{\frac{1}{\epsilon}})$
$\gamma > 1$	$\widetilde{\mathcal{O}}((nd)^{5/4}H^{7/4}H^{3\gamma/2}T^{3/4}\sqrt{\frac{1}{\epsilon}}\sqrt{\frac{1}{\epsilon^\gamma}})$

Table 2: Regret bound for the Bounded Laplace (BL) mechanism. MA-LDP algorithm with BL mechanism offers the same order of regret as that of the Laplace mechanism when $B = O(b^\gamma)$ for $\gamma \in [0, 1]$.

For the case of $\gamma > 1$, we have that $R_K^{BL}/R_K^L = \left(H^3/\epsilon\right)^{\frac{\gamma}{2}}$. If $\left(H^3/\epsilon\right)^{\frac{\gamma}{2}} > 1$ and $\gamma > 1$, then the regret of MA-LDP with BL mechanism is more than that of Laplace mechanism. So, the regret order for the BL mechanism will be the same as that of the Laplace mechanism if $\gamma \in [0, 1]$. Further, if $\gamma > 1$ and $\left(H^3/\epsilon\right)^{\frac{\gamma}{2}} < 1$, then BL mechanism will have lower regret than Laplace. In a very restrictive setting where $\gamma > 1$ and $\left(H^3/\epsilon\right)^{\frac{\gamma}{2}} > 1$, the regret from BL mechanism is more than the Laplace mechanism.

Theorem 12. *If* $B = O(b^\gamma)$*, where* $\gamma \le 1$*, then the BL mechanism has the same regret order as that of the Laplace mechanism.*

Thus, we have that, though the BL mechanism injects noise from bounded support, the regret of the MA-LDP algorithm with the BL mechanism is either on par or lower with that of the Laplace mechanism in most of the cases, i.e., when $B = O(b^\gamma)$ with $\gamma \in [0, 1]$.

Remark 1. *For each noise-adding mechanism, the regret is sub-linear, K, but it is super-linear, n, while it could have been quadratic, as interactions among n agents (in the worst case) are $O(n^2)$. In designing algorithms for multi-agent systems, scaling with respect to n is a critical question. However, we didn't aim for an optimal scaling of our algorithm wrt n; this is an important direction to explore.*

9 Computational Experiments

This section presents computational results to demonstrate our MA-LDP algorithm's efficacy. More details are given in Appendix F.

Setup. We analyze a network comprising $q+2$ nodes, where $q \geq 1$, denoted as $\{s_{in}, 1, 2, \ldots, q, g\}$, with s_{in} and g are initial and goal nodes, respectively. The number of global states is $(q+2)^n$. In each state, the available actions for each agent are $\mathcal{A}^i = \{-1, 1\}^{d-1}$, where $d \geq 2$. Consequently, the total number of feasible actions is $2^{n(d-1)}$. Each agent receives a reward of 5/1000 units for any action taken in the initial state s_{in}, a reward of 1000 units for action taken in the goal state g, and a reward of 0 units for actions in any other nodes. The collective objective shared by the agents is to achieve decentralized navigation to the goal node while maximizing the overall reward.

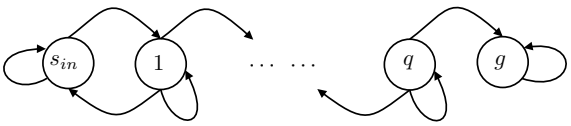

Figure 1: The network we consider.

Why this network is challenging. The network we consider is one of the most complex instances for multi-agent problems, with combinatorial state and action spaces. This is a generalization of the instance known to achieve the lower bound for regret of the algorithm that learns the stochastic shortest path (Vial et al., 2021). To address humongous state and action space, we parameterize the transition probability as $\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a}) = \langle \phi(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a}), \boldsymbol{\theta}(\boldsymbol{s}) \rangle$. The feature design is novel to our work and is available in Appendix F.1.

Observations. Figure 2 shows the cumulative regret with $n = 2$, the number of nodes in the network as 3, and the planning horizon $H = 5$. All the values are averaged over ten runs; each run has $K = 8500$ episodes. The cumulative regret is inversely proportional to the privacy losses; for instance, $R_K(\epsilon_L = 0.1) > R_K(\epsilon_L = 0.2)$, $R_K(\epsilon_G = 0.05) > R_K(\epsilon_G = 0.2)$, and $R_K^G(\epsilon = 0.2) > R_K^L(\epsilon = 0.2)$. This validates Theorems 10 and 11.

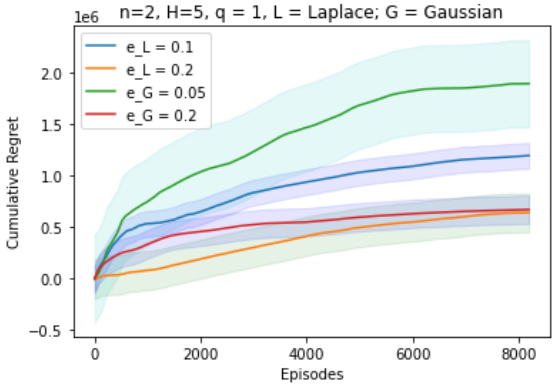

Figure 2: Cumulative regret vs. episodes for the Laplace and Gaussian mechanism.

10 Discussions

Conclusions. We define the LDP for MARL and propose an MA-LDP algorithm. We show the privacy and regret guarantees of the MA-LDP algorithm for four different noise-adding mechanisms. Next, we compare the noise mechanisms with bounded support with that of unbounded support. Our key observation is that bounded noise, with appropriately chosen endpoints, results in lower regret than unbounded noise, showing that bounded noise is often sufficient for LDP without significantly impacting regret. We illustrate our results on a networked MDP with two states and actions that are exponential in number of agents.

Future Works. This work offers a rich set of further possibilities. We mention some of these here. Firstly, our regret bound is super-linear in the number of agents and feature dimension; towards this, one can propose a better update rule for the optimistic estimators of the state-action value function. Secondly, the regret bounds we show are of their first kind, so an attempt to get a better sub-linear regret bound is possible. Moreover, a matching lower bound can also be tried. A careful study of other bounded support mechanisms that lead to lower regret bounds with low noise values would be interesting.

Limitations. While we have sub-linear regret guarantees, linear function approximations of the transition probabilities often fail to capture the complexity of many real-world applications. So, providing the regret and privacy guarantees with the non-linear dynamics presents an intriguing direction for future research.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

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

Barrett, S. and Stone, P. (2015). Cooperating with unknown teammates in complex domains: A robot soccer case study of ad hoc teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Borkar, V. S. (2022). *Stochastic approximation: a dynamical systems viewpoint. Second Edition*, volume 48. Springer.

Cyffers, E. and Bellet, A. (2022). Privacy amplification by decentralization. In *International Conference on Artificial Intelligence and Statistics*, pages 5334–5353. PMLR.

Dobbe, R., Pu, Y., Zhu, J., Ramchandran, K., and Tomlin, C. (2018). Customized local differential privacy for multi-agent distributed optimization. *arXiv preprint arXiv:1806.06035*.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. (2013). Local privacy, data processing inequalities, and statistical minimax rates. *arXiv preprint arXiv:1302.3203*.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer.

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.

Garcelon, E., Perchet, V., Pike-Burke, C., and Pirotta, M. (2021). Local differential privacy for regret minimization in reinforcement learning. *Advances in Neural Information Processing Systems*, 34:10561–10573.

Geng, Q. and Viswanath, P. (2015). Optimal noise adding mechanisms for approximate differential privacy. *IEEE Transactions on Information Theory*, 62(2):952–969.

Hu, W. and Fang, H. (2022). Decentralized matrix factorization with heterogeneous differential privacy. *arXiv preprint arXiv:2212.00306*.

Jia, Z., Yang, L., Szepesvari, C., and Wang, M. (2020). Model-based reinforcement learning with value-targeted regression. In *Learning for Dynamics and Control*, pages 666–686. PMLR.

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

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. (2011). What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826.

Kushner, H. and Yin, G. G. (2003). *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media.

Li, T., Zhu, K., Luong, N. C., Niyato, D., Wu, Q., Zhang, Y., and Chen, B. (2022). Applications of multi-agent reinforcement learning in future internet: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 24(2):1240–1279.

Liao, C., He, J., and Gu, Q. (2021). Locally differentially private reinforcement learning for linear mixture markov decision processes. *arXiv preprint arXiv:2110.10133*.

Metivier, M. and Priouret, P. (1984). Applications of a Kushner and Clark lemma to general classes of stochastic algorithms. *IEEE Transactions on Information Theory*, 30(2):140–151.

Min, Y., He, J., Wang, T., and Gu, Q. (2022). Learning stochastic shortest path with linear function approximation. In *International Conference on Machine Learning*, pages 15584–15629. PMLR.

Ross, S. M. (2022). *Simulation*. Academic Press.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE.

Tao, T. (2012). *Topics in random matrix theory*, volume 132. American Mathematical Soc.

Trivedi, P. and Hemachandra, N. (2022). Multi-agent natural actor-critic reinforcement learning algorithms. *Dynamic Games and Applications, Special Issue on Multi-agent Dynamic Decision Making and Learning, edited by Konstantin Avrachenkov, Vivek S. Borkar and U. Jayakrishnan Nair*, pages 1–31.

Trivedi, P. and Hemachandra, N. (2023). Multi-agent congestion cost minimization with linear function approximations. In *International Conference on Artificial Intelligence and Statistics*, pages 7611–7643. PMLR.

Vial, D., Parulekar, A., Shakkottai, S., and Srikant, R. (2021). Regret bounds for stochastic shortest path problems with linear function approximation. *arXiv preprint arXiv:2105.01593*.

Vial, D., Parulekar, A., Shakkottai, S., and Srikant, R. (2022). Regret bounds for stochastic shortest path problems with linear function approximation. In *International Conference on Machine Learning*, pages 22203–22233. PMLR.

Yuan, T., Chung, H.-M., and Fu, X. (2022). Ppmarl: Efficient privacy-preserving marl for cooperative intelligence in communication. *arXiv preprint arXiv:2204.12064*.

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

Zhou, D., Gu, Q., and Szepesvari, C. (2021). Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes/No/Not Applicable] Yes.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/No/Not Applicable] Yes.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable] Yes, see supplementary material.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes/No/Not Applicable] Yes.
   (b) Complete proofs of all theoretical results. [Yes/No/Not Applicable] Yes.
   (c) Clear explanations of any assumptions. [Yes/No/Not Applicable] Yes.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/No/Not Applicable] Yes, anonymous link given at the end of the abstract.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/Not Applicable] Yes.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes/No/Not Applicable] Yes.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/No/Not Applicable] Yes, details available within the codes..

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes/No/Not Applicable] Not Applicable.
   (b) The license information of the assets, if applicable. [Yes/No/Not Applicable] Not Applicable.
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/Not Applicable] Not Applicable.
   (d) Information about consent from data providers/curators. [Yes/No/Not Applicable] Not Applicable.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/Not Applicable] Not Applicable.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Yes/No/Not Applicable] Not Applicable.
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/Not Applicable] Not Applicable.
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/Not Applicable] Not Applicable.

Appendix

A Proof of Lemma 1

Recall Lemma 1: Under assumption 1 for the sequence $\{\boldsymbol{w}^i_{k,h}\}$, we have $lim_k\ \boldsymbol{w}^i_{k,h} = \boldsymbol{w}^\star$ a.s. for each agent $i \in N$, and for all $h \in [H]$, where $\boldsymbol{w}^\star$ is unique solution to $\Psi^\top D^{s,a}(\Psi \boldsymbol{w}^\star - \bar{r}) = 0$.

To prove this lemma, we require the following standard assumption the consensus matrix (Zhang et al., 2018; Trivedi and Hemachandra, 2022).

Assumption 3. *The consensus matrices $\{L_t\}_{t\geq 0} \subseteq \mathbb{R}^{n\times n}$ satisfies*

- *L_t is row stochastic, and $\mathbb{E}(L_t)$ is column stochastic. Further, there exists a constant $\kappa \in (0,1)$ such that for any $l_t(i,j) > 0$, we have $l_t(i,j) \geq \kappa$,*
- *Consensus matrix L_t respects $\mathcal{G}_t$, i.e., $l_t(i,j) = 0$, if $(i,j) \notin \mathcal{E}_t$,*
- *The spectral norm of $\mathbb{E}[L_t^\top(I - \mathbb{1}\mathbb{1}^\top/n)L_t]$ is smaller than one.*

Proof. Let $t = kh$, therefore, as $k \to \infty$ we have $t \to \infty$. The proof of this result is on the same lines to Zhang et al. (2018); Trivedi and Hemachandra (2023, 2022). We briefly give the proof details here.

To prove the convergence of the reward function parameters, we use the following proposition to give bounds on $\boldsymbol{w}^i_t$ for all $i \in N$. For proof, we refer to Zhang et al. (2018).

Proposition 1. *Under assumptions 3, and 1 the sequence $\{\boldsymbol{w}^i_t\}$ satisfy $sup_t\ ||\boldsymbol{w}^i_t|| < \infty$ a.s., for all $i \in N$.*

Let $\mathcal{F}_t = \sigma(r_\tau, \boldsymbol{w}_\tau, \boldsymbol{s}_\tau, \boldsymbol{a}_\tau, L_{\tau-1}, \tau \leq t)$ be the filtration which is an increasing σ-algebra over time t. Define the following for notation convenience. Let $r_t = [r^1_t, \ldots, r^n_t]^\top \in \mathbb{R}^n$, and $\boldsymbol{w}_t = [(\boldsymbol{w}^1_t)^\top, \ldots, (\boldsymbol{w}^n_t)^\top]^\top \in \mathbb{R}^{np}$. Moreover, let $A \otimes B$ represent the Kronecker product of any two matrices A and B. Let $y_t = [(y^1_t)^\top, \ldots, (y^n_t)^\top]^\top$, where $y^i_{t+1} = [(r^i_{t+1} - \psi_t^\top \boldsymbol{w}^i_t)\psi_t^\top]^\top$. Recall, $\psi_t = \psi(\boldsymbol{s}_t, \boldsymbol{a}_t)$. Let I be the identity matrix of the dimension $p \times p$. Then update of $\boldsymbol{w}_t$ can be written as

$$\boldsymbol{w}_{t+1} = (L_t \otimes I)(\boldsymbol{w}_t + \gamma_t \cdot y_{t+1}). \tag{13}$$

Let $\mathbb{1} = (1, \ldots, 1)$ represents the vector of all 1's. We define the operator $\langle \boldsymbol{w} \rangle = \frac{1}{n}(\mathbb{1}^\top \otimes I)\boldsymbol{w} = \frac{1}{n}\sum_{i\in N} \boldsymbol{w}^i$. This $\langle \boldsymbol{w} \rangle \in \mathbb{R}^p$ represents the average of the vectors in $\{\boldsymbol{w}^1, \boldsymbol{w}^2, \ldots, \boldsymbol{w}^n\}$. Moreover, let $\mathcal{J} = (\frac{1}{n}\mathbb{1}\mathbb{1}^\top) \otimes I \in \mathbb{R}^{np\times np}$ is the projection operator that projects a vector into the consensus subspace $\{\mathbb{1} \otimes u : u \in \mathbb{R}^p\}$. Thus $\mathcal{J}\boldsymbol{w} = \mathbb{1} \otimes \langle \boldsymbol{w} \rangle$. Now define the disagreement vector $\boldsymbol{w}_\perp = \mathcal{J}_\perp \boldsymbol{w} = \boldsymbol{w} - \mathbb{1} \otimes \langle \boldsymbol{w} \rangle$, where $\mathcal{J}_\perp = I - \mathcal{J}$. Here I is $np \times np$ dimensional identity matrix. The iteration $\boldsymbol{w}_t$ can be decomposed as the sum of a vector in disagreement space and a vector in consensus space, i.e., $\boldsymbol{w}_t = \boldsymbol{w}_{\perp,t} + \mathbb{1} \otimes \langle \boldsymbol{w}_t \rangle$. The proof of convergence consists of two steps.

