LOCAL DIFFERENTIAL PRIVACY PRESERVING MECHANISMS FOR MULTI-AGENT REINFORCEMENT LEARNING

Prashant Trivedi
Industrial Engineering and Operations Research
Indian Institute of Technology Bombay India
trivedi.prashant15@iitb.ac.in

Nandyala Hemachandra
Industrial Engineering and Operations Research
Indian Institute of Technology Bombay India
nh@iitb.ac.in

ABSTRACT

This work considers the local differential privacy (LDP) aspects of multi-agent reinforcement learning (MARL). We design a fully decentralized and generic multi-agent locally differential private (MA-LDP) algorithm that can handle any noise adding mechanisms. It takes the episodic form, where the data from each agent is anonymized using noise adding mechanism in each episode. MA-LDP uses the linear function approximations of the transition probabilities and the reward functions. We prove that the MA-LDP algorithm preserves the user's data privacy and attains the sub-linear regret for four noise mechanisms with different noise supports. Further, we compare the noise adding mechanisms with unbounded support to those with bounded support. A key observation is that for a suitably chosen bounded noise support, the regret of the MA-LDP algorithms is on-par or lower than the noise mechanism with unbounded support. We validate our theoretical findings on an network MDP with a large state and action spaces.

Keywords Differential privacy · Decentralized models · Noise mechanisms · Privacy loss · Sub-linear regret · Finite horizon MDP · Network MDP · Linear function approximations · Reinforcement learning · Bellman equations

1 Introduction

Multi-agent reinforcement learning (MARL) is one of the powerful tools used widely in many real-life applications. However, most of the algorithms rely on the user given data. Thus, the privacy of the users data is an utmost concern of any user. To this end, Dwork et al. (2006) introduced the notion of the differential privacy. The idea is to work with the anonymized user data, yet achieve the same performance and the user experience. Therefore, the notion of Locally Differentially Private (LDP) is introduced in Dwork and Roth (2014); Kasiviswanathan et al. (2011), which enhanced the performance while maintaining the privacy. However, to our knowledge the notion of LDP is not considered in MARL settings, where the privacy of user data is more serious. LDP in MARL can be used to protect the sensitive information that is shared or collected among the agents. Some of the applications include the health data analysis where a patients data is sensitive, interactions of agents within a network, say a social media platform etc.

In the MARL setup the agents usually have a common objective; however, the decisions are taken sequentially by each agent in a decentralised fashion. In this work, we introduce the notion of differential privacy for the MARL problem. In particular, we address the following questions. 1) Can we design a MARL algorithm that can preserve the user-data privacy while attaining the same outcome? We answer this question affirmatively and introduce a novel and generic Multi-Agent Local Differential Privacy (MA-LDP) MARL algorithm. This algorithm can work with any noise mechanism. In this work, we consider four noise adding mechanisms, Gaussian, Laplace, uniform and bounded Laplace (BL). Two of which have the unbounded support, whereas other two have bounded support. For each noise adding mechanism we investigate 3 aspects of our MA-LDP algorithm 1) its ability to preserve user sensitive information, 2) the effect on the regret, and 3) how its regret fares with that of the other mechanisms. The privacy from these noise mechanisms ranges from one extreme of no privacy loss ($\epsilon = 0$) and less confidence ($\delta > 0$) to some privacy loss ($\epsilon > 0$) and high confidence ($\delta = 0$).

Next, we compare the regret and the privacy guarantees of our MA-LDP algorithm across these noise mechanisms. This comparison is motivated by the following question: can the noise mechanisms with bounded support attain the same or better privacy and the regret guarantee as that of the noise mechanisms with unbounded support. Again we answer this question affirmatively. In particular, we show that the regret for the BL mechanism is of the same order as that of the Laplace mechanism, if the end points of the BL mechanism are chosen suitably. Our contributions are
(a) To address the LDP aspects in the MARL we introduce four novel noise adding mechanisms. Two of these mechanisms have unbounded noise support, whereas the remaining two have bounded support. We propose a generic algorithm MA-LDP that can work with any noise mechanism. Moreover, with these four noise mechanisms our MA-LDP achieves a sub-linear regret (Sections 5, 6).
(b) These four different noise adding mechanisms yield different privacy guarantees; Gaussian mechanism achieves (ϵ, δ) privacy, Laplace mechanism achieves $(\epsilon, 0)$ privacy, uniform mechanism achieves $(0, \delta)$, and finally for bounded Laplace noise mechanism we get $(\epsilon, 0)$ privacy. Thus, the privacy guarantees range from one extreme $(\epsilon, 0)$ to $(0, \delta)$ (Section 5, 6)
(c) While the regret of MA-LDP algorithm is sub-linear in total numbers of steps, it is super-linear in number of agents and the feature dimension for all the noise mechanisms. But this order with respect to number of agents and feature dimension is the same across all the four noise mechanisms.
(d) We compare the regret of MA-LDP standard (unbounded support) noise mechanisms with those of the bounded support mechanisms. If the end points of the support of bounded Laplace distribution is of the same order as that of the distribution parameter .. the regret of BL mechanism is on par with the Laplace mechanism. So, instead of injecting the potentially unbounded noise values, a suitably chosen bounded support can attain the lower regret (see Table 2, Section 7).

2 LDP for Decentralized MARL

In this Section, we first introduce the episodic decentralized MARL, and then provide the notion of LDP for the MARL.

2.1 Episodic Decentralized MARL

Let $N = \{1, 2, \ldots, n\}$ be the set of agents. An instance of multi-agent time inhomogeneous episodic Markov Decision Process is given by $(N, \mathcal{S}, \{\mathcal{A}^i\}_{i\in N}, H, \{r_h^i\}_{i\in N, h\in H}, \{\mathbb{P}_h\}_{h\in H}, \{\mathcal{G}_t\}_{t\geq 0})$. Here $\mathcal{S}$ is the finite set of global state-space. $\mathcal{A}^i$ is the finite set of local actions available to agent $i \in N$, i.e., the local action-space of agent $i \in N$. The global action at any global state $\boldsymbol{s} \in \mathcal{S}$ is $\mathcal{A}(\boldsymbol{s}) = \prod_{i\in N} \mathcal{A}^i(\boldsymbol{s})$. We use product of the local action-space because the agents are independently taking the actions. A typical element in $\mathcal{A}(\boldsymbol{s})$ is a vector of size n, one for each agent as $(\boldsymbol{a}^1(\boldsymbol{s}), \boldsymbol{a}^2(\boldsymbol{s}), \ldots, \boldsymbol{a}^n(\boldsymbol{s}))$, where $\boldsymbol{a}^i(\boldsymbol{s}) \in \mathcal{A}^i(\boldsymbol{s})$ represents the action taken by agent i when the global state is $\boldsymbol{s}$. Let K be the total number of episodes; each episode consists of a fixed planning horizon H. We use h to denote an intermediate stage in planning horizon H. Moreover, let $T = KH$ be the total number of interactions with the MDP. For a global state $\boldsymbol{s}$, global action $\boldsymbol{a}$, each agent $i \in N$ realizes a deterministic local reward $r_h^i(\boldsymbol{s}, \boldsymbol{a}) \in [0, 1]$ at the stage h of the planning horizon. It is important to note that the reward of agent i depends on the action taken by other agents and the global state. Moreover, this reward r_h^i is a private information of the agent i and hence not known to other agents. Once an action $\boldsymbol{a}$ is taken in state $\boldsymbol{s}$ at stage h, the state change to $\boldsymbol{s}'$ with probability $\mathbb{P}_h(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a})$. In the MARL, this transition probability is generally unknown. Finally, $\mathcal{G}_t$ is the time varying network that allows the sharing of some information across the agents at time $t = kh$. We describe more details about it later.

One of the goals in MARL is to learn a policy $\pi = \{\pi_h\}_{h=1}^H$, that is the collection of H functions, where each $\pi_h(\boldsymbol{s})$ is the global action taken in the global state $\boldsymbol{s}$ at stage h. In particular, the aim is to learn an optimal policy, in a decentralized way, by interacting with the environment and observing the past information. To this end, we define the global state-action value function for the policy π at stage h $Q_h^\pi(\boldsymbol{s}, \boldsymbol{a}) = \bar{r}_h(\boldsymbol{s}, \boldsymbol{a}) + \mathbb{E}_\pi \left[\sum_{h'=h+1}^H \bar{r}_{h'}(\boldsymbol{s}_{h'}, \pi_{h'}(\boldsymbol{s}'_{h}))\right]$. Here $\bar{r}_h = \frac{1}{n}\sum_{i\in N} r_h^i$ and $\boldsymbol{s}_h = \boldsymbol{s}$, $\boldsymbol{a}_h = \boldsymbol{a}$ and $\boldsymbol{s}_{h'+1} \sim \mathbb{P}_{h'}(\cdot|\boldsymbol{s}_{h'}, \boldsymbol{a}_{h'})$. Note that this state-action value function represents the value of policy π when action $\boldsymbol{a}$ is taken at stage h in the state $\boldsymbol{s}_h = \boldsymbol{s}$, and the policy $\pi_h(\boldsymbol{s})$ is followed thereafter. Moreover, we also define the global state value function at stage h as $V_h^\pi(\boldsymbol{s}) = Q_h^\pi(\boldsymbol{s}, \pi_h(\boldsymbol{s}))$. As opposed to the state action value function the state value function represents the value of a state $\boldsymbol{s}$ when policy π_h is followed starting for state $\boldsymbol{s}$ at stage h.

For simplicity of notation, given any function $V : \mathcal{S} \to [0, H]$, for all $(\boldsymbol{s}, \boldsymbol{a}) \in \mathcal{S} \times \mathcal{A}$, we define the cost-to-go function as $\mathbb{P}_h V(\boldsymbol{s}, \boldsymbol{a}) = \sum_{\boldsymbol{s}'\in\mathcal{S}} \mathbb{P}_h(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a}) V(\boldsymbol{s}')$. Using above we write the Bellman equation for the policy π for all $(\boldsymbol{s}, \boldsymbol{a}) \in \mathcal{S} \times \mathcal{A}$ as $Q_h^\pi(\boldsymbol{s}, \boldsymbol{a}) = \bar{r}_h(\boldsymbol{s}, \boldsymbol{a}) + \mathbb{P}_h V_{h+1}^\pi(\boldsymbol{s}, \boldsymbol{a})$; $V_h^\pi(\boldsymbol{s}) = Q_h^\pi(\boldsymbol{s}, \pi_h(\boldsymbol{a}))$; $V_{H+1}^\pi(\boldsymbol{s}) = 0$. Let the optimal state-action value be $Q_h^\star(\boldsymbol{s}, \boldsymbol{a}) = \max_\pi Q_h^\pi(\boldsymbol{s}, \boldsymbol{a})$, and the optimal state value function be $V_h^\star(\boldsymbol{s}) = \max_\pi V_h^\pi(\boldsymbol{s})$.

The Bellman optimality equation for all $(\boldsymbol{s},\boldsymbol{a}) \in \mathcal{S}\times\mathcal{A}$ satisfies $Q_h^\star(\boldsymbol{s},\boldsymbol{a}) = \bar{r}_h(\boldsymbol{s},\boldsymbol{a}) + \mathbb{P}_h V_{h+1}^\star(\boldsymbol{s},\boldsymbol{a})$; $V_{h+1}^\star(\boldsymbol{s}) = \max_{\boldsymbol{a}\in\mathcal{A}} Q_h^\star(\boldsymbol{s},\boldsymbol{a})$; $V_{H+1}^\star(\boldsymbol{s}) = 0$. Since $\bar{r}_h(\cdot,\cdot)$ is a bounded function, for any policy π, both $V^\pi(\cdot)$ and $Q^\pi(\cdot,\cdot)$ are also bounded. This work assumes that the transition probability function $\mathbb{P}_h$ is written as the linear mixture of given basis functions (Min et al. (2022); Vial et al. (2022); Liao et al. (2021)). In particular, we make the following standard assumption about the transition probability function.

Assumption 1 (Transition probability approximation)**.** *Suppose the feature mapping $\phi : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to \mathbb{R}^{nd}$ is known and pre-given. There exists a $\boldsymbol{\theta}_h^\star \in \mathbb{R}^{nd}$ with $||\boldsymbol{\theta}_h^\star||_2 \le \sqrt{nd}$ such that $\mathbb{P}_h(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) = \langle \phi(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}), \boldsymbol{\theta}_h^\star\rangle$ for any triplet $(\boldsymbol{s}',\boldsymbol{a},\boldsymbol{s}) \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$ and stage h. Also, for a bounded function $V : \mathcal{S} \mapsto [0,H]$, it holds that $||\phi_V(\boldsymbol{s},\boldsymbol{a})||_2 \le H$, where $\phi_V(\boldsymbol{s},\boldsymbol{a}) = \sum_{\boldsymbol{s}'\in\mathcal{S}} \phi(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})V(\boldsymbol{s}')$.*

Using above we have, $\mathbb{P}_h V(\boldsymbol{s},\boldsymbol{a}) = \sum_{\boldsymbol{s}'\in\mathcal{S}} \langle \phi(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}), \boldsymbol{\theta}_h^\star\rangle V(\boldsymbol{s}') = \langle \phi_V(\boldsymbol{s},\boldsymbol{a}), \boldsymbol{\theta}_h^\star\rangle$. Moreover, recall the Bellman optimality equation use the averaged reward $\bar{r}_h(\cdot,\cdot)$, however in our decentralized model this averaged reward is not known to any agent. To this end, each agent maintains an estimate the globally averaged reward function $\bar{r}_h(\cdot,\cdot)$.

