Local Differentially Private Regret Minimization in Reinforcement Learning

Evrard Garcelon
Facebook AI Research

Vianney Perchet
Crest, ENSAE

Ciara Pike-Burke
Imperial College London

Matteo Pirotta
Facebook AI Research

Abstract

Reinforcement learning algorithms are widely used in domains where it is desirable to provide a personalized service. In these domains it is common that user data contains sensitive information that needs to be protected from third parties. Motivated by this, we study privacy in the context of finite-horizon Markov Decision Processes (MDPs) by requiring information to be obfuscated on the user side. We formulate this notion of privacy for RL by leveraging the local differential privacy (LDP) framework. We present an optimistic algorithm that simultaneously satisfies LDP requirements, and achieves sublinear regret. We also establish a lower bound for regret minimization in finite-horizon MDPs with LDP guarantees. These results show that while LDP is appealing in practical applications, the setting is inherently more complex. In particular, our results demonstrate that the cost of privacy is multiplicative when compared to non-private settings.

1 Introduction

Reinforcement learning (RL) is a fundamental sequential decision-making problem for learning in uncertain environments. In each round, the learner observes the state of the environment, and then selects an action. This action will provide them with a reward and take them to a new state via some unknown dynamics. The learner receives feedback in the form of trajectories (i.e., sequences of states, actions and rewards) which encode information about the unknown dynamics and objectives, and allow the agent to learn about the environment. The agent's aim is to select actions to maximize their total reward. To succeed in this problem, an agent needs to trade-off exploration to gather information about the environment (reward and dynamics) and exploitation of available information to maximize the cumulative reward. In this paper, we consider finite-horizon RL problems (Puterman, 1994, Chp. 4) with S states and A actions, where the agent interacts with the environment in a sequence of K episodes of length H ($T = KH$). The agent's performance is measured by the regret, the difference between the expected total cost of the optimal policy and agent's policy. Many algorithms have been developed for regret minimization in this setting (e.g., Jaksch et al., 2010; Azar et al., 2017).

Reinforcement learning algorithms have become ubiquitous in many settings such as digital marketing, healthcare and finance, where it is desirable to provide a personalized service. Nowadays most users understand that it is necessary to provide a certain amount of personal information to receive a service tailored to their specific needs. At the same time, there is increasing concern about protecting users' privacy and personal data. In particular, in many of the aforementioned domains, the data obtained by the RL algorithm are highly sensitive. For example, in healthcare, the state encodes personal information such as gender, age, vital signs etc. We assume that the user wants to prevent this information from being discovered by external parties or malicious agents. Unfortunately, Pan et al. (2019) have shown that, unless sufficient precautions are taken, the RL agent may leak information about the environment, putting user privacy at jeopardy.

Differential privacy (DP) (Dwork et al., 2006) is a standard mechanism for preserving data privacy, widely studied for supervised learning. Differential privacy has also been studied in a simpler version of the RL problem known as multi-armed bandits (e.g., Mishra and Thakurta, 2015; Tossou and Dimitrakakis, 2016). However, Shariff and Sheffet (2018) showed that the standard definition of DP is incompatible with regret minimization in the contextual bandit problem. Weaker or different notions of privacy have thus been considered (e.g., Shariff and Sheffet, 2018; Boursier and Perchet, 2020). Recently, Vietri et al. (2020) transferred some of these techniques to RL presenting the first private algorithm for regret minimization in finite-horizon problems. They considered a relaxed definition of DP called *joint differential privacy* (JDP) and proposed an ϵ-JDP algorithm based on randomized response which achieves a regret of $\widetilde{O}(\sqrt{H^4SAT} + SAH^3(S+H)/\varepsilon)$. This shows that the cost of JDP privacy is an additive term of $\widetilde{O}(SAH^3(S+H)/\varepsilon)$. In the JDP setting, the privacy burden lies with the learning algorithm which directly observes user trajectories containing sensitive data. In particular, this means that the data itself is not private and could potentially be used by the owner of the application to train other algorithms which do not necessarily guarantee privacy. An alternative definition of privacy is *Local Differential Privacy* (LDP) (Duchi et al., 2013). This requires that the user's data is protected at collection time before the learning agent has access to it. Intuitively, LDP ensures that each collected trajectory is DP when observed by the learning agent while DP requires computation on the entire set of trajectories to be DP. LDP is in general a stronger definition of privacy[1]. It is also simpler to understand and more user friendly. These characteristics make LDP more suited for real-world applications.

In this paper, we study LDP for regret minimization in finite horizon reinforcement learning. We first provide a lower bound of $\widetilde{\Omega}\big(\sqrt{HSAT}/\min\{\exp(\varepsilon)-1,1\}\big)$. This shows that LDP is inherently harder than JPD in RL, where the lower-bound is $\widetilde{\Omega}\big(\sqrt{HSAT} + SAH\log(T)/\varepsilon\big)$ (Vietri et al., 2020). Then, we propose the first LDP algorithm for regret minimization in tabular finite-horizon problems. Inspired by (Ren et al., 2020), we use a privacy-preserving mechanism (e.g. Laplace mechanism, see Dwork et al. (2006)) to perturb the information associated to each trajectory and derive LDP-OBI, an ε-LDP algorithm with a sublinear regret bound. This algorithm is compatible with several different privacy-preserving mechanisms. The main challenge in this setting is to find a level of noise in the mechanism that guarantees privacy but also allows the algorithm to learn about the environment. For the Laplace mechanism, we show a regret bound of $\widetilde{O}\big(\max\big\{H^{3/2}S^2A\sqrt{T}/\varepsilon, HS\sqrt{AT}\big\}\big)$. This matches the lower bound up to a $H^2S^{3/2}\sqrt{A}$ factor when $\varepsilon \to 0$. We also perform preliminary numerical simulations to evaluate the impact of LDP on the learning process. For comparison, we derive LDP-PSRL, a locally private version of posterior sampling (Osband et al., 2013).

1.1 Related Work

The notion of DP was introduced in (Dwork et al., 2006) and is now a standard in machine learning (e.g., Abowd, 2018; Erlingsson et al., 2014; Dwork and Roth, 2014). In stochastic multi-armed bandits, ϵ-DP algorithms have been extensively studied (see e.g., Mishra and Thakurta, 2015; Tossou and Dimitrakakis, 2016). Recently, Sajed and Sheffet (2019) proposed an ϵ-DP algorithm for stochastic MABs that achieves the private lower-bound presented in (Shariff and Sheffet, 2018). In contextual bandits, Shariff and Sheffet (2018) derived an impossibility result for learning under DP by showing a regret lower-bound $\Omega(T)$ for any (ϵ, δ)-DP algorithm. As a consequence, they considered the relaxed JDP setting and proposed an optimistic algorithm with sublinear regret and ϵ-JDP guarantees. Recently, *local differential privacy* (Duchi et al., 2013) has attracted increasing interest in the bandit literature. Gajane et al. (2018) were the first to study LDP in stochastic MABs. They proposed an optimistic and Bayesian algorithm with sublinear regret. Chen et al. (2020) extended LDP to combinatorial bandits, and Zheng et al. (2020) focused on LDP for MAB and contextual bandit. Private algorithms for regret minimization have also been investigated in multi-agent bandits (a.k.a. federated learning) both in centralized and decentralized settings (see e.g., Tossou and Dimitrakakis, 2015; Dubey and Pentland, 2020b,a). Empirically, private approaches in centralized bandits have been investigated in (Malekzadeh et al., 2020; Hannun et al., 2019).

In RL, Balle et al. (2016) proposed the first private algorithm for policy evaluation with linear function approximation. Wang and Hegde (2019) considered the RL problem in continuous space, where reward information is protected. They designed a private version of Q-learning with function approximation where privacy is achieved by injecting noise in the value function. In both cases, those works considered a standard definition of ε-DP, but did not focus on regret minimization. The first paper ensures privacy with respect to the change of trajectories collected off-policy and the second paper ensures privacy with respect to different reward functions. Ono and Takahashi (2020) recently studied LDP for actor-critic methods in the context of distributed RL. Regret minimization with privacy guarantees has only been considered in RL recently. Vietri et al. (2020) designed a private optimistic algorithm for regret minimization with JDP guarantees. They proposed a variation of UBEV (Dann et al., 2017) using a binary mechanism with parameter ϵ/H. Their algorithm PUCB achieves a regret bound

[1] LDP and JDP are not directly comparable in contextual bandits and RL. As discussed in (Zheng et al., 2020), LDP seems a more appropriate privacy definition for contextual bandit and we believe this is the case in RL too.

$\widetilde{O}(\sqrt{H^4SAT} + SAH^3(S+H)/\varepsilon)$ while enjoying ε-JDP guarantees. Compared to the worst case regret of UBEV, the penalty for JDP privacy is only additive as shown by their lower-bound of $\widetilde{\Omega}\big(\sqrt{HSAT} + SAH/\varepsilon\big)$.

2 Preliminaries

In this section, we recall the basics of finite-horizon Markov Decision Processes (MDPs) and introduce the definition of local differential privacy for MDPs.

2.1 Finite-Horizon MDPs

We consider a finite-horizon Markov Decision Process (MDP) (Puterman, 1994, Chp. 4) $M = (\mathcal{S}, \mathcal{A}, p, r, H)$ with state space $\mathcal{S}$, action space $\mathcal{A}$ and horizon $H \in \mathbb{N}$. Every state-action pair is characterized by a reward distribution with mean $r(s,a)$ supported in $[0,1]$ and a transition distribution $p(\cdot|s,a)$ over next state.[2] We denote by $S = |\mathcal{S}|$ and $A = |\mathcal{A}|$ the number of states and actions. We define a non-stationary policy as a collection $\pi = (\pi_1, \dots, \pi_H)$ of Markovian and deterministic policies $\pi_t : \mathcal{S} \to \mathcal{A}$. For any $t \in [H] := \{1, \dots, H\}$ and state $s \in \mathcal{S}$, the value function of a non-stationary policy π is defined as $V_t^\pi(s) = \mathbb{E}\left[\sum_{l=t}^H r_l(s_l, a_l) \mid s_t = s, a_t = \pi_t(s_t)\right]$. There exists an optimal Markovian and deterministic policy $\pi^\star$ (Puterman, 1994, Sec. 4.4) for which $V_t^\star(s) = V_t^{\pi^\star}(s) = \max_{\pi \in \Pi^{\mathrm{MD}}} V_t^\pi(s)$ with Π^{MD} the space of deterministic Markovian policies. Both V^π and $V^\star$ satisfy the Bellman equations:

$$V_t^\pi(s) = r(s, \pi_t(s)) + p(\cdot|s, \pi_t(s))^\mathsf{T} V_{t+1}^\pi := L_t^\pi V_{t+1}^\pi(s)$$
$$V_t^\star(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + p(\cdot|s,a)^\mathsf{T} V_{t+1}^\star \right\} := L_t^\star V_{t+1}^\star$$

where $V_{H+1}^\star(s) = 0$ for any state $s \in \mathcal{S}$. Note that by boundness of the reward, all value functions V_t^π are bounded in $[0, H-t+1]$ for any t and s.

The learning problem. The learning agent interacts with the MDP in a sequence of episodes $k \in [K]$ of fixed length H by playing a non-stationary policy $\pi_k = (\pi_{1,k}, \dots, \pi_{H,k})$. In each episode, the initial state $s_{1,k}$ is randomly selected by some user u_k. We assume that the learning agent knows $\mathcal{S}$, $\mathcal{A}$ and the support of the reward distribution, while the reward and dynamics are *unknown* and need to be estimated online. We evaluate the performance of a learning algorithm $\mathfrak{A}$ which plays policies $\pi_1, \dots, \pi_K$ by its cumulative regret after K episodes

$$R(\mathfrak{A}, K) = \sum_{k=1}^K (V_1^\star(s_{k,1}) - V_1^{\pi_k}(s_{k,1})). \tag{1}$$

2.2 Local Differential Privacy

When modeling a decision problem as a finite horizon MDP, it is natural to view each episode $k \in [K]$ as a trajectory associated to a specific user. In this paper, we assume that the sensitive information is contained in the states and rewards of the trajectory, and we wish to take actions that keep these quantities private. This is reasonable in many settings such as healthcare, where the trajectory of each episode corresponds to the evolution of the health status (state) of a patient in response to the actions decided by the doctor, and the reward is some measure of the patient's health. We want to make sure that no sensitive health data is leaked by the executed actions. This poses a fundamental challenge since in many cases, the information about the actions taken in each state is essential for learning and creating a personalized experience for the user. The goal of a private RL algorithm is thus to ensure that the sensitive information remains private, while preserving the learnability of the problem.