Step 01: To show $lim_t\ \boldsymbol{w}_{\perp,t} = 0$ a.s. From Proposition 1 we have $\mathbb{P}[sup_t ||\boldsymbol{w}_t|| < \infty] = 1$, i.e., $\mathbb{P}[\cup_{K_1\in\mathbb{Z}^+}\ \{sup_t ||\boldsymbol{w}_t|| < K_1\}] = 1$. It suffices to show that $lim_t\ \boldsymbol{w}_{\perp,t}\mathbb{1}_{\{sup_t ||\boldsymbol{w}_t|| < K_1\}} = 0$ for any $K_1 \in \mathbb{Z}^+$. Lemma 5.5 in Zhang et al. (2018) proves the boundedness of $\mathbb{E}\left[||\beta_t^{-1}\boldsymbol{w}_{\perp,t}||^2\right]$ over the set $\{sup_t ||\boldsymbol{w}_t|| \leq K_1\}$, for any $K_1 > 0$. We state the lemma here.

Proposition 2 (Lemma 5.5 in Zhang et al. (2018))**.** *Under assumptions 3, and 1 for any $K_1 > 0$, we have*

$$sup_t\ \mathbb{E}[||\beta_t^{-1}\boldsymbol{w}_{\perp,t}||^2 \mathbb{1}_{\{sup_t ||\boldsymbol{w}_t|| \leq K\}}] < \infty.$$

From Proposition 2 we obtain that for any $K_1 > 0$, $\exists\ K_2 < \infty$ such that for any $t \geq 0$, $\mathbb{E}[||\boldsymbol{w}_{\perp,t}||^2] < K_2\gamma_t^2$ over the set $\{sup_t\ ||\boldsymbol{w}_t|| < K_1\}$. Since $\sum_t \gamma_t^2 < \infty$, by Fubini's theorem we have $\sum_t \mathbb{E}(||\boldsymbol{w}_{\perp,t}||^2\ \mathbb{1}_{\{sup_t\ ||\boldsymbol{w}_t|| < K_1\}}) < \infty$. Thus, $\sum_t ||\boldsymbol{w}_{\perp,t}||^2 \mathbb{1}_{\{sup_t\ ||\boldsymbol{w}_t|| < K_1\}} < \infty$ a.s. Therefore, $lim_t\ \boldsymbol{w}_{\perp,t}\ \mathbb{1}_{\{sup_t ||\boldsymbol{w}_t|| < K_1\}} = 0$ a.s. Since $\{sup_t ||\boldsymbol{w}_t|| < \infty\}$ with probability 1, thus $lim_t\ \boldsymbol{w}_{\perp,t} = 0$ a.s. This ends the proof of Step 01.

Step 02: To show the convergence of the consensus vector $\mathbb{1}\otimes\langle \boldsymbol{w}_t\rangle$, first note that the iteration of $\langle \boldsymbol{w}_t\rangle$ (Equation (13)) can be written as

$$\begin{aligned}\langle \boldsymbol{w}_{t+1}\rangle &= \frac{1}{N}(\mathbb{1}^\top\otimes I)(L_t\otimes I)(\mathbb{1}\otimes\langle \boldsymbol{w}_t\rangle + \boldsymbol{w}_{\perp,t} + \gamma_t\ y_{t+1})\\ &= \langle \boldsymbol{w}_t\rangle + \gamma_t\langle (L_t\otimes I)(y_{t+1}+\gamma_t^{-1}\boldsymbol{w}_{\perp,t})\rangle\\ &= \langle \boldsymbol{w}_t\rangle + \gamma_t\ \mathbb{E}(\langle y_{t+1}\rangle|\mathcal{F}_t) + \beta_t\xi_{t+1},\end{aligned} \tag{14}$$

where

$$\begin{aligned}\xi_{t+1} &= \langle (L_t\otimes I)(y_{t+1}+\gamma_t^{-1}\boldsymbol{w}_{\perp,t})\rangle - \mathbb{E}(\langle y_{t+1}\rangle|\mathcal{F}_t),\quad \textit{and}\\ \langle y_{t+1}\rangle &= [(\bar{r}_{t+1}-\psi_t^\top\langle \boldsymbol{w}_t\rangle)\psi_t^\top]^\top.\end{aligned}$$

Note that $\mathbb{E}(\langle y_{t+1}\rangle|\mathcal{F}_t)$ is Lipschitz continuous in $\langle \boldsymbol{w}_t\rangle$. Moreover, ξ_{t+1} is a martingale difference sequence and satisfies

$$\mathbb{E}[||\xi_{t+1}||^2\mid\mathcal{F}_t] \le \mathbb{E}[||y_{t+1}+\gamma_t^{-1}\boldsymbol{w}_{\perp,t}||^2_{R_t}\mid\mathcal{F}_t] + ||\mathbb{E}(\langle y_{t+1}\rangle\mid\mathcal{F}_t)||^2, \tag{15}$$

where $R_t = \frac{L_t^\top\mathbb{1}\mathbb{1}^\top L_t\otimes I}{n^2}$ has bounded spectral norm. Bounding first and second terms in RHS of Equation (15), we have, for any $K_1>0$

$$\mathbb{E}(||\xi_{t+1}||^2|\mathcal{F}_t) \le K_3(1+||\langle \boldsymbol{w}_t\rangle||^2),$$

over the set $\{sup_t\ ||\boldsymbol{w}_t||\le K_1\}$ for some $K_3<\infty$. Thus condition (3) of assumption 4 is satisfied. The ODE associated with the Equation (14) has the form

$$\langle \dot{\boldsymbol{w}}\rangle = -\Psi^\top D^{\boldsymbol{s},\boldsymbol{a}}\Psi\langle \boldsymbol{w}\rangle + \Psi^\top D^{\boldsymbol{s},\boldsymbol{a}}\bar{r} \tag{16}$$

Let the RHS of Equation (16) be $h(\langle \boldsymbol{w}\rangle)$. Note that $h(\langle \boldsymbol{w}\rangle)$ is Lipschitz continuous in $\langle \boldsymbol{w}\rangle$. Also, recall that $D^{\boldsymbol{s},\boldsymbol{a}} = diag[d(\boldsymbol{s})\cdot\pi(\boldsymbol{s},\boldsymbol{a}), \boldsymbol{s}\in\mathcal{S},\boldsymbol{a}\in\mathcal{A}]$. Hence the ODE given in Equation (16) has unique globally asymptotically stable equilibrium $\boldsymbol{w}^\star$ satisfying

$$\Psi^\top D^{\boldsymbol{s},\boldsymbol{a}}(\bar{r}-\Psi\boldsymbol{w}^\star) = 0.$$

Moreover, from Propositions 1, and 2, the sequence $\{\boldsymbol{w}_t\}$ is bounded almost surely, so is the sequence $\{\langle \boldsymbol{w}_t\rangle\}$. Specializing Corollary 8 and Theorem 9 on page 74-75 in Borkar (2022) we have $lim_t\ \langle \boldsymbol{w}_t\rangle = \boldsymbol{w}^\star$ a.s. over the set $\{sup_t\ ||\boldsymbol{w}_t||\le K_1\}$ for any $K_1>0$. This concludes the proof of Step 02.

The proof of the Theorem follows from Proposition 1 and results from Step 01. Thus, we have $lim_t\ \boldsymbol{w}_t^i = \boldsymbol{w}^\star$ a.s. for each $i\in N$. This implies, $lim_k\ \boldsymbol{w}_{kh}^i = \boldsymbol{w}^\star$ a.s. for each $i\in N$ and for all $h\in[H]$. □

B Proof of Lemma 2

Recall Lemma 2: Let $Q_{k,h}^i$ and $V_{k,h}^i$ be the estimate of the global state-action value and global state value functions by agent $i\in N$. Then, for any pairs $(\boldsymbol{s},\boldsymbol{a},k,h)\in\mathcal{S}\times\mathcal{A}\times[K]\times[H]$, we have $Q_h^{\star,i}(\boldsymbol{s},\boldsymbol{a})\le Q_{k,h}^i(\boldsymbol{s},\boldsymbol{a})$ and $V_h^{\star,i}(\boldsymbol{s})\le V_{k,h}^i(\boldsymbol{s})$.

Proof. The proof of this lemma is by induction over h. Consider the basic case $h=H+1$. By assumption we have that $Q_{k,H+1}^i(\cdot,\cdot)=0=Q_{H+1}^{\star,i}(\cdot,\cdot)$, and $V_{k,H+1}^i(\cdot)=0=V_{H+1}^{\star,i}(\cdot)$. Now suppose that this statement is true for all $h+1$, so we have $Q_{k,h+1}^i(\cdot,\cdot)\ge Q_{k,h+1}^{\star,i}(\cdot,\cdot)$, and $V_{k,h+1}^i(\cdot)\ge V_{k,h+1}^{\star,i}(\cdot)$. For any stage h and $(\boldsymbol{s},\boldsymbol{a})$, if $Q_{k,h}^i(\boldsymbol{s},\boldsymbol{a})\ge H$, then, the statement is also true for stage h, i.e., $Q_{k,h}^i(\boldsymbol{s},\boldsymbol{a})\ge H\ge Q_h^{\star,i}(\boldsymbol{s},\boldsymbol{a})$. However, if $Q_{k,h}^i(\boldsymbol{s},\boldsymbol{a})\le H$, then consider the following:

$$\begin{aligned}Q_{k,h}^i(\boldsymbol{s},\boldsymbol{a}) - Q_h^{\star,i}(\boldsymbol{s},\boldsymbol{a}) &\overset{(i)}{=} \bar{r}_h(\boldsymbol{s},\boldsymbol{a};\boldsymbol{w}_{k,h}^i) + \left\langle \hat{\boldsymbol{\theta}}_{k,h}^i, \phi_{V_{k,h+1}^i}(\boldsymbol{s},\boldsymbol{a})\right\rangle + \beta_k||\Sigma_{k,h}^{i^{-1/2}}\phi_{V_{k,h+1}^i}(\boldsymbol{s},\boldsymbol{a})||_2\\ &\quad - \bar{r}_h(\boldsymbol{s},\boldsymbol{a};\boldsymbol{w}_{k,h}^i) - \mathbb{P}_hV_{h+1}^{\star,i}(\boldsymbol{s},\boldsymbol{a})\\ &= \left\langle \hat{\boldsymbol{\theta}}_{k,h}^i, \phi_{V_{k,h+1}^i}(\boldsymbol{s},\boldsymbol{a})\right\rangle + \beta_k||\Sigma_{k,h}^{i^{-1/2}}\phi_{V_{k,h+1}^i}(\boldsymbol{s},\boldsymbol{a})||_2 - \mathbb{P}_hV_{h+1}^{\star,i}(\boldsymbol{s},\boldsymbol{a})\end{aligned}$$

$$
\begin{aligned}
&\overset{(ii)}{=} \left\langle \hat{\boldsymbol{\theta}}^i_{k,h}, \phi_{V^i_{k,h+1}}(\boldsymbol{s},\boldsymbol{a}) \right\rangle + \beta_k ||\Sigma^{i^{-1/2}}_{k,h} \phi_{V^i_{k,h+1}}(\boldsymbol{s},\boldsymbol{a})||_2 - \mathbb{P}_h V^{\star,i}_{h+1}(\boldsymbol{s},\boldsymbol{a}) \\
&\quad - \left\langle \boldsymbol{\theta}^\star_h, \phi_{V^i_{k,h+1}}(\boldsymbol{s},\boldsymbol{a}) \right\rangle + \left\langle \boldsymbol{\theta}^\star_h, \phi_{V^i_{k,h+1}}(\boldsymbol{s},\boldsymbol{a}) \right\rangle \\
&= \beta_k ||\Sigma^{i^{-1/2}}_{k,h} \phi_{V^i_{k,h+1}}(\boldsymbol{s},\boldsymbol{a})||_2 - \left\langle \hat{\boldsymbol{\theta}}^i_{k,h} - \boldsymbol{\theta}^\star_h, \phi_{V^i_{k,h+1}}(\boldsymbol{s},\boldsymbol{a}) \right\rangle \\
&\quad - \mathbb{P}_h V^{\star,i}_{h+1}(\boldsymbol{s},\boldsymbol{a}) + \mathbb{P}_h V^i_{k,h+1}(\boldsymbol{s},\boldsymbol{a}) \\
&\overset{(iii)}{\geq} \beta_k ||\Sigma^{i^{-1/2}}_{k,h} \phi_{V^i_{k,h+1}}(\boldsymbol{s},\boldsymbol{a})||_2 - ||\Sigma^{i^{1/2}}_{k,h} (\hat{\boldsymbol{\theta}}^i_{k,h} - \boldsymbol{\theta}^\star_h)||_2 ||\Sigma^{i^{-1/2}}_{k,h} \phi_{V^i_{k,h+1}}(\boldsymbol{s},\boldsymbol{a})||_2 \\
&\quad - \mathbb{P}_h V^{\star,i}_{h+1}(\boldsymbol{s},\boldsymbol{a}) + \mathbb{P}_h V^i_{k,h+1}(\boldsymbol{s},\boldsymbol{a}) \\
&\overset{(iv)}{\geq} -\mathbb{P}_h V^{\star,i}_{h+1}(\boldsymbol{s},\boldsymbol{a}) + \mathbb{P}_h V^i_{k,h+1}(\boldsymbol{s},\boldsymbol{a}) \\
&\overset{(v)}{\geq} 0
\end{aligned}
$$

In (i) we use the update of $Q^i_{k,h}(\boldsymbol{s},\boldsymbol{a})$ as in line 3 of user sub-routine. In (ii), we add and subtract an inner product term. Inequality (iii) follows from the Cauchy-Schwartz inequality. The inequality (iv) follows from the fact proved in next Lemma 3. Finally, the last inequality (v) uses the monotone property of $\mathbb{P}_h$ with respect to the partial ordering of the function. □

Remark 2. *Note that we have used $\boldsymbol{w}^i_{k,h}$ in the above lemma because of the following reasons. Firstly, in each episode k, we employ backward induction for each agent to obtain $Q^{\star,i}_{k,1}$, and each agent $i \in N$ utilizes this value throughout the episode to take action. Consequently, while updating $Q^i_{k,h}$, we incorporate the most recent available reward function parameters $\boldsymbol{w}^i_{k,h}$ that converges to the same $\boldsymbol{w}^\star$ for all agents $i \in N$ almost surely (Lemma 1). Therefore, using these recent parameters, each agent updates its optimistic estimators of the state-action value function. Nevertheless, as $k \to \infty$, $Q^{\star,i}_{k,1}$ converges to the same $Q^\star$ for all agents $i \in N$, that is the true optimal value. It is important to highlight that, unlike the single-agent infinite horizon RL model where the convergence is desirable, in our decentralized MARL model with a finite planning horizon problem, the relevant performance metric is regret over K episodes. Moreover, the state-action value function, Q^i is a continuous function of reward parameters $\boldsymbol{w}^i$ (which asymptotically converge to the same $\boldsymbol{w}^\star$ a.s. for each agent $i \in N$), and as $K \to \infty$, the average regret approaches zero as claimed in Theorem 1 (since the reward function parameters are recovered).*

C Proof of Lemma 3

Recall Lemma 3: If $\eta = 1$, then for any fixed policy π and all pairs (s,a,h,k), with probability at least $1-\alpha/2$, we have $||(\Sigma^i_{k,h})^{1/2}(\hat{\boldsymbol{\theta}}^i_{k,h} - \boldsymbol{\theta}^\star_h)|| \leq \beta_k,\ \forall\ i \in N$.