Let $\bar{r}_h(\cdot,\cdot;\boldsymbol{w}) : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ be the class of parameterized functions where $\boldsymbol{w} \in \mathbb{R}^p$ for some $p << |\mathcal{S}||\mathcal{A}|$. To obtain the estimate $\bar{r}_h(\cdot,\cdot;\boldsymbol{w})$ we seek to minimize the following $min_{\boldsymbol{w}}\ \mathbb{E}_{\boldsymbol{s},\boldsymbol{a}}[\bar{r}_h(\boldsymbol{s},\boldsymbol{a};\boldsymbol{w}) - \bar{r}_h(\boldsymbol{s},\boldsymbol{a})]^2$. This optimization problem can be equivalently characterized as (both have the same stationary points) $min_{\boldsymbol{w}} \sum_{i=1}^n \mathbb{E}_{\boldsymbol{s},\boldsymbol{a}}[\bar{r}_h(\boldsymbol{s},\boldsymbol{a};\boldsymbol{w}) - r_h^i(\boldsymbol{s},\boldsymbol{a})]^2$. The details of this equivalence of the optimization problems is given in the Appendix G.1. This motivates the following updates for parameters of the global reward function parameters $\boldsymbol{w}^i$ by agent $i \in N$, $\widetilde{\boldsymbol{w}}_{k,h}^i \leftarrow \boldsymbol{w}_{k,h}^i + \gamma_{k,h}\cdot[r_h^i(\cdot,\cdot) - \bar{r}_h(\cdot,\cdot;\boldsymbol{w}_{k,h}^i)]\cdot\nabla_{\boldsymbol{w}}\bar{r}_h(\cdot,\cdot;\boldsymbol{w}_{k,h}^i)$; $\boldsymbol{w}_{k+1,h}^i = \sum_{j\in N} l_{k,h}(i,j)\widetilde{\boldsymbol{w}}_{k,h}^j$. 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 $\sum_{k,h}\gamma_{k,h}^2 < \infty$, and $\bar{r}_h(\cdot,\cdot;\boldsymbol{w}_{k,h}^i)$ is the estimate of global reward function by agent i in the stage h of the episode k. We make following standard assumptions on $\{L_{k,h}\}_{k,h\ge 0}$ and the features for the reward function approximation Zhang et al. (2018); Trivedi and Hemachandra (2022, 2023).

Assumption 2 (Consensus matrix)**.** *The consensus matrices $\{L_t\}_{t\ge 0} \subseteq \mathbb{R}^{n\times n}$ satisfies (i) 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 $l_t(i,j) \ge \kappa$; (ii) Consensus matrix L_t respects $\mathcal{G}_t$, i.e., $l_t(i,j) = 0$, if $(i,j) \notin \mathcal{E}_t$; (iii) The spectral norm of $\mathbb{E}[L_t^\top(I - \mathbb{1}\mathbb{1}^\top/n)L_t]$ is smaller than one.*

Assumption 3 (Full rank)**.** *For each agent $i \in N$, the reward function $\bar{r}(\boldsymbol{s},\boldsymbol{a})$ is parameterized as $\bar{r}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{w}) = \langle \psi(\boldsymbol{s},\boldsymbol{a}), \boldsymbol{w}\rangle$. Here $\psi(\boldsymbol{s},\boldsymbol{a}) = [\psi_1(\boldsymbol{s},\boldsymbol{a}),\ldots,\psi_k(\boldsymbol{s},\boldsymbol{a})] \in \mathbb{R}^p$ are the features associated with pair $(\boldsymbol{s},\boldsymbol{a})$. We assume that features are uniformly bounded. Let the feature matrix $\Psi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times p}$ have $[\psi_m(\boldsymbol{s},\boldsymbol{a}), \boldsymbol{s}\in\mathcal{S}, \boldsymbol{a}\in\mathcal{A}]^\top$ as its m-th column, then Ψ has full column rank.*

Since the global reward is unknown, each agent uses the parameterized reward and maintains its estimate of $V_h(\cdot)$ and $Q_h(\cdot,\cdot)$. Let $V_h^i(\cdot)$ and $Q_h^i(\cdot,\cdot)$ be the estimate of these functions by agent i at stage h. So, the modified Bellman optimality equation for all $(\boldsymbol{s},\boldsymbol{a})$ and for all agents $i \in N$ is $Q_h^{\star,i}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{w}_{k,h}^i) = \bar{r}_h(\boldsymbol{s},\boldsymbol{a};\boldsymbol{w}_{k,h}^i) + \mathbb{P}_h V_{h+1}^{\star,i}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{w}_{k,h}^i)$; $V_{h+1}^{\star,i}(\boldsymbol{s};\boldsymbol{w}_{k,h}^i) = \max_{\boldsymbol{a}\in\mathcal{A}} Q_h^{\star,i}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{w}^i)$; $V_{H+1}^{\star,i}(\boldsymbol{s};\boldsymbol{w}_{k,H+1}^i) = 0$. We later show that as the number of episodes increases the reward function parameters converge, i.e., as $k\to\infty$, $\boldsymbol{w}_{k,h}^i \to \boldsymbol{w}^\star$ *a.s.* for all $i\in N$ and for all $h\in[H]$. Hence, $\bar{r}_h(\boldsymbol{s},\boldsymbol{a};\boldsymbol{w}_{k,h}^i) \to \bar{r}_h(\boldsymbol{s},\boldsymbol{a};\boldsymbol{w}^\star)$, $Q_h^{\star,i}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{w}_{k,h}^i) \to Q_h^{\star,i}(\boldsymbol{s},\boldsymbol{a})$ and $V_h^{\star,i}(\boldsymbol{s};\boldsymbol{w}_{k,h}^i) \to V_h^{\star,i}(\boldsymbol{s})$ as $\bar{r}_h(\boldsymbol{s},\boldsymbol{a};\boldsymbol{w}_{k,h}^i)$, $Q_h^{\star,i}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{w}_{k,h}^i)$ and $V_h^{\star,i}(\boldsymbol{s};\boldsymbol{w}_{k,h}^i)$ are continuous functions of $\boldsymbol{w}^i$, where $Q_h^{\star,i}(\boldsymbol{s},\boldsymbol{a})$ and $V_h^{\star,i}(\boldsymbol{s})$ are defined as $Q^{\star,i}(\boldsymbol{s},\boldsymbol{a}) = \bar{r}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{w}^\star) + \mathbb{P}V^{\star,i}(\boldsymbol{s},\boldsymbol{a})$; $V^{\star,i}(\boldsymbol{s}) = \max_{\boldsymbol{a}\in\mathcal{A}} Q^{\star,i}(\boldsymbol{s},\boldsymbol{a})$.

At each episode k, with the initial state $\boldsymbol{s}_{k,1}$ agents choose a policy π_k. Moreover, in each stage $h\in[H]$ of the episode k, for the observed state $\boldsymbol{s}_{k,h}$, agents take a action according to the policy π_k, i.e., $\boldsymbol{a}_{k,h} = \pi_{k,h}(\boldsymbol{s}_{k,h})$. Further, they also observe a next state $\boldsymbol{s}_{k,h+1} \sim \mathbb{P}_h(\cdot|\boldsymbol{s}_{k,h},\boldsymbol{a}_{k,h})$. The expected regret in the k-th episode is $R_k = \frac{1}{n}\sum_{i\in N}\{V_1^{\star,i}(\boldsymbol{s}_1^k) - V_1^i(\boldsymbol{s}_1^k)\}$. Hence our objective is to design an algorithm with the sub-linear regret. The total expected regret in K episodes is defined as $R_K = \sum_{k=1}^K \left(\frac{1}{n}\sum_{i\in N}\{V_1^{\star,i}(\boldsymbol{s}_1^k) - V_1^i(\boldsymbol{s}_1^k)\}\right)$. Note that we are using $V_1^{\star,i}(\boldsymbol{s}_1^k)$ instead of $V_1^\star(\boldsymbol{s}_1^k)$. This is because $\boldsymbol{w}_{k,h}^i \to \boldsymbol{w}^\star$ and hence $V_1^{\star,i}(\boldsymbol{s}_1^k) = V_1^\star(\boldsymbol{s}_1^k)$ for all $i\in N$. Such an independence of agents is desirable and is a sign of good decentralised algorithms.

2.2 LDP for MARL

In this Section, we introduce the notion of the multi-agent local differential privacy (MA-LDP). This definition is inspired by the single agent DP introduce in Dwork et al. (2006) and LDP introduced in Kasiviswanathan et al.

(2011); Duchi et al. (2013). Throughout, we use the following notations. 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. Here for each agent $i \in N$, D_h^i and $D_h^{i'}$ differs at exactly one component. Let $\mathbf{D}_{1:h-1}$ be the information collected from stage 1 to stage h, i.e., $\mathbf{D}_{1:h-1} = (\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_{h-1})$.

In online RL, each episode $k \in [K]$ is viewed as the trajectory associated with a specific user. Thus, there are K users and n agents. Note that in MA-LDP the agent is different from the user. So, for MA-LDP we guarantee that for any user the information of any agent $i \in N$ send to the server is privatized. Therefore, the server is agnostic to the sensitive data.

Definition 1 (MA-LDP)**.** *For any $\epsilon \geq 0$, and $\delta \geq 0$, a randomized mechanism $\mathcal{M}$ preserves (ϵ, δ) MA-LDP if for any two users u and u' and their corresponding data $\boldsymbol{D}_u = (D_u^1, D_u^2, \cdots D_u^n) \in \mathcal{U}$ and $\boldsymbol{D}_{u'} = (D_{u'}^1, D_{u'}^2, \ldots, D_{u'}^n) \in \mathcal{U}$, it satisfies $\mathbb{P}(\mathcal{M}(\boldsymbol{D}_u) \in U) \leq e^{\epsilon}\mathbb{P}(\mathcal{M}(\boldsymbol{D}_{u'}) \in U) + \delta$, $U \in \mathcal{U}$. Here for each agent $i \in N$, the D_u^i and $D_{u'}^i$ differs at exactly one component.*

3 MA-LDP algorithm

In this Section, we introduce the multi-agent locally differential private (MA-LDP) algorithm that achieves the sub-linear regret. Our MA-LDP algorithm is inspired from the UCRL-VTR algorithm of Jia et al. (2020) and also use some structure of UCRL-VTR-LDP algorithm for the single agent RL in Liao et al. (2021). Let K be the number of episodes, and each episode consists of a fixed planning horizon H. Initially, for each agent $i \in N$, the estimate of the global state value function V^i and the global state-action value function Q^i is taken as 0 at the $H+1$-th stage. At every stage h, the optimistic estimator of the state action value function for each agent $i \in N$ is obtained via the backward induction algorithm.

For every stage $h \in [H]$ and agent $i \in N$, we initially receive $\Lambda_{1,h}^i = \Sigma_{1,h}^i = \lambda\mathbf{I}$ and $\widehat{\boldsymbol{\theta}}_{1,h}^i = \mathbf{0}_{nd}$ information from the server for the local user $k = 1$. For the local user k and the received information $\Lambda_{k,h}^i, u_{k,h}^i$, each agent $i \in N$ uses the backward induction algorithm along with an additional UCB bonus term to get the optimistic estimator of the optimal state-action value function. The update is $Q_{k,h}^i(\cdot,\cdot) \leftarrow \min\{H+1-h, \beta_{k,h}||\Sigma_{k,h}^{i^{-1/2}}\phi_{V_{k,h+1}^i}(\cdot,\cdot)||_2 + \bar{r}_h(\cdot,\cdot;\boldsymbol{w}_{k,h}^i) + \langle\hat{\boldsymbol{\theta}}_{k,h}^i, \phi_{V_{k,h+1}^i}(\cdot,\cdot)\rangle\}$ here $\beta_{k,h}$ is identified for each noise mechanism separately. We use $\beta_{k,h}^G, \beta_{k,h}^L, \beta_{k,h}^{BU}, \beta_{k,h}^{BL}$ for each of the noise mechanisms respectively. These are identified in the next section.

In episode k, each agent $i \in N$ realizes the initial global state $\boldsymbol{s}_{k,1}$ and then for all stages $h \in [H]$ it takes action using the current optimistic estimator of the global state-action value $Q_{k,h}^i$. In particular, the action taken by agent $i \in N$ is according to the $\max\min$ criteria that captures its best action against the worst possible action by other agents (lines 11-17). Moreover, each agent maintains an intermediate reward function parameters for each stage $h \in [H]$. This intermediate reward function parameter is used to compute the reward function parameters for the next episode.