Privacy in RL has been tackled in (Vietri et al., 2020) through the lens of *joint differential privacy* (JDP). Intuitively, JDP requires that when changing an user, the information seen by the other $K-1$ users will not change too much (Vietri et al., 2020). The privacy burden lies within the RL algorithm. The algorithm has access to all the information about the user (i.e., trajectories) containing sensitive data. The algorithm has to provide guarantees about the security of the data and carefully select the policies to execute in order to guarantee JDP. This approach to privacy requires the user to trust the RL algorithm to securely handle the data and not to expose or share sensitive information.

In contrast to prior work, in this paper, we consider *local differential privacy* (LDP) in RL. This removes the requirement that the RL algorithm observes the true sensitive data. At high level, LDP considers that the user interacts locally with the environment and only the outcome of this interaction –i.e., trajectory– could be observed by a third party or a malicious agent. This is different than JDP or DP where the sequences of actions from all the users can be observed and thus should be privatized. Thus, LDP requires that an algorithm has access to user information only through samples that have been secured before being stored. Information is secured locally by the user using a private randomizer $\mathcal{M}$, before being sent to the RL agent. The appeal of this local model is that *all privacy computations are done locally on the user-side.* Because

[2] We can simply modify the algorithm to handle step dependent transitions and rewards. The regret is then multiplied by a factor $H\sqrt{H}$.

Algorithm 1: Locally Private Episodic RL

```
Input: Agent: $\mathfrak{A}$, Local Randomizer: $\mathcal{M}$, Users: $u_1, \cdots, u_K$
for $k = 1, \ldots, K$ do
    Agent $\mathfrak{A}$ computes policy $\pi_k$ using $\{\mathcal{M}(X_{u_k}) \mid h \in [K-1]\}$
    User $u_k$ receives policy $\pi_k$ from agent $\mathfrak{A}$ and observes $s_{1,k} \sim \rho_{0,u_k}$
    User $u_k$ executes policy $\pi_k$ and observes a trajectory $X_{u_k} = \{(s_{h,k}, a_{h,k}, r_{h,k}) \mid h \in [H]\}$
    User $u_k$ sends privatized information $\mathcal{M}(X_{u_k})$ to agent $\mathfrak{A}$
```

nobody other than the owner has ever access to any piece of non private data, this local setting is far more secure. In particular, it does not require a trusted party, and there is no central agent who might be subject to an intrusion.

We now formally define local differential privacy in finite horizon reinforcement learning. Following the definition in (Vietri et al., 2020), we say a user u is characterized by a starting state distribution $\rho_{0,u}$ (i.e., for user u, $s_1 \sim \rho_{0,u}$) and a tree of depth H, which describes the state and rewards corresponding to all possible sequences of actions. Alg. 1 describes the LDP private interaction protocol between K unique users $\{u_1, \ldots, u_K\} \subset \mathcal{U}^K$,with $\mathcal{U}$ the set of all users, and an RL algorithm $\mathfrak{A}$. For any $k \in [K]$, let $s_{1,k} \sim \rho_{0,u_k}$ be the initial state for user u_k and denote by $X_{u_k} = \{(s_{k,h}, a_{k,h}, r_{k,h}) \mid h \in [H]\} \in \mathcal{X}_{u_k}$ the trajectory observed by user u_k. We write $\mathcal{M}(X_{u_k})$ to denote the privatized data generated by the randomizer for user u_k. The goal of mechanism $\mathcal{M}$ is to secure sensitive information while encoding sufficient information for learning. With these notions in mind, we state the formal definition of LDP for RL.

Definition 1. *For any $\varepsilon \geq 0$ and $\delta \geq 0$, a privacy preserving mechanism $\mathcal{M}$ is said to be (ε, δ)-Locally Differential Private (LDP) if and only if for all users $u, u' \in \mathcal{U}$ and trajectories $(X_u, X_{u'}) \in \mathcal{X}_u \times \mathcal{X}_{u'}$:*

$$\mathbb{P}\left(\mathcal{M}(X_u) \in S\right) \leq e^{\varepsilon}\, \mathbb{P}\left(\mathcal{M}(X_{u'}) \in S\right) + \delta \qquad (2)$$

where $\mathcal{X}_u$ is the space of trajectories associated to user u.

Note that Definition 1 is coherent with the definition given in the contextual bandits (see Zheng et al., 2020; Chen et al., 2020), where a user is simply identified by its context.

3 Optimism with Local Privacy

Our primary goal in this work is to provide an algorithm for finite-horizon MDPs which has bounded regret and satisfies local differential privacy. In order to do this, we combine "optimism" (e.g., Jaksch et al., 2010; Azar et al., 2017; Zanette and Brunskill, 2019) with a privacy mechanism. A key challenge here is to ensure that the privacy mechanism does not prevent the RL algorithm from learning. In this work, we build on several well-known privacy mechanisms like the Laplace mechanism or the Gaussian mechanism (e.g., Dwork and Roth, 2014) to generate private counters that allow us to construct "reasonable" estimates of the unknown rewards and transitions. These private estimates are then used to define confidence intervals from which optimistic policies for exploration can be derived. We call the resulting algorithm LDP-OBI. In the following, we provide further details of LDP-OBI, and prove that is an LDP algorithm for regret minimization in finite-horizon MDPs.

3.1 Privacy-Preserving Mechanism

At the end of each episode $k \in [K]$, the user u_k uses a private randomizer $\mathcal{M}$ to generate a private statistic $\mathcal{M}(X_{u_k})$ that is sent to the RL algorithm $\mathfrak{A}$. This statistic should encode sufficient information for the RL algorithm to improve the policy while maintaining the user's privacy. For a given trajectory X, let $R_X(s,a) = \sum_{h=1}^{H} r_h \mathbb{1}_{\{s_h=s, a_h=a\}}$, $N_X^r(s,a) = \sum_{h=1}^{H} \mathbb{1}_{\{s_h=s,a_h=a\}}$ and $N_X^p(s,a,s') = \sum_{h=1}^{H-1} \mathbb{1}_{\{s_h=s,a_h=a,s_{h+1}=s'\}}$ be the true non-private statistics (which the agent will never observe). We design the mechanism $\mathcal{M}$ so that for a given trajectory $X = \{(s_h, a_h, r_h) \mid h \leq H\} \in (\mathcal{S}\times\mathcal{A}\times\mathbb{R})^H$, $\mathcal{M}$ returns private versions of these statistics. That is, the output $\mathcal{M}(X) = (\widetilde{R}_X, \widetilde{N}_X^r, \widetilde{N}_X^p) \in \mathbb{R}^{S\times A} \times \mathbb{R}^{S\times A} \times \mathbb{R}^{S\times A\times S}$ is a perturbed aggregate statistic. Here, $\widetilde{R}_X(s,a)$ is a noisy version of the cumulative reward $R_X(s,a)$ in trajectory X, $\widetilde{N}_X^r$ and $\widetilde{N}_X^p$ are perturbed counters of visits to state-action and state-action-next state tuples, respectively. At the beginning of episode k, the algorithm has access to the aggregated private statistics:

$$\begin{aligned} \widetilde{R}_k(s,a) &= \sum_{l<k} \widetilde{R}_{X_{u_l}}(s,a), \\ \widetilde{N}_k^r(s,a) &= \sum_{l<k} \widetilde{N}^r_{X_{u_l}}(s,a) \\ \text{and } \widetilde{N}_k^p(s,a,s') &= \sum_{l<k} \widetilde{N}^p_{X_{u_l}}(s,a,s') \end{aligned} \qquad (3)$$

We denote the non-private counterparts of these aggregated statistics as $R_k(s,a) = \sum_{l<k} R_{X_{u_l}}(s,a)$, $N_k^r(s,a) = \sum_{l<k} N^r_{X_{u_l}}(s,a)$ and $N_k^p(s,a,s') = \sum_{l<k} N^p_{X_{u_l}}(s,a,s')$ (these are also unknown to the RL agent).

Using these private statistics, we can define conditions that a private randomizer must satisfy in order for the RL agent to be able to learn the reward and dynamics of the MDP. In particular, we require that the randomizer must return private statistics that enable a "reasonable" estimate of the MDP to be constructed. The exact definition we require is given below:

Definition 2 (Private Randomizer for RL)**.** *The private randomizer $\mathcal{M}$ should satisfy $(\varepsilon_0, \delta_0)$-LDP, Def. 1, with $\varepsilon_0, \delta_0 \geq 0$. Moreover, for any $\delta > 0$ and $k \geq 0$, there must exist four finite strictly positive function, $c_{k,1}(\varepsilon_0, \delta_0, \delta), c_{k,2}(\varepsilon_0, \delta_0, \delta), c_{k,3}(\varepsilon_0, \delta_0, \delta) \in \mathbb{R}_+^\star$ and $c_{k,4}(\varepsilon_0, \delta_0, \delta) \in \mathbb{R}_+^\star$ such that with probabilty at least $1-\delta$ for all $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$:*

$$\left|\widetilde{R}_k(s,a) - R_k(s,a)\right| \leq c_{k,1}(\varepsilon_0, \delta_0, \delta) \tag{4}$$

$$\left|\widetilde{N}_k^r(s,a) - N_k^r(s,a)\right| \leq c_{k,2}(\varepsilon_0, \delta_0, \delta) \tag{5}$$

$$\left|\sum_{s'} N_k^p(s,a,s') - \sum_{s\text{‘}} \widetilde{N}_k^p(s,a,s')\right| \leq c_{k,3}(\varepsilon_0, \delta_0, \delta) \tag{6}$$

$$\left|N_k^p(s,a,s') - \widetilde{N}_k^p(s,a,s')\right| \leq c_{k,4}(\varepsilon_0, \delta_0, \delta) \tag{7}$$

The functions $c_{k,1}(\varepsilon_0, \delta_0, \delta)$, $c_{k,2}(\varepsilon_0, \delta_0, \delta)$, $c_{k,3}(\varepsilon_0, \delta_0, \delta)$ and $c_{k,4}(\varepsilon_0, \delta_0, \delta)$ must be increasing functions of k and decreasing functions of δ. We also write $c_{k,1}(\varepsilon_0, \delta)$, $c_{k,2}(\varepsilon_0, \delta)$, $c_{k,3}(\varepsilon_0, \delta)$ and $c_{k,4}(\varepsilon_0, \delta)$ when $\delta_0 = 0$.

The problem of constructing counters which satisfy conditions of Def. 2 has been studied extensively in the differential privacy literature (Dwork et al., 2006; Geng et al., 2020).

In Dwork et al. (2010), it was shown for DP that supervised learning in databases of size T, there exists a private randomizer such that the difference between the private and non-private counters is of order $\mathcal{O}(\ln(T)/\varepsilon)$. For local differential privacy, this difference is of order $\mathcal{O}(\sqrt{T}/\varepsilon)$ using the Laplace mechanism for databases of size T (see Dwork and Roth, 2014, Sec. 12.3). In the non-private case the regret scales with $\mathcal{O}(\sqrt{T})$. The effect of privacy on estimating a model of the environment is of order $\mathcal{O}(\sqrt{T}/\varepsilon)$, so we expect the regret to be of order $\mathcal{O}(\sqrt{T}(1+1/\varepsilon))$.