Proof. To prove this lemma, we first give the details of β_k for each noise adding mechanism. To this end, we decompose $||(\Sigma^i_{k,h})^{1/2}(\hat{\boldsymbol{\theta}}^i_{k,h} - \boldsymbol{\theta}^\star_h)||$ in three terms. Note that this decomposition is common to all the noise adding mechanisms. However, we bound each term differently for different noise mechanisms to get the exact expression. Consider the difference

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}^i_{k,h} - \boldsymbol{\theta}^\star_h &\overset{(i)}{=} (\Sigma^i_{k,h})^{-1} \sum_{\tau=1}^{k-1} \{\phi_{V^i_{\tau,h+1}}(\boldsymbol{s}_{\tau,h},\boldsymbol{a}_{\tau,h}) V^i_{\tau,h+1}(s_{\tau,h+1}) + \boldsymbol{\xi}^i_{\tau,h}\} - \boldsymbol{\theta}^\star_h \\
&= (\Sigma^i_{k,h})^{-1} \left\{ -\Sigma^i_{k,h}\boldsymbol{\theta}^\star_h + \sum_{\tau=1}^{k-1} \{\phi_{V^i_{\tau,h+1}} V^i_{\tau,h+1} + \boldsymbol{\xi}^i_{\tau,h}\} \right\} \\
&\overset{(ii)}{=} (\Sigma^i_{k,h})^{-1} \left\{ -(\lambda \mathbf{I} + \sum_{\tau=1}^{k-1} \phi_{V^i_{\tau,h+1}} \phi^\top_{V^i_{\tau,h+1}} + \mathbf{W}^i_h)\boldsymbol{\theta}^\star_h + \sum_{\tau=1}^{k-1} \{\phi_{V^i_{\tau,h+1}} V^i_{\tau,h+1} + \boldsymbol{\xi}^i_{\tau,h}\} \right\} \\
&= (\Sigma^i_{k,h})^{-1} \left\{ -\lambda\boldsymbol{\theta}^\star_h - \sum_{\tau=1}^{k-1} \phi_{V^i_{\tau,h+1}} \phi^\top_{V^i_{\tau,h+1}} \boldsymbol{\theta}^\star_h - \mathbf{W}^i_h \boldsymbol{\theta}^\star_h + \sum_{\tau=1}^{k-1} \{\phi_{V^i_{\tau,h+1}} V^i_{\tau,h+1} + \boldsymbol{\xi}^i_{\tau,h}\} \right\}
\end{aligned}
$$

$$\overset{(iii)}{=} (\Sigma^i_{k,h})^{-1}\left\{(-\lambda\mathbf{I}-\mathbf{W}^i_h)\boldsymbol{\theta}^\star_h + \sum_{\tau=1}^{k-1}\phi_{V^i_{\tau,h+1}}\left[V^i_{\tau,h+1}-\mathbb{P}_hV^i_{\tau,h+1}\right] + \sum_{\tau=1}^{k-1}\boldsymbol{\xi}^i_{\tau,h}\right\}$$

where $\mathbf{W}^i_h = \sum_{\tau=1}^{k-1}\mathbf{W}^i_{\tau,h}$. Here (i) uses the definition of $\hat{\boldsymbol{\theta}}^i_{k,h}$ given in line 6 of the server sub-routine. The (ii) uses the update definition of $\Sigma^i_{k,h}$. In (iii), we combine some terms and use the linear function approximation of the transition probability. So, from the above equation, we can write the following:

$$||(\Sigma^i_{k,h})^{1/2}(\hat{\boldsymbol{\theta}}^i_{k,h}-\boldsymbol{\theta}^\star_h)|| = ||(\Sigma^i_{k,h})^{-1/2}(\mathbf{q}_1+\mathbf{q}_2+\mathbf{q}_3)|| \tag{17}$$

where $\mathbf{q}_1 = (-\lambda\mathbf{I}-\mathbf{W}^i_h)\boldsymbol{\theta}^\star_h$; $\mathbf{q}_2 = \sum_{\tau=1}^{k-1}\phi_{V^i_{\tau,h+1}}\left[V^i_{\tau,h+1}-\mathbb{P}_hV^i_{\tau,h+1}\right]$; and $\mathbf{q}_3 = \sum_{\tau=1}^{k-1}\boldsymbol{\xi}^i_{\tau,h}$.

To complete the proof we need to bound each $\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3$. Here, we give the general framework for bounding each of these terms, and later we will specialize them to each of the noise adding mechanisms.

Bounding $\mathbf{q}_1$

To bound $\mathbf{q}_1$, we need to give the upper and the lower bound on the eigenvalues of the symmetric matrix $\mathbf{W}^i_h$. Recall that each entry of matrix $\mathbf{W}^i_{j,h}$ is sampled from the corresponding distribution (Gaussian, Laplace, uniform, and bounded Laplace). So, the variance of the matrix $\mathbf{W}^i_h$ is $(k-1)\sigma^2$, where σ^2 is the variance of the corresponding noise distribution. From the known concentration results of Tao (2012), we have that

$$\mathbb{P}\left(\left|\left|\sum_{\tau=1}^{k-1}\mathbf{W}^i_{\tau,h}\right|\right| \geq \sigma\sqrt{k-1}(\sqrt{4nd}+2\log(6H/\alpha))\right) \leq \frac{\alpha}{6H} \tag{18}$$

that is

$$\mathbb{P}\left(\left|\left|\mathbf{W}^i_h\right|\right| \geq \Gamma\right) \leq \frac{\alpha}{6H} \tag{19}$$

where $\Gamma = \sigma\sqrt{k-1}(\sqrt{4nd}+2\log(6H/\alpha))$.

For the symmetric matrix $\mathbf{W}^i_h$ the PSD property might not be preserved, so we add a matrix with positive entries $2\Gamma\mathbf{I}$ to the matrix $W^i_h = \sum_{\tau=1}^{k-1}\mathbf{W}^i_{\tau,h}$ for each stage $h \in [H]$. Thus, the eigenvalues of the shifted matrix are bounded in the interval $[\Gamma, 3\Gamma]$ with probability $1-\alpha/6$.

Note that we use this updated $\mathbf{W}^i_h$ throughout the proof that ensures, with probability at least $1-\alpha/6$ all the eigenvalues and entries of the modified W^i_h remain positive. Define the following event

$$\mathcal{E}_1 := \{\forall\ h\in[H],\ \forall\ j\in[nd],\ \Gamma\leq\sigma_j\leq 3\Gamma\}, \tag{20}$$

where σ_j's are the eigenvalues of the matrix $\mathbf{W}^i_h$, and we have $\mathbb{P}(\mathcal{E}_1) \geq 1-\alpha/6$. Let $\rho_{\max} = 3\Gamma+\lambda$, and $\rho_{\min} = \Gamma+\lambda$. Then for the term $\mathbf{q}_1$, we have

$$\begin{aligned}
||(\Sigma^i_{k,h})^{-1/2}\mathbf{q}_1|| &\overset{(i)}{\leq} ||(\mathbf{W}^i_h+\lambda\mathbf{I})^{-1/2}\mathbf{q}_1|| \\
&= ||(\mathbf{W}^i_h+\lambda\mathbf{I})^{-1/2}(-\mathbf{W}^i_h-\lambda\mathbf{I})\boldsymbol{\theta}^\star_h|| \\
&= ||(\mathbf{W}^i_h+\lambda\mathbf{I})^{1/2}\boldsymbol{\theta}^\star_h|| \\
&\overset{(ii)}{\leq} \sqrt{\rho_{\max}}||\boldsymbol{\theta}^\star_h|| \\
&\overset{(iii)}{\leq} \sqrt{\rho_{\max}nd} \\
&= \sqrt{\{3\sigma\sqrt{k-1}(\sqrt{4nd}+2\log(6H/\alpha))+\lambda\}nd} \\
&= \sqrt{3nd\sigma\sqrt{k-1}(\sqrt{4nd}+2\log(6H/\alpha))+\lambda nd}.
\end{aligned}$$

Since $\rho_{\max} = 3\Gamma+\lambda = 3\sigma\sqrt{k-1}(\sqrt{4nd}+2\log(6H/\alpha))+\lambda$ and $||\boldsymbol{\theta}^\star_h|| \leq \sqrt{nd}$. The first inequality holds because $(\Sigma^i_{k,h})^{-1/2} \succeq (\mathbf{W}^i_h+\lambda\mathbf{I})^{-1/2}$. The second inequality holds due to event $\mathcal{E}_1$. Moreover, the last inequality (iii) holds because of the assumption **X**. 2.

Bounding $\mathbf{q}_2$

For the term $\mathbf{q}_2$, we have the following

$$||(\Sigma^i_{k,h})^{-1/2}\mathbf{q}_2|| = ||\sum_{\tau=1}^{k-1}\phi_{V^i_{\tau,h+1}}[V^i_{\tau,h+1} - \mathbb{P}_h V^i_{\tau,h+1}]||_{(\Sigma^i_{k,h})^{-1}} \tag{21}$$

$$\leq ||\sum_{\tau=1}^{k-1}\phi_{V^i_{\tau,h+1}}[V^i_{\tau,h+1} - \mathbb{P}_h V^i_{\tau,h+1}]||_{(\mathbf{Z})^{-1}}, \tag{22}$$

where $\mathbf{Z} = \lambda\mathbf{I} + \sum_{\tau=1}^{k-1}\phi_{V^i_{\tau,h+1}}\phi^\top_{V^i_{\tau,h+1}}$. The inequality holds because $\Sigma^i_{k,h} \succeq \mathbf{Z} = \lambda\mathbf{I} + \sum_{\tau=1}^{k-1}\phi_{V^i_{\tau,h+1}}\phi^\top_{V^i_{\tau,h+1}}$ with probability at least $1-\alpha/6$. Let $\eta^i_{\tau,h+1} = V^i_{\tau,h+1} - \mathbb{P}_h V^i_{\tau,h+1} = V^i_{\tau,h+1} - \phi^\top_{V^i_{\tau,h+1}}\boldsymbol{\theta}^\star_h$. Moreover, let $\{\mathcal{G}_t\}_{t=1}^\infty$ be a filtration, $\{\phi_{V^i_{\tau,t}}, \eta^i_{\tau,t}\}_{t=1}^\infty$ a stochastic process so that $\phi_{V^i_{\tau,t}}$ is $\mathcal{G}_t$-measurable and $\eta^i_{\tau,t}$ is $\mathcal{G}_{t+1}$-measurable. With the above notations, we have

$$|\eta^i_{\tau,h}| = |V^i_{\tau,h} - \mathbb{P}_h V^i_{\tau,h}| \leq H. \tag{23}$$

The above is true because $V^i_{\tau,h} > 0$, and $P_h V^i_{\tau,h}$ is the average of $V^i_{\tau,h}$, we have $V^i_{\tau,h} - P_h V^i_{\tau,h} \leq V^i_{\tau,h}$. Thus, $|V^i_{\tau,h} - P_h V^i_{\tau,h}| < |V^i_{\tau,h}| \leq H$ as all per-period rewards are assumed to lie within $(0,1)$. Moreover, if $V^i_{\tau,h} \leq 0$ (stochasticity of the estimates), in that case also, above equation is valid as follows: as $|V^i_{\tau,h} - P_h V^i_{\tau,h}| \leq |P_h V^i_{\tau,h}| \leq H$ as $V^i_{\tau,h} < 0$, hence $P_h V^i_{\tau,h} < 0$. Further, we have

$$\mathbb{E}[(\eta^i_{\tau,h})^2|\mathcal{G}_h] \leq \mathbb{E}[(V^i_{\tau,h})^2|\mathcal{G}_h] \leq H^2. \tag{24}$$

Moreover, we define the following event

$$\mathcal{E}_2 = \left\{\forall\ h \in [H],\ ||\mathbf{q}_2||_{\mathbf{Z}^{-1}} \leq 4H\left(2\sqrt{and\log\left(1 + \frac{(k-1)H^2}{nd\lambda}\right)\log\left(\frac{24(k-1)^2}{\alpha}\right)} + \log\left(\frac{24(k-1)^2}{\alpha}\right)\right)\right\}. \tag{25}$$

Theorem 2 of Zhou et al. (2021) shows that the probability of the above event is at least $1-\alpha/6$.