Once the action is taken, a new state is realized according to the unknown distribution $\mathbb{P}_h(\cdot|\boldsymbol{s}_h, \boldsymbol{a}_h)$ (line 21). To get the estimate of the true transition probability parameters each agent requires to send some information to the server. In MA-LDP algorithm data that is shared with the server from each agent is obtained via the ridge regression based minimization of the transition probability parameters. Thus, the server requires the true information $\Delta\widetilde{\Lambda}_{k,h}^i = \phi_{V_{k,h+1}^i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})\phi_{V_{k,h+1}^i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})^\top$ and $\Delta\widetilde{u}_{k,h}^i = \phi_{V_{k,h+1}^i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})V_{k,h+1}^i(s_{k,h+1})$ for each agent $i \in N$ and and stage h in episode k. However, to preserve the privacy, we privatize the above true information using different noise adding mechanisms. In particular, to the true information $\Delta\widetilde{\Lambda}_{k,h}^i$ we add a matrix $\mathbf{W}_{k,h}^i$, where each entry of the matrix $\mathbf{W}_{k,h}^i$ is drawn according to a distribution corresponding the noise-adding mechanism used. Moreover, to the true information $\Delta\widetilde{u}_{k,h}^i$, we add $\boldsymbol{\xi}_{k,h}^i$. Again each entry of $\boldsymbol{\xi}_{k,h}^i$ is drawn from different distributions corresponding to different noise adding mechanisms. Let $\Delta\Lambda_{k,h}^i$ and $\Delta u_{k,h}^i$ be the anonymized information that is shared to the server (lines 26-32), i.e., $\Delta\Lambda_{k,h}^i \leftarrow \Delta\widetilde{\Lambda}_{k,h}^i + \mathbf{W}_{k,h}^i$; $\Delta u_{k,h}^i \leftarrow \Delta\widetilde{u}_{k,h}^i + \boldsymbol{\xi}_{k,h}^i$ Server on the other hand collects this information and use it to updates $\Sigma_{k,h}^i$ and $u_{k,h}^i$ and hence gives the next estimate of the transition probability parameters. However, merely adding the noise might not preserve the PSD property, so we shift this matrix by adding an $\eta\mathbf{I}$ to guarantee the PSD property. Finally, each agent sends $\Lambda_{k,h}^i$ and $u_{k,h}^i$ to the $k+1$-th user. Apart from executing a policy that uses an optimistic estimator, each agent $i \in N$ also updates the $\Sigma_{k,h}$ and $u_{k,h}$. These updates are used to estimate the true model parameters $\hat{\boldsymbol{\theta}}_{k+1,h}^i$. They are inspired from the minimizer to a regularized linear regression problem similar to Zhou et al. (2021). Lastly, to ensure the privacy of the reward function, the parameters $\boldsymbol{w}^i$ of the reward functions are updated according to the consensus matrix. It is important to note that we use the stochastic

approximation based rule to update the reward function parameters $\boldsymbol{w}^i$. We want to emphasize that our algorithm uses the most recent reward function parameters. These parameters are updated in each episode, and stage of the episode via the consensus matrix in line 34 of the MA-LDP algorithm. The convergence of the reward function parameters ensures the convergence of state-action value function and these are used in regret analysis of MA-LDP algorithm.

Algorithm 1 MA-Gaussian/Laplace-LDP

Require: Privacy parameters ϵ, δ; failure probability α; parameter η, $\boldsymbol{w}^i_{0,0} = 0$ for all $i \in N$
Set $\sigma = 4H^3\sqrt{2\log(2.5H/\delta)}/\epsilon$ for Gaussian; Set $b = 4H^3\sqrt{nd}/\epsilon$ for Laplace
for user $k = 1, \ldots, K$ **do**
  for $i = 1, 2, \ldots, n$ **do**
    For local user k:
    Receive $\{\Sigma^i_{k,1}, \ldots, \Sigma^i_{k,H}, \hat{\boldsymbol{\theta}}^i_{k,1}, \ldots, \hat{\boldsymbol{\theta}}^i_{k,H}\}$
  end for
  for $i = 1, \ldots, n$ **do**
    for $h = H, \ldots, 1$ **do**
      $Q^i_{k,h}(\cdot,\cdot) \leftarrow \min\{H + 1 - h, \bar{r}_h(\cdot,\cdot;\boldsymbol{w}^i_{k,h}) + \langle \hat{\boldsymbol{\theta}}^i_{k,h}, \phi_{V^i_{k,h+1}}(\cdot,\cdot)\rangle + \beta_{k,h}||\Sigma^{i^{-1/2}}_{k,h}\phi_{V^i_{k,h+1}}(\cdot,\cdot)||_2\}$
      $V^i_{k,h} \leftarrow \max_{a^i \in \mathcal{A}^i} Q^i_{k,h}(\cdot, a^i, \boldsymbol{a}^{-i}_{k,h})$
    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 $\boldsymbol{a}^i_{k,h} \leftarrow arg\max_{\boldsymbol{a} \in \mathcal{A}^i} \min_{\boldsymbol{a}^{-i} \in \mathcal{A}^{-i}} Q^i_{k,h}(\boldsymbol{s}_{k,h}, \boldsymbol{a}, \boldsymbol{a}^{-i})$
      Set $\widetilde{\boldsymbol{w}}^i_{k,h} \leftarrow \boldsymbol{w}^i_{k,h} + \gamma_{k,h} \cdot [r^i_h(\cdot,\cdot) - \bar{r}_h(\cdot,\cdot;\boldsymbol{w}^i_{k,h})] \cdot \nabla_{\boldsymbol{w}} \bar{r}(\cdot,\cdot;\boldsymbol{w}^i_{k,h})$
    end for
    Set $\boldsymbol{a}_{k,h} = (\boldsymbol{a}^1_{k,h}, \boldsymbol{a}^2_{k,h}, \ldots, \boldsymbol{a}^n_{k,h})$
    Observe the next state $s_{k,h+1}$
    for $i = 1, \ldots, n$ **do**
      $\Delta\widetilde{\Lambda}^i_{k,h} = \phi_{V^i_{k,h+1}}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})\phi_{V^i_{k,h+1}}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})^\top$
      $\Delta\tilde{u}^i_{k,h} = \phi_{V^i_{k,h+1}}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})V^i_{k,h+1}(s_{k,h+1})$
      Set $\Delta\Lambda^i_{k,h} \leftarrow \Delta\widetilde{\Lambda}^i_{k,h} + \mathbf{W}^i_{k,h}$
      Set $\Delta u^i_{k,h} \leftarrow \Delta\tilde{u}^i_{k,h} + \boldsymbol{\xi}^i_{k,h}$
    end for
  end for
  for $i = 1, \ldots, n$ **do**
    Set $D^i_k = \{\Delta\Lambda^i_{k,1}, \ldots, \Delta\Lambda^i_{k,H}, \Delta u^i_{k,1}, \ldots, \Delta u^i_{k,H}\}$
  end for
  Send $\mathbf{D}_k = (D^1_k, D^2_k, \ldots, D^n_k)$ to the server
  For the server:
  for $h = 1, \ldots, H$ **do**
    for $i = 1, \ldots, n$ **do**
      $\Lambda^i_{k+1,h} \leftarrow \Lambda^i_{k,h} + \Delta\Lambda^i_{k,h}$
      $u^i_{k+1,h} \leftarrow u^i_{k,h} + \Delta u^i_{k,h}$
      $\Sigma^i_{k+1,h} \leftarrow \Lambda^i_{k+1,h} + \eta I$
      $\hat{\boldsymbol{\theta}}^i_{k+1,h} \leftarrow (\Sigma^i_{k+1,h})^{-1} u^i_{k+1,h}$
    end for
  end for
  Send $\{\Sigma^i_{k+1,1}, \ldots, \Sigma^i_{k+1,H}, \hat{\boldsymbol{\theta}}^i_{k+1,1}, \ldots, \hat{\boldsymbol{\theta}}^i_{k+1,H}\}$ to the user $k+1$
  Update $\boldsymbol{w}^i_{k+1,h} = \sum_{j \in N} l_{k,h}(i,j)\widetilde{\boldsymbol{w}}^j_{k,h}, \forall\, h \in [H]$
end for

4 Preliminary Results

In this Section, we state the results and the definition that are useful in regret analysis and are common for all the noise adding mechanisms. To this end, we first define the notion of privacy loss.

Definition 2 (Privacy loss Dwork et al. (2006)). *For any neighboring datasets d, d', a mechanism $\mathcal{M}$, auxiliary input $\boldsymbol{aux}$, and an outcome $o \in \mathbb{R}$, the privacy loss at outcome o is defined as* $c(o; \mathcal{M}, \boldsymbol{aux}, d, d') := \log \frac{\mathbb{P}(\mathcal{M}(\boldsymbol{aux}, d)=o)}{\mathbb{P}(\mathcal{M}(\boldsymbol{aux}, d')=o)}$.

The privacy loss defined above represents the loss incurred by the outcome o, when the data d is replaced by d'. Note that this loss might be positive or negative. The (ϵ, δ) DP ensures that for neighbouring data d, d' the absolute value of the above privacy loss is bounded by ϵ with probability at least $1-\delta$. When $\delta = 0$ we get $(\epsilon, 0)$ DP, that is, the absolute privacy loss is at most ϵ *a.s.* The $(0, \delta)$ DP implies that the absolute privacy loss is 0 with confidence $1-\delta$.

In out decentralized multi-agent LDP setting, the proof of the regret bounds for each of the noise adding mechanism relies on the asymptotic convergence of reward function parameters. In particular, we show that the reward function parameters converge independent of agent as k (number of episodes) goes to infinity. To this end, let $d(\boldsymbol{s})$ be the stationary distribution of the Markov chain $\{\boldsymbol{s}_t\}_{t\geq 0}$ under policy π, and $\pi(\boldsymbol{s}, \boldsymbol{a})$ be the probability of taking action $\boldsymbol{a}$ in state $\boldsymbol{s}$. Moreover, let $D^{s,a} = diag[d(\boldsymbol{s})\pi(\boldsymbol{s}, \boldsymbol{a})]$ be the diagonal matrix with $d(\boldsymbol{s})\pi(\boldsymbol{s}, \boldsymbol{a})$ as diagonal entries.

Lemma 1. *Under assumptions 2 and 3, 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 $\Psi^\top D^{s,a}(\Psi \boldsymbol{w}^\star - \bar{r}) = 0$.*

We would like to emphasize that this convergence is specifically needed for the decentralized multi-agent setting, and hence novel to our work. The proof is deferred to the Appendix A and it follows on the same lines as in Trivedi and Hemachandra (2023, 2022); Zhang et al. (2018) that uses stochastic approximations methods of Borkar (2022). Next, we show that in our MA-LDP algorithm each agent uses an optimistic estimator of the state-action value function in each episode. To this end, we have the following lemma common to all the noise adding mechanisms except a term $\beta_{k,h}$ (see line 10 of the MA-LDP algorithm) that is separately identified in the Appendix C. The proof is via induction over h, see Appendix B.

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 respectively 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 $Q^{i\star}_h(\boldsymbol{s}, \boldsymbol{a}) \leq Q^i_{k,h}(\boldsymbol{s}, \boldsymbol{a})$ and $V^{\star,i}_h(\boldsymbol{s}) \leq V^i_{k,h}(\boldsymbol{s})$.*

In our regret analysis we also require the following bound on the estimated model parameters.

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

The proof of this Lemma uses a decomposition of $||(\Sigma^i_{k,h})^{1/2}(\hat{\boldsymbol{\theta}}^i_{k,h} - \boldsymbol{\theta}^\star_h)||$ in three different terms, we call them $\mathbf{q}_1, \mathbf{q}_2$, and $\mathbf{q}_3$. We give this decomposition in Appendix C. Each of these terms are bounded differently for every noise mechanism that helps in identifying the respective β_k.

5 Gaussian and Laplace noise adding mechanisms (unbounded support)

In this Section, we consider two most popular noise adding mechanisms, Gaussian and Laplace. To our knowledge, the multi-agent version of these are not studied so far. We first show that our MA-LDP algorithm preserves (ϵ, δ) and $(\epsilon, 0)$ LDP for Gaussian and Laplace mechanisms respectively. We also show that MA-LDP algorithm for these mechanisms achieves the sub-linear regret.

5.1 Gaussian Noise Adding Mechanism

For the Gaussian mechanism, we add $\mathbf{W}^i_{k,h}$ to $\Lambda^i_{k,h}$ in line (21) of MA-LDP algorithm. Recall, $\mathbf{W}^i_{k,h}$ is a symmetric matrix, so for $l \leq m$, each entry (l, m) of $\mathbf{W}^i_{k,h}$ is sampled from the Gaussian distribution $\mathcal{N}(0, \sigma^2)$ distribution. 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.

Theorem 1. *If we choose 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 Gaussian mechanism $\mathcal{M}_G$ will satisfy (ϵ, δ) MA-LDP property.*

The proof of this Theorem is referred to Appendix D and D.1. and use Theorem A.2 of Liao et al. (2021) (for details see Appendix G.2). Next, we show that the above mentioned Gaussian mechanism with privacy parameters ϵ, δ achieves a sub-linear regret.

Theorem 2. *Let $\alpha \in (0,1)$, $\eta = 1$, privacy parameter $\epsilon > 0, \delta > 0$, and $\beta^G_{k,h} = c_g(nd)^{3/4}(H-h+1)^{3/2}k^{1/4}\log(ndT/\alpha)((\log(H-h+1)/\delta))^{1/4}\sqrt{1/\epsilon}$ with c_g being an absolute constant. Consider the Gaussian noise mechanism with parameter σ as in Theorem 1. Then, for any user k, with probability at least $1-\alpha$, the total regret of MA-LDP algorithm in the first $T = KH$ steps with BL noise mechanism is at most $\widetilde{\mathcal{O}}(n^{5/4}d^{5/4}H^{7/4}T^{3/4}\log(ndT/\alpha)(\log(H/\delta))^{1/4}\sqrt{1/\epsilon})$.*

Proof. (Sketch Only). Here we provide the outline of the proof. The proof involve four major steps: 1) to show that the model parameters obtained from the MA-LDP algorithm converges to the true model parameters $\boldsymbol{\theta}^\star$ within the confidence radius β^g_k (Lemma 3 in previous Section). 2) Using induction argument to show that in each step MA-LDP algorithm uses an optimistic estimator of the state-action value function (Lemma 2 in the previous Section). 3) The convergence of the reward function parameters that is shown in the previous subsection in Lemma 1. Finally, 4) bounding the regret via concentration inequalities. The details of this step are given in Appendix E.1. □

5.2 Laplace Noise Adding Mechanism

Unlike Gaussian mechanism for the Laplace noise adding mechanisms, we add $\mathbf{W}^i_{k,h}$ to $\Lambda^i_{k,h}$ in line (21) of the MA-LDP algorithm. Recall, $\mathbf{W}^i_{k,h}$ is a symmetric matrix, so for $l \leq m$, each entry (l,m) of $\mathbf{W}^i_{k,h}$ is sampled from the Laplace distribution $\mathcal{L}(b)$ distribution. Moreover, to each entry of $u^i_{k,h}$ we add $\boldsymbol{\xi}^i_{k,h}$ sampled from $\mathcal{L}(b;)$ distribution.