3.2 LDP-OBI

In this section, we introduce LDP-OBI, an LDP algorithm for exploration. As commonly done in the literature, LDP-OBI is based on the optimism-in-the-face-of-uncertainty principle. When developing optimistic algorithms is it necessary to define confidence intervals using the estimated model that are broad enough to capture the true model with high probability, but narrow enough to ensure low regret. This is made more complicated in the LDP setting, since the estimated model is defined using randomized counters. In particular, this means we cannot use standard concentration inequalities such as those used in (Azar et al., 2017; Zanette and Brunskill, 2019). Moreover, when working with randomized counters, classical estimators like the empirical mean can even be ill-defined as the number of visits to a state-action pair, for example, can be negative.

Nevertheless, we show that by exploiting the properties of the mechanism $\mathcal{M}$ in Def. 2, it is still possible to define an empirical model which can be shown to be close to the true model with high probability. To construct this empirical estimator, we rely on the fact that for each state-action pair (s,a), $\widetilde{N}_k^r(s,a) + c_{k,2}(\varepsilon_0, \delta_0, \delta) \geq N_k^r(s,a) \geq 0$ with high probability where the precision $c_{k,2}(\varepsilon_0, \delta_0, \delta)$ ensures the positivity of the noisy number of visits to a state action-pair. A similar argument holds for the transitions. Formally, we define the estimated (*private*) rewards and transitions before episode k as follows:

$$\begin{aligned} \widetilde{r}_k(s,a) &= \frac{\widetilde{R}_k(s,a)}{\widetilde{N}_k^r(s,a) + \alpha c_{k,2}(\varepsilon_0, \delta_0, \delta)}, \\ \widetilde{p}_k(s' \mid s,a) &= \frac{\widetilde{N}_k^p(s,a,s')}{\widetilde{N}_k^p(s,a) + \alpha c_{k,3}(\varepsilon_0, \delta_0, \delta)} \end{aligned} \tag{8}$$

By leveraging properties of Def. 2 and standard concentrations, we can show the following result. The proof is provided in App. A.

Proposition 1. *For any $\varepsilon_0 > 0$, $\delta_0 \geq 0$, $\delta > 0$, $\alpha > 1$ and episode k, using mechanism $\mathcal{M}$, we have that with probability at least $1 - 2\delta$, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$*

$$\begin{aligned} |r(s,a) - \widetilde{r}_k(s,a)| &\leq \sqrt{\frac{2\ln\left(\frac{4\pi^2 SAHk^3}{3\delta}\right)}{\widetilde{N}_k^r(s,a) + \alpha c_{k,2}(\varepsilon_0, \delta_0, \delta)}} \\ &+ \frac{(\alpha+1)c_{k,2}(\varepsilon_0, \delta_0, \delta) + c_{k,1}(\varepsilon_0, \delta_0, \delta)}{\widetilde{N}_k^r(s,a) + \alpha c_{k,2}(\varepsilon_0, \delta_0, \delta)} := \beta_k^r(s,a) \end{aligned} \tag{9}$$

$$\begin{aligned} ||p(\cdot \mid s,a) - \widetilde{p}_k(\cdot \mid s,a)||_1 &\leq \sqrt{\frac{14S\ln\left(\frac{4\pi^2 SAHk^3}{3\delta}\right)}{\widetilde{N}_k^p(s,a) + \alpha c_{k,3}(\varepsilon_0, \delta_0, \delta)}} \\ &+ \frac{Sc_{k,4}(\varepsilon_0, \delta_0, \delta) + (\alpha+1)c_{k,3}(\varepsilon_0, \delta_0, \delta)}{\widetilde{N}_k^p(s,a) + \alpha c_{k,3}(\varepsilon_0, \delta_0, \delta)} := \beta_k^p(s,a) \end{aligned} \tag{10}$$

The shape of the bonuses in Prop. 1 highlights two terms. The first term is reminiscent of Hoeffding bonuses in optimistic RL algorithms as it scales as $\mathcal{O}\left(1/\sqrt{\widetilde{N}_k^p}\right)$. The remaining term is of order $\mathcal{O}\left(1/\widetilde{N}_k^p\right)$ and accounts for the variance (and potentially bias) of the noise added by the privacy-preserving mechanism.

As commonly done in the literature (e.g., Azar et al., 2017; Qian et al., 2019; Neu and Pike-Burke, 2020), we use these concentration results to define a bonus function

$$b_{h,k}(s,a) := (H-h+1)\cdot \beta_k^p(s,a) + \beta_k^r(s,a). \qquad (11)$$

At each episode k, LDP-OBI builds an estimated MDP $M_k = (\mathcal{S}, \mathcal{A}, \widetilde{p}_k, \widetilde{r}_k + b_k, H)$ and computes the optimal value function V_k (and associated policy) using truncated backward induction (e.g., Azar et al., 2017):

$$\begin{aligned} Q_{h,k}(s,a) &= \widetilde{r}_k(s,a) + b_{h,k}(s,a) + \widetilde{p}_k(\cdot|s,a)^\intercal V_{h+1,k} \\ \pi_{h,k}(s) &= \arg\max_a Q_{h,k}(s,a) \end{aligned} \qquad (12)$$

where $V_{h,k}(s) = \min\{H-h+1, \max_a Q_{h,k}(s,a)\}$ and $V_{H+1,k}(s) = 0$. It can be shown that $V_{k,h}(s) \geq V_h^\star(s)$ and thus the greedy policy w.r.t. V_k, $\pi_k = (\pi_{h,k})_h$, is optimistic.

Algorithm 2: LDP-OBI ($\mathcal{M}$)

```
om Input: number of episodes $K$, horizon $H$, failure
        probability $\delta \in (0,1)$, bias $\alpha > 1$, private
        randomizer $\mathcal{M}$ with parameters $(\epsilon_0, \delta_0)$
Set $\mathcal{H}_0 = \emptyset$
for $k = 1, \dots, K$ do
    Compute $\widetilde{p}_k$ and $\widetilde{r}_k$ as in Eq. 8 using $\mathcal{H}_{k-1}$
    Compute $\beta_k^r$ and $\beta_k^p$ as in Eq. 9-10 using
      $c_{k,1}(\varepsilon_0,\delta_0,\delta')$, $c_{k,2}(\varepsilon_0,\delta_0,\delta')$, $c_{k,3}(\varepsilon_0,\delta_0,\delta')$ and
      $c_{k,4}(\varepsilon_0,\delta_0,\delta')$ with $\delta' = \frac{3\delta}{2k^2\pi^2}$
    Compute exploration bonus $b_{h,k}$ as in Eq. 11
    Compute $\pi_k$ as in Eq. 12
    Send policy $\pi_k$ to user $u_k$
    User $u_k$ executes policy $\pi_k$ in the environment,
      collects trajectory $X_k = \{(s_{k,h}, a_{k,h}, r_{k,h})_{h \leq H}\}$
      and sends back privatized information $\mathcal{M}(X_k)$
    Update historical data $\mathcal{H}_k = \mathcal{H}_{k-1} \cup \mathcal{M}(X_k)$
```

4 Regret Guarantees

We start this section by presenting a lower bound on the regret of any algorithm in the LDP setting. We then state the regret of LDP-OBI algorithm associated to any private mechanism satisfying Def. 2.

4.1 Lower Bound

To provide a lower bound for LDP setting, we build a hard instance of finite-horizon MDP leveraging the idea in (Auer et al., 2002; Lattimore and Szepesvári, 2020). To handle privacy, we rely on the fact that local differential privacy acts as Lipschitz function, with respect to the KL-divergence, in the space of probability distribution (see Duchi et al., 2013, Thm. 1). The proof of Thm. 1 is provided in App. B.

Theorem 1. *For any algorithm $\mathfrak{A}$ associated to a ε-LDP mechanism, any number of states $S \geq 3$ and actions $A \geq 2$ and $H \geq 2\log_A(S-2)+2$ there exists an MDP M with S states and A actions such that:*

$$\mathbb{E}_M(R(\mathfrak{A}, K)) \geq \Omega\left(\frac{H\sqrt{SAK}}{\min\{\exp(\varepsilon)-1, 1\}}\right) \qquad (13)$$

The recent work of Vietri et al. (2020) states that for joint differential privacy the regret in finite-horizon MDPs is lower-bounded by $\Omega\left(H\sqrt{SAK} + \frac{1}{\varepsilon}\right)$. Thm. 1 shows that the local differential privacy setting is inherently harder than the joint differential privacy one for small ϵ as our lower-bound scales with $\sqrt{K}/\varepsilon$ when $\varepsilon \approxeq 0$. Both bounds scale with $\sqrt{K}$ when $\varepsilon \to +\infty$.

4.2 Regret Upper Bound

We now show a high probability upper bound on the regret of LDP-OBI associated with any LDP mechanism $\mathcal{M}$ satisfying Def. 2.

Theorem 2. *For any privacy mechanism $\mathcal{M}$ satisfying Def. 1 and Def. 2 with $\varepsilon > 0$, $\delta_0 \geq 0$ and bounds $c_{k,1}(\varepsilon,\delta_0,.)$, $c_{k,2}(\varepsilon,\delta_0,.)$, $c_{k,3}(\varepsilon,\delta_0,.)$ and $c_{k,4}(\varepsilon,\delta_0,.)$, for any $\delta > 0$ the regret of* LDP-OBI *is bounded with probability at least $1-\delta$ by:*

$$\begin{aligned} R(\text{LDP-OBI}, K) \leq \tilde{\mathcal{O}}\Bigg(\max\Bigg\{ & HS\sqrt{AT}, \\ & SAH^2 c_{K,3}\left(\varepsilon, \delta_0, \frac{3\delta}{2\pi^2K^2}\right), \\ & H^2S^2Ac_{K,4}\left(\varepsilon, \delta_0, \frac{3\delta}{2\pi^2K^2}\right), \\ & SAHc_{K,2}\left(\varepsilon, \delta_0, \frac{3\delta}{2\pi^2K^2}\right), \\ & SAHc_{K,1}\left(\varepsilon, \delta_0, \frac{3\delta}{2\pi^2K^2}\right)\Bigg\}\Bigg) \end{aligned} \qquad (14)$$

In addition, the combination of $\mathcal{M}$ and LDP-OBI *is (ε, δ_0)-LDP.*

Theorem 2 shows that the regret of LDP-OBI depends directly on the precision of the privacy mechanism used though $c_{K,1}, \dots, c_{K,4}$. Thus improving the precision, that is to say reducing the amount of noise that needs to be added to the data to guarantee LDP of the privacy mechanism directly improves the regret bounds of LDP-OBI.

5 Choice of Randomizer

In this section, we provide a practical implementation of LDP-OBI based on different randomizers. We start by providing a detailed discussion of the Laplace mechanism (Dwork and Roth, 2014), including accuracy ($c_{k,i}$), privacy and regret guarantees. We then compare these results to those achievable with other mechanisms.

```
Algorithm 3: Laplace mechanism for LDP
Input: Trajectory: $X = \{(s_h, a_h, r_h) \mid h \leq H\}$,
       Privacy Parameter: $\varepsilon_0$
Draw $(Y_{i,X}(s,a))_{(s,a)\in\mathcal{S}\times\mathcal{A}, i\leq 2}$ i.i.d Lap$(1/\varepsilon_0)$ and
  $(Z_X(s,a,s'))_{(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}}$ i.i.d Lap$(1/\varepsilon_0)$ and
  independent from $Y_{i,X}$ for $i \in \{1,2\}$
for $(s,a) \in \mathcal{S}\times\mathcal{A}$ do
    $\widetilde{R}_X(s,a) = \sum_{h=1}^{H} r_h \mathbb{1}_{\{s_h=s, a_h=a\}} + Y_{1,X}(s,a)$
    $\widetilde{N}_X^r(s,a) = \sum_{h=1}^{H} \mathbb{1}_{\{s_h=s, a_h=a\}} + Y_{2,X}(s,a)$
    for $s' \in \mathcal{S}$ do
        $\widetilde{N}_X^p(s,a,s') =$
          $\sum_{h=1}^{H-1} \mathbb{1}_{\{s_h=s, a_h=a, s_{h+1}=s'\}} + Z_X(s,a,s')$
Output: $(\widetilde{R}_X, \widetilde{N}_X^r, \widetilde{N}_X^p) \in \mathbb{R}^{S\times A} \times \mathbb{R}^{S\times A} \times \mathbb{R}^{S\times A\times S}$
```

The detailed derivations for these other mechanisms is deferred to App. E.