Bounding $\mathbf{q}_3$

The term $\mathbf{q}_3$ can be bounded as

$$||\sum_{\tau=1}^{k-1}\boldsymbol{\xi}^i_{\tau,h}||_{(\Sigma^i_{k,h})^{-1}} \leq ||\sum_{\tau=1}^{k-1}\boldsymbol{\xi}^i_{\tau,h}||_{(\mathbf{W}^i_h+\lambda\mathbf{I})^{-1}} \leq \frac{1}{\sqrt{\rho_{\min}}}||\sum_{\tau=1}^{k-1}\boldsymbol{\xi}^i_{\tau,h}||_2 \tag{26}$$

where the first inequality holds because $\Sigma^i_{k,h} \succeq \mathbf{W}^i_h + \lambda\mathbf{I}$, and the second inequality holds due to the definition of event $\mathcal{E}_1$. So, with probability $1-\alpha/6H$, we have

$$||\sum_{\tau=1}^{k-1}\boldsymbol{\xi}^i_{\tau,h}||_2 \leq \sigma\sqrt{(k-1)nd\log\frac{12ndH}{\alpha}}. \tag{27}$$

We define the event $\mathcal{E}_3$ as

$$\mathcal{E}_3 = \left\{\forall\ h \in [H] :\ ||\sum_{\tau=1}^{k-1}\boldsymbol{\xi}^i_{\tau,h}||_2 \leq \sigma\sqrt{(k-1)nd\log\frac{12ndH}{\alpha}}\right\}. \tag{28}$$

Taking the union bound on all the stage $h \in [H]$, we have

$$||\sum_{\tau=1}^{k-1}\boldsymbol{\xi}^i_{\tau,h}||_{(\Sigma^i_{k,h})^{-1}} \overset{(i)}{\leq} \frac{1}{\sqrt{\rho_{\min}}}||\sum_{\tau=1}^{k-1}\boldsymbol{\xi}^i_{\tau,h}||_2$$

$$\overset{(ii)}{\leq} \frac{\sigma\sqrt{(k-1)nd\log\frac{12ndH}{\alpha}}}{\sqrt{\rho_{\min}}}$$
$$\overset{(iii)}{=} \frac{\sigma\sqrt{(k-1)nd\log\frac{12ndH}{\alpha}}}{\sqrt{\sigma\sqrt{k-1}(\sqrt{4nd}+2\log(6H/\alpha))+\lambda}}$$
$$\overset{(iv)}{\leq} \frac{\sqrt{\sigma}(k-1)^{1/4}\sqrt{nd\log\frac{12ndH}{\alpha}}}{\sqrt{\sqrt{4nd}+2\log(6H/\alpha)}}$$
$$\overset{(v)}{\leq} (nd)^{1/4}\sqrt{\sigma}(k-1)^{1/4}\sqrt{\log(12ndH/\alpha)}. \tag{29}$$

The inequality (i) follows from the relation between the l_2 norm and the norm on $(\Sigma^i_{k,h})^{-1}$ and hence bounded by the $\frac{1}{\sqrt{\rho_{\min}}}$. The inequality (ii) follows from the definition of event $\mathcal{E}_3$. In (iii) we use the definition of $\rho_{\min}$. Inequality (iv) follows after dropping $\lambda \geq 0$ in the previous expression. Finally, (v) follows by combining the common terms.

Combining the above bounds for $\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3$ from Equations equation 21, equation 25, and equation 29, we have that with probability at least $1-\alpha/2$, we have for each $i \in N$ and for each $h \in [H]$,

$$||(\Sigma^i_{k,h})^{1/2}(\hat{\boldsymbol{\theta}}^i_{k,h} - \boldsymbol{\theta}^\star_h)|| \leq \beta_k \tag{30}$$

here,

$$\beta_k = c(nd)^{3/4}\sqrt{\sigma}k^{1/4}\log(ndT/\alpha) \tag{31}$$

where c is an absolute constant (different for different mechanisms.) □

The next section identifies the β_k for each noise adding mechanism.

C.1 β^G_k for the Gaussian mechanism

For the (ϵ, δ) LDP for the Gaussian noise adding mechanism, the σ is taken as $\sigma = \frac{2H\sqrt{2\log(2.5H/\delta)}\Delta f}{\epsilon}$. This is obtained using the Lemma G.2 given in Appendix G.2. Moreover, Δf is the l_2 sensitivity identified as $2H^2$. Using this $\sigma = \frac{4H^3\sqrt{2\log(2.5H/\delta)}}{\epsilon}$ in equation 31, we have β^G_k for the Gaussian mechanism is

$$\beta^G_k = c_g(nd)^{3/4}H^{3/2}k^{1/4}\log(ndT/\alpha)(\log(H/\delta))^{1/4}\sqrt{1/\epsilon}. \tag{32}$$

C.2 β^L_k for the Laplace Mechanism

The variance of the Laplace mechanism is $2b^2$, and for $(\epsilon, 0)$ LDP we identify $b = \frac{2H\Delta f}{\epsilon}$, where Δf is the l_1 sensitivity. The l_1 sensitivity in MA-LDP is $2H^2\sqrt{nd}$. Therefore, $b = \frac{4H^3\sqrt{nd}}{\epsilon}$. Substituting this in equation 31, we have β^L_k for the Laplace mechanism as

$$\beta^L_k = c_l(nd)^{3/4}H^{3/2}k^{1/4}\log(ndT/\alpha)\sqrt{1/\epsilon}. \tag{33}$$

C.3 β^U_k for the Uniform Mechanism

The variance of the uniform mechanism is $a^2/3$, and for $(0, \delta)$ LDP we identify $a = 4H^3\sqrt{\log\left(\frac{2H}{\delta}\right)}$. Substituting this in equation 31, we have β^L_k for the Laplace mechanism as

$$\beta^U_k = c_u(nd)^{3/4}H^{3/2}k^{1/4}\log(ndT/\alpha)(\log(H/\delta))^{1/4}. \tag{34}$$

C.4 β_k^{BL} for the Bounded Laplace Mechanism

The variance of the bounded Laplace mechanism is $\frac{2b^2}{1-\exp(\frac{-B}{b})} - \kappa$, where $\kappa = \frac{((B+b)^2+b^2)\times\exp(\frac{-B}{b})}{1-\exp(\frac{-B}{b})}$. Thus, for $(\epsilon, 0)$-LDP, similar to Laplace with unbounded support, we identify $b = \frac{4H^3\sqrt{nd}}{\epsilon}$. Substituting this in equation 31, we have β_k^{BL} for the bounded Laplace mechanism as

$$\beta_k^{BL} = c_{bl}(nd)^{3/4}\zeta^{1/4}k^{1/4}\log(ndT/\alpha). \tag{35}$$

D Proof of Privacy Guarantee of Noise Adding Mechanisms

In this section, we prove the privacy guarantees of the MA-LDP algorithm for all the noise adding mechanisms. To this end, we first find the l_1 and l_2 sensitivity of the information shared to the server by each agent $i \in N$. To this end, we first compute the l_2 sensitivity coefficient for the MA-DP algorithm. Let $\Delta\tilde{u}_{k,h}^i$ and $\Delta\tilde{\Lambda}_{k,h}^i$ be the noise-free information of agent $i \in N$ at the h-th stage of k-th episode. That is,

$$\begin{aligned}
\Delta\tilde{u}_{k,h}^i &= \phi_{V_{k,h+1}^i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})V_{k,h+1}^i(s_{k,h+1}),\\
\Delta\tilde{\Lambda}_{k,h}^i &= \phi_{V_{k,h+1}^i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})\phi_{V_{k,h+1}^i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})^\top.
\end{aligned}$$

For $\Delta\tilde{u}_{k,h}^i$, the ℓ_2 sensitivity coefficient is upper bounded as

$$||\Delta\tilde{u}_{k,h}^i - (\Delta\tilde{u}_{k,h}^i)'||_2 \le ||\phi_{V_{k,h+1}^i}|| \cdot |V_{k,h+1}^i| + ||\phi'_{V_{k,h+1}^i}|| \cdot |V_{k,h+1}'^i| \le 2H^2 \tag{36}$$

Similarly, the ℓ_2 sensitivity of $\Delta\tilde{\Lambda}_{k,h}^i$ is upper bounded as

$$\begin{aligned}
||\phi_{V^i}\phi_{V^i}^\top - \phi'_{V^i}\phi'^\top_{V^i}||_F &\le ||\phi_{V^i}\phi_{V^i}^\top||_F + ||\phi'_{V^i}\phi'^\top_{V^i}||_F\\
&= \sqrt{tr[\phi_{V^i}\phi_{V^i}^\top\phi_{V^i}\phi_{V^i}^\top]} + \sqrt{tr[\phi'_{V^i}\phi'^\top_{V^i}\phi'_{V^i}\phi'^\top_{V^i}]}\\
&= \phi_{V^i}^\top\phi_{V^i} + \phi'^\top_{V^i}\phi'_{V^i}\\
&\le 2H^2.
\end{aligned}$$

Thus, the l_2 sensitivity of both the information is $2H^2$. Next, we find the l_1 sensitivities. Recall, for any matrix $A \in \mathbb{R}^{l\times l}$, we have that $||A||_1 \le \sqrt{l}||A||_2$. Similarly, for any vector $\mathbf{x} \in \mathbb{R}^l$, we have $||\mathbf{x}||_1 \le \sqrt{l}||\mathbf{x}||_2$. Using this property on the information we have

$$||\Delta\tilde{u}_{k,h}^i - (\Delta\tilde{u}_{k,h}^i)'||_2 \le \sqrt{nd}||\Delta\tilde{u}_{k,h}^i - (\Delta\tilde{u}_{k,h}^i)'||_1 \le 2\sqrt{nd}H^2. \tag{37}$$

Similarly, we have

$$\begin{aligned}
||\phi_{V^i}\phi_{V^i}^\top - \phi'_{V^i}\phi'^\top_{V^i}||_F &\le ||\phi_{V^i}\phi_{V^i}^\top||_F + ||\phi'_{V^i}\phi'^\top_{V^i}||_F\\
&\le \sqrt{nd}||\phi_{V^i}\phi_{V^i}^\top||_1 + \sqrt{nd}||\phi'_{V^i}\phi'^\top_{V^i}||_1 \le 2\sqrt{nd}H^2
\end{aligned}$$

Thus, the l_1 sensitivity of both the information is $2\sqrt{nd}H^2$. Let $\mathbf{D}_h = (D_h^1, D_h^2, \ldots, D_h^n)$ and $\mathbf{D}'_h = (D_h'^1, D_h'^2, \ldots, D_h'^n)$ are the different datasets collected by the server at stage h. For simplicity of notation, let $\mathbf{M} = (\mathbf{M}^1, \mathbf{M}^2, \ldots, \mathbf{M}^n)$ and let $\boldsymbol{\alpha} = (\boldsymbol{\alpha}^1, \boldsymbol{\alpha}^2, \ldots, \boldsymbol{\alpha}^n)$. Moreover, let $(\mathbf{M}, \boldsymbol{\alpha})$ be a possible outcome of the algorithm. Further, let $\Delta\boldsymbol{\Lambda}_{k,h} = (\Delta\Lambda_{k,h}^1, \Delta\Lambda_{k,h}^2, \ldots, \Delta\Lambda_{k,h}^n)$ and $\Delta\mathbf{u}_{k,h} = (\Delta u_{k,h}^1, \Delta u_{k,h}^2, \ldots, \Delta u_{k,h}^n)$ be the information from all the agents. Let $\mathbf{D}_{1:h-1}$ be the information collected from stage 1 to stage h, i.e., $\mathbf{D}_{1:h-1} = (\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_{h-1})$. Also let $\mathbf{W}_{k,h} = (W_{k,h}^1, W_{k,h}^2, \ldots, W_{k,h}^n)$ and $\boldsymbol{\xi}_{k,h} = (\xi_{k,h}^1, \xi_{k,h}^2, \ldots, \xi_{k,h}^n)$. Then, we have

$$\begin{aligned}
&\frac{\mathbb{P}(\forall\ h \in [H], (\Delta\boldsymbol{\Lambda}_{k,h}, \Delta\mathbf{u}_{k,h}) = (\mathbf{M}, \boldsymbol{\alpha}) \mid \mathbf{D}_{1:h-1})}{\mathbb{P}(\forall\ h \in [H], ((\Delta\boldsymbol{\Lambda}_{k,h})', (\Delta\mathbf{u}_{k,h})') = (\mathbf{M}, \boldsymbol{\alpha}) \mid \mathbf{D}'_{1:h-1})}\\
&= \prod_{h=1}^{H} \frac{\mathbb{P}((\mathbf{W}_{k,h}, \boldsymbol{\xi}_{k,h}) = (\mathbf{M} - \Delta\widetilde{\boldsymbol{\Lambda}}_{k,h}, \boldsymbol{\alpha} - \Delta\widetilde{\mathbf{u}}_{k,h}) \mid \mathbf{D}_{1:h-1})}{\mathbb{P}(((\mathbf{W}_{k,h})', (\boldsymbol{\xi}_{k,h})') = (\mathbf{M} - (\Delta\widetilde{\boldsymbol{\Lambda}}_{k,h})', \boldsymbol{\alpha} - (\Delta\widetilde{\mathbf{u}}_{k,h})') \mid \mathbf{D}'_{1:h-1})}
\end{aligned}$$

$$= \prod_{h=1}^{H} \frac{\mathbb{P}((\mathbf{W}_{k,h}, \boldsymbol{\xi}_{k,h}) = (\mathbf{M} - \Delta\widetilde{\boldsymbol{\Lambda}}_{k,h}, \boldsymbol{\alpha} - \Delta\widetilde{\mathbf{u}}_{k,h}) \mid \mathbf{D}_{h-1})}{\mathbb{P}(((\mathbf{W}_{k,h})', (\boldsymbol{\xi}_{k,h})') = (\mathbf{M} - (\Delta\widetilde{\boldsymbol{\Lambda}}_{k,h})', \boldsymbol{\alpha} - (\Delta\widetilde{\mathbf{u}}_{k,h})') \mid \mathbf{D}'_{h-1})}$$
$$= \prod_{h=1}^{H} \frac{\mathbb{P}(\mathbf{W}_{k,h} = \mathbf{M} - \Delta\widetilde{\boldsymbol{\Lambda}}_{k,h} \mid \mathbf{D}_{h-1}) \times \mathbb{P}(\boldsymbol{\xi}_{k,h} = \boldsymbol{\alpha} - \Delta\widetilde{\mathbf{u}}_{k,h} \mid \mathbf{D}_{h-1})}{\mathbb{P}((\mathbf{W}_{k,h})' = \mathbf{M} - (\Delta\widetilde{\boldsymbol{\Lambda}}_{k,h})' \mid \mathbf{D}'_{h-1}) \times \mathbb{P}((\boldsymbol{\xi}_{k,h})' = \boldsymbol{\alpha} - (\Delta\widetilde{\mathbf{u}}_{k,h})' \mid \mathbf{D}'_{h-1})}. \tag{38}$$

The first and the second equations again use the Markov property, and the last inequality is true because of the independence of the two informations, $\mathbf{W}_{k,h}$ and $\boldsymbol{\xi}_{k,h}$.

D.1 (ϵ, δ) Privacy for Gaussian Mechanism

Recall Theorem 2: If we chose the parameter σ of the Gaussian distribution $\mathcal{N}(0, \sigma^2)$ such that $\sigma = 4H^3\sqrt{2\log(2.5H/\delta)}/\epsilon$, then, MA-LDP algorithm with Gaussian mechanism $\mathcal{M}_G$ preserves (ϵ, δ)-MA-LDP privacy.