Theorem 3. *If we choose the parameter b of the Laplace distribution $\mathcal{L}(b)$ such that $b = \frac{4H^3\sqrt{nd}}{\epsilon}$, then, the Laplace noise adding mechanism $\mathcal{M}_L$ will satisfy $(\epsilon, 0)$ MA-LDP property.*

The proof of this theorem is deferred to Appendix D and D.2. Next, we provide the upper bound on the regret incurred by MA-LDP algorithm with Laplace mechanism.

Theorem 4. *Let $\alpha \in (0,1)$, $\eta = 1$, privacy parameter $\epsilon > 0$, and $\beta^L_{k,h} = c_l(nd)^{3/4}(H-h+1)^{3/2}k^{1/4}\log(ndT/\alpha)\sqrt{1/\epsilon}$ with c_l being an absolute constant. Consider the Laplace noise mechanism with parameter b as in Theorem 3. Then, for any user k, with probability at least $1-\alpha$, the total regret of MA-LDP algorithm in the first $T = KH$ steps with Laplace noise mechanism is at most $\widetilde{\mathcal{O}}(n^{5/4}d^{5/4}H^{7/4}T^{3/4}\log(ndT/\alpha)\sqrt{1/\epsilon})$.*

6 Uniform and bounded Laplace noise adding mechanisms

In this Section, we consider the noise adding mechanism with the bounded support. The motivation comes from the fact that noise mechanisms with unbounded support allows all the noise values to attain the LDP with given privacy parameters ϵ, δ; some of these noise values can be arbitrarily large. So, the central question is: does restricting the noise values to a bounded support change the privacy property or the regret or both? If so, how? If it doesn't affect the privacy property, then is it that adding a bounded noise attains the same regret with the same privacy parameters. In this later case, it means that we need not inject huge noises to the system. In the this section, we show that this is indeed possible, and for a certain relation between the distribution parameters and the support of the Laplace mechanism, our MA-LDP algorithm attains the same regret up to the constants; this is one of our main result. Moreover, we compare these regrets with the most natural choice of the uniform noise mechanism. We show that the privacy guarantees of the uniform and Laplace mechanisms with bounded support are two extremes, $(0,\delta)$ and $(\epsilon, 0)$ LDP respectively.

6.1 Uniform Noise Adding Mechanism

The uniform (U) mechanism with MA-LDP algorithm works as follows; for a given $a > 0$ we denote by $\mathcal{U}[-a,a]$ the uniform distribution over the interval $[-a,a]$. For the uniform mechanism, we add $\mathbf{W}^i_{k,h}$ to $\Lambda^i_{k,h}$ in line (21) of the MA-LDP algorithm. Recall, $\mathbf{W}^i_{k,h}$ is a symmetric matrix, so for $l \leq m$, each entry (l,m) of $\mathbf{W}^i_{k,h}$ is sampled from the uniform distribution $\mathcal{U}[-a,a]$ distribution. Moreover, to $u^i_{k,h}$ we add $\boldsymbol{\xi}^i_{k,h}$ whose each entry is sampled from $\mathcal{U}[-a,a]$ distribution.

Theorem 5. *If we choose the parameter a of the uniform distribution $\mathcal{U}[-a,a]$ such that $a = 4H^3\sqrt{\log(2H/\delta)}$, then MA-LDP algorithm with mechanism $\mathcal{M}_U$ as preserves $(0,\delta)$ LDP property.*

The proof of this theorem is deferred to the Appendix D adn D.3 The regret of MA-LDP algorithm with BU mechanism is given in the following theorem. The proof is given in Appendix E and E.3

Theorem 6. *Let $\alpha \in (0,1)$, $\eta = 1$, privacy parameter $\delta > 0$, and $\beta^U_{k,h} = c_u(nd)^{3/4}(H-h+1)^{3/2}k^{1/4}\log(ndT/\alpha)(\log(H-h+1/\delta))^{1/4}$, with c_u being an absolute constant. Consider the uniform noise mechanism with parameter a as in Theorem 5. Then, for any user k, with probability at least $1-\alpha$, the total regret of MA-LDP algorithm in the first $T = KH$ steps with uniform noise mechanism is at most $\widetilde{\mathcal{O}}(n^{5/4}d^{5/4}H^{7/4}\ T^{3/4}\log(ndT/\alpha)(\log(H/\delta))^{1/4})$.*

6.2 Bounded Laplace Noise Adding Mechanism

In this Section, we design another noise adding mechanism with bounded support. We call it bounded Laplace (BL) mechanism. The BL mechanism is obtained by restricting the support of the Laplace distribution to $[-B, B]$ for a given $B > 0$. The probability density of the bounded Laplace distribution, $\mathcal{BL}(x; b)$, with parameter b is $f_{\mathcal{BL}}(x; b) = (2b\left(1-\exp(\frac{-B}{b})\right))^{-1}\exp\left(\frac{-|x|}{b}\right), \forall\ x \in [-B, B]$, and 0 otherwise. The variance of this distribution is $\zeta = 2b^2(1-\exp(\frac{-B}{b}))^{-1} - \kappa$ where $\kappa = ((B+b)^2+b^2)\times\exp(\frac{-B}{b})\times(1-\exp(\frac{-B}{b}))^{-1}$. For the BL mechanism, we add $\mathbf{W}^i_{k,h}$ to $\Lambda^i_{k,h}$ in line (21) of the MA-LDP algorithm. Recall, $\mathbf{W}^i_{k,h}$ is a symmetric matrix, so for $l \le m$, each entry (l,m) of $\mathbf{W}^i_{k,h}$ is sampled from $\mathcal{BL}(b; B)$ distribution. Moreover, to $u^i_{k,h}$ we add $\boldsymbol{\xi}^i_{k,h}$ whose each entry is sampled from $\mathcal{BL}(b, B)$ distribution.

Theorem 7. *If we choose the parameter b of the bounded Laplace distribution $\mathcal{BL}(b; B)$ such that $b = \frac{4H^3\sqrt{nd}}{\epsilon}$, then, the BL mechanism $\mathcal{M}_{BL}$ will satisfy $(\epsilon, 0)$ MA-LDP property.*

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

Theorem 8. *Let $\alpha \in (0,1)$, $\eta = 1$, privacy parameter $\epsilon > 0$, and $\beta^{BL}_{k,h} = c_{bl}(nd)^{3/4}(H-h+1)^{3/2}k^{1/4}\log(ndT/\alpha)\sqrt{1/\epsilon}$, with c_{bl} being an absolute constant. Consider the BL noise mechanism with parameter b as in Theorem 7. Then, for any user k, with probability at least $1-\alpha$, the total regret of MA-LDP algorithm in the first $T = KH$ steps with BL noise mechanism is at most $\widetilde{\mathcal{O}}(n^{5/4}d^{5/4}\zeta^{1/4}H^{1/4}T^{3/4}\log(ndT/\alpha))$*

Mechanism	Type	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: Privacy guarantees and the order of regret for different noise adding mechanisms.

7 Comparison of regret for different noise mechanisms

First, from the regret expressions of Gaussian and Laplace mechanism in Theorems 2 and 4 it follows that $R_K(\epsilon_1)/R_K(\epsilon_2) = \sqrt{\epsilon_2/\epsilon_1}$. Thus, we have:

Theorem 9. *If privacy parameters ϵ_1 and ϵ_2 are such that $\epsilon_1 > \epsilon_2$. Then for both the Gaussian and Laplace mechanisms we have that $R_K(\epsilon_1) < R_K(\epsilon_2)$.*

Moreover, we have the following Theorem for the cumulative regret between the Gaussian and Laplace mechanism for the same privacy parameter ϵ.

Theorem 10. *Let $R^G_K(\epsilon), R^L_K(\epsilon)$ be the cumulative regret of the Gaussian and Laplace mechanism respectively with privacy parameters ϵ, δ, and $H > 2$. Then, $R^G_K(\epsilon) > R^L_K(\epsilon)$.*

The proof follows from the fact that the Gaussian noise mechanism gives the approximate LDP, i.e., $\delta > 0$, so $\log(H/\delta) > 1$ for all $\delta \in (0,1)$ and $H > 2$. Therefore, $R^G_K(\epsilon)/R^L_K(\epsilon) = \log(H/\delta) > 1$.

Apart from the above, recall ζ, the variance of the bounded Laplace distribution and is a function of B and b. So, depending on whether B has the same order as that of b or not, we get different expressions of ζ, thus different

order of the regret (Theorem 8). So, we compare the regret of the BL mechanism with the Laplace mechanism when $B = O(b^\gamma)$ for different possible values of γ. For the case of $\gamma > 1$, we have that $R_K^{BL}/R_K^L = \left(H^3/\epsilon\right)^{\frac{\gamma}{2}}$. If

B	R_K^{BL}
$O(b^\gamma), 0 \leq \gamma \leq 1$	$\widetilde{\mathcal{O}}((nd)^{5/4}H^{7/4}T^{3/4}\log(ndT/\alpha))\sqrt{1/\epsilon}$
$O(b^\gamma), \gamma > 1$	$\widetilde{\mathcal{O}}((nd)^{5/4}H^{7/4}H^{3\gamma/2}T^{3/4}\log(ndT/\alpha))\sqrt{1/\epsilon^{\gamma+1}}$

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 $\gamma \in [0, 1]$.

$\left(H^3/\epsilon\right)^{\frac{\gamma}{2}} > 1$ and $\gamma > 1$, then the regret of MA-LDP with BL mechanism is more than that of Laplace mechanism. So, for the given problem instance, the regret of 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 will have lower regret than that of Laplace. Though the BL mechanism injects noise from a bounded support, the regret of MA-LDP algorithm with 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 $\gamma \in [0, 1]$. 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 that of the Laplace mechanism.

Theorem 11. *If the end point B of the support of bounded Laplace, $\mathcal{BL}$, distribution (with parameter b) is of order $O(b^\gamma)$, where $\gamma \leq 1$, then the order of the regret of the MA-LDP algorithm with BL mechanism is the same as that of the Laplace mechanism.*

8 Computational Experiments

In this section, we give computational results to validate the usefulness of our MA-LDP algorithm. To this end, we consider a network with $q + 2$ nodes shown in Figure 1 below, $q \geq 1$, i.e., $\{s_{in}, 1, 2, \ldots, q, g\}$, where s_{in}, g are the initial and goal nodes respectively. The number of global states are $(q + 2)^n$.

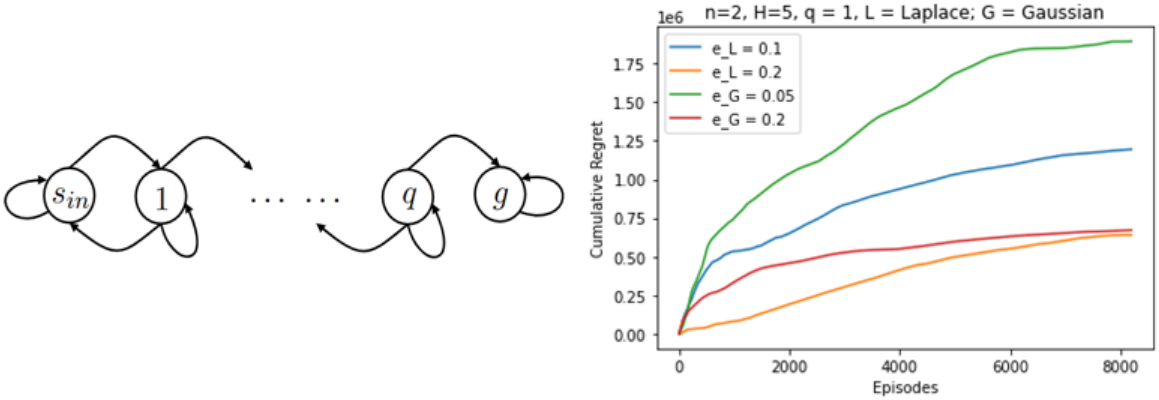

Figure 1: (Left) The MDP problem instance that we consider. (Right) Cumulative regret with number of episodes for the Laplace and Gaussian mechanism.