The Laplace mechanism works by injecting Laplace noise to the aggregated statistic computed from trajectory X.[3] The pseudocode is reported in Alg. 3. In the below theorem, we show that by properly tuning the parameter ε_0 of the Laplace noise, we can prove that this mechanism is LDP (proof in App. A.1):

Theorem 3. *For any $\varepsilon > 0$, the Laplace mechanism described by Alg. 3 with parameter $\varepsilon_0 = \varepsilon/6H$ is $(\varepsilon, 0)$-LDP (and thus (ε, δ_0)-LDP for every $\delta_0 \geq 0$).*

Moreover, due to the sub-exponential nature of Laplace distribution, we can use the Chernoff concentration bound to show that Alg. 3 satisfies the requirements in Def. 2 (see App. A for the proof):

Proposition 2. *For any $\varepsilon > 0$, the Laplace mechnism, Alg. 3, with parameter $\varepsilon_0 = \varepsilon/6H$ satisfies Def. 2 for any $\delta > 0$ and $k \in \mathbb{N}$ with $c_{k,1}(\varepsilon,\delta) = c_{k,2}(\varepsilon,\delta)$, $c_{k,3}(\varepsilon,\delta) = c_{k,4}(\varepsilon,\delta)$ and:*

$$c_{k,1}(\varepsilon,\delta) = \max\left\{\sqrt{k}, \ln\left(\frac{6SA}{\delta}\right)\right\} \frac{\sqrt{8\ln\left(\frac{6SA}{\delta}\right)}}{\varepsilon/6H},$$

$$c_{k,3}(\varepsilon,\delta) = \max\left\{\sqrt{k}, \ln\left(\frac{6S^2A}{\delta}\right)\right\} \frac{\sqrt{8\ln\left(\frac{6S^2A}{\delta}\right)}}{\varepsilon/6H}$$

As a corollary of Thm. 2, we obtain the following regret bound for LDP-OBI with Laplace mechanism.

Corollary 1. *For any $\delta' > 0$ the regret of* LDP-OBI *using the Laplace mechanism with $\varepsilon_0 = \varepsilon/6H$, Alg. 3, is bounded with probability at least $1 - \delta'$ by:*

$$\tilde{\mathcal{O}}\left(\max\left\{\frac{H^3S^2A\sqrt{K}}{\varepsilon}, H^{3/2}S\sqrt{AK}\right\}\right) \tag{15}$$

and the algorithm is $(\varepsilon, 0)$-LDP.

$\mathcal{M}$	LDP	$R(\mathfrak{A}, T)$
L	$(\varepsilon, 0)$	$\widetilde{O}(H^3S^2A\sqrt{K}/\varepsilon)$
G	(ε, δ_0)	$\widetilde{O}(H^3S^2A\sqrt{K\ln(1/\delta_0)}/\varepsilon)$
B	$(\varepsilon, 0)$	$\widetilde{O}(\frac{H^{5/2}S^2A}{e^{\varepsilon/H}-1}\sqrt{K})$

Table 1: Summary of the guarantees of LDP-OBI with Laplace (L), Gaussian (G) and Bernoulli (B) mechanisms. We assume $\varepsilon \in (0, 6H)$ and $\delta_0 > 0$. Full regret bounds are reported in appendix.

Note that the leading term of the regret is $\tilde{\mathcal{O}}(\sqrt{T}/\epsilon)$ that is compatible with the rates obtained for LDP bandits (see e.g., Zheng et al., 2020; Ren et al., 2020). While this shows an optimal dependency w.r.t. the number of episodes and privacy parameter, it is suboptimal in terms of state size, action size and horizon. Compared to a non-private version of OBI (with a regret of order $\tilde{\mathcal{O}}(H^{3/2}\sqrt{SAK})$) (e.g., Azar et al., 2017) , aside from the $1/\epsilon$ term which is to be expected, there is an additional factor of $H^{3/2}S^{3/2}\sqrt{A}$ in the leading term of the regret. We believe that this is due to the fact that we are having to make S^2A terms private (i.e., the counters $N_X^p(s,a,s')$ must be private for all $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$). We think the extra dependence on H comes from diving ε by H to ensure privacy over the whole trajectory. An other $\sqrt{SH}$ factor also comes the fact that we do not use variance-aware concentration inequalities contrary to UCB-VI. We investigated the use of such inequalities but it did not improve the dependency on S, A and H in the leading terms of the regret. It is an open question whether this dependence can be improved or not.

PAC Guarantees. Following the discussion of Sec. 3.1 in (Jin et al., 2018), we get that our algorithm LDP-OBI with the Laplace mechanism finds α-optimal policies in the PAC setting using $\widetilde{O}\left(\frac{H^2S^2A\max\{1, H^3S^2A/\varepsilon^2\}}{\alpha^2}\right)$ samples, for any $\alpha \in (0, H]$. Whereas in the non-private case, OBI finds α -optimal policies using at most $\widetilde{O}\left(\frac{H^2SA}{\alpha^2}\right)$ samples.

5.1 Alternative Mechanisms

There are other privacy-preserving mechanisms which can be used in LDP-OBI, for example, the Bernoulli mechanism (Erlingsson et al., 2014; Kairouz et al., 2016) and the Gaussian mechanism (Wang et al., 2019). We summarize the properties of the different variants of LDP-OBI in Tab. 1.

From looking at Tab. 1, we note that the Gaussian mechanism provides a slightly worse privacy guarantee. While Laplace and Bernoulli can guarantee $(\varepsilon, 0)$-LDP, the Gaussian mechanism only (ε, δ_0)-LDP for some

[3] A random variable $X \sim \text{Lap}(b)$ a Laplace distribution with parameter b if and only if: $\forall x \in \mathbb{R}, p_X(x) = \frac{1}{2b}\exp(-|x|/b)$.

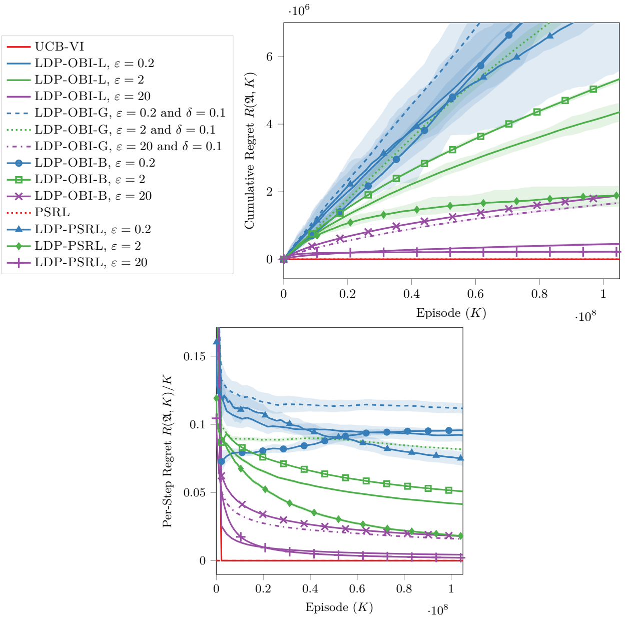

Figure 1: Evaluation of the algorithms in the RandomMDP environment. *Top)* Cumulative regret. *Bottom)* per-step regret ($k \mapsto R_k/k$). Results are averaged over 20 runs and the the confidence intervals are the minimum and maximum runs. While the regret looks almost linear for $\varepsilon = 0.2$, the decreasing trend of the per-step regret shows that the algorithms are learning.

$\delta_0 > 0$. On the other hand, all the mechanisms achieve a regret bound of order $\widetilde{O}(\sqrt{K})$. While Laplace and Gaussian variants are equally affected by the privacy level ε (whose impact is ε^{-1}), the Bernoulli mechanism has an exponential dependence in ε similar to the lower bound. However, this improvement comes at the price of worse dependency in H when ε is small, and a worse multiplicative constant in the regret. This is due to the fact that the Bernoulli mechanism needs to perturb the counters for each stage $h \in [H]$, leading to up to HS^2A obfuscated elements (see App. E.2 for details). This worse dependence is also observed in our numerical simulations. All the details about the Gaussian and Bernoulli mechanism can be found in App. E and App. E.2, respectively.

6 Numerical Evaluation

In this section, we illustrate the empirical performance of the proposed algorithms on a toy MDP. To the best of our knowledge there is no other LDP algorithm for regret minimization in MDPs in the literature. We thus compare LDP-OBI with the non-private algorithm UCB-VI (Azar et al., 2017). Since randomized algorithm proved to be effective in many situations, we also investigate a Thompson sampling approach. We introduce and evaluate LDP-PSRL, an LDP variant of PSRL (Osband et al., 2013). LDP-PSRL is detailed in App. D where we also show that it is LDP. We leave the regret proof as an open question.

We consider the RandomMDP environment described in (Dann et al., 2017) where for each state-action pair transition probabilities are sampled from a Dirichlet(α) distribution (with $\alpha_{s,a,s'} = 0.1$ for all (s,a,s')) and rewards are deterministic in $\{0,1\}$ where $r(s,a) = \mathbb{1}_{\{U_{s,a} \leq 0.5\}}$ and $(U_{s,a})_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sim \mathcal{U}([0,1])$ are sampled once when generating the MDP. We set the number of states $S = 3$, number of actions $A = 2$ and horizon $H = 2$. We evaluate the regret of our algorithm for $\varepsilon \in \{0.2, 2, 20\}$ and $K = 1 \times 10^8$ episodes. For each ε we run 20 simulations.

Figure 1 shows that the learning speed of the optimistic algorithm is severely impacted by the LDP constraint. This is consistent with our theoretical results. The reason for this is the very large confidence intervals that are needed in order to take into account the noise from the privacy preserving mechanism that is necessary to guarantee privacy. This is not necessarily the case with LDP-PSRL as in general posterior sampling algorithms are empirically quite robust to noise in the parameters of the posterior distributions. Although these experimental results only consider a small MDP, we expect that many of the observations will carry across to larger, more practical settings. However, further experiments are needed to conclusively assess the impact of LDP in large MDPs. In App. F, we show the impact of the LDP constraint on the collected trajectories using the Laplace mechanism.

7 Conclusion

In this work, we have introduced the definition of local differential privacy in RL and designed the first LDP algorithm, LDP-OBI, for regret minimization in finite-horizon MDPs. By leveraging new confidence intervals accounting for the noise introduced by the private mechanism, we have derived an upper-bound on the regret of LDP-OBI. This approach leads to suboptimal dependency on S, A and H compared to the lower bound we established in Sec. 4.1. Closing this gap would be an interesting technical question for future works. Additionally, while we have shown that LDP-PSRL achieves LDP guarantees and good empirical performance, a direction for future work is to provide regret guarantees for this approach.

The study of differential privacy in RL is still recent and a lot of questions are still not answered. In particular, we think a promising direction would be to study model-free techniques for DP that could be used to design deep RL approaches.

References

John M Abowd. The us census bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2867–2867, 2018.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1): 48–77, 2002.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 263–272. PMLR, 2017.

Borja Balle, Maziar Gomrokchi, and Doina Precup. Differentially private policy evaluation. In *ICML*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 2130–2138. JMLR.org, 2016.

Etienne Boursier and Vianney Perchet. Utility/privacy trade-off through the lens of optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pages 591–601, 2020.