Proof. Consider $\mathbb{P}(\mathbf{W}_{k,h} = \mathbf{M} - \Delta\widetilde{\boldsymbol{\Lambda}}_{k,h} \mid \mathbf{D}_{h-1})$. Since the Gaussian noise in the information is independent across the agents, we have that

$$\mathbb{P}(\mathbf{W}_{k,h} = \mathbf{M} - \Delta\widetilde{\boldsymbol{\Lambda}}_{k,h} \mid \mathbf{D}_{h-1}) = \prod_{i \in N} \mathbb{P}(W^i_{k,h} = \mathbf{M}^i - \Delta\widetilde{\Lambda}^i_{k,h} \mid \mathbf{D}_{h-1}) \tag{39}$$

Now from Lemma G.2, and the sensitivity definition of $\Delta\widetilde{\Lambda}^i_{k,h}$, for $\sigma = 4H^3\sqrt{2\log(2.5H/\delta)}/\epsilon$, then with probability at least $1 - \delta/2nH$ for each $W^i_{k,h}$ for each agent $i \in N$, we have that

$$\mathbb{P}(W^i_{k,h} = \mathbf{M}^i - \Delta\widetilde{\Lambda}^i_{k,h} \mid \mathbf{D}_{h-1}) \leq \exp\left(\frac{\epsilon}{2nH}\right) \times \mathbb{P}((W^i_{k,h})' = \mathbf{M}^i - (\Delta\widetilde{\Lambda}^i_{k,h})' \mid \mathbf{D}'_{h-1}). \tag{40}$$

Taking the union bound on the agents and applying the composition theorem, we have that with probability at least $1 - \delta/(2H)$

$$\mathbb{P}(W_{k,h} = \mathbf{M} - \Delta\widetilde{\Lambda}_{k,h} \mid \mathbf{D}_{h-1}) \leq \exp\left(\frac{\epsilon}{2H}\right) \times \mathbb{P}((W_{k,h})' = \mathbf{M} - (\Delta\widetilde{\Lambda}_{k,h})' \mid \mathbf{D}'_{h-1}) \tag{41}$$

Moreover, for the other term $\mathbb{P}(\boldsymbol{\xi}_{k,h} = \boldsymbol{\alpha} - \Delta\widetilde{\mathbf{u}}_{k,h} \mid \mathbf{D}_{h-1})$ note again that by the independence of the noise distribution, we have

$$\mathbb{P}(\boldsymbol{\xi}_{k,h} = \boldsymbol{\alpha} - \Delta\widetilde{\mathbf{u}}_{k,h} \mid \mathbf{D}_{h-1}) = \prod_{i \in N} \mathbb{P}(\xi^i_{k,h} = \boldsymbol{\alpha}^i - \Delta\widetilde{u}^i_{k,h} \mid \mathbf{D}_{h-1}) \tag{42}$$

Using the property of the Gaussian distribution, we have

$$\begin{aligned}
\frac{\mathbb{P}(\xi^i_{k,h} = \boldsymbol{\alpha}^i - \Delta\widetilde{u}^i_{k,h} \mid \mathbf{D}_{h-1})}{\mathbb{P}((\xi^i_{k,h})' = \boldsymbol{\alpha}^i - (\Delta\widetilde{u}^i_{k,h})' \mid \mathbf{D}'_{h-1})} &= \frac{\exp\left(-||\boldsymbol{\alpha}^i - \Delta\widetilde{u}^i_{k,h}||^2/2\sigma^2\right)}{\exp\left(-||\boldsymbol{\alpha}^i - \Delta\widetilde{u}^i_{k,h} + (\Delta\widetilde{u}^i_{k,h} - (\Delta\widetilde{u}^i_{k,h})')||^2/2\sigma^2\right)} \\
&= \exp\left(-\frac{||\boldsymbol{\alpha}^i - \Delta\widetilde{u}^i_{k,h}||^2}{2\sigma^2} + \frac{||\boldsymbol{\alpha}^i - \Delta\widetilde{u}^i_{k,h} + (\Delta\widetilde{u}^i_{k,h} - (\Delta\widetilde{u}^i_{k,h})')||^2}{2\sigma^2}\right) \\
&\leq \exp\left(-\frac{||\boldsymbol{\alpha}^i - \Delta\widetilde{u}^i_{k,h}||^2}{2\sigma^2} + \frac{||\boldsymbol{\alpha}^i - \Delta\widetilde{u}^i_{k,h}||^2}{2\sigma^2}\right. \\
&\qquad \left. + \frac{||\Delta\widetilde{u}^i_{k,h} - (\Delta\widetilde{u}^i_{k,h})'||^2}{2\sigma^2}\right) \\
&= \exp\left(\frac{||\Delta\widetilde{u}^i_{k,h} - (\Delta\widetilde{u}^i_{k,h})'||^2}{2\sigma^2}\right).
\end{aligned}$$

Here the first equation is due to the Gaussian mechanism. The first inequality follows from the triangle inequality. Again using the Lemma A.2 of Liao et al. (2021), with probability at least $1-\delta/(2H)$ for each agent $i \in N$, we have that

$$\mathbb{P}(\xi^i_{k,h} = \boldsymbol{\alpha}^i - \Delta\widetilde{u}^i_{k,h} \mid \mathbf{D}_{h-1}) \leq \exp\left(\frac{\epsilon}{2nH}\right) \times \mathbb{P}((\xi^i_{k,h})' = \boldsymbol{\alpha}^i - (\Delta\widetilde{u}^i_{k,h})' \mid \mathbf{D}'_{h-1}) \tag{43}$$

Therefore, with probability $1-\delta/(2H)$, we have

$$\mathbb{P}(\boldsymbol{\xi}_{k,h} = \boldsymbol{\alpha} - \Delta\widetilde{\mathbf{u}}_{k,h} \mid \mathbf{D}_{h-1}) \leq \exp\left(\frac{\epsilon}{2H}\right) \times \mathbb{P}((\boldsymbol{\xi}_{k,h})' = \boldsymbol{\alpha} - (\Delta\widetilde{\mathbf{u}}_h)' \mid \mathbf{D}'_{h-1}) \tag{44}$$

Now taking the union bound for $\mathbf{W}_{k,h}, \boldsymbol{\xi}_{k,h}$ terms and all the stages $h \in [H]$, with probability at least $1-(2H)\times\delta/(2H) = 1-\delta$, we have

$$\log\left[\frac{\mathbb{P}(\forall\ h \in [H], (\Delta\boldsymbol{\Lambda}_{k,h}, \Delta\mathbf{u}_{k,h}) = (\mathbf{M}, \boldsymbol{\alpha}) \mid \mathbf{D}_{1:h-1})}{\mathbb{P}(\forall\ h \in [H], ((\Delta\boldsymbol{\Lambda}_{k,h})', (\Delta\mathbf{u}_{k,h})') = (\mathbf{M}, \boldsymbol{\alpha}) \mid \mathbf{D}'_{1:h-1})}\right] \leq \epsilon. \tag{45}$$

Therefore, from Theorem G.2, we conclude that our MA-LDP algorithm preserves (ϵ, δ)-LDP property with the Gaussian mechanism. □

D.2 $(\epsilon, 0)$ Privacy for Laplace Mechanism

Recall Theorem 4: If we choose the parameter b of the Laplace distribution $\mathcal{L}(b)$ such that $b = \frac{4H^3\sqrt{nd}}{\epsilon}$, then, MA-LDP algorithm with Laplace mechanism $\mathcal{M}_L$ preserves $(\epsilon, 0)$-MA-LDP privacy.

Proof. To prove that algorithm 1 with Laplace noise adding mechanism achieves $(\epsilon, 0)$-LDP, we need to show that equation 38 is upper bounded by e^ϵ with probability 1. Using the independence of the information across the agents, we have that

$$\mathbb{P}(\mathbf{W}_{k,h} = \mathbf{M} - \Delta\widetilde{\boldsymbol{\Lambda}}_{k,h} \mid \mathbf{D}_{h-1}) = \prod_{i\in N} \mathbb{P}(W^i_{k,h} = \mathbf{M}^i - \Delta\widetilde{\Lambda}^i_{k,h} \mid \mathbf{D}_{h-1}) \tag{46}$$

Now from Theorem 2, and the sensitivity definition of $\Delta\widetilde{\Lambda}^i_{k,h}$, if we set $b = \frac{4H^3\sqrt{nd}}{\epsilon}$, then for each $W^i_{k,h}$, for each agent $i \in N$, we have

$$\mathbb{P}(W^i_{k,h} = \mathbf{M}^i - \Delta\widetilde{\Lambda}^i_{k,h} \mid \mathbf{D}_{h-1}) \leq \exp\left(\frac{\epsilon}{2nH}\right) \times \mathbb{P}((W^i_{k,h})' = \mathbf{M}^i - (\Delta\widetilde{\Lambda}^i_{k,h})' \mid \mathbf{D}'_{h-1}). \tag{47}$$

Taking the union bound on the agents and applying the composition theorem, we have that with probability 1,

$$\mathbb{P}(W_{k,h} = \mathbf{M} - \Delta\widetilde{\Lambda}_{k,h} \mid \mathbf{D}_{h-1}) \leq \exp\left(\frac{\epsilon}{2H}\right) \times \mathbb{P}((W_{k,h})' = \mathbf{M} - (\Delta\widetilde{\Lambda}_{k,h})' \mid \mathbf{D}'_{h-1}). \tag{48}$$

Moreover, for the other term $\mathbb{P}(\boldsymbol{\xi}_{k,h} = \boldsymbol{\alpha} - \Delta\widetilde{\mathbf{u}}_{k,h} \mid \mathbf{D}_{h-1})$ note again that by the independence of the noise distribution, we have

$$\mathbb{P}(\boldsymbol{\xi}_{k,h} = \boldsymbol{\alpha} - \Delta\widetilde{\mathbf{u}}_{k,h} \mid \mathbf{D}_{h-1}) = \prod_{i\in N} \mathbb{P}(\xi^i_{k,h} = \boldsymbol{\alpha}^i - \Delta\widetilde{u}^i_{k,h} \mid \mathbf{D}_{h-1}). \tag{49}$$

Using the property of the Laplace distribution, we have

$$\begin{aligned}
\frac{\mathbb{P}(\xi^i_{k,h} = \boldsymbol{\alpha}^i - \Delta\widetilde{u}^i_{k,h} \mid \mathbf{D}_{h-1})}{\mathbb{P}((\xi^i_{k,h})' = \boldsymbol{\alpha}^i - (\Delta\widetilde{u}^i_{k,h})' \mid \mathbf{D}'_{h-1})} &= \prod_{j=1}^{nd} \frac{\exp\left(\frac{-\epsilon|(\boldsymbol{\alpha}^i)_j - (\Delta\widetilde{u}^i_{k,h})_j|}{2nH||\Delta\widetilde{u}^i_{k,h} - (\Delta\widetilde{u}^i_{k,h})'||_1}\right)}{\exp\left(\frac{-\epsilon|(\boldsymbol{\alpha}^i)_j - ((\Delta\widetilde{u}^i_{k,h})')_j|}{2nH||\Delta\widetilde{u}^i_{k,h} - (\Delta\widetilde{u}^i_{k,h})'||_1}\right)} \\
&= \prod_{j=1}^{nd} \exp\left(\frac{-\epsilon|(\boldsymbol{\alpha}^i)_j - (\Delta\widetilde{u}^i_{k,h})_j| + \epsilon|(\boldsymbol{\alpha}^i)_j - ((\Delta\widetilde{u}^i_{k,h})')_j|}{2nH||\Delta\widetilde{u}^i_{k,h} - (\Delta\widetilde{u}^i_{k,h})'||_1}\right)
\end{aligned}$$

$$\leq \prod_{j=1}^{nd} \exp\left(\frac{\epsilon|(\Delta\widetilde{u}^i_{k,h})_j - ((\Delta\widetilde{u}^i_{k,h})')_j|}{2nH||\Delta\widetilde{u}^i_{k,h} - (\Delta\widetilde{u}^i_{k,h})'||_1}\right)$$
$$= \exp\left(\frac{\epsilon}{2nH}\right).$$

In the above, we use subscript j to denote the j-th element of the corresponding vector. Here the first equation is due to the Laplace mechanism. The first inequality follows from the triangle inequality. So with probability 1 for each agent $i \in N$, we have that

$$\mathbb{P}(\xi^i_{k,h} = \boldsymbol{\alpha}^i - \Delta\widetilde{u}^i_{k,h} \mid \mathbf{D}_{h-1}) \leq \exp\left(\frac{\epsilon}{2nH}\right) \times \mathbb{P}((\xi^i_{k,h})' = \boldsymbol{\alpha}^i - (\Delta\widetilde{u}^i_{k,h})' \mid \mathbf{D}'_{h-1}) \tag{50}$$

Therefore, we have with probability 1 that

$$\mathbb{P}(\boldsymbol{\xi}_{k,h} = \boldsymbol{\alpha} - \Delta\widetilde{\mathbf{u}}_{k,h} \mid \mathbf{D}_{h-1}) \leq \exp\left(\frac{\epsilon}{2H}\right) \times \mathbb{P}((\boldsymbol{\xi}_{k,h})' = \boldsymbol{\alpha} - (\Delta\widetilde{\mathbf{u}}_h)' \mid \mathbf{D}'_{h-1}) \tag{51}$$

Now taking the union bound for $\mathbf{W}_{k,h}, \boldsymbol{\xi}_{k,h}$ terms and all the stages $h \in [H]$, with probability 1, we have

$$\log\left[\frac{\mathbb{P}(\forall\ h \in [H], (\Delta\boldsymbol{\Lambda}_{k,h}, \Delta\mathbf{u}_{k,h}) = (\mathbf{M}, \boldsymbol{\alpha}) \mid \mathbf{D}_{1:h-1})}{\mathbb{P}(\forall\ h \in [H], ((\Delta\boldsymbol{\Lambda}_{k,h})', (\Delta\mathbf{u}_{k,h})') = (\mathbf{M}, \boldsymbol{\alpha}) \mid \mathbf{D}'_{1:h-1})}\right] \leq \epsilon. \tag{52}$$

Therefore, the MA-LDP algorithm preserves $(\epsilon, 0)$-LDP property with the Laplace mechanism. □

D.3 $(0, \delta)$ Privacy for Uniform Mechanism

Recall Theorem 6: If we choose $a = 4H^3\sqrt{\log(2H/\delta)}$ for the uniform distribution $\mathcal{U}[-a, a]$ then, MA-LDP algorithm with uniform mechanism $\mathcal{M}_U$ preserves $(0, \delta)$-MA-LDP privacy.