In each state the actions available to each agent are $\mathcal{A}^i = \{-1, 1\}^{d-1}$, for given $d \geq 2$. So, the total number of actions is $2^{n(d-1)}$. Each agent $i \in N$ receives a reward of $5/1000$ units for taking any action in s_{in}, a reward of 1000 for action in g, and the reward of 0 unit for action in other nodes. The collective objective of the agents is to reach the goal node in a decentralized way while maximizing the overall reward. To address humongous state, action space, we parameterize the transition probability as $\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a}) = \langle \phi(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a}), \boldsymbol{\theta}(\boldsymbol{s}) \rangle$ for each $(\boldsymbol{s}', \boldsymbol{a}, \boldsymbol{s}) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. The feature design for this transition probability is novel to our work and is given below. Let $S(\boldsymbol{s})$ is the set all feasible states from state $\boldsymbol{s}$. We define the global features as: $\phi(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a}) = (\phi(s'^1|s^1, \boldsymbol{a}^1), \ldots, \phi(s'^n|s^n, \boldsymbol{a}^n))$, 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 $\boldsymbol{s} = \boldsymbol{g}, \boldsymbol{s}' \neq \boldsymbol{g}$, and $(\mathbf{0}_{nd-1}, \alpha(\boldsymbol{s}))$, if $\boldsymbol{s} = \boldsymbol{g}, \boldsymbol{s}' = \boldsymbol{g}$. 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 in Appendix F. 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 $\Delta < \delta$. More details of the experiments and the MDP involved are given in Appendix F.1. The result below shows that the above choice of feature design $\phi(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})$ and model parameters $\boldsymbol{\theta}(\boldsymbol{s})$ a yields a valid MDP.

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) $\langle \phi(\boldsymbol{s}' \neq \boldsymbol{g}|\boldsymbol{s} = \boldsymbol{g},\boldsymbol{a}), \boldsymbol{\theta}(\boldsymbol{s})\rangle = 0, \ \forall\, \boldsymbol{a}$.*

The proof of this Lemma is deferred to the Appendix F.1. We implement our MA-LDP algorithm with various privacy parameters on the Gaussian and Laplace mechanisms with unbounded support. Figure 1 shows the cumulative regret with $n = 2$, number of nodes in the network as 3, planning horizon $H = 5$. All the values are averaged over 10 runs; each run has $K = 8500$ episodes. Here are some observations from the experiments. Firstly, for a given mechanism, the cumulative regret with lower privacy losses are higher than the higher privacy losses. That is $R_K(\epsilon_L = 0.1) > R_K(\epsilon_L = 0.2)$; and $R_K(\epsilon_G = 0.05) > R_K(\epsilon_G = 0.2)$. This illustrates the results of Theorems 9 and 10 that compare cumulative regrets of various privacy losses.

9 Related Work

The idea of DP is first introduced in Dwork et al. (2006). DP is motivated by the designing the algorithms that preserve the user's sensitive data. It indicates that changing or removing a data point has little influence on any observable output. However, DP has risk of data leakage, and is vulnerable to membership influence attacks Shokri et al. (2017). Hence a stronger notion 'Locally DP' is introduced Kasiviswanathan et al. (2011); Duchi et al. (2013). Under LDP, users send privatized data to the server and each individual user maintains its own sensitive data. The server, on the other hand, is totally agnostic about the sensitive data. The LDP problem in the multi-agent distributed optimization is studied in Dobbe et al. (2018). The amplification of the privacy in machine learning is studied in Cyffers and Bellet (2022). The optimal noise adding mechanism are introduced in Geng and Viswanath (2015). Apart from the learning and decision making the notion of DP is also common 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 distributionally trust model of the DP that guarantees the privacy without trustworthy server. Recently, Jia et al. (2020); Jin et al. (2020) propose single agent RL scheme with linear function approximation of the transition probability function. Recently, Liao et al. (2021) use the UCRL-VTR algorithm of Jia et al. (2020) and incorporate the notion of DP into it. However, these consider a single agent taking the decisions in the finite horizon models. In this work, we introduce MA-LDP algorithm in a fully decentralized MARL framework that use different noise mechanisms and identify conditions when bounded support mechanism attains regret that is comparable to that is offered by conventional mechanisms.

10 Discussion

In this work, we consider the notion of the LDP in the fully decentralized MARL setting. We first define LDP for MARL and then propose a generic MA-LDP algorithm which can handle any noise adding mechanism. We show that the MA-LDP algorithm preserves the privacy for four different noise adding mechanisms, then prove that it also achieves the sub-linear regret. Next, we compare the noise mechanisms with bounded support with that of unbounded support. Our key observation is that if the support of bounded noise distribution is picked appropriately, the regret is lower than the unbounded support noise mechanism. Thus, injecting a bounded noise is often sufficient for LDP without substantially affecting the nature of the regret. We illustrate our results on a networked MDP with many states and actions.

The work we consider 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, a nice update rule for the optimistic estimators of the state-action value function can be attempted. Secondly, to our knowledge, the regret bound we show are of its 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 bounded support noise mechanism that leads to the lower regret bounds with low noise values would be interesting.

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.

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.

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.

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*, pages 1–31.

Trivedi, P. and Hemachandra, N. (2023). Multi-agent congestion cost minimization with linear function approximation. *arXiv preprint arXiv:2301.10993*.

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.

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.

A Proofs of convergence of reward function parameters (Lemma 1)

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 $\boldsymbol{w}_t^i$ for all $i \in N$. For proof, we refer to Zhang et al. (2018).

Proposition 1. *Under assumptions 2, and 3 the sequence* $\{\boldsymbol{w}_t^i\}$ *satisfy* $sup_t\ ||\boldsymbol{w}_t^i|| < \infty$ *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 $r_t = [r_t^1, \ldots, r_t^n]^\top \in \mathbb{R}^n$, and $\boldsymbol{w}_t = [(\boldsymbol{w}_t^1)^\top, \ldots, (\boldsymbol{w}_t^n)^\top]^\top \in \mathbb{R}^{np}$. Moreover, let $A \otimes B$ represent the Kronecker product of any two matrices A and B. Let $y_t = [(y_t^1)^\top, \ldots, (y_t^n)^\top]^\top$, where $y_{t+1}^i = [(r_{t+1}^i - \psi_t^\top \boldsymbol{w}_t^i)\psi_t^\top]^\top$. 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

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

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 2, and 3 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 (1)) can be written as

$$\begin{aligned}
\langle \boldsymbol{w}_{t+1} \rangle &= \frac{1}{N}(\mathbb{1}^\top \otimes I)(L_t \otimes I)(\mathbb{1} \otimes \langle \boldsymbol{w}_t \rangle + \boldsymbol{w}_{\perp,t} + \gamma_t\ y_{t+1}) \\
&= \langle \boldsymbol{w}_t \rangle + \gamma_t \langle (L_t \otimes I)(y_{t+1} + \gamma_t^{-1}\boldsymbol{w}_{\perp,t}) \rangle \\
&= \langle \boldsymbol{w}_t \rangle + \gamma_t\ \mathbb{E}(\langle y_{t+1} \rangle | \mathcal{F}_t) + \beta_t \xi_{t+1},
\end{aligned} \tag{2}$$

where

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

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

$$\mathbb{E}[||\xi_{t+1}||^2 \mid \mathcal{F}_t] \leq \mathbb{E}[||y_{t+1} + \gamma_t^{-1}\boldsymbol{w}_{\perp,t}||_{R_t}^2 \mid \mathcal{F}_t] + ||\mathbb{E}(\langle y_{t+1} \rangle \mid \mathcal{F}_t)||^2, \tag{3}$$

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 (3), we have, for any $K_1 > 0$

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

over the set $\{sup_t\ ||\boldsymbol{w}_t|| \leq K_1\}$ for some $K_3 < \infty$. Thus condition (3) of assumption 4 is satisfied. The ODE associated with the Equation (2) has the form

$$\langle \dot{\boldsymbol{w}} \rangle = -\Psi^\top D^{\boldsymbol{s},\boldsymbol{a}} \Psi \langle \boldsymbol{w} \rangle + \Psi^\top D^{\boldsymbol{s},\boldsymbol{a}} \bar{r} \tag{4}$$

Let the RHS of Equation (4) 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 (4) 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 8.1 and Theorem 8.3 on page 114-115 in Borkar (2022) we have $lim_t \langle \boldsymbol{w}_t \rangle = \boldsymbol{w}^\star$ a.s. over the set $\{sup_t ||\boldsymbol{w}_t|| \leq K_1\}$ 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

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 $Q_{k,h+1}^i(\cdot,\cdot) \geq Q_{k,h+1}^{\star,i}(\cdot,\cdot)$, and $V_{k,h+1}^i(\cdot) \geq V_{k,h+1}^{\star,i}(\cdot)$. For any stage h and $(\boldsymbol{s},\boldsymbol{a})$, if $Q_{k,h}^i(\boldsymbol{s},\boldsymbol{a}) \geq H$, then the statement is also true for stage h, i.e., $Q_{k,h}^i(\boldsymbol{s},\boldsymbol{a}) \geq H \geq Q_h^{\star,i}(\boldsymbol{s},\boldsymbol{a})$. However, if $Q_{k,h}^i(\boldsymbol{s},\boldsymbol{a}) \leq H$, 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}_h V_{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}_h V_{h+1}^{\star,i}(\boldsymbol{s},\boldsymbol{a}) \\
&\overset{(ii)}{=} \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}_h V_{h+1}^{\star,i}(\boldsymbol{s},\boldsymbol{a}) \\
&\quad - \left\langle \boldsymbol{\theta}_h^\star, \phi_{V_{k,h+1}^i}(\boldsymbol{s},\boldsymbol{a}) \right\rangle + \left\langle \boldsymbol{\theta}_h^\star, \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 - \left\langle \hat{\boldsymbol{\theta}}_{k,h}^i - \boldsymbol{\theta}_h^\star, \phi_{V_{k,h+1}^i}(\boldsymbol{s},\boldsymbol{a}) \right\rangle \\
&\quad - \mathbb{P}_h V_{h+1}^{\star,i}(\boldsymbol{s},\boldsymbol{a}) + \mathbb{P}_h V_{k,h+1}^i(\boldsymbol{s},\boldsymbol{a}) \\
&\overset{(iii)}{\geq} \beta_k ||\Sigma_{k,h}^{i^{-1/2}} \phi_{V_{k,h+1}^i}(\boldsymbol{s},\boldsymbol{a})||_2 - ||\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},\boldsymbol{a})||_2 \\
&\quad - \mathbb{P}_h V_{h+1}^{\star,i}(\boldsymbol{s},\boldsymbol{a}) + \mathbb{P}_h V_{k,h+1}^i(\boldsymbol{s},\boldsymbol{a}) \\
&\overset{(iv)}{\geq} -\mathbb{P}_h V_{h+1}^{\star,i}(\boldsymbol{s},\boldsymbol{a}) + \mathbb{P}_h V_{k,h+1}^i(\boldsymbol{s},\boldsymbol{a}) \\
&\overset{(v)}{\geq} 0
\end{aligned}$$

In (i) we use the update of $Q_{k,h}^i(\boldsymbol{s},\boldsymbol{a})$ as in line 7 of the MA-LDP algorithm 1. In (ii) we add and subtract a inner product term. Inequality (iii) follows from the Cauchy-Schwartz inequality. The inequality (iii) follows from Lemma 3, and (iv) is by induction assumption. Finally, the last inequality (v) uses the monotone property of $\mathbb{P}_h$ with respect to the partial ordering of the function. □

C Finding β_k (Proof of Lemma 3)

In this Section, we give the details of β_k for each of the noise adding mechanism. To this end, we first decompose $||(\Sigma_{k,h}^i)^{1/2}(\hat{\boldsymbol{\theta}}_{k,h}^i - \boldsymbol{\theta}_h^\star)|| = ||(\Sigma_{k,h}^i)^{-1/2}(\mathbf{q}_1 + \mathbf{q}_2 + \mathbf{q}_3)||$ into three terms. Note that this decomposition is common to all the noise adding mechanism. However, to get the exact expression we bound each term differently for different noise mechanisms. Consider the difference

$$\hat{\boldsymbol{\theta}}_{k,h}^i - \boldsymbol{\theta}_h^\star \overset{(i)}{=} (\Sigma_{k,h}^i)^{-1} \sum_{\tau=1}^{k-1} \{\phi_{V_{\tau,h+1}^i}(\boldsymbol{s}_{\tau,h},\boldsymbol{a}_{\tau,h}) V_{\tau,h+1}^i(s_{\tau,h+1}) + \boldsymbol{\xi}_{\tau,h}^i\} - \boldsymbol{\theta}_h^\star$$

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

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 30 of the algorithm 1. 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:

$$
||(\Sigma^i_{k,h})^{1/2}(\hat{\boldsymbol{\theta}}^i_{k,h} - \boldsymbol{\theta}^\star_h)|| = ||(\Sigma^i_{k,h})^{-1/2}(\mathbf{q}_1 + \mathbf{q}_2 + \mathbf{q}_3)|| \tag{5}
$$

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 thee noise adding mechanism.