Xiaoyu Chen, Kai Zheng, Zixin Zhou, Yunchang Yang, Wei Chen, and Liwei Wang. (locally) differentially private combinatorial semi-bandits. In *ICML*, 2020.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5713–5723, 2017.

Abhimanyu Dubey and Alex Pentland. Differentially-private federated linear bandits. Technical report, Massachusetts Institute of Technology, 2020a.

Abhimanyu Dubey and Alex Pentland. Private and byzantine-proof cooperative decision-making. In *AAMAS*, pages 357–365. International Foundation for Autonomous Agents and Multiagent Systems, 2020b.

John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy, data processing inequalities, and statistical minimax rates. *arXiv preprint arXiv:1302.3203*, 2013.

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

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer, 2006.

Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N Rothblum. Differential privacy under continual observation. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 715–724, 2010.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067, 2014.

Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Improved analysis of ucrl2 with empirical bernstein inequality. *arXiv preprint arXiv:2007.05456*, 2020.

Pratik Gajane, Tanguy Urvoy, and Emilie Kaufmann. Corrupt bandits for preserving local privacy. In *ALT*, volume 83 of *Proceedings of Machine Learning Research*, pages 387–412. PMLR, 2018.

Quan Geng, Wei Ding, Ruiqi Guo, and Sanjiv Kumar. Tight analysis of privacy and utility tradeoff in approximate differential privacy. In *International Conference on Artificial Intelligence and Statistics*, pages 89–99, 2020.

Awni Y. Hannun, Brian Knott, Shubho Sengupta, and Laurens van der Maaten. Privacy-preserving multi-party contextual bandits. *CoRR*, abs/1910.05299, 2019.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

Chi Jin, Zeyuan Allen-Zhu, Sébastien Bubeck, and Michael I. Jordan. Is q-learning provably efficient? In *NeurIPS*, pages 4868–4878, 2018.

Peter Kairouz, Keith Bonawitz, and Daniel Ramage. Discrete distribution estimation under local privacy. *arXiv preprint arXiv:1602.07387*, 2016.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Mohammad Malekzadeh, Dimitrios Athanasakis, Hamed Haddadi, and Benjamin Livshits. Privacy-preserving bandits. In *MLSys*. mlsys.org, 2020.

Nikita Mishra and Abhradeep Thakurta. (nearly) optimal differentially private stochastic multi-arm bandits. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, UAI'15, page 592–601, Arlington, Virginia, USA, 2015. AUAI Press. ISBN 9780996643108.

Gergely Neu and Ciara Pike-Burke. A unifying view of optimism in episodic reinforcement learning. *arXiv preprint arXiv:2007.01891*, 2020.

Hajime Ono and Tsubasa Takahashi. Locally private distributed reinforcement learning. *arXiv preprint arXiv:2001.11718*, 2020.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *NIPS*, pages 3003–3011, 2013.

Xinlei Pan, Weiyao Wang, Xiaoshuai Zhang, Bo Li, Jinfeng Yi, and Dawn Song. How you act tells a lot: Privacy-leaking attack on deep reinforcement learning. In *AAMAS*, pages 368–376. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley & Sons, Inc., New York, NY, USA, 1994. ISBN 0471619779.

Jian Qian, Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Exploration bonus for regret minimization in discrete and continuous average reward mdps. In *NeurIPS*, pages 4891–4900, 2019.

Wenbo Ren, Xingyu Zhou, Jia Liu, and Ness B Shroff. Multi-armed bandits with local differential privacy. *arXiv preprint arXiv:2007.03121*, 2020.

Touqir Sajed and Or Sheffet. An optimal private stochastic-mab algorithm based on optimal private stopping rule. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 5579–5588. PMLR, 2019.

Roshan Shariff and Or Sheffet. Differentially private contextual linear bandits. In *NeurIPS*, pages 4301–4311, 2018.

Aristide Tossou and Christos Dimitrakakis. Differentially private multi-agent multi-armed bandits. In *European Workshop on Reinforcement Learning (EWRL-15)*, 2015.

Aristide C. Y. Tossou and Christos Dimitrakakis. Algorithms for differentially private multi-armed bandits. In *AAAI*, pages 2087–2093. AAAI Press, 2016.

Giuseppe Vietri, Borja de Balle Pigem, Akshay Krishnamurthy, and Steven Wu. Private reinforcement learning with pac and regret guarantees. In *ICML*, 2020.

Baoxiang Wang and Nidhi Hegde. Privacy-preserving q-learning with functional noise in continuous spaces. In *NeurIPS*, pages 11323–11333, 2019.

Teng Wang, Jun Zhao, Xinyu Yang, and Xuebin Ren. Locally differentially private data collection and analysis. *arXiv preprint arXiv:1906.01777*, 2019.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l1 deviation of the empirical distribution. 2003.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 7304–7312. PMLR, 2019.

Kai Zheng, Tianle Cai, Weiran Huang, Zhenguo Li, and Liwei Wang. Locally differentially private (contextual) bandits learning. *CoRR*, abs/2006.00701, 2020.

Local Differentially Private Regret Minimization in Reinforcement Learning: Supplementary Material

A Local Differential Privacy

A.1 The Laplace mechanism (Alg. 3) satisfies local differential privacy (Def. 2)

We first prove Thm. 3 which states that using Alg. 3 with parameter $\varepsilon_0 = \varepsilon/6H$ guarantee (ε, δ)-LDP. Formally, we need to show that, for any two trajectories X and X' and tuple (r, n, n'), the following inequality holds

$$\mathbb{P}\Big(\mathcal{M}(X) = (r, n, n')\Big) \leq e^{\varepsilon}\mathbb{P}\Big(\mathcal{M}(X') = (r, n, n')\Big) + \delta \tag{16}$$

where r, n, n' are vectors of dimension SA, SA and S^2A, respectively. See LDP definition in Def. 1.

Proof of Thm. 3. Let's consider two trajectories $X = \{(s_h, a_h, r_h) \mid h \leq H\}$ and $X' = \{(s'_h, a'_h, r'_h) \mid h \leq H\}$. We denote the output of the private randomizer $\mathcal{M}$ by $\mathcal{M}(X) = (\widetilde{R}_X, \widetilde{N}^r_X, \widetilde{N}^p_X)$ and $\mathcal{M}(X') = (\widetilde{R}_{X'}, \widetilde{N}^r_{X'}, \widetilde{N}^p_{X'})$. Recall that $\widetilde{R}_X(s,a) := \sum_{h=1}^H r_h \mathbb{1}_{\{s_h = s, a_h = a\}} + Y_{1,X}(s,a)$ where $(Y_{1,X}(s,a))_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ are independent Laplace variables with parameter $\varepsilon/(6H)$. Consider a vector $r \in \mathbb{R}^{S\times A}$, then:

$$\frac{\mathbb{P}\left(\forall (s,a), \widetilde{R}_X(s,a) = r_{s,a} \mid X\right)}{\mathbb{P}\left(\forall (s,a), \widetilde{R}_{X'}(s,a) = r_{s,a} \mid X'\right)} = \prod_{s,a} \frac{\mathbb{P}\left(Y_{1,X}(s,a) = \sum_{h=1}^H r_h \mathbb{1}_{\{s_h = s, a_h = a\}} - r_{s,a} \mid X\right)}{\mathbb{P}\left(Y_{1,X'}(s,a) = \sum_{h=1}^H r'_h \mathbb{1}_{\{s'_h = s, a'_h = a\}} - r_{s,a} \mid X'\right)} \tag{17}$$

since the Laplace distribution is symmetric. But $Y_{1,X}(s,a)$ and $Y_{1,X'}(s,a)$ are independent random variables for any state-action pair. Thus:

$$\begin{aligned}
\prod_{s,a} \frac{\mathbb{P}\left(Y_{1,X}(s,a) = \sum_{h=1}^H r_h \mathbb{1}_{\{s_h = s, a_h = a\}} - r_{s,a} \mid X\right)}{\mathbb{P}\left(Y_{1,X'}(s,a) = \sum_{h=1}^H r'_h \mathbb{1}_{\{s'_h = s, a'_h = a\}} - r_{s,a} \mid X'\right)} &= \prod_{s,a} \frac{\exp\left(\varepsilon_0 \left|\sum_{h=1}^H (r_h \mathbb{1}_{\{s_h = s, a_h = a\}} - r_{s,a}\right|\right)}{\exp\left(\varepsilon_0 \left|\sum_{h=1}^H (r'_h \mathbb{1}_{\{s'_h = s, a'_h = a\}} - r_{s,a}\right|\right)} \\
&\leq \exp\left(\varepsilon_0 \sum_{s,a} \left|\sum_{h=1}^H (r_h \mathbb{1}_{\{s_h = s, a_h = a\}} - r'_h \mathbb{1}_{\{s'_h = s, a'_h = a\}})\right|\right) \\
&\leq \exp\left(\varepsilon_0 \sum_{s,a,h} (|r_h| \mathbb{1}_{\{s_h = s, a_h = a\}} + |r'_h| \mathbb{1}_{\{s'_h = s, a'_h = a\}})\right) \\
&= \exp\left(\varepsilon_0 \sum_{h} (|r_h| + |r'_h|)\right) \leq \exp\left(2H\varepsilon_0\right) = \exp\left(\frac{\varepsilon}{3}\right)
\end{aligned} \tag{18}$$

where we used the definition of Laplace distribution, $x \mapsto \frac{1}{2b}\exp(|x|/b)$. Let $n \in \mathbb{R}^{S\times A}$ and $n' \in \mathbb{R}^{S\times A\times S}$. Similarly, since $\widetilde{N}^r_X(s,a) = \sum_{h=1}^H \mathbb{1}_{\{s_h = s, a_h = a\}} + Y_{2,X}(s,a)$ and $\widetilde{N}^p_X(s,a,s') = \sum_{h=1}^{H-1} \mathbb{1}_{\{s_h = s, a_h = a, s_{h+1} = s'\}} + Z_X(s,a,s')$, we have:

$$\frac{\mathbb{P}\left(\forall (s,a), \widetilde{N}^r_X(s,a) = n_{s,a} \mid X\right)}{\mathbb{P}\left(\forall (s,a), \widetilde{N}^r_{X'}(s,a) = n_{s,a} \mid X'\right)} \leq \exp\left(\frac{\varepsilon}{3}\right) \tag{19}$$

and:

$$\frac{\mathbb{P}\left(\forall (s,a,s'), \widetilde{N}^p_X(s,a,s') = n'_{s,a,s'} \mid X\right)}{\mathbb{P}\left(\forall (s,a,s'), \widetilde{N}^p_{X'}(s,a,s') = n'_{s,a,s'} \mid X'\right)} \leq \exp\left(\frac{\varepsilon}{3}\right) \tag{20}$$

Then because $(Y_{i,X}(s,a))_{i\leq 2,(s,a)\in\mathcal{S}\times\mathcal{A}}$, $(Z_X(s,a,s'))_{(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}}$ are independent it holds that:

$$\mathbb{P}\left(\widetilde{R}_X = r, \widetilde{N}^r_X = n, \widetilde{N}^p_X = n' \mid X\right) = \mathbb{P}\left(\widetilde{R}_X = r \mid X\right)\mathbb{P}\left(\widetilde{N}^r_X = n \mid X\right)\mathbb{P}\left(\widetilde{N}^p_X = n' \mid X\right)$$

Thus for any $(r,n,n') \in \mathbb{R}^{S\times A}\times\mathbb{R}^{S\times A}\times\mathbb{R}^{S\times A\times S}$ and any two trajectories X and X':

$$\begin{aligned}\mathbb{P}\Big(\mathcal{M}(X) = (r,n,n') \mid X\Big) &= \mathbb{P}\left(\widetilde{R}_X = r, \widetilde{N}^r_X = n, \widetilde{N}^p_X = n' \mid X\right) && (21)\\ &= \mathbb{P}\left(\widetilde{R}_X = r \mid X\right)\mathbb{P}\left(\widetilde{N}^r_X = n \mid X\right)\mathbb{P}\left(\widetilde{N}^p_X = n' \mid X\right) && (22)\end{aligned}$$

because $(Y_{1,X}(s,a))_{(s,a)\in\mathcal{S}\times\mathcal{A}}$, $(Y_{2,X}(s,a))_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ and $(Z_X(s,a,s'))_{(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}}$ are independent. Therefore using inequalities (18), (19) and (20) in (22), we have:

$$\begin{aligned}\mathbb{P}\Big(\mathcal{M}(X) = (r,n,n') \mid X\Big) &= \mathbb{P}\left(\widetilde{R}_X = r \mid X\right)\mathbb{P}\left(\widetilde{N}^r_X = n \mid X\right)\mathbb{P}\left(\widetilde{N}^p_X = n' \mid X\right)\\ &\leq \exp(\varepsilon)\mathbb{P}\left(\widetilde{R}_{X'} = r \mid X'\right)\mathbb{P}\left(\widetilde{N}^r_{X'} = n \mid X'\right)\mathbb{P}\left(\widetilde{N}^p_{X'} = n' \mid X'\right)\\ &= \exp(\varepsilon)\mathbb{P}\left(\widetilde{R}_{X'} = r, \widetilde{N}^r_{X'} = n, \widetilde{N}^p_{X'} = n' \mid X'\right)\\ &= \exp(\varepsilon)\mathbb{P}\left(\mathcal{M}(X') = (r,n,n') \mid X'\right)\end{aligned}$$

This concludes the proof. □

Before proving Prop. 2 we state the following concentration inequality for the sum of Laplace variables.