Proof. For the uniform distribution, $\epsilon = 0$, thus setting $a = 4H^3\log(2H/\delta)$ satisfies the following with probability at least $1 - \delta/(2H)$

$$\mathbb{P}(W^i_{k,h} = \mathbf{M}^i - \Delta\widetilde{\Lambda}^i_{k,h} \mid \mathbf{D}_{h-1}) \leq \mathbb{P}((W^i_{k,h})' = \mathbf{M}^i - (\Delta\widetilde{\Lambda}^i_{k,h})' \mid \mathbf{D}'_{h-1}). \tag{53}$$

Moreover, for the other term $\mathbb{P}(\boldsymbol{\xi}_{k,h} = \boldsymbol{\alpha} - \Delta\widetilde{\mathbf{u}}_{k,h} \mid \mathbf{D}_{h-1})$ note again that by the independence of the noise distribution, we have

$$\mathbb{P}(\boldsymbol{\xi}_{k,h} = \boldsymbol{\alpha} - \Delta\widetilde{\mathbf{u}}_{k,h} \mid \mathbf{D}_{h-1}) = \prod_{i \in N} \mathbb{P}(\xi^i_{k,h} = \boldsymbol{\alpha}^i - \Delta\widetilde{u}^i_{k,h} \mid \mathbf{D}_{h-1}). \tag{54}$$

Using the property of uniform distribution, we have that

$$\frac{\mathbb{P}(\xi^i_{k,h} = \boldsymbol{\alpha}^i - \Delta\widetilde{u}^i_{k,h} \mid \mathbf{D}_{h-1})}{\mathbb{P}((\xi^i_{k,h})' = \boldsymbol{\alpha}^i - (\Delta\widetilde{u}^i_{k,h})' \mid \mathbf{D}'_{h-1})} = \prod_{j=1}^{nd} \frac{1/2a}{1/2a}$$
$$= 1 = \exp(0).$$

So, for each agent $i \in N$, we have with probability at least $1 - \delta/(2nH)$, the following

$$\mathbb{P}(\xi^i_{k,h} = \boldsymbol{\alpha}^i - \Delta\widetilde{u}^i_{k,h} \mid \mathbf{D}_{h-1}) \leq \exp(0) \times \mathbb{P}((\xi^i_{k,h})' = \boldsymbol{\alpha}^i - (\Delta\widetilde{u}^i_{k,h})' \mid \mathbf{D}'_{h-1}). \tag{55}$$

Therefore, we have with probability $1 - \delta/(2H)$ that

$$\mathbb{P}(\boldsymbol{\xi}_{k,h} = \boldsymbol{\alpha} - \Delta\widetilde{\mathbf{u}}_{k,h} \mid \mathbf{D}_{h-1}) \leq \exp(0) \times \mathbb{P}((\boldsymbol{\xi}_{k,h})' = \boldsymbol{\alpha} - (\Delta\widetilde{\mathbf{u}}_h)' \mid \mathbf{D}'_{h-1}). \tag{56}$$

Now taking the union bound for $\mathbf{W}_{k,h}, \boldsymbol{\xi}_{k,h}$ terms and all the stages $h \in [H]$, with probability $1 - \delta$, we have

$$\log\left[\frac{\mathbb{P}(\forall\ h \in [H], (\Delta\boldsymbol{\Lambda}_{k,h}, \Delta\mathbf{u}_{k,h}) = (\mathbf{M}, \boldsymbol{\alpha}) \mid \mathbf{D}_{1:h-1})}{\mathbb{P}(\forall\ h \in [H], ((\Delta\boldsymbol{\Lambda}_{k,h})', (\Delta\mathbf{u}_{k,h})') = (\mathbf{M}, \boldsymbol{\alpha}) \mid \mathbf{D}'_{1:h-1})}\right] = 0. \tag{57}$$

Therefore, the MA-LDP algorithm preserves $(0, \delta)$-LDP property with the uniform mechanism. □

D.4 $(\epsilon, 0)$ Privacy for Bounded Laplace Mechanism

Recall Theorem 8: If we choose the parameter $b = \frac{4H^3\sqrt{nd}}{\epsilon}$ for the bounded Laplace distribution $\mathcal{BL}(b; B)$ then, MA-LDP algorithm with BL mechanism $\mathcal{M}_{BL}$ preserves $(\epsilon, 0)$-MA-LDP privacy.

Proof. The proof of $(\epsilon, 0)$ LDP for the bounded Laplace is the same as the Laplace as given in Appendix D.2, so we avoid writing it. □

E Proof of Regret Bounds - Theorem 1

In this section, we give proof of the upper bound on the regret of the MA-LDP algorithm with different noise adding mechanisms. We first bound the regret for generic noise adding mechanisms and use the specific mechanism to give the explicit bound.

Recall Theorem 1: Under the assumptions 1, 2, and 3, for any user k, with probability at least $1-\alpha$, the total regret of MA-LDP algorithm in the first T steps is at most

$$R_K \leq H\beta_K\sqrt{2ndK\log(1+K/\lambda)} + 4H\sqrt{2T\log(2H/\alpha)}. \tag{58}$$

Proof. Consider the following difference

$$\begin{aligned}
V_h^{\star,i}(\boldsymbol{s}_{k,h}) - V_h^{\pi_k,i}(\boldsymbol{s}_{k,h}) &\overset{(i)}{\leq} V_{k,h}^i(\boldsymbol{s}_{k,h}) - V_h^{\pi_k,i}(\boldsymbol{s}_{k,h}) \\
&= \max_{\boldsymbol{a}} Q_{k,h}^i(\boldsymbol{s}_{k,h}, \boldsymbol{a}) - \max_{\boldsymbol{a}} Q_h^{\pi_k,i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}) \\
&\overset{(ii)}{\leq} Q_{k,h}^i(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) - Q_h^{\pi_k,i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) \\
&\overset{(iii)}{\leq} \bar{r}_h(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}; \boldsymbol{w}_{k,h}^i) + \left\langle \hat{\boldsymbol{\theta}}_{k,h}^i, \phi_{V_{k,h+1}^i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) \right\rangle \\
&\quad + \beta_k ||\Sigma_{k,h}^{i^{-1/2}} \phi_{V_{k,h+1}^i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})||_2 \\
&\quad - \bar{r}_h(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}; \boldsymbol{w}_{k,h}^i) - \mathbb{P}_h V_{h+1}^{\pi_k,i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) \\
&= \left\langle \hat{\boldsymbol{\theta}}_{k,h}^i, \phi_{V_{k,h+1}^i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) \right\rangle + \beta_k ||\Sigma_{k,h}^{i^{-1/2}} \phi_{V_{k,h+1}^i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})||_2 \\
&\quad - \mathbb{P}_h V_{h+1}^{\pi_k,i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) + \mathbb{P}_h V_{k,h+1}^i(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) - \mathbb{P}_h V_{k,h+1}^i(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) \\
&= \left\langle \hat{\boldsymbol{\theta}}_{k,h}^i - \boldsymbol{\theta}_h^\star, \phi_{V_{k,h+1}^i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) \right\rangle + \beta_k ||\Sigma_{k,h}^{i^{-1/2}} \phi_{V_{k,h+1}^i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})||_2 \\
&\quad - \mathbb{P}_h V_{h+1}^{\pi_k,i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) + \mathbb{P}_h V_{k,h+1}^i(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) \\
&\overset{(iv)}{\leq} ||\Sigma_{k,h}^{i^{1/2}} \hat{\boldsymbol{\theta}}_{k,h}^i - \boldsymbol{\theta}_h^\star||_2 ||\Sigma_{k,h}^{i^{-1/2}} \phi_{V_{k,h+1}^i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})||_2 - \mathbb{P}_h V_{h+1}^{\pi_k,i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) \\
&\quad + \beta_k ||\Sigma_{k,h}^{i^{-1/2}} \phi_{V_{k,h+1}^i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})||_2 + \mathbb{P}_h V_{k,h+1}^i(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) \\
&\overset{(v)}{\leq} 2\beta_k ||\Sigma_{k,h}^{i^{-1/2}} \phi_{V_{k,h+1}^i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})||_2 \\
&\quad - \mathbb{P}_h V_{h+1}^{\pi_k,i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) + \mathbb{P}_h V_{k,h+1}^i(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})
\end{aligned} \tag{59}$$

$$\tag{60}$$

(i) follows from the previous Lemma 2, in (ii) we replace the max over all the actions by $\boldsymbol{a}_{k,h}$. The inequality (iii) uses the update of the state-action value function from line 3 of user sub-routine. In (iv), we use the Cauchy-Schwartz inequality. Finally (v) follows from the Lemma 3. Apart from the above, we also have the following

$$V_{h+1}^i(\boldsymbol{s}_{k,h}) - V_{h+1}^{\pi_k,i}(\boldsymbol{s}_{k,h}) \leq V_{h+1}^i(\boldsymbol{s}_{k,h}) \leq H. \tag{61}$$

Combining the Equations equation 60 and equation 61 we have the following:

$$\begin{aligned}
&V^i_{h+1}(\boldsymbol{s}_{k,h}) - V^{\pi_k,i}_{h+1}(\boldsymbol{s}_{k,h}) \\
&\leq \min\{H, 2\beta_k ||\Sigma^{i^{-1/2}}_{k,h} \phi_{V^i_{k,h+1}}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})||_2 - \mathbb{P}_h V^{\pi_k,i}_{h+1}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) + \mathbb{P}_h V^i_{k,h+1}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})\} \\
&\leq \min\{H, 2\beta_k ||\Sigma^{i^{-1/2}}_{k,h} \phi_{V^i_{k,h+1}}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})||_2\} - \mathbb{P}_h V^{\pi_k,i}_{h+1}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) + \mathbb{P}_h V^i_{k,h+1}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}),
\end{aligned}$$

where the second inequality holds because $V^i_{k,h+1} \geq V^{\star,i}_{h+1} \geq V^{\pi_k,i}_{h+1}$. Adding $V^{\pi_k,i}_{h+1}(\boldsymbol{s}_{k,h+1}) - V^i_{k,h+1}(\boldsymbol{s}_{k,h+1})$ to both sides in the above equation we have the following:

$$\begin{aligned}
&V^i_{h+1}(\boldsymbol{s}_{k,h}) - V^{\pi_k,i}_{h+1}(\boldsymbol{s}_{k,h}) + [V^{\pi_k,i}_{h+1}(\boldsymbol{s}_{k,h+1}) - V^i_{k,h+1}(\boldsymbol{s}_{k,h+1})] \\
&\leq \min\{H, 2\beta_k ||\Sigma^{i^{-1/2}}_{k,h} \phi_{V^i_{k,h+1}}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})||_2\} - \mathbb{P}_h V^{\pi_k,i}_{h+1}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) + \mathbb{P}_h V^i_{k,h+1}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) \\
&\quad + [V^{\pi_k,i}_{h+1}(\boldsymbol{s}_{k,h+1}) - V^i_{k,h+1}(\boldsymbol{s}_{k,h+1})].
\end{aligned} \tag{62}$$

Summing these inequalities for $k = 1, 2, \ldots, K$ and stages $h = h', \ldots, H$, we have

$$\begin{aligned}
\sum_{k=1}^{K} [V^i_{k,h'}(\boldsymbol{s}_{k,h'}) - V^{\pi_k,i}_{h'}(\boldsymbol{s}_{k,h'})] \leq \sum_{k=1}^{K} \sum_{h=h'}^{H} \Big[&2\beta_k \min\{1, ||\Sigma^{i^{-1/2}}_{k,h} \phi_{V^i_{k,h+1}}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})||_2\} \\
&+ [\mathbb{P}_h(V^i_{k,h+1} - V^{\pi_k,i}_{h+1})](\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) - [V^i_{k,h+1} - V^{\pi_k,i}_{h+1}](\boldsymbol{s}_{k,h+1}) \Big].
\end{aligned} \tag{63}$$

Define the following event $\mathcal{E}_4$ as

$$\begin{aligned}
\mathcal{E}_4 = \Big\{ \forall\ h' \in [H],\ \sum_{k=1}^{K} \sum_{h=h'}^{H} &[\mathbb{P}_h(V^i_{k,h+1} - V^{\pi_k,i}_{h+1})](\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) - [V^i_{k,h+1} - V^{\pi_k,i}_{h+1}](\boldsymbol{s}_{k,h+1}) \\
&\leq 4H\sqrt{2T \log(2H/\alpha)} \Big\}
\end{aligned} \tag{64}$$

Since, $[\mathbb{P}_h(V^i_{k,h+1} - V^{\pi_k,i}_{h+1})](\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) - [V^i_{k,h+1} - V^{\pi_k,i}_{h+1}](\boldsymbol{s}_{k,h+1})$ forms the martingale difference sequence and it is less than $4H$, i.e.,

$$[[\mathbb{P}_h(V^i_{k,h+1} - V^{\pi_k,i}_{h+1})](\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) - [V^i_{k,h+1} - V^{\pi_k,i}_{h+1}](\boldsymbol{s}_{k,h+1})] \leq 4H \tag{65}$$

Applying the Azuma-Hoeffdings inequality, we have that $\mathcal{E}_4$ holds with probability at least $1 - \alpha/2$. That is $\mathbb{P}(\mathcal{E}_4) \geq 1 - \alpha/2$. Recall, $\Sigma \succeq \lambda I$, and choosing $h' = 1$ we have

$$\begin{aligned}
&\sum_{k=1}^{K} \sum_{h=1}^{H} \beta_k \min\{1, ||\Sigma^{i^{-1/2}}_{k,h} \phi_{V^i_{k,h+1}}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})||_2\} \\
&\quad \leq \beta_K \sum_{k=1}^{K} \sum_{h=1}^{H} \min\{1, ||\Sigma^{i^{-1/2}}_{k,h} \phi_{V^i_{k,h+1}}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})||_2\} \\
&\quad \leq H\beta_K \sqrt{2ndK \log(1 + K/\lambda)}
\end{aligned}$$

The above inequalities hold because of the Cauchy-Schwartz and the Theorem G.3. Finally, on the events $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_4$, we conclude with probability at least $1 - \alpha$, we have the exact expression of the regret as follows

$$R_K \leq H\beta_K \sqrt{2ndK \log(1 + K/\lambda)} + 4H\sqrt{2T \log(2H/\alpha)}. \tag{66}$$

This ends the proof. □

E.1 Regret Bound for Gaussian Mechanism

Recall Theorem 3: Consider the Gaussian noise-adding mechanism with parameter σ as in Theorem 2. For any user k, with probability at least $1 - \alpha$, the total regret of MA-LDP algorithm in the first T steps is at most $\widetilde{\mathcal{O}}((nd)^{5/4} H^{7/4} T^{3/4} \log(ndT/\alpha)(\log(H/\delta))^{1/4} \sqrt{1/\epsilon})$.

Proof. To complete the regret bound for the MA-LDP algorithm with the Gaussian mechanism, we substitute β_K^G given in equation 32 in the regret expression given in equation 58. Recall,

$$\beta_K^G = c_g(nd)^{3/4}H^{3/2}K^{1/4}\log(ndT/\alpha)(\log(H/\delta))^{1/4}\sqrt{1/\epsilon}. \tag{67}$$

Thus, the regret of the MA-LDP algorithm with the Gaussian noise adding mechanism is given by

$$\begin{aligned} R_K^G &\leq c_g H(nd)^{3/4}H^{3/2}K^{1/4}\log(ndT/\alpha)(\log(H/\delta))^{1/4}\sqrt{1/\epsilon} \\ &\quad \cdot \sqrt{2ndK\log(1+K/\lambda)} + 4H\sqrt{2T\log(2H/\alpha)} \\ &\leq \widetilde{\mathcal{O}}(n^{5/4}d^{5/4}H^{7/4}T^{3/4}\log(ndT/\alpha)(\log(H/\delta))^{1/4}\sqrt{1/\epsilon}). \end{aligned} \tag{68}$$

This ends the proof of the regret upper bound for the Gaussian mechanism. □

E.2 Regret Bound for Laplace Mechanism

Recall Theorem 5: Consider the Laplace noise-adding mechanism with parameter b as in Theorem 4. For any user k, with probability at least $1-\alpha$, the total regret of MA-LDP algorithm in the first T steps is at most $\widetilde{\mathcal{O}}((nd)^{5/4}H^{7/4}T^{3/4}\log(ndT/\alpha)\sqrt{1/\epsilon})$.