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, each entry of matrix $\mathbf{W}^i_{j,h}$ is sampled from the a 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. So, from the known concentration results of Tao (2012), we have that

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

that is

$$
\mathbb{P}\left(||\mathbf{W}^i_h|| \geq \Gamma\right) \leq \frac{\alpha}{6H} \tag{7}
$$

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 basic matrix $2\Gamma\mathbf{I}$ 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 $1 - \alpha/6$. Define the following event

$$
\mathcal{E}_1 \coloneqq \{\forall\, h \in [H],\ \forall\, j \in [nd],\ \Gamma \leq \sigma_j \leq 3\Gamma\}, \tag{8}
$$

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

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

$$= \sqrt{3nd\sigma\sqrt{k-1}(\sqrt{4nd} + 2\log(6H/\alpha)) + \lambda nd}.$$

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

Bounding $\mathbf{q}_2$

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

$$||(\Sigma_{k,h}^i)^{-1/2}\mathbf{q}_2|| = ||\sum_{\tau=1}^{k-1} \phi_{V_{\tau,h+1}^i}[V_{\tau,h+1}^i - \mathbb{P}_h V_{\tau,h+1}^i]||_{(\Sigma_{k,h}^i)^{-1}} \tag{9}$$

$$\leq ||\sum_{\tau=1}^{k-1} \phi_{V_{\tau,h+1}^i}[V_{\tau,h+1}^i - \mathbb{P}_h V_{\tau,h+1}^i]||_{(\mathbf{Z})^{-1}}, \tag{10}$$

where $\mathbf{Z} = \lambda\mathbf{I} + \sum_{\tau=1}^{k-1} \phi_{V_{\tau,h+1}^i}\phi_{V_{\tau,h+1}^i}^\top$. The inequality holds because, $\Sigma_{k,h}^i \succeq \mathbf{Z} = \lambda\mathbf{I} + \sum_{\tau=1}^{k-1} \phi_{V_{\tau,h+1}^i}\phi_{V_{\tau,h+1}^i}^\top$. Let $\eta_{\tau,h+1}^i = V_{\tau,h+1}^i - \mathbb{P}_h V_{\tau,h+1}^i = V_{\tau,h+1}^i - \phi_{V_{\tau,h+1}^i}^\top \boldsymbol{\theta}_h^\star$. Moreover, let $\{\mathcal{G}_t\}_{t=1}^\infty$ be a filtration, $\{\phi_{V_{\tau,t}^i}, \eta_{\tau,t}^i\}_{t=1}^\infty$ a stochastic process so that $\phi_{V_{\tau,t}^i}$ is $\mathcal{G}_t$-measurable and $\eta_{\tau,t}^i$ is $\mathcal{G}_{t+1}$-measurable. With above notations, we have

$$|\eta_{\tau,h}^i| = |V_{\tau,h}^i - \mathbb{P}_h V_{\tau,h}^i| \leq H. \tag{11}$$

The above is true because $V_{\tau,h}^i \leq H$. Further, we have

$$\mathbb{E}[(\eta_{\tau,h}^i)^2|\mathcal{G}_h] \leq \mathbb{E}[(V_{\tau,h}^i)^2|\mathcal{G}_h] \leq H^2. \tag{12}$$

Moreover, we define the following event

$$\mathcal{E}_2 = \Bigg\{\forall\, h \in [H],\ ||\mathbf{q}_2||_{\mathbf{Z}^{-1}} \leq 4H\Bigg(2\sqrt{nd\log\left(1 + \frac{(k-1)H^2}{nd\lambda}\right)\log\left(\frac{24(k-1)^2}{\alpha}\right)} + \log\left(\frac{24(k-1)^2}{\alpha}\right)\Bigg)\Bigg\}. \tag{13}$$

From Theorem 2 of Zhou et al. (2021) we have 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

$$||\sum_{\tau=1}^{k-1} \boldsymbol{\xi}_{\tau,h}^i||_{(\Sigma_{k,h}^i)^{-1}} \leq ||\sum_{\tau=1}^{k-1} \boldsymbol{\xi}_{\tau,h}^i||_{(\mathbf{W}_h^i + \lambda\mathbf{I})^{-1}} \leq \frac{1}{\sqrt{\rho_{\min}}}||\sum_{\tau=1}^{k-1} \boldsymbol{\xi}_{\tau,h}^i||_2 \tag{14}$$

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

$$||\sum_{\tau=1}^{k-1} \boldsymbol{\xi}_{\tau,h}^i||_2 \leq \sigma\sqrt{(k-1)nd\log\frac{12ndH}{\alpha}}. \tag{15}$$

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

$$\mathcal{E}_3 = \left\{\forall\, h \in [H] :\ ||\sum_{\tau=1}^{k-1} \boldsymbol{\xi}_{\tau,h}^i||_2 \leq \sigma\sqrt{(k-1)nd\log\frac{12ndH}{\alpha}}\right\}. \tag{16}$$

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

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

$$\overset{(ii)}{\leq} \frac{\sigma\sqrt{(k-1)nd\log\frac{12ndH}{\alpha}}}{\sqrt{\rho_{\min}}}$$

$$\overset{(iii)}{=} \frac{\sigma\sqrt{(k-1)nd\log\frac{12ndH}{\alpha}}}{\sqrt{\sigma\sqrt{k-1}(\sqrt{4nd}+2\log(6H/\alpha))+\lambda}}$$

$$\overset{(iv)}{\leq} \frac{\sqrt{\sigma}(k-1)^{1/4}\sqrt{nd\log\frac{12ndH}{\alpha}}}{\sqrt{\sqrt{4nd}+2\log(6H/\alpha)}}$$

$$\overset{(v)}{\leq} (nd)^{1/4}\sqrt{\sigma}(k-1)^{1/4}\sqrt{\log(12ndH/\alpha)}. \tag{17}$$

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 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 (9), (13), and (17), we have that with probability at least $1-\alpha/2$, we have for each $i \in N$ and for each $h \in [H]$,

$$||(\Sigma^i_{k,h})^{1/2}(\hat{\boldsymbol{\theta}}^i_{k,h} - \boldsymbol{\theta}^\star_h)|| \leq \beta_k \tag{18}$$

here,

$$\beta_k = c(nd)^{3/4}\sqrt{\sigma}k^{1/4}\log(ndT/\alpha) \tag{19}$$

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

In the next Section, we identify 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 5 given in Appendix G.2. Moreover, Δf is the l_2 sensitivity which is identified as $2H^2$. Using this $\sigma = \frac{4H^3\sqrt{2\log(2.5H/\delta)}}{\epsilon}$ in Equation (19), we have β^G_k for the Gaussian mechanism is

$$\beta^G_k = c_g(nd)^{3/4}H^{3/2}k^{1/4}\log(ndT/\alpha)(\log(H/\delta))^{1/4}\sqrt{1/\epsilon}. \tag{20}$$

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 (19), we have β^L_k for the Laplace mechanism as

$$\beta^L_k = c_l(nd)^{3/4}H^{3/2}k^{1/4}\log(ndT/\alpha)\sqrt{1/\epsilon}. \tag{21}$$

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 (19), we have β^L_k for the Laplace mechanism as

$$\beta^U_k = c_u(nd)^{3/4}H^{3/2}k^{1/4}\log(ndT/\alpha)(\log(H/\delta))^{1/4}. \tag{22}$$

C.4 β^{BL}_k 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 (19), we have β^{BL}_k for the bounded Laplace mechanism as

$$\beta^{BL}_k = c_{bl}(nd)^{3/4}\zeta^{1/4}k^{1/4}\log(ndT/\alpha). \tag{23}$$

D Proof of privacy guarantee of noise adding mechanism

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 $\Delta\tilde{u}^i_{k,h}$ and $\Delta\tilde{\Lambda}^i_{k,h}$ be the noise-free information of agent $i \in N$ at the h-th stage of k-th episode. That is,

$$\Delta\tilde{u}^i_{k,h} = \phi_{V^i_{k,h+1}}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})V^i_{k,h+1}(s_{k,h+1})$$
$$\Delta\tilde{\Lambda}^i_{k,h} = \phi_{V^i_{k,h+1}}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})\phi_{V^i_{k,h+1}}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})^\top$$

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

$$||\Delta\tilde{u}^i_{k,h} - (\Delta\tilde{u}^i_{k,h})'||_2 \leq ||\phi_{V^i_{k,h+1}}|| \cdot |V^i_{k,h+1}| + ||\phi'_{V^i_{k,h+1}}|| \cdot |V'^i_{k,h+1}| \leq 2H^2 \tag{24}$$

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

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

Thus, 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 $||A||_1 \leq \sqrt{l}||A||_2$. Similarly, for any vector $\mathbf{x} \in \mathbb{R}^l$, we have $||\mathbf{x}||_1 \leq \sqrt{l}||\mathbf{x}||_2$. Using this property on the information's we have

$$||\Delta\tilde{u}^i_{k,h} - (\Delta\tilde{u}^i_{k,h})'||_2 \leq \sqrt{nd}||\Delta\tilde{u}^i_{k,h} - (\Delta\tilde{u}^i_{k,h})'||_1 \leq 2\sqrt{nd}H^2 \tag{25}$$

Similarly, we have

$$||\phi_{V^i}\phi^\top_{V^i} - \phi'_{V^i}\phi'^\top_{V^i}||_F \leq ||\phi_{V^i}\phi^\top_{V^i}||_F + ||\phi'_{V^i}\phi'^\top_{V^i}||_F \leq \sqrt{nd}||\phi_{V^i}\phi^\top_{V^i}||_1 + \sqrt{nd}||\phi'_{V^i}\phi'^\top_{V^i}||_1 \leq 2\sqrt{nd}H^2$$

Thus, the l_1 sensitivity of both the information is $2\sqrt{nd}H^2$. Let $\mathbf{D}_h = (D^1_h, D^2_h, \ldots, D^n_h)$ and $\mathbf{D}'_h = (D'^1_h, D'^2_h, \ldots, D'^n_h)$ 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 $\Delta\boldsymbol{\Lambda}_{k,h} = (\Delta\Lambda^1_{k,h}, \Delta\Lambda^2_{k,h}, \ldots, \Delta\Lambda^n_{k,h})$ and $\Delta\mathbf{u}_{k,h} = (\Delta u^1_{k,h}, \Delta u^2_{k,h}, \ldots, \Delta u^n_{k,h})$ 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 $\mathbf{W}_{k,h} = (W^1_{k,h}, W^2_{k,h}, \ldots, W^n_{k,h})$ and $\boldsymbol{\xi}_{k,h} = (\xi^1_{k,h}, \xi^2_{k,h}, \ldots, \xi^n_{k,h})$. Then, we have

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

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

D.1 (ϵ, δ) privacy for Gaussian mechanism (Theorem 1)

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

$$\mathbb{P}(\mathbf{W}_{k,h} = \mathbf{M} - \Delta\widetilde{\boldsymbol{\Lambda}}_{k,h} \mid \mathbf{D}_{h-1}) = \prod_{i\in N} \mathbb{P}(W^i_{k,h} = \mathbf{M}^i - \Delta\widetilde{\Lambda}^i_{k,h} \mid \mathbf{D}_{h-1}) \tag{26}$$

Now from Lemma 5, 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

$$\mathbb{P}(W^i_{k,h} = \mathbf{M}^i - \Delta\widetilde{\Lambda}^i_{k,h} \mid \mathbf{D}_{h-1}) \leq \exp\left(\frac{\epsilon}{2nH}\right) \times \mathbb{P}((W^i_{k,h})' = \mathbf{M}^i - (\Delta\widetilde{\Lambda}^i_{k,h})' \mid \mathbf{D}'_{h-1}) \tag{27}$$

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

$$\mathbb{P}(W_{k,h} = \mathbf{M} - \Delta\widetilde{\Lambda}_{k,h} \mid \mathbf{D}_{h-1}) \leq \exp\left(\frac{\epsilon}{2H}\right) \times \mathbb{P}((W_{k,h})' = \mathbf{M} - (\Delta\widetilde{\Lambda}_{k,h})' \mid \mathbf{D}'_{h-1}) \tag{28}$$

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

$$\mathbb{P}(\boldsymbol{\xi}_{k,h} = \boldsymbol{\alpha} - \Delta\widetilde{\mathbf{u}}_{k,h} \mid \mathbf{D}_{h-1}) = \prod_{i\in N} \mathbb{P}(\xi^i_{k,h} = \boldsymbol{\alpha}^i - \Delta\widetilde{u}^i_{k,h} \mid \mathbf{D}_{h-1}) \tag{29}$$

Using the property of the Gaussian distribution, we have

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

Here the first equation is due to the Gaussian mechanism. The first inequality follows from the triangle inequality. Again using the Lemma 5, with probability at least $1-\delta/(2H)$ for each agent $i \in N$, we have that

$$\mathbb{P}(\xi^i_{k,h} = \boldsymbol{\alpha}^i - \Delta\widetilde{u}^i_{k,h} \mid \mathbf{D}_{h-1}) \leq \exp\left(\frac{\epsilon}{2nH}\right) \times \mathbb{P}((\xi^i_{k,h})' = \boldsymbol{\alpha}^i - (\Delta\widetilde{u}^i_{k,h})' \mid \mathbf{D}'_{h-1}) \tag{30}$$

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

$$\mathbb{P}(\boldsymbol{\xi}_{k,h} = \boldsymbol{\alpha} - \Delta\widetilde{\mathbf{u}}_{k,h} \mid \mathbf{D}_{h-1}) \leq \exp\left(\frac{\epsilon}{2H}\right) \times \mathbb{P}((\boldsymbol{\xi}_{k,h})' = \boldsymbol{\alpha} - (\Delta\widetilde{\mathbf{u}}_h)' \mid \mathbf{D}'_{h-1}) \tag{31}$$

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

$$\log\left[\frac{\mathbb{P}(\forall\, h \in [H], (\Delta\boldsymbol{\Lambda}_{k,h}, \Delta\mathbf{u}_{k,h}) = (\mathbf{M}, \boldsymbol{\alpha}) \mid \mathbf{D}_{1:h-1})}{\mathbb{P}(\forall\, h \in [H], ((\Delta\boldsymbol{\Lambda}_{k,h})', (\Delta\mathbf{u}_{k,h})') = (\mathbf{M}, \boldsymbol{\alpha}) \mid \mathbf{D}'_{1:h-1})}\right] \leq \epsilon. \tag{32}$$

Therefore, from Theorem 12, we conclude that Algorithm 1 preserves (ϵ, δ)-LDP property with the Gaussian mechanism.