Proposition 3. *(Dwork and Roth, 2014, Cor. 12.3) Let $Y_1,\ldots,Y_k$ be independent Lap(b) random variables with $b>0$ and $\delta\in(0,1)$ then for any $\nu > b\max\left\{\sqrt{k}, \sqrt{\ln(2/\delta)}\right\}$ then:*

$$\mathbb{P}\left(\left|\sum_{l=1}^{k} Y_l\right| > \nu\sqrt{8\ln(2/\delta)}\right) \leq \delta$$

We can now prove Prop. 2 that shows that Alg. 3 satisfies Def. 2.

Proof of Prop. 2. Let $X_1,\ldots,X_{k-1}$ be the $k-1$ trajectories generated before episode $k\geq 1$. Consider the private statistic $\widetilde{R}_k(s,a)$ generated by the private randomizer before episode k. Then for any state-action pair $(s,a)\in\mathcal{S}\times\mathcal{A}$:

$$\begin{aligned}\left|\widetilde{R}_k(s,a) - R_k(s,a)\right| &= \left|\sum_{l<k}(\widetilde{R}_{X_l}(s,a) - R_{X_l}(s,a))\right| && (23)\\ &= \left|\sum_{l<k}\left(Y_{1,X_l}(s,a) + \sum_{h=1}^{H} r_h\mathbb{1}_{\{s_{l,h}=s,a_{l,h}=a\}}\right) - \sum_{l<k}\sum_{h=1}^{H} r_h\mathbb{1}_{\{s_{l,h}=s,a_{l,h}=a\}}\right| && (24)\\ &= \left|\sum_{l=1}^{k-1} Y_{1,X_l}(s,a)\right| && (25)\end{aligned}$$

which is the sum of independent Laplace variables. Let $\delta>0$. By Prop. 3 we have that with probability at least $1-\delta/(3SA)$

$$\left|\sum_{l=1}^{k-1} Y_{1,X_l}(s,a)\right| \leq \frac{1}{\varepsilon_0}\max\left\{\sqrt{k-1}, \ln\left(\frac{6SA}{\delta}\right)\right\}\sqrt{8\ln\left(\frac{6SA}{\delta}\right)} \quad (26)$$

The same property holds for $\widetilde{N}^r_k$ and $\widetilde{N}^p_k$ and we again apply Prop. 3. Properties in Def. 2 follow from union bounds. □

A.2 Concentration under Local Differential Privacy (Proof of Prop. 1):

In this subsection, we proceed with the proof of Prop. 1 (recalled below).

Proposition. *For any $\varepsilon_0 > 0$, $\delta_0 \geq 0$, $\delta > 0$, $\alpha > 1$ and episode k, using mechanism $\mathcal{M}$, we have that with probability at least $1 - 2\delta$, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$*

$$|r(s,a) - \widetilde{r}_k(s,a)| \leq \frac{(\alpha+1)c_{k,2}(\varepsilon_0,\delta_0,\delta) + c_{k,1}(\varepsilon_0,\delta_0,\delta)}{\widetilde{N}^r_k(s,a) + \alpha c_{k,2}(\varepsilon_0,\delta_0,\delta)} + \sqrt{\frac{2\ln(4\pi^2 SAHk^3/3\delta)}{\widetilde{N}^r_k(s,a) + \alpha c_{k,2}(\varepsilon_0,\delta_0,\delta)}} \tag{27}$$

$$||\widetilde{p}_k(\cdot \mid s,a) - p(\cdot \mid s,a)||_1 \leq \frac{(\alpha+1)c_{k,3}(\varepsilon_0,\delta_0,\delta) + Sc_{k,4}(\varepsilon_0,\delta_0,\delta)}{\widetilde{N}^p_k(s,a) + \alpha c_{k,3}(\varepsilon_0,\delta_0,\delta)} + \sqrt{\frac{14S\ln(4\pi^2 SAHk^3/3\delta)}{\widetilde{N}^p_k(s,a) + \alpha c_{k,3}(\varepsilon_0,\delta_0,\delta)}} \tag{28}$$

with $\widetilde{r}_k(s,a) = \frac{\widetilde{R}_k(s,a)}{\widetilde{N}^r_k(s,a) + \alpha c_{k,2}(\varepsilon_0,\delta_0,\delta)}$ *and* $\widetilde{p}_k(s' \mid s,a) = \frac{\widetilde{N}^p_k(s,a,s')}{\widetilde{N}^p_k(s,a) + \alpha c_{k,3}(\varepsilon_0,\delta_0,\delta)}$.

Proof. On the event that all inequalities of Def. 2 holds, we have:

$$\left| \frac{\widetilde{R}_k(s,a)}{\widetilde{N}^r_k(s,a) + \alpha c_{k,2}(\varepsilon_0,\delta_0,\delta)} - \frac{R_k(s,a)}{\widetilde{N}^r_k(s,a) + \alpha c_{k,2}(\varepsilon_0,\delta_0,\delta)} \right| \leq \frac{c_{k,1}(\varepsilon_0,\delta_0,\delta)}{\widetilde{N}^r_k(s,a) + \alpha c_{k,2}(\varepsilon_0,\delta_0,\delta)} \tag{29}$$

since $\widetilde{N}^r_k(s,a) + \alpha c_{k,2}(\varepsilon_0,\delta_0,\delta) > N^k_k(s,a) \geq 0$ with $\alpha > 1$. But, we also have that with probability $1-\delta$:

$$\begin{aligned} \left| \frac{R_k(s,a)}{\widetilde{N}^r_k(s,a) + \alpha c_{k,2}(\varepsilon_0,\delta_0,\delta)} - r(s,a) \right| &\leq \left| r(s,a) \left(\frac{N^r_k(s,a)}{\widetilde{N}^r_k(s,a) + \alpha c_{k,2}(\varepsilon_0,\delta_0,\delta)} - 1 \right) \right| \\ &\quad + \left| \frac{N^r_k(s,a)}{\widetilde{N}^r_k(s,a) + \alpha c_{k,2}(\varepsilon_0,\delta_0,\delta)} \times \underbrace{\left(\frac{R_k(s,a)}{N^r_k(s,a)} - r(s,a) \right)}_{:=\overline{r}_k(s,a) - r(s,a)} \right| \end{aligned} \tag{30}$$

$$\leq \frac{N^r_k(s,a)}{\widetilde{N}^r_k(s,a) + \alpha c_{k,2}(\varepsilon_0,\delta_0,\delta)} \frac{L(\delta)}{\sqrt{N^r_k(s,a)}} + r(s,a) \left| 1 - \frac{N^r_k(s,a)}{\widetilde{N}^r_k(s,a) + \alpha c_{k,2}(\varepsilon_0,\delta_0,\delta)} \right| \tag{31}$$

$$\leq \frac{L(\delta)\sqrt{N^r_k(s,a)}}{\widetilde{N}^r_k(s,a) + \alpha c_{k,2}(\varepsilon_0,\delta_0,\delta)} + \frac{(\alpha+1)c_{k,2}(\varepsilon_0,\delta_0,\delta)}{\widetilde{N}^r_k(s,a) + \alpha c_{k,2}(\varepsilon_0,\delta_0,\delta)} \tag{32}$$

where the second inequality follows from Chernoff-Hoeffding bound on the empirical non-private rewards and $L(\delta) = \sqrt{2\ln(4\pi^2 SAHk^3/3\delta)}$, and we use Def. 2 for the last. Furthermore:

$$\frac{L(\delta)\sqrt{N^r_k(s,a)}}{\widetilde{N}^r_k(s,a) + \alpha c_{k,2}(\varepsilon_0,\delta_0,\delta)} \leq \frac{L(\delta)\sqrt{\widetilde{N}^r_k(s,a) + c_{k,2}(\varepsilon_0,\delta_0,\delta)}}{\widetilde{N}^r_k(s,a) + \alpha c_{k,2}(\varepsilon_0,\delta_0,\delta)} \leq \frac{L(\delta)}{\sqrt{\widetilde{N}^r_k(s,a) + \alpha c_{k,2}(\varepsilon_0,\delta_0,\delta)}} \tag{33}$$

Therefore combining Eq. (29), (32) and (33), we have:

$$\left| \frac{\widetilde{R}_k(s,a)}{\widetilde{N}^r_k(s,a) + \alpha c_{k,2}(\varepsilon_0,\delta_0,\delta)} - r(s,a) \right| \leq \frac{c_{k,1}(\varepsilon_0,\delta_0,\delta) + (\alpha+1)c_{k,2}(\varepsilon_0,\delta_0,\delta)}{\widetilde{N}^r_k(s,a) + \alpha c_{k,2}(\varepsilon_0,\delta_0,\delta)} + \frac{L(\delta)}{\sqrt{\widetilde{N}^r_k(s,a) + \alpha c_{k,2}(\varepsilon_0,\delta_0,\delta)}}$$

thus proving the first statement of the proposition. Now, let's bound the deviation between the private estimate $\widetilde{p}_k$ and the true transition dynamics p. First, because $\alpha > 1$, we have that $\sum_{s'} \widetilde{N}^p_k(s,a,s') + \alpha c_{k,3}(\varepsilon_0,\delta_0,\delta) \geq \sum_{s'} N^p_k(s,a,s') + (\alpha - 1)c_{k,3}(\varepsilon_0,\delta_0,\delta) > 0$. We start by decomposing the error as

$$\sum_{s' \in \mathcal{S}} |\widetilde{p}(s'|s,a) - p(s'|s,a)| = \sum_{s' \in \mathcal{S}} \left| \frac{\widetilde{N}^p_k(s,a,s')}{\sum_{s'} \widetilde{N}^p_k(s,a,s') + \alpha c_{k,3}(\varepsilon_0,\delta_0,\delta)} - p(s'|s,a) \right| \tag{34}$$

$$\leq \underbrace{\sum_{s' \in \mathcal{S}} \left| \frac{N^p_k(s,a,s')}{\sum_{s'} \widetilde{N}^p_k(s,a,s') + \alpha c_{k,3}(\varepsilon_0,\delta_0,\delta)} - p(s' \mid s,a) \right|}_{\textcircled{1}} + \underbrace{\sum_{s' \in \mathcal{S}} \left| \frac{\widetilde{N}^p_k(s,a,s') - N^p_k(s,a,s')}{\sum_{s'} \widetilde{N}^p_k(s,a,s') + \alpha c_{k,3}(\varepsilon_0,\delta_0,\delta)} \right|}_{\textcircled{2}} \tag{35}$$