Proof. To complete the regret bound for the MA-LDP algorithm with the Laplace mechanism, we substitute β_K^L given in equation 33 in the regret expression given in equation 58. Recall,

$$\beta_K^L = c_l(nd)^{3/4}H^{3/2}K^{1/4}\log(ndT/\alpha)\sqrt{1/\epsilon}. \tag{69}$$

Thus, the regret of the MA-LDP algorithm with the Gaussian noise adding mechanism is given by

$$\begin{aligned} R_K^L &\leq c_l H(nd)^{3/4}H^{3/2}K^{1/4}\log(ndT/\alpha)\sqrt{1/\epsilon} \\ &\quad \cdot \sqrt{2ndK\log(1+K/\lambda)} + 4H\sqrt{2T\log(2H/\alpha)} \\ &\leq \widetilde{\mathcal{O}}(n^{5/4}d^{5/4}H^{7/4}T^{3/4}\log(ndT/\alpha)\sqrt{1/\epsilon}). \end{aligned} \tag{70}$$

This ends the proof of the regret upper bound for the Laplace mechanism. □

E.3 Regret Bound for Uniform Mechanism

Recall Theorem 7: Consider the uniform noise-adding mechanism with parameter a as in Theorem 6. For any user k, with probability at least $1-\alpha$, the total regret of MA-LDP algorithm in the first T steps is at most $\widetilde{\mathcal{O}}(n^{5/4}d^{5/4}H^{7/4}\ T^{3/4}\ \log(ndT/\alpha)(\log(H/\delta))^{1/4})$.

Proof. To complete the regret bound for the MA-LDP algorithm with the uniform mechanism, we substitute β_K^U given in equation 34 in the regret expression given in equation 58. Recall,

$$\beta_K^U = c_u(nd)^{3/4}H^{3/2}k^{1/4}\log(ndT/\alpha)(\log(H/\delta))^{1/4}. \tag{71}$$

Thus, the regret of the MA-LDP algorithm with uniform noise adding mechanism is given by

$$\begin{aligned} R_K^L &\leq c_u H(nd)^{3/4}H^{3/2}k^{1/4}\log(ndT/\alpha)(\log(H/\delta))^{1/4} \\ &\quad \cdot \sqrt{2ndK\log(1+K/\lambda)} + 4H\sqrt{2T\log(2H/\alpha)} \\ &\leq \widetilde{\mathcal{O}}(n^{5/4}d^{5/4}H^{7/4}T^{3/4}\log(ndT/\alpha)(\log(1/\delta))^{1/4}). \end{aligned} \tag{72}$$

This ends the proof of the regret upper bound for the uniform mechanism. □

E.4 Regret Bound for Bounded Laplace Mechanism

Recall Theorem 9: Consider the BL noise mechanism with parameter b as in Theorem 8. For any user k, with probability at least $1-\alpha$, the total regret of MA-LDP algorithm in the first T steps is at most $\widetilde{\mathcal{O}}(n^{5/4}d^{5/4}\zeta^{1/4}H^{1/4}T^{3/4}\log(ndT/\alpha))$.

Proof. To complete the regret bound for the MA-LDP algorithm with the Bounded Laplace mechanism, we substitute β_K^{BL} given in equation 35 in the regret expression given in equation 58. Recall,

$$\beta_K^{BL} = c_{bl}(nd)^{3/4}\zeta^{1/4}K^{1/4}\log(ndT/\alpha). \tag{73}$$

and $\zeta = \frac{2b^2}{1-\exp(\frac{-B}{b})} - \kappa$, where $\kappa = \frac{((B+b)^2+b^2)\times\exp(\frac{-B}{b})}{1-\exp(\frac{-B}{b})}$. Thus, the regret of the MA-LDP algorithm with the Bounded Laplace noise adding mechanism is given by

$$\begin{aligned} R_K^{BL} &\leq c_{bl}H(nd)^{3/4}\zeta^{1/4}K^{1/4}\log(ndT/\alpha) \\ &\quad \cdot \sqrt{2ndK\log(1+K/\lambda)} + 4H\sqrt{2T\log(2H/\alpha)} \\ &\leq \widetilde{\mathcal{O}}(n^{5/4}d^{5/4}H^{1/4}T^{3/4}\zeta^{1/4}\log(ndT/\alpha)). \end{aligned} \tag{74}$$

This ends the proof of the regret upper bound for the Bounded Laplace mechanism. □

F More details of experiments

Here we give more details of the experiments and the other results stated in the main paper.

The following figure shows the MDP we consider in the experiments.

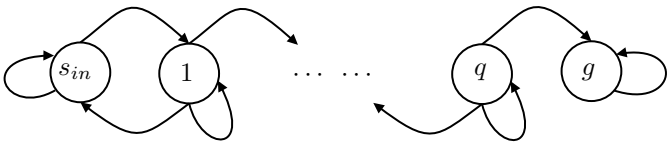

Figure 3: Network with $q+2$ nodes.

F.1 Feature design

The features we use in the experiments are as follows: Let $S(\boldsymbol{s})$ is the set all feasible states from state $\boldsymbol{s}$.

$$\phi(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) = \begin{cases} (\phi(s'^1|s^1,\boldsymbol{a}^1),\ldots,\phi(s'^n|s^n,\boldsymbol{a}^n)), & if\ \boldsymbol{s}\neq\boldsymbol{g},\ \boldsymbol{s}'\in S(\boldsymbol{s}) \\ \mathbf{0}_{nd}, & if\ \boldsymbol{s}\neq\boldsymbol{g},\ \ \boldsymbol{s}'\notin S(\boldsymbol{s}) \\ \mathbf{0}_{nd}, & if\ \boldsymbol{s}=\boldsymbol{g},\ \boldsymbol{s}'\neq\boldsymbol{g} \\ (\mathbf{0}_{nd-1},\alpha(\boldsymbol{s})), & if\ \boldsymbol{s}=\boldsymbol{g},\ \boldsymbol{s}'=\boldsymbol{g}, \end{cases} \tag{75}$$

where we identify $\alpha(\boldsymbol{s})$ as $\alpha(\boldsymbol{s}) = \frac{|S(\boldsymbol{s})|}{n}\left\{\frac{x_0}{2} + x_{q+1} + \sum_{j=1}^{q}\frac{x_j}{3}\right\}$. Here $x_0, x_1, \ldots, x_q, x_{q+1}$ are the number of agents at the nodes $s_{in}, 1, \ldots, q, g$ respectively in the state $\boldsymbol{s}$. The local features $\phi(s'^i|s^i, \boldsymbol{a}^i)$ are defined as

$$\phi(s'^i|s^i,\boldsymbol{a}^i) = \begin{cases} \left(-\boldsymbol{a}^i, \frac{1-\delta}{n}\right)^\top, & if\ s^i = s'^i = s_{in} \\ \left(\boldsymbol{a}^i, \frac{\delta}{n}\right)^\top, & if\ s^i = s_{in},\ s'^i = g \\ \left(-\boldsymbol{a}^i, \frac{1-\delta_{j,j}}{n}\right)^\top, & if\ s^i = s^{i'} = j \in \{1,2,\ldots,q\}, \\ \left(\boldsymbol{a}^i, \frac{\delta_{j,j+1}}{n}\right)^\top, & if\ s^i = j,\ s^{i'} = j+1,\ \forall\ j \in \{1,2,\ldots,q\}, \\ \left(\boldsymbol{0}_{d-1}, \frac{\delta_{j,j}-\delta_{j,j+1}}{n}\right)^\top, & if\ s^i = j,\ s^{i'} = j-1,\ \forall\ j \in \{1,2,\ldots,q\}, \\ \boldsymbol{0}_d^\top, & if\ s^i = g,\ s'^i = s_{in} \\ \left(\boldsymbol{0}_{d-1}, \frac{1}{n}\right)^\top, & if\ s^i = g,\ s'^i = g. \end{cases}$$

Here $\delta_{j,j} \geq \delta_{j,j+1}$, and $\boldsymbol{0}_d^\top = (0, 0, \ldots, 0)^\top$ of d dimension. Moreover, the transition probability parameters for any state $\boldsymbol{s}$ are taken as $\boldsymbol{\theta}(\boldsymbol{s}) = \left(\boldsymbol{\theta}^1, \frac{1}{\alpha(\boldsymbol{s})}, \boldsymbol{\theta}^2, \frac{1}{\alpha(\boldsymbol{s})} \ldots, \boldsymbol{\theta}^n, \frac{1}{\alpha(\boldsymbol{s})}\right)$ where $\boldsymbol{\theta}^i \in \left\{-\frac{\Delta}{n(d-1)}, \frac{\Delta}{n(d-1)}\right\}^{d-1}$, and $\Delta < \delta$.

We have following lemma that shows the validity of our feature design.

Lemma 4. *For every $\boldsymbol{\theta}(\boldsymbol{s})$, features $\phi(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a})$ satisfies the following: (a) $\sum_{\boldsymbol{s}'}\langle\phi(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a}), \boldsymbol{\theta}(\boldsymbol{s})\rangle = 1,\ \forall\ \boldsymbol{s}, \boldsymbol{a}$; (b) $\langle\phi(\boldsymbol{s}' = \boldsymbol{g}|\boldsymbol{s} = \boldsymbol{g}, \boldsymbol{a}), \boldsymbol{\theta}(\boldsymbol{s})\rangle = 1,\ \forall\ \boldsymbol{a}$; (c) $\langle\phi(\boldsymbol{s}' \neq \boldsymbol{g}|\boldsymbol{s} = \boldsymbol{g}, \boldsymbol{a}), \boldsymbol{\theta}(\boldsymbol{s})\rangle = 0,\ \forall\ \boldsymbol{a}$.*

Proof. We consider two cases. In case 1, $\boldsymbol{s} \neq \boldsymbol{g}$ and case 2, $\boldsymbol{s} = \boldsymbol{g}$.

Case 01: ($\boldsymbol{s} \neq \boldsymbol{g}$). Without loss of generality we consider the following state $\boldsymbol{s} = (\underbrace{s_{init}, \ldots, s_{init}}_{x_0\ times}, \underbrace{1, 1, \ldots, 1}_{x_1\ times}, \underbrace{2, 2, \ldots, 2}_{x_2\ times}, \underbrace{q-1, q-1, \ldots, q-1}_{x_{q-1}\ times}, \underbrace{q, q, \ldots, q}_{x_q\ times}, \underbrace{g, g, \ldots, g}_{x_{q+1}\ times})$, i.e., x_0 agents are at s_{init}, x_1 agents are at node 1, x_2 agents at node 2, and so on, finally remaining $x_{q+1} = n - x_0 - x_1 - \cdots - x_q$ agents are at g. Consider an agent i at s_{init} node. Let $|S(\boldsymbol{s})|$ denotes the number of next states feasible from state $\boldsymbol{s}$. A simple calculation shows that $|S(\boldsymbol{s})| = 2^{x_0} \times 3^{x_1} \times 3^{x_2} \cdots \times \ldots 3^{x_q} \times 1^{x_{q+1}}$. Out of these possible next states, there are exactly $\frac{|S(\boldsymbol{s})|}{2}$ states in which agent i will remain at s_{init}, and in $\frac{|S(\boldsymbol{s})|}{2}$ states the agent i moves to node 1. The probability that the next node of agent i is s_{init} given that the current node of agent i is s_{init} is given by $-\langle\boldsymbol{a}^i, \boldsymbol{\theta}^i\rangle + \frac{1-\delta}{n} \times \frac{1}{\alpha(\boldsymbol{s})}$. And the probability that the next node of agent i is 1 given that the current node of agent i is s_{init} is $\langle\boldsymbol{a}^i, \boldsymbol{\theta}^i\rangle + \frac{\delta}{n} \times \frac{1}{\alpha(\boldsymbol{s})}$. These probabilities are obtained using the corresponding features. Since this is true for all the agents $1, 2, \ldots, x_0$, which are at s_{init}. So, the contribution to the probability term from these x_0 agents who are at s_{init} is

$$\sum_{i=1}^{x_0}\left\{\left(-\langle\boldsymbol{a}^i, \boldsymbol{\theta}^i\rangle + \frac{1-\delta}{n} \times \frac{1}{\alpha(\boldsymbol{s})}\right) \times \frac{|S(\boldsymbol{s})|}{2}\right\} + \sum_{i=1}^{x_0}\left\{\left(\langle\boldsymbol{a}^i, \boldsymbol{\theta}^i\rangle + \frac{\delta}{n} \times \frac{1}{\alpha(\boldsymbol{s})}\right) \times \frac{|S(\boldsymbol{s})|}{2}\right\} = \frac{|S(\boldsymbol{s})|}{2} \times \frac{1}{n} \times \frac{x_0}{\alpha(\boldsymbol{s})}. \quad (76)$$

Next consider an agent i who is at node $j \in \{1, 2, \ldots, q\}$. Out of the next possible states, the number of next possible states where agent i will remain at node j is $\frac{|S(\boldsymbol{s})|}{3}$, move to node $j+1$ is $\frac{|S(\boldsymbol{s})|}{3}$ states and moves to node $j-1$ is $\frac{|S(\boldsymbol{s})|}{3}$. The probability of staying at node j is $-\langle\boldsymbol{a}^i, \boldsymbol{\theta}^i\rangle + \frac{1-\delta_{j,j}}{n} \times \frac{1}{\alpha(\boldsymbol{s})}$; moving to node $j+1$ is $\langle\boldsymbol{a}^i, \boldsymbol{\theta}^i\rangle + \frac{\delta_{j,j+1}}{n} \times \frac{1}{\alpha(\boldsymbol{s})}$; and probability of going to node $j-1$ is $\frac{\delta_{j,j}-\delta_{j,j+1}}{n} \times \frac{1}{\alpha(\boldsymbol{s})}$. This is true for all the agents at node j in the state $\boldsymbol{s}$. Therefore, the contribution to the overall probability from this agent is

$$\begin{aligned} &\sum_{i=1}^{x_j}\left\{\left(-\langle\boldsymbol{a}^i, \boldsymbol{\theta}^i\rangle + \frac{1-\delta_{j,j}}{n} \times \frac{1}{\alpha(\boldsymbol{s})}\right) \times \frac{|S(\boldsymbol{s})|}{3}\right\} + \sum_{i=1}^{x_j}\left\{\left(\langle\boldsymbol{a}^i, \boldsymbol{\theta}^i\rangle + \frac{\delta_{j,j+1}}{n} \times \frac{1}{\alpha(\boldsymbol{s})}\right) \times \frac{|S(\boldsymbol{s})|}{3}\right\} \\ &+ \sum_{i=1}^{x_j}\left(\frac{\delta_{j,j}-\delta_{j,j+1}}{n} \times \frac{1}{\alpha(\boldsymbol{s})}\right) \times \frac{|S(\boldsymbol{s})|}{3} = \frac{|S(\boldsymbol{s})|}{3} \times \frac{1}{n} \times \frac{x_j}{\alpha(\boldsymbol{s})}. \end{aligned} \quad (77)$$

The above expression is valid for any node $j \in \{1, 2, \ldots, q\}$. Finally, consider the agent who is at node g, the number of next states in which the agent stays at node g is $|S(\boldsymbol{s})|$. The probability of this event is $\frac{1}{n}\frac{x_{q+1}}{\alpha(\boldsymbol{s})}$. Therefore, the contribution in the probability from the agent who is at node g is

$$\sum_{i=1}^{x_{q+1}} |S(\boldsymbol{s})| \times \frac{1}{n} \times \frac{1}{\alpha(\boldsymbol{s})} = |S(\boldsymbol{s})| \times \frac{1}{n} \times \frac{x_{q+1}}{\alpha(\boldsymbol{s})}. \tag{78}$$

Adding Equations equation 76, equation 77 and equation 78, we have

$$\begin{aligned}\sum_{\boldsymbol{s}' \neq \boldsymbol{g}} \langle \phi(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}), \boldsymbol{\theta}(\boldsymbol{s})\rangle &= \left(\frac{|S(\boldsymbol{s})|}{2} \times \frac{1}{n} \times \frac{x_0}{\alpha(\boldsymbol{s})}\right) \\ &\quad + \sum_{j=1}^{q} \left(\frac{|S(\boldsymbol{s})|}{3} \times \frac{1}{n} \times \frac{x_j}{\alpha(\boldsymbol{s})}\right) + \left(|S(\boldsymbol{s})| \times \frac{1}{n} \times \frac{x_{q+1}}{\alpha(\boldsymbol{s})}\right) \\ &= \frac{|S(\boldsymbol{s})|}{n\alpha(\boldsymbol{s})} \left\{\frac{x_0}{2} + x_{q+1} + \sum_{j=1}^{q} \frac{x_j}{3}\right\}.\end{aligned}$$

Since, we set $\alpha(\boldsymbol{s}) = \frac{|S(\boldsymbol{s})|}{n} \left\{\frac{x_0}{2} + x_{q+1} + \sum_{j=1}^{q} \frac{x_j}{3}\right\}$, we have that the above summation as 1.