D.2 $(\epsilon, 0)$ privacy for Laplace mechanism (Theorem 3)

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

$$\mathbb{P}(\mathbf{W}_{k,h} = \mathbf{M} - \Delta\widetilde{\boldsymbol{\Lambda}}_{k,h} \mid \mathbf{D}_{h-1}) = \prod_{i\in N} \mathbb{P}(W^i_{k,h} = \mathbf{M}^i - \Delta\widetilde{\Lambda}^i_{k,h} \mid \mathbf{D}_{h-1}) \tag{33}$$

Now from Theorem 1, 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 that

$$\mathbb{P}(W^i_{k,h} = \mathbf{M}^i - \Delta\widetilde{\Lambda}^i_{k,h} \mid \mathbf{D}_{h-1}) \leq \exp\left(\frac{\epsilon}{2nH}\right) \times \mathbb{P}((W^i_{k,h})' = \mathbf{M}^i - (\Delta\widetilde{\Lambda}^i_{k,h})' \mid \mathbf{D}'_{h-1}) \tag{34}$$

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

$$\mathbb{P}(W_{k,h} = \mathbf{M} - \Delta\widetilde{\Lambda}_{k,h} \mid \mathbf{D}_{h-1}) \leq \exp\left(\frac{\epsilon}{2H}\right) \times \mathbb{P}((W_{k,h})' = \mathbf{M} - (\Delta\widetilde{\Lambda}_{k,h})' \mid \mathbf{D}'_{h-1}) \tag{35}$$

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

$$\mathbb{P}(\boldsymbol{\xi}_{k,h} = \boldsymbol{\alpha} - \Delta\widetilde{\mathbf{u}}_{k,h} \mid \mathbf{D}_{h-1}) = \prod_{i \in N} \mathbb{P}(\xi^i_{k,h} = \boldsymbol{\alpha}^i - \Delta\widetilde{u}^i_{k,h} \mid \mathbf{D}_{h-1}) \tag{36}$$

Using the property of the Laplace distribution, we have

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

In 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

$$\mathbb{P}(\xi^i_{k,h} = \boldsymbol{\alpha}^i - \Delta\widetilde{u}^i_{k,h} \mid \mathbf{D}_{h-1}) \leq \exp\left(\frac{\epsilon}{2nH}\right) \times \mathbb{P}((\xi^i_{k,h})' = \boldsymbol{\alpha}^i - (\Delta\widetilde{u}^i_{k,h})' \mid \mathbf{D}'_{h-1}) \tag{37}$$

Therefore, we have with probability 1 that

$$\mathbb{P}(\boldsymbol{\xi}_{k,h} = \boldsymbol{\alpha} - \Delta\widetilde{\mathbf{u}}_{k,h} \mid \mathbf{D}_{h-1}) \leq \exp\left(\frac{\epsilon}{2H}\right) \times \mathbb{P}((\boldsymbol{\xi}_{k,h})' = \boldsymbol{\alpha} - (\Delta\widetilde{\mathbf{u}}_h)' \mid \mathbf{D}'_{h-1}) \tag{38}$$

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

$$\log\left[\frac{\mathbb{P}(\forall\, h \in [H], (\Delta\boldsymbol{\Lambda}_{k,h}, \Delta\mathbf{u}_{k,h}) = (\mathbf{M}, \boldsymbol{\alpha}) \mid \mathbf{D}_{1:h-1})}{\mathbb{P}(\forall\, h \in [H], ((\Delta\boldsymbol{\Lambda}_{k,h})', (\Delta\mathbf{u}_{k,h})') = (\mathbf{M}, \boldsymbol{\alpha}) \mid \mathbf{D}'_{1:h-1})}\right] \leq \epsilon \tag{39}$$

Therefore, Algorithm 1 preserves $(\epsilon, 0)$-LDP property with the Laplace mechanism.

D.3 $(0, \delta)$ privacy for uniform mechanism (Theorem 5)

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

$$\mathbb{P}(W^i_{k,h} = \mathbf{M}^i - \Delta\widetilde{\Lambda}^i_{k,h} \mid \mathbf{D}_{h-1}) \leq \mathbb{P}((W^i_{k,h})' = \mathbf{M}^i - (\Delta\widetilde{\Lambda}^i_{k,h})' \mid \mathbf{D}'_{h-1}) \tag{40}$$

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

$$\mathbb{P}(\boldsymbol{\xi}_{k,h} = \boldsymbol{\alpha} - \Delta\widetilde{\mathbf{u}}_{k,h} \mid \mathbf{D}_{h-1}) = \prod_{i \in N} \mathbb{P}(\xi^i_{k,h} = \boldsymbol{\alpha}^i - \Delta\widetilde{u}^i_{k,h} \mid \mathbf{D}_{h-1}) \tag{41}$$

Using the property of the 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

$$\mathbb{P}(\xi^i_{k,h} = \boldsymbol{\alpha}^i - \Delta\widetilde{u}^i_{k,h} \mid \mathbf{D}_{h-1}) \leq \exp(0) \times \mathbb{P}((\xi^i_{k,h})' = \boldsymbol{\alpha}^i - (\Delta\widetilde{u}^i_{k,h})' \mid \mathbf{D}'_{h-1}) \tag{42}$$

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

$$\mathbb{P}(\boldsymbol{\xi}_{k,h} = \boldsymbol{\alpha} - \Delta\widetilde{\mathbf{u}}_{k,h} \mid \mathbf{D}_{h-1}) \leq \exp(0) \times \mathbb{P}((\boldsymbol{\xi}_{k,h})' = \boldsymbol{\alpha} - (\Delta\widetilde{\mathbf{u}}_h)' \mid \mathbf{D}'_{h-1}) \tag{43}$$

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

$$\log\left[\frac{\mathbb{P}(\forall\, h \in [H], (\Delta\boldsymbol{\Lambda}_{k,h}, \Delta\mathbf{u}_{k,h}) = (\mathbf{M}, \boldsymbol{\alpha}) \mid \mathbf{D}_{1:h-1})}{\mathbb{P}(\forall\, h \in [H], ((\Delta\boldsymbol{\Lambda}_{k,h})', (\Delta\mathbf{u}_{k,h})') = (\mathbf{M}, \boldsymbol{\alpha}) \mid \mathbf{D}'_{1:h-1})}\right] = 0 \tag{44}$$

Therefore, Algorithm 1 preserves $(0, \delta)$-LDP property with the uniform mechanism.

D.4 $(\epsilon, 0)$ privacy for bounded Laplace mechanism (Theorem 7)

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

E Proof of Regret Bounds

In this Section, we give the proof the upper bound on the regret of MA-LDP with different noise adding mechanisms. We first bound the regret for any noise mechanism, and then use the noise mechanism to give the explicit bound for each noise mechanism. Consider the following difference

$$\begin{aligned}
V^{\star,i}_h(\boldsymbol{s}_{k,h}) - V^{\pi_k,i}_h(\boldsymbol{s}_{k,h}) &\overset{(i)}{\leq} V^i_{k,h}(\boldsymbol{s}_{k,h}) - V^{\pi_k,i}_h(\boldsymbol{s}_{k,h}) \\
&= \max_{\boldsymbol{a}} Q^i_{k,h}(\boldsymbol{s}_{k,h}, \boldsymbol{a}) - \max_{\boldsymbol{a}} Q^{\pi_k,i}_h(\boldsymbol{s}_{k,h}, \boldsymbol{a}) \\
&\overset{(ii)}{\leq} Q^i_{k,h}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) - Q^{\pi_k,i}_h(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) \\
&\overset{(iii)}{\leq} \bar{r}_h(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}; \boldsymbol{w}^i_{k,h}) + \left\langle \hat{\boldsymbol{\theta}}^i_{k,h}, \phi_{V^i_{k,h+1}}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) \right\rangle \\
&\quad + \beta_k ||\Sigma^{i^{-1/2}}_{k,h} \phi_{V^i_{k,h+1}}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})||_2 \\
&\quad - \bar{r}_h(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}; \boldsymbol{w}^i_{k,h}) - \mathbb{P}_h V^{\pi_k,i}_{h+1}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) \\
&= \left\langle \hat{\boldsymbol{\theta}}^i_{k,h}, \phi_{V^i_{k,h+1}}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) \right\rangle + \beta_k ||\Sigma^{i^{-1/2}}_{k,h} \phi_{V^i_{k,h+1}}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})||_2 \\
&\quad - \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}) - \mathbb{P}_h V^i_{k,h+1}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) \\
&= \left\langle \hat{\boldsymbol{\theta}}^i_{k,h} - \boldsymbol{\theta}^\star_h, \phi_{V^i_{k,h+1}}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) \right\rangle + \beta_k ||\Sigma^{i^{-1/2}}_{k,h} \phi_{V^i_{k,h+1}}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})||_2 \\
&\quad - \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}) \\
&\overset{(iv)}{\leq} ||\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}_{k,h}, \boldsymbol{a}_{k,h})||_2 - \mathbb{P}_h V^{\pi_k,i}_{h+1}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) \\
&\quad + \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^i_{k,h+1}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) \\
&\overset{(v)}{\leq} 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} \tag{45, 46}$$

(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 state-action value function from line 10 of the MA-DP algorithm. In (iv) we use the Cauchy-Schwartz inequality. Finally (v) follows from the Lemma 3. Apart from the above, we also have the following

$$V^i_{h+1}(\boldsymbol{s}_{k,h}) - V^{\pi_k,i}_{h+1}(\boldsymbol{s}_{k,h}) \leq V^i_{h+1}(\boldsymbol{s}_{k,h}) \leq H \tag{47}$$

Combining the Equations (46) and (47) we have the following:

$$V^i_{h+1}(\boldsymbol{s}_{k,h}) - V^{\pi_k,i}_{h+1}(\boldsymbol{s}_{k,h})$$

$$\leq \min\{H, 2\beta_k ||\Sigma_{k,h}^{i^{-1/2}} \phi_{V_{k,h+1}^i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})||_2 - \mathbb{P}_h V_{h+1}^{\pi_k,i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) + \mathbb{P}_h V_{k,h+1}^i(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})\}$$
$$\leq \min\{H, 2\beta_k ||\Sigma_{k,h}^{i^{-1/2}} \phi_{V_{k,h+1}^i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})||_2\} - \mathbb{P}_h V_{h+1}^{\pi_k,i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) + \mathbb{P}_h V_{k,h+1}^i(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}),$$

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

$$\begin{aligned} &V_{h+1}^i(\boldsymbol{s}_{k,h}) - V_{h+1}^{\pi_k,i}(\boldsymbol{s}_{k,h}) + [V_{h+1}^{\pi_k,i}(\boldsymbol{s}_{k,h+1}) - V_{k,h+1}^i(\boldsymbol{s}_{k,h+1})] \\ &\leq \min\{H, 2\beta_k ||\Sigma_{k,h}^{i^{-1/2}} \phi_{V_{k,h+1}^i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})||_2\} - \mathbb{P}_h V_{h+1}^{\pi_k,i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) + \mathbb{P}_h V_{k,h+1}^i(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) \\ &\quad + [V_{h+1}^{\pi_k,i}(\boldsymbol{s}_{k,h+1}) - V_{k,h+1}^i(\boldsymbol{s}_{k,h+1})] \end{aligned} \tag{48}$$

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

$$\begin{aligned} \sum_{k=1}^{K} [V_{k,h'}^i(\boldsymbol{s}_{k,h'}) - V_{h'}^{\pi_k,i}(\boldsymbol{s}_{k,h'})] \leq 2 \sum_{k=1}^{K} \sum_{h=h'}^{H} \beta_k \min\{1, ||\Sigma_{k,h}^{i^{-1/2}} \phi_{V_{k,h+1}^i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})||_2\} \\ + \sum_{k=1}^{K} \sum_{h=h'}^{H} [[\mathbb{P}_h(V_{k,h+1}^i - V_{h+1}^{\pi_k,i})](\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) - [V_{k,h+1}^i - V_{h+1}^{\pi_k,i}](\boldsymbol{s}_{k,h+1})] \end{aligned} \tag{49}$$

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

$$\mathcal{E}_4 = \Bigg\{ \forall\, h' \in [H],\ \sum_{k=1}^{K} \sum_{h=h'}^{H} [[\mathbb{P}_h(V_{k,h+1}^i - V_{h+1}^{\pi_k,i})](\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) - [V_{k,h+1}^i - V_{h+1}^{\pi_k,i}](\boldsymbol{s}_{k,h+1})] \\ \leq 4H\sqrt{2T\log(2H/\alpha)} \Bigg\} \tag{50}$$

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

$$[[\mathbb{P}_h(V_{k,h+1}^i - V_{h+1}^{\pi_k,i})](\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h}) - [V_{k,h+1}^i - V_{h+1}^{\pi_k,i}](\boldsymbol{s}_{k,h+1})] \leq 4H \tag{51}$$

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

$$\begin{aligned} \sum_{k=1}^{K} \sum_{h=1}^{H} \beta_k \min\{1, ||\Sigma_{k,h}^{i^{-1/2}} \phi_{V_{k,h+1}^i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})||_2\} \leq \beta_K \sum_{k=1}^{K} \sum_{h=1}^{H} \min\{1, ||\Sigma_{k,h}^{i^{-1/2}} \phi_{V_{k,h+1}^i}(\boldsymbol{s}_{k,h}, \boldsymbol{a}_{k,h})||_2\} \\ \leq H\beta_K \sqrt{2ndK\log(1 + K/\lambda)} \end{aligned} \tag{52}$$