Recall that $\sum_{s'} \widetilde{N}_k^p(s,a,s') = \widetilde{N}_k^p(s,a)$ and $\sum_{s'} N_k^p(s,a,s') = N_k^p(s,a)$. Therefore:

$$
\begin{aligned}
① &= \sum_{s' \in \mathcal{S}} \left| \frac{N_k^p(s,a,s')}{N_k^p(s,a)} \frac{N_k^p(s,a)}{\widetilde{N}_k^p(s,a) + \alpha c_{k,3}(\varepsilon_0,\delta_0,\delta)} - p(s' \mid s,a) \right| \\
&= \sum_{s'} \left| \left(\frac{N_k^p(s,a,s')}{N_k^p(s,a)} - p(s'|s,a) \right) \underbrace{\frac{N_k^p(s,a)}{\widetilde{N}_k^p(s,a) + \alpha c_{k,3}(\varepsilon_0,\delta_0,\delta)}}_{>0} + p(s'|s,a) \left(\frac{N_k^p(s,a)}{\widetilde{N}_k^p(s,a) + \alpha c_{k,3}(\varepsilon_0,\delta_0,\delta)} - 1 \right) \right| \\
&\le \sum_{s'} \left(p(s'|s,a) \frac{(\alpha+1) c_{k,3}(\varepsilon_0,\delta_0,\delta)}{\widetilde{N}_k^p(s,a) + \alpha c_{k,3}(\varepsilon_0,\delta_0,\delta)} \right) + \frac{N_k^p(s,a)}{\widetilde{N}_k^p(s,a) + \alpha c_{k,3}(\varepsilon_0,\delta_0,\delta)} ||\overline{p}_k(\cdot|s,a) - p(\cdot|s,a)||_1 \\
&\overset{(a)}{\le} \frac{(\alpha+1) c_{k,3}(\varepsilon_0,\delta_0,\delta)}{\widetilde{N}_k^p(s,a) + \alpha c_{k,3}(\varepsilon_0,\delta_0,\delta)} + \frac{N_k^p(s,a)}{\widetilde{N}_k^p(s,a) + \alpha c_{k,3}(\varepsilon_0,\delta_0,\delta)} \frac{L(\delta)}{\sqrt{N_k^p(s,a)}} \\
&\le \frac{(\alpha+1) c_{k,3}(\varepsilon_0,\delta_0,\delta)}{\widetilde{N}_k^p(s,a) + \alpha c_{k,3}(\varepsilon_0,\delta_0,\delta)} + \frac{L(\delta)}{\sqrt{\widetilde{N}_k^p(s,a) + \alpha c_{k,3}(\varepsilon_0,\delta_0,\delta)}} \qquad (36)
\end{aligned}
$$

where $L(\delta) = \sqrt{14 S \ln(4\pi^2 SAHk^3/3\delta)}$ and inequality (a) follows from the Weissman inequality (Weissman et al., 2003), and we have again used the fact that the inequalities in Def. 2 hold.

In addition,we have:

$$
② \le \sum_{s' \in \mathcal{S}} \frac{|c_{k,4}(\varepsilon_0,\delta_0,\delta)|}{\widetilde{N}_k^p(s,a) + \alpha c_{k,3}(\varepsilon_0,\delta_0,\delta)} = \frac{S c_{k,4}(\varepsilon_0,\delta_0,\delta)}{\widetilde{N}_k^p(s,a) + \alpha c_{k,3}(\varepsilon_0,\delta_0,\delta)} \qquad (37)
$$

Hence putting together Eq. (37) and Eq. (36), we have:

$$
\begin{aligned}
\sum_{s' \in \mathcal{S}} \left| \frac{\widetilde{N}_k^p(s,a,s')}{\widetilde{N}_k^p(s,a) + \alpha c_{k,3}(\varepsilon_0,\delta_0,\delta)} - p(s' \mid s,a) \right| &\le \frac{S c_{k,4}(\varepsilon_0,\delta_0,\delta) + (\alpha+1) c_{k,3}(\varepsilon_0,\delta_0,\delta)}{\widetilde{N}_k^p(s,a) + \alpha c_{k,3}(\varepsilon_0,\delta_0,\delta)} \\
&\quad + \frac{L(\delta)}{\sqrt{\widetilde{N}_k^p(s,a) + \alpha c_{k,3}(\varepsilon_0,\delta_0,\delta)}} \qquad (38)
\end{aligned}
$$

which gives the result.

□

B Regret Lower Bound (Proof of Thm. 1)

Let's consider the following MDP for a given number of states S and actions A. The initial state 0 has A actions which deterministically lead the next state. The MDP is a tree with A children for each node and exactly $S-2$ states.

We denote by $x_1, \cdots, x_L$ the leaves of this tree. There exists a unique action $a^\star$ and leaf $x_{i^\star}$ such that: $\mathbb{P}(+ \mid x_{i^\star}, a^\star) = 1/2 + \Delta$ for a chosen Δ. Each other leaf transitions with equal probability to two states $+$ and $-$ where each has a reward of 1 and 0. All other states have a reward of 0 and every other transition is deterministic.

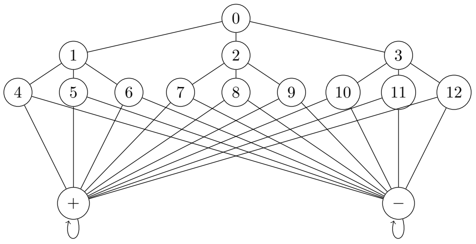

Figure 2: Example of an MDP described in this section with $S = 15$ and $A = 3$

Once the agent arrives at $+$ or $-$, they stay there until the end of the episode. In addition, we assume that $H \geq 2\ln(S-2)/\ln(A) + 2$. Let $d > 0$ be the depth of the tree, i.e., the depth of the tree with $S-2$ nodes is $d-1$ and nodes $+$,$-$ are at depth d. Then leaves $x_1, \ldots, x_L$ are at depth either $d-1$ or $d-2$. Without loss of generality we assume that all $x_1, \ldots, x_L$ are at depth $d-1$, i.e., the number of leaves is $L = A^{d-1} \geq (S-2)/2$ that is to say the tree without the nodes $+$ and $-$ is a perfect A-ary tree. In the general case we have that $L \geq (S-2)/2$.

For a policy π, the value function can be written:

$$V^\pi(0) = (H-d)\mathbb{P}(s_d = +) = (H-d)(1/2 + \Delta \mathbb{P}\left(s_{d-1} = x_{i^\star}, a_{d-1} = a^\star\right)) \tag{39}$$

Thus the regret can be written as:

$$R(K, I) = (H-d)\Delta\Big(K - \underbrace{\sum_{k=1}^{K} \mathbb{P}\left(s_{k,d-1} = x_{i^\star}, a_{k,d-1} = a^\star\right)}_{:=\mathbb{E}(T(K,I))}\Big) \tag{40}$$

where $I = (x_{i^\star}, a^\star)$ is the optimal state action pair and we define $T(K, I)$ as:

$$T(K, I) = \sum_{k=1}^{K} \mathbb{1}_{\{s_{k,d-1} = x_{i^\star}, a_{k,d-1} = a^\star\}}. \tag{41}$$

$T(K, I)$ is a function of the history observed by the algorithm. Since we are in the LDP setting, we can write this history as:

$$\mathcal{M}(\mathcal{H}_K) = \{\mathcal{M}(X_l) \mid l \leq K\} \tag{42}$$

where $X_l = \{(s_{l,h}, a_{l,h}, r_{l,h}) \mid h \leq H\}$ is the trajectory observed by the user for episode l and $\mathcal{M}$ is a privacy mechanism which maintains ε-LDP. Thus $T(K, I)$ is a function of $\mathcal{M}(\mathcal{H}_K)$. By Lem. A.1 in (Auer et al., 2002):

$$\mathbb{E}(T(K, I)) \leq \mathbb{E}_0(T(K, I)) + K\sqrt{\mathrm{KL}\Big(\mathbb{P}_0(\mathcal{M}(\mathcal{H}_K)) \mid\mid \mathbb{P}(\mathcal{M}(\mathcal{H}_K))\Big)} \tag{43}$$

where $\mathbb{E}_0$ is the expectation when $\Delta = 0$. But, because we can see $T(K, I)$ as a function on the history only, thus we can use Exercise 14.4 in (Lattimore and Szepesvári, 2020) which states that for any random variable

$Y : \Omega \to [a, b]$ with $(\Omega, \mathcal{F})$ a measurable space, $a < b$ and two distributions P and Q on $\mathcal{F}$, then:

$$\left| \int_{w \in \Omega} Y(w) dP(w) - \int_{w \in \Omega} Y(w) dQ(w) \right| \leq (b - a) \sqrt{\frac{\mathrm{KL}(P||Q)}{2}} \tag{44}$$

In our case the random variable Y is the combination of $T(K, I)$ and the privacy mechanism $\mathcal{M}$ so we have:

$$\mathbb{E}(T(K, I)) \leq \mathbb{E}_0(T(K, I)) + K \sqrt{\mathrm{KL}\Big(\mathbb{P}_0(\mathcal{H}_K) \mid\mid \mathbb{P}(\mathcal{H}_K)\Big)} \tag{45}$$

Putting together Eq. (43) and (45) we have:

$$\mathbb{E}(T(K, I)) \leq \mathbb{E}_0(T(K, I)) + K \min \left\{ \underbrace{\sqrt{\mathrm{KL}\Big(\mathbb{P}_0(\mathcal{M}(\mathcal{H}_K)) \mid\mid \mathbb{P}(\mathcal{M}(\mathcal{H}_K))\Big)}}_{\textcircled{1}}, \underbrace{\sqrt{\mathrm{KL}\Big(\mathbb{P}_0(\mathcal{H}_K) \mid\mid \mathbb{P}(\mathcal{H}_K)\Big)}}_{\textcircled{2}} \right\} \tag{46}$$

Bounding ①. Now we bound the KL-divergence between the two measures for the history. Using the chain rule we have:

$$\mathrm{KL}\left(\mathbb{P}_0(\mathcal{M}(\mathcal{H}_K)) \mid\mid \mathbb{P}(\mathcal{M}(\mathcal{H}_K))\right) = \sum_{k=1}^{K} \mathbb{E}_{\mathcal{H}_{k-1} \sim \mathbb{P}_0} \left(\mathrm{KL}\left(\mathbb{P}_0(\cdot|\mathcal{M}(\mathcal{H}_{k-1})) \mid\mid \mathbb{P}(\cdot|\mathcal{M}(\mathcal{H}_{k-1}))\right)\right) \tag{47}$$

But because $\mathcal{M}$ is an ε-LDP mechanism, Thm. 1 in (Duchi et al., 2013) ensures that:

$$\mathrm{KL}\left(\mathbb{P}_0(\cdot|\mathcal{M}(\mathcal{H}_{k-1})) \mid\mid \mathbb{P}(\cdot|\mathcal{M}(\mathcal{H}_{k-1}))\right) \leq 4(\exp(\varepsilon) - 1)^2 \mathrm{KL}\left(\mathbb{P}_0(\cdot|\mathcal{H}_{k-1}) \mid\mid \mathbb{P}(\cdot|\mathcal{H}_{k-1})\right) \tag{48}$$

Additionally, the KL-divergence can be written as:

$$\mathrm{KL}\left(\mathbb{P}_0(\cdot|\mathcal{H}_{k-1}) \mid\mid \mathbb{P}(\cdot|\mathcal{H}_{k-1})\right) = \sum_{h=1}^{H} \mathbb{E}_{X_k \sim \mathbb{P}_0} \left(\ln \left(\frac{\mathbb{P}_0(s_{k,h}, a_{k,h}, r_{k,h} \mid \mathcal{H}_{k-1}, (s_{k,j}, a_{k,j}, r_{k,j})_{j \leq h-1})}{\mathbb{P}(s_{k,h}, a_{k,h}, r_{k,h} \mid \mathcal{H}_{k-1}, (s_{k,j}, a_{k,j}, r_{k,j})_{j \leq h-1})} \right) \right) \tag{49}$$

where $X_k = \{(s_{k,h}, a_{k,h}, r_{k,h}) \mid h \leq H\}$ is a trajectory sampled from the MDP with the transitions distributed according to $\mathbb{P}_0$ and for each step h, $s_{k,h}$ is a state, $a_{k,h}$ an action and $r_{k,h}$ the reward associated with $(s_{k,h}, a_{k,h})$.