Case 02: $(\boldsymbol{s} = \boldsymbol{g})$. For this case, the probability is

$$\begin{aligned}\sum_{\boldsymbol{s}'} \langle \phi(\boldsymbol{s}'|\boldsymbol{s} = \boldsymbol{g}, \boldsymbol{a}), \boldsymbol{\theta}(\boldsymbol{s})\rangle &= \sum_{\boldsymbol{s}' \neq \boldsymbol{g}} \langle \phi(\boldsymbol{s}'|\boldsymbol{s} = \boldsymbol{g}, \boldsymbol{a}), \boldsymbol{\theta}(\boldsymbol{s})\rangle + \langle \phi(\boldsymbol{s}' = \boldsymbol{g}|\boldsymbol{s} = \boldsymbol{g}, \boldsymbol{a}), \boldsymbol{\theta}(\boldsymbol{s})\rangle \\ &= \langle \mathbf{0}, \boldsymbol{\theta}(\boldsymbol{s})\rangle + \langle (\mathbf{0}_{nd-1}, \alpha(\boldsymbol{s})), \boldsymbol{\theta}(\boldsymbol{s})\rangle = 1\end{aligned}$$

Therefore, in both cases, we have

$$\sum_{\boldsymbol{s}'} \langle \phi(\boldsymbol{s}'|\boldsymbol{s} = \boldsymbol{g}, \boldsymbol{a}), \boldsymbol{\theta}(\boldsymbol{s})\rangle = 1, \quad \forall\ \boldsymbol{s}, \boldsymbol{a}.$$

The other two statements of the lemma follow by feature design and model parameter space. □

G Some useful results

G.1 Equivalence of the optimization problems

To start with we show the equivalence of the optimization problems we obtain from the least square minimizer of the global reward function. Recall the optimization problem is

$$min_{\boldsymbol{w}} \ \ \mathbb{E}_{\boldsymbol{s},\boldsymbol{a}}[\bar{r}(\boldsymbol{s},\boldsymbol{a}) - \bar{r}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{w})]^2. \tag{OP 1}$$

We prove the following key Proposition which enables the decentralized working of our algorithm.

Proposition 3 (Zhang et al. (2018), Trivedi and Hemachandra (2022))**.** *The optimization problem OP 1 is equivalently characterized as (both have the same stationary points)*

$$min_{\boldsymbol{w}} \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{s},\boldsymbol{a}}[r^i(\boldsymbol{s},\boldsymbol{a}) - \bar{r}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{w})]^2. \tag{OP 2}$$

Proof. Taking the first order derivative of the objective function in optimization problem (OP 1) w.r.t. $\boldsymbol{w}$, we have:

$$\begin{aligned}-2 \times \mathbb{E}_{\boldsymbol{s},\boldsymbol{a}}[\bar{r}(\boldsymbol{s},\boldsymbol{a}) - \bar{r}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{w})] \times \nabla_{\boldsymbol{w}} \bar{r}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{w}) &= -2 \times \mathbb{E}_{\boldsymbol{s},\boldsymbol{a}} \left[\frac{1}{n} \sum_{i \in N} r^i(\boldsymbol{s},\boldsymbol{a}) - \bar{r}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{w})\right] \times \nabla_{\boldsymbol{w}} \bar{r}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{w}), \\ &= -\frac{2}{n} \times \mathbb{E}_{\boldsymbol{s},\boldsymbol{a}} \left[\sum_{i \in N} r^i(\boldsymbol{s},\boldsymbol{a}) - n \cdot \bar{r}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{w})\right] \times \nabla_{\boldsymbol{w}} \bar{r}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{w}), \\ &= -\frac{2}{n} \times \mathbb{E}_{\boldsymbol{s},\boldsymbol{a}} \left[\sum_{i \in N} \left(r^i(\boldsymbol{s},\boldsymbol{a}) - \bar{r}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{w})\right)\right] \times \nabla_{\boldsymbol{w}} \bar{r}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{w}).\end{aligned}$$

Ignoring the factor $\frac{1}{n}$ in the above equation, we exactly have the first order derivative of the objective function in OP 2. Thus, both optimization problems have the same stationary points. Hence, OP 1 is an *equivalent characterization* of OP 2. □

G.2 Lemma for Gaussian and Laplace mechanisms

For the Gaussian mechanism, we can only hope for the (ϵ, δ) LDP. In this subsection, we show that our MA-LDP algorithm with Gaussian mechanism indeed preserves the differential privacy. To this edn, we have following Theorem (Theorem A.2 in Liao et al. (2021))

(Theorem A.2 Liao et al. (2021)). If the privacy loss c satisfy $\mathbb{P}_{o\sim\mathcal{M}(d)}[c(o;\mathcal{M},\mathbf{aux}, d, d') > \epsilon] \leq \delta$ for all auxiliary input **aux** and neighboring data sets d, d', then the mechanism $\mathcal{M}$ satisfies (ϵ, δ)-LDP property.

The basic idea involved in the above Theorem is that the if we set the privacy parameter δ then for any auxiliary input ***aux*** the probability that any outcome o obtained using the privacy preserving mechanism incurs at least ϵ privacy loss then the mechanism $\mathcal{M}$ satisfies the (ϵ, δ)-LDP. The proof of this Theorem is available in Liao et al. (2021); Abadi et al. (2016) and is based on the construction of an event containing all possible outcomes for which the absolute privacy loss is at least ϵ.

(Gaussian Mechanism Dwork et al. (2006); Liao et al. (2021)): Let $f : \mathcal{N}^{\mathcal{X}} \to \mathbb{R}^d$ be an arbitrary d-dimensional function (a query), and define the l_2 sensitivity as $\Delta_2 f = \max_{adj(x,y)} ||f(x) - f(y)||_2$, where $adj(x, y)$ indicates that x, y are different at one entry only. For any $0 \leq \epsilon \leq 1$ and $c^2 > 2\log(1.25/\delta)$, the Gaussian mechanism with parameter $\sigma \geq c\Delta_2 f/\epsilon$ is (ϵ, δ)-LDP.

The following Analogous Lemma for the Laplace Mechanism is also available in Theorem 3.6 of Dwork and Roth (2014).

(Laplace Mechanism; Theorem 3.6 Dwork and Roth (2014)). Let $f : \mathcal{N}^{\mathcal{X}} \to \mathbb{R}^d$ be an arbitrary d-dimensional function (a query), and define the l_1 sensitivity as $\Delta f = \max_{||x-y||_1=1} ||f(x) - f(y)||_1$. For any $0 \leq \epsilon \leq 1$ the Laplace mechanism with parameter $b = \frac{\Delta f}{\epsilon}$ preserves $(\epsilon, 0)$ differential privacy.

G.3 Other important results

(Lemma 11, Abbasi-Yadkori et al. (2011)): Let $\{\phi_t\}_{t\geq 0}$ be the bounded sequence in $\mathbb{R}^d$ satisfying $sup_{t\geq 0}||\phi_t|| \leq 1$. Let $\mathbf{\Lambda}_0 \in \mathbb{R}^{d\times d}$ be a positive definite matrix. For any $t \geq 0$, we define $\mathbf{\Lambda}_t = \mathbf{\Lambda}_0 + \sum_{j=0}^{t} \phi_j^\top \phi_j$. Then, if the smallest eigenvalue of $\mathbf{\Lambda}_0$ satisfies $\lambda_{\min}(\mathbf{\Lambda}) \geq 1$, we have

$$\log\left[\frac{det(\mathbf{\Lambda}_t)}{det(\mathbf{\Lambda}_0)}\right] \leq \sum_{j=1}^{t} \phi_j^\top \mathbf{\Lambda}_{j-1}^{-1} \phi_j \leq 2\log\left[\frac{det(\mathbf{\Lambda}_t)}{det(\mathbf{\Lambda}_0)}\right] \tag{79}$$

(Theorem 2 of Zhou et al. (2021)). Let $\{\mathcal{G}_t\}_{t=1}^{\infty}$ be the filtration. Let $\{x_t, \eta_t\}_{t\geq 1}$ be a stochastic process so that $x_t \in \mathbb{R}^d$ is $\mathcal{G}_t$-measurable and η_t be $\mathcal{G}_{t+1}$ measurable. Fix, $R, L, \sigma, \lambda, \mu^\star \in \mathbb{R}^d$. For $t \geq 1$ let $y_t =< \mu^\star, x_t > +\eta_t$ and suppose that η_t, x_t also satisfy

$$|\eta_t| \leq R,\ \mathbb{E}[\eta_t^2|\mathcal{G}_t] \leq \sigma^2,\ ||x_t||_2 \leq L \tag{80}$$

Then, for any $0 < \delta < 1$, with probability at least $1-\delta$ we have

$$\forall\ t \geq 0,\ ||\sum_{i=1}^{t} x_i\eta_i||_{Z_t^{-1}} \leq \beta_t,\ ||\mu_t - \mu^\star||_{Z_t} \leq \beta_t + \sqrt{\lambda}||\mu^\star||_2, \tag{81}$$

where for $t \geq 1$, $\mu_t = Z_t^{-1}b_t, Z_t = \lambda I + \sum_{i=1}^{t} x_i x_i^\top, b_t = \sum_{i=1}^{t} y_i x_i$ and

$$\beta_t = 8\sigma\sqrt{d\log(1 + tL^2/(d\lambda))\log(4t^2/\delta)} + 4R\log(4t^2/\delta) \tag{82}$$

(Kushner-Clark Lemma (Kushner and Yin, 2003; Metivier and Priouret, 1984)). Let $\mathcal{X} \subseteq \mathbb{R}^p$ be a compact set and let $h : \mathcal{X} \to \mathbb{R}^p$ be a continuous function. Consider the following recursion in p-dimensions

$$x_{t+1} = \Gamma\{x_t + \gamma_t[h(x_t) + \zeta_t + \beta_t]\}. \tag{83}$$

Let $\hat{\Gamma}(\cdot)$ be transformed projection operator defined for any $x \in \mathcal{X} \subseteq \mathbb{R}^p$ as

$$\hat{\Gamma}(h(x)) = lim_{0<\eta\to 0}\left\{\frac{\Gamma(x + \eta h(x)) - x}{\eta}\right\},$$

then the ODE associated with Equation (83) is $\dot{x} = \hat{\Gamma}(h(x))$.

Assumption 4. *Kushner-Clark lemma requires the following assumptions*

1. *Stepsize $\{\gamma_t\}_{t\geq 0}$ satisfy $\sum_t \gamma_t = \infty$, and $\gamma_t \to 0$ as $t \to \infty$.*
2. *The sequence $\{\beta_t\}_{t\geq 0}$ is a bounded random sequence with $\beta_t \to 0$ almost surely as $t \to \infty$.*
3. *For any $\epsilon > 0$, the sequence $\{\zeta_t\}_{t\geq 0}$ satisfy*

$$\lim_t\ \mathbb{P}\left(sup_{p\geq t}\left\|\sum_{\tau=t}^{p}\gamma_\tau\zeta_\tau\right\| \geq \epsilon\right) = 0.$$

Kushner-Clark lemma is as follows: suppose that ODE $\dot{x} = \hat{\Gamma}(h(x))$ has a compact set $\mathcal{K}^\star$ as its asymptotically stable equilibria, then under Assumption 4, x_t in Equation (83) converges almost surely to $\mathcal{K}^\star$ as $t \to \infty$.

H Generating a bounded Laplace distribution, $\mathcal{BL}$ (Ross, 2022)

The cdf of the bounded Laplace random variable, $\mathcal{BL}$ can be obtained as follows:

$$\begin{aligned}
F_{\mathcal{BL}}(x) &= \int_{-B}^{x} \frac{1}{2b\left(1 - \exp(\frac{-B}{b})\right)} \exp\left(\frac{-|x|}{b}\right) && (84)\\
&= \int_{-B}^{0} \frac{1}{2b\left(1 - \exp(\frac{-B}{b})\right)} \exp\left(\frac{x}{b}\right) + \int_{0}^{x} \frac{1}{2b\left(1 - \exp(\frac{-B}{b})\right)} \exp\left(\frac{-x}{b}\right) && (85)\\
&= \frac{1}{2b\left(1 - \exp(\frac{-B}{b})\right)}[b - b\exp(\frac{-B}{b})] - \frac{1}{2b\left(1 - \exp(\frac{-B}{b})\right)}[b\exp(\frac{-x}{b}) - b] && (86)
\end{aligned}$$

$$= \frac{2 - \exp(\frac{-B}{b}) - \exp(\frac{-x}{b})}{2\left(1 - \exp(\frac{-B}{b})\right)} \tag{87}$$

To simulate this, let $u = F_{\mathcal{BL}}(x)$, where $u \sim \mathcal{U}(0,1)$, then

$$u = \frac{2 - \exp(\frac{-B}{b}) - \exp(\frac{-x}{b})}{2\left(1 - \exp(\frac{-B}{b})\right)} \tag{88}$$

$$\exp(\frac{-x}{b}) = 2 - \exp(\frac{-B}{b}) - 2u\left(1 - \exp(\frac{-B}{b})\right) \tag{89}$$

$$\frac{-x}{b} = \log\left(2 - \exp(\frac{-B}{b}) - 2u\left(1 - \exp(\frac{-B}{b})\right)\right) \tag{90}$$

$$x = -b\log\left(2 - \exp(\frac{-B}{b}) - 2u\left(1 - \exp(\frac{-B}{b})\right)\right) \tag{91}$$

Note, this x will follow the Laplace distribution with bounded support