The above inequalities hold because of the Cauchy-Schwartz and the Theorem 7. 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:

$$R_K \leq H\beta_K \sqrt{2ndK\log(1 + K/\lambda)} + 4H\sqrt{2T\log(2H/\alpha)}. \tag{53}$$

E.1 Regret bound for Gaussian mechanism (Theorem 2)

To complete the regret bound for MA-LDP algorithm with the Gaussian mechanism, we substitute β_K^G given in Equation (20) in the regret expression given in Equation (53). Recall,

$$\beta_K^G = c_g (nd)^{3/4} H^{3/2} K^{1/4} \log(ndT/\alpha)(\log(H/\delta))^{1/4} \sqrt{1/\epsilon}. \tag{54}$$

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

$$\begin{aligned} R_K^G &\leq c_g H (nd)^{3/4} H^{3/2} K^{1/4} \log(ndT/\alpha)(\log(H/\delta))^{1/4} \sqrt{1/\epsilon} \\ &\quad \cdot \sqrt{2ndK\log(1 + K/\lambda)} + 4H\sqrt{2T\log(2H/\alpha)} \\ &\leq \widetilde{\mathcal{O}}(n^{5/4} d^{5/4} H^{7/4} T^{3/4} \log(ndT/\alpha)(\log(H/\delta))^{1/4} \sqrt{1/\epsilon}) \end{aligned} \tag{55}$$

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

E.2 Regret bound for Laplace mechanism (Theorem 4)

To complete the regret bound for MA-LDP algorithm with the Laplace mechanism, we substitute β_K^L given in Equation (21) in the regret expression given in Equation (53). Recall,

$$\beta_K^L = c_l(nd)^{3/4}H^{3/2}K^{1/4}\log(ndT/\alpha)\sqrt{1/\epsilon}. \tag{56}$$

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

$$\begin{aligned} R_K^L &\leq c_l H(nd)^{3/4}H^{3/2}K^{1/4}\log(ndT/\alpha)\sqrt{1/\epsilon} \\ &\quad \cdot \sqrt{2ndK\log(1+K/\lambda)} + 4H\sqrt{2T\log(2H/\alpha)} \\ &\leq \widetilde{\mathcal{O}}(n^{5/4}d^{5/4}H^{7/4}T^{3/4}\log(ndT/\alpha)\sqrt{1/\epsilon}) \end{aligned} \tag{57}$$

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

E.3 Regret bound for uniform mechanism (Theorem 6)

To complete the regret bound for MA-LDP algorithm with the uniform mechanism, we substitute β_K^U given in Equation (22) in the regret expression given in Equation (53). Recall,

$$\beta_K^U = c_u(nd)^{3/4}H^{3/2}k^{1/4}\log(ndT/\alpha)(\log(H/\delta))^{1/4}. \tag{58}$$

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

$$\begin{aligned} R_K^L &\leq c_u H(nd)^{3/4}H^{3/2}k^{1/4}\log(ndT/\alpha)(\log(H/\delta))^{1/4} \\ &\quad \cdot \sqrt{2ndK\log(1+K/\lambda)} + 4H\sqrt{2T\log(2H/\alpha)} \\ &\leq \widetilde{\mathcal{O}}(n^{5/4}d^{5/4}H^{7/4}T^{3/4}\log(ndT/\alpha)(\log(1/\delta))^{1/4}) \end{aligned} \tag{59}$$

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

E.4 Regret bound for bounded Laplace mechanism (Theorem 8)

To complete the regret bound for MA-LDP algorithm with the Bounde Laplace mechanism, we substitute β_K^{BL} given in Equation (23) in the regret expression given in Equation (53). Recall,

$$\beta_K^{BL} = c_{bl}(nd)^{3/4}\zeta^{1/4}K^{1/4}\log(ndT/\alpha). \tag{60}$$

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 MA-LDP algorithm with Bounded Laplace noise adding mechanism is given by

$$\begin{aligned} R_K^{BL} &\leq c_{bl} H(nd)^{3/4}\zeta^{1/4}K^{1/4}\log(ndT/\alpha) \\ &\quad \cdot \sqrt{2ndK\log(1+K/\lambda)} + 4H\sqrt{2T\log(2H/\alpha)} \\ &\leq \widetilde{\mathcal{O}}(n^{5/4}d^{5/4}H^{1/4}T^{3/4}\zeta^{1/4}\log(ndT/\alpha)) \end{aligned} \tag{61}$$

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.

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

$$\phi(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) = \begin{cases} (\phi(s'^1|s^1,\boldsymbol{a}^1),\ldots,\phi(s'^n|s^n,\boldsymbol{a}^n)), & if\ \boldsymbol{s}\neq\boldsymbol{g},\ \boldsymbol{s}'\in S(\boldsymbol{s}) \\ \mathbf{0}_{nd}, & if\ \boldsymbol{s}\neq\boldsymbol{g},\ \ \boldsymbol{s}'\notin S(\boldsymbol{s}) \\ \mathbf{0}_{nd}, & if\ \boldsymbol{s}=\boldsymbol{g},\ \boldsymbol{s}'\neq\boldsymbol{g} \\ (\mathbf{0}_{nd-1},\alpha(\boldsymbol{s})), & if\ \boldsymbol{s}=\boldsymbol{g},\ \boldsymbol{s}'=\boldsymbol{g}, \end{cases} \tag{62}$$

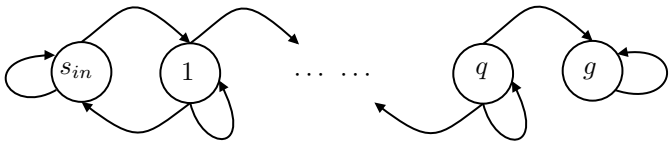

Figure 2: Network with $q+2$ nodes.

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

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

Here $\delta_{j,j} \geq \delta_{j,j+1}$, and $\mathbf{0}_d^\top = (0,0,\ldots,0)^\top$ 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 $\Delta < \delta$.

F.1 Proof of Lemma 4

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 $\boldsymbol{s} = (\underbrace{s_{init}, \ldots, s_{init}}_{x_0\ times}, \underbrace{1,1,\ldots,1}_{x_1\ times}, \underbrace{2,2,\ldots,2}_{x_2\ times}, \underbrace{q-1,q-1,\ldots,q-1}_{x_{q-1}\ times}, \underbrace{q,q,\ldots,q}_{x_q\ times} \underbrace{g,g,\ldots,g}_{(n-x_0-x_1-\ldots,-x_q)\ times})$, 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 $n - x_0 - x_1 - \cdots - x_q$ agents are at g. Consider an agent i, who is at s_{init} node. Let $|S(\boldsymbol{s})|$ denotes the number of next states feasible from state $\boldsymbol{s}$. A simple calculation shows that $|S(\boldsymbol{s})| = 2^{x_0} \times 3^{x_1} \times 3^{x_2} \cdots \times \ldots 3^{x_q} \times 1^{n-x_0-x_1-\ldots,-x_q}$. 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 features defined in Equation (62). 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

$$\sum_{i=1}^{x_0}\left\{\left(-\langle \boldsymbol{a}^i, \boldsymbol{\theta}^i\rangle + \frac{1-\delta}{n} \times \frac{1}{\alpha(\boldsymbol{s})}\right) \times \frac{|S(\boldsymbol{s})|}{2}\right\} + \sum_{i=1}^{x_0}\left\{\left(\langle \boldsymbol{a}^i, \boldsymbol{\theta}^i\rangle + \frac{\delta}{n} \times \frac{1}{\alpha(\boldsymbol{s})}\right) \times \frac{|S(\boldsymbol{s})|}{2}\right\}$$
$$= \frac{|S(\boldsymbol{s})|}{2} \times \frac{1}{n} \times \frac{x_0}{\alpha(\boldsymbol{s})}. \tag{63}$$

Next consider an agent i who is at node $j \in \{1,2,\ldots,q\}$. Out of 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 who are at node j in the state $\boldsymbol{s}$.

Therefore, the contribution in the overall probability from this agent is

$$\sum_{i=1}^{x_j}\left\{\left(-\langle \boldsymbol{a}^i,\boldsymbol{\theta}^i\rangle+\frac{1-\delta_{j,j}}{n}\times\frac{1}{\alpha(\boldsymbol{s})}\right)\times\frac{|S(\boldsymbol{s})|}{3}\right\}+\sum_{i=1}^{x_j}\left\{\left(\langle \boldsymbol{a}^i,\boldsymbol{\theta}^i\rangle+\frac{\delta_{j,j+1}}{n}\times\frac{1}{\alpha(\boldsymbol{s})}\right)\times\frac{|S(\boldsymbol{s})|}{3}\right\}$$
$$+\sum_{i=1}^{x_j}\left(\frac{\delta_{j,j}-\delta_{j,j+1}}{n}\times\frac{1}{\alpha(\boldsymbol{s})}\right)\times\frac{|S(\boldsymbol{s})|}{3}=\frac{|S(\boldsymbol{s})|}{3}\times\frac{1}{n}\times\frac{x_j}{\alpha(\boldsymbol{s})}. \tag{64}$$

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 $|S(\boldsymbol{s})|$. Let $x_{q+1}=n-x_0-x_1-\cdots-x_q$. The probability of this is $\frac{1}{n}\frac{x_{q+1}}{\alpha(\boldsymbol{s})}$, so Therefore, the contribution in the probability from the agent who is at node g is

$$\sum_{i=1}^{x_{q+1}}|S(\boldsymbol{s})|\times\frac{1}{n}\times\frac{1}{\alpha(\boldsymbol{s})}=|S(\boldsymbol{s})|\times\frac{1}{n}\times\frac{x_{q+1}}{\alpha(\boldsymbol{s})} \tag{65}$$

Adding Equations (63), (64) and (65), we have

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

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

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

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 in Eq. (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:

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

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

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

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 mechanism

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 12 (Theorem A.2 Liao et al. (2021)). *If the privacy loss c satisfy $\mathbb{P}_{o \sim \mathcal{M}(d)}[c(o; \mathcal{M}, \boldsymbol{aux}, d, d') > \epsilon] \leq \delta$ for all auxiliary input $\boldsymbol{aux}$ 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 $\boldsymbol{aux}$ 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.

Lemma 5 (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).

Lemma 6 (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 7 (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*

$$\log \left[\frac{det(\mathbf{\Lambda}_t)}{det(\mathbf{\Lambda}_0)} \right] \leq \sum_{j=1}^{t} \phi_j^\top \mathbf{\Lambda}_{j-1}^{-1} \phi_j \leq 2 \log \left[\frac{det(\mathbf{\Lambda}_t)}{det(\mathbf{\Lambda}_0)} \right] \tag{66}$$

Theorem 13 (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*

$$|\eta_t| \leq R,\ \mathbb{E}[\eta_t^2 | \mathcal{G}_t] \leq \sigma^2,\ ||x_t||_2 \leq L \tag{67}$$

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

$$\forall\, t \geq 0,\ ||\sum_{i=1}^{t} x_i \eta_i||_{Z_t^{-1}} \leq \beta_t,\ ||\mu_t - \mu^\star||_{Z_t} \leq \beta_t + \sqrt{\lambda}||\mu^\star||_2, \tag{68}$$

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

$$\beta_t = 8\sigma \sqrt{d \log(1 + tL^2/(d\lambda)) \log(4t^2/\delta)} + 4R \log(4t^2/\delta) \tag{69}$$

Lemma 8 (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*

$$x_{t+1} = \Gamma\{x_t + \gamma_t[h(x_t) + \zeta_t + \beta_t]\}. \tag{70}$$

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 (70) 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 (70) converges almost surely to $\mathcal{K}^\star$ as $t \to \infty$.

G.4 Generating a bounded Laplace distribution, $\mathcal{BL}$ (Ross, 2022)

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

$$\begin{aligned}
F_{\mathcal{BL}}(x) &= \int_{-B}^{x} \frac{1}{2b\left(1-\exp(\frac{-B}{b})\right)}\exp\left(\frac{-|x|}{b}\right) && (71)\\
&= \int_{-B}^{0} \frac{1}{2b\left(1-\exp(\frac{-B}{b})\right)}\exp\left(\frac{x}{b}\right) + \int_{0}^{x} \frac{1}{2b\left(1-\exp(\frac{-B}{b})\right)}\exp\left(\frac{-x}{b}\right) && (72)\\
&= \frac{1}{2b\left(1-\exp(\frac{-B}{b})\right)}[b - b\exp(\frac{-B}{b})] - \frac{1}{2b\left(1-\exp(\frac{-B}{b})\right)}[b\exp(\frac{-x}{b}) - b] && (73)\\
&= \frac{2-\exp(\frac{-B}{b}) - \exp(\frac{-x}{b})}{2\left(1-\exp(\frac{-B}{b})\right)} && (74)
\end{aligned}$$

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

$$\begin{aligned}
u &= \frac{2-\exp(\frac{-B}{b}) - \exp(\frac{-x}{b})}{2\left(1-\exp(\frac{-B}{b})\right)} && (75)\\
\exp(\frac{-x}{b}) &= 2 - \exp(\frac{-B}{b}) - 2u\left(1-\exp(\frac{-B}{b})\right) && (76)\\
\frac{-x}{b} &= \log\left(2 - \exp(\frac{-B}{b}) - 2u\left(1-\exp(\frac{-B}{b})\right)\right) && (77)\\
x &= -b\log\left(2 - \exp(\frac{-B}{b}) - 2u\left(1-\exp(\frac{-B}{b})\right)\right) && (78)
\end{aligned}$$

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