Therefore for a step $h \geq 1$,

$$\begin{aligned} \ln\left(\mathbb{P}_0(s_{k,h}, a_{k,h}, r_{k,h} \mid \mathcal{H}_{k-1}, (s_{k,j}, a_{k,j}, r_{k,j})_{j \leq h-1})\right) &= \ln\left(\mathbb{P}_0(s_{k,h} \mid \mathcal{H}_{k-1}, (s_{k,j}, a_{k,j}, r_{k,j})_{j \leq h-1})\right) \\ &\quad + \ln\left(\mathbb{P}_0(a_{k,h} \mid \mathcal{H}_{k-1}, (s_{k,j}, a_{k,j}, r_{k,j})_{j \leq h-1}, s_{k,h})\right) \\ &\quad + \ln\left(\mathbb{P}_0(r_{k,h} \mid \mathcal{H}_{k-1}, (s_{k,j}, a_{k,j}, r_{k,j})_{j \leq h-1}, s_{k,h}, a_{k,h})\right) \end{aligned}$$

By the Markov property of the environment:

$$\ln\left(\mathbb{P}_0(s_{k,h} \mid \mathcal{H}_{k-1}, (s_{k,j}, a_{k,j}, r_{k,j})_{j \leq h-1})\right) = \ln\left(\mathbb{P}_0(s_{k,h} \mid s_{k,h-1}, a_{k,h-1})\right) \tag{50}$$

Also, since the reward only depends on the current state-action pair:

$$\ln\left(\mathbb{P}_0(r_{k,h} \mid \mathcal{H}_{k-1}, (s_{k,j}, a_{k,j}, r_{k,j})_{j \leq h-1}, s_{k,h}, a_{k,h})\right) = \ln\left(\mathbb{P}_0(r_{k,h} \mid s_{k,h}, a_{k,h})\right). \tag{51}$$

The same results holds for $\mathbb{P}$, thus:

$$\mathrm{KL}\left(\mathbb{P}_0(\cdot|\mathcal{H}_{k-1}) \mid\mid \mathbb{P}(\cdot|\mathcal{H}_{k-1})\right) = \sum_{h=1}^{H} \mathbb{E}_{X_k \sim \mathbb{P}_0} \Bigg(\ln \left(\frac{\mathbb{P}_0(s_{k,h} \mid s_{k,h-1}, a_{k,h-1})}{\mathbb{P}_0(s_{k,h} \mid s_{k,h-1}, a_{k,h-1})} \right) \tag{52}$$

$$+ \ln \left(\frac{\mathbb{P}_0(a_{k,h} \mid \mathcal{H}_{k-1}, (s_{k,j}, a_{k,j}, r_{k,j})_{j \leq h-1}, s_{k,h})}{\mathbb{P}(a_{k,h} \mid \mathcal{H}_{k-1}, (s_{k,j}, a_{k,j}, r_{k,j})_{j \leq h-1}, s_{k,h})} \right) + \ln \left(\frac{\mathbb{P}_0(r_{k,h} \mid s_{k,h}, a_{k,h})}{\mathbb{P}(r_{k,h} \mid s_{k,h}, a_{k,h})} \right) \Bigg) \tag{53}$$

But for $\mathbb{P}$ and $\mathbb{P}_0$ the rewards are distributed accordingly to the same distribution hence $\ln\left(\frac{\mathbb{P}_0(r_{k,h}|s_{k,h},a_{k,h})}{\mathbb{P}(r_{k,h}|s_{k,h},a_{k,h})}\right) = 0$ for each $h \leq H$. Also, the action taken at each step depends only the history of data and the current state, thus

$\ln\left(\frac{\mathbb{P}_0(a_{k,h}|\mathcal{H}_{k-1},(s_{k,j},a_{k,j},r_{k,j})_{j\leq h-1})}{\mathbb{P}(a_{k,h}|\mathcal{H}_{k-1},(s_{k,j},a_{k,j},r_{k,j})_{j\leq h-1})}\right) = 0$. Lastly, transition dynamics between $\mathbb{P}$ and $\mathbb{P}_0$ only differ when at step $d-1$ thus for all $h \neq d-1$, $\ln\left(\frac{\mathbb{P}_0(s_{k,h}|s_{k,h-1},a_{k,h-1})}{\mathbb{P}_0(s_{k,h}|s_{k,h-1},a_{k,h-1})}\right) = 0$. Overall, we get:

$$\mathrm{KL}\left(\mathbb{P}_0(\cdot|\mathcal{H}_{k-1}) \mid\mid \mathbb{P}(\cdot|\mathcal{H}_{k-1})\right) = \sum_{l=1}^{L}\sum_{a=1}^{A}\sum_{j\in\{-,+\}} \mathbb{E}_{X_k\sim\mathbb{P}_0}\left(\ln\left(\frac{\mathbb{P}_0(j \mid x_l, a)}{\mathbb{P}(j \mid x_l, a)}\right)\mathbb{1}_{\{s_{k,d-1}=x_l, a_{k,d-1}=a, s_{k,d}=j\}}\right)$$

Finally, for $j \in \{-,+\}$, $x_l \neq x_{i^\star}$ and $a \neq a^\star$, $\mathbb{P}(j \mid x_l, a) = \mathbb{P}_0(j \mid x_l, a)$. Hence,

$$\mathrm{KL}\left(\mathbb{P}_0(\cdot|\mathcal{H}_{k-1}) \mid\mid \mathbb{P}(\cdot|\mathcal{H}_{k-1})\right) = \frac{1}{2}\ln\left(\frac{1}{1-4\Delta^2}\right)\mathbb{E}_{X_k\sim\mathbb{P}_0}\left(\mathbb{1}_{\{s_{k,d-1}=x_{i^\star}, a_{k,d-1}=a^\star,\}}\right) \tag{54}$$

where we have used $\mathbb{P}(+ \mid x_{i^\star}, a^\star) = \frac{1}{2}+\Delta$, $\mathbb{P}_0(+ \mid x_{i^\star}, a^\star) = \frac{1}{2}$, $\mathbb{P}(- \mid x_{i^\star}, a^\star) = \frac{1}{2}-\Delta$ and $\mathbb{P}_0(- \mid x_{i^\star}, a^\star) = \frac{1}{2}$. Therefore summing over the episodes, we get:

$$\begin{aligned}\mathrm{KL}\Big(\mathbb{P}_0(\mathcal{M}(\mathcal{H}_K)) \mid\mid \mathbb{P}(\mathcal{M}(\mathcal{H}_K))\Big) &\leq 2(\exp(\varepsilon)-1)^2\ln\left(\frac{1}{1-4\Delta^2}\right)\sum_{k=1}^{K}\mathbb{P}_0\left(s_{k,d-1}=x_{i^\star}, a_{k,d-1}=a^\star\right) \\ &= 2(\exp(\varepsilon)-1)^2\ln\left(\frac{1}{1-4\Delta^2}\right)\mathbb{E}_0(T(K,I))\end{aligned} \tag{55}$$

Bounding ②. Using again the chain rule of the KL-divergence, we have that:

$$\mathrm{KL}\left(\mathbb{P}_0(\mathcal{H}_K) \mid\mid \mathbb{P}(\mathcal{H}_K)\right) = \sum_{k=1}^{K}\mathbb{E}_{\mathcal{H}_{k-1}\sim\mathbb{P}_0}\left(\mathrm{KL}\left(\mathbb{P}_0(\cdot|\mathcal{H}_{k-1}) \mid\mid \mathbb{P}(\cdot|\mathcal{H}_{k-1})\right)\right) \tag{56}$$

Therefore, using Eq. (54), we have:

$$\begin{aligned}\mathrm{KL}\left(\mathbb{P}_0(\mathcal{H}_K) \mid\mid \mathbb{P}(\mathcal{H}_K)\right) &= \sum_{k=1}^{K}\mathbb{E}_{\mathcal{H}_{k-1}\sim\mathbb{P}_0}\left(\frac{1}{2}\ln\left(\frac{1}{1-4\Delta^2}\right)\mathbb{E}_{X_k\sim\mathbb{P}_0}\left(\mathbb{1}_{\{s_{k,d-1}=x_{i^\star}, a_{k,d-1}=a^\star,\}}\right)\right) \\ &= \frac{1}{2}\ln\left(\frac{1}{1-4\Delta^2}\right)\mathbb{E}_0(T(K,I))\end{aligned} \tag{57}$$

Finishing the proof. Hence using Eq. (55) and Eq. (57) in Eq. (46):

$$\mathbb{E}(T(K,I)) \leq \mathbb{E}_0(T(K,I)) + K\min\left\{\sqrt{2}(\exp(\varepsilon)-1), \frac{1}{\sqrt{2}}\right\}\sqrt{\mathbb{E}_0(T(K,I))\ln\left(\frac{1}{1-4\Delta^2}\right)} \tag{58}$$

Now, let's assume that $I = (x_{i^\star}, a^\star)$ is distributed uniformly over $\{x_1, \ldots, x_L\} \times [\![1, A]\!]$. That is to say, that the leaf $i^\star \sim \mathcal{U}([\![1, L]\!])$ and given the realization of $i^\star$, $a^\star$ is drawn uniformly in the action set of node $x_{i^\star}$ i.e., $a^\star \sim \mathcal{U}([\![1, A]\!])$. We denote the expectation over the random variable $(x_{i^\star}, a^\star)$ by $\mathbb{E}_I$. It then holds that:

$$\mathbb{E}_I\mathbb{E}_0(T(K,I)) = \mathbb{E}_0\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{a=1}^{A}\frac{1}{LA}\mathbb{1}_{\{s_{k,d-1}=s, a_{k,d-1}=a\}} = \frac{K}{LA} \tag{59}$$

Therefore thanks to Jensen's inequality the regret is lower-bounded by:

$$\mathbb{E}_I R(K,I) \geq (H-d)\Delta K\left(1 - \frac{1}{LA} - \min\left\{\sqrt{2}(\exp(\varepsilon)-1), \frac{1}{\sqrt{2}}\right\}\sqrt{\frac{K}{LA}\ln\left(1+\frac{4\Delta^2}{1-4\Delta^2}\right)}\right) \tag{60}$$

Therefore for $LA \geq 2$, $K \geq \frac{LA}{\min\{8(\exp(\varepsilon)-1),4\}^2}$ and choosing $\Delta = \sqrt{\frac{LA}{K}} \times \frac{1}{16\sqrt{2}\min\left\{(\exp(\varepsilon)-1),\frac{1}{2}\right\}}$ we get that:

$$\min\left\{\sqrt{2}(\exp(\varepsilon)-1), \frac{1}{\sqrt{2}}\right\}\sqrt{\frac{K}{LA}\ln\left(1+\frac{4\Delta^2}{1-4\Delta^2}\right)} \leq \frac{1}{4}$$

Hence:

$$\max_{I\in\{x_1,\ldots,x_L\}\times[\![1,A]\!]} R(K,I) \geq \mathbb{E}_I R(K,I) \geq \frac{(H-d)\sqrt{KLA}}{64\min\left\{(\exp(\varepsilon)-1),\frac{1}{2}\right\}} \tag{61}$$

And because I is a finite random variable there exist $I^\star$ such that $\max_{I\in\{x_1,\ldots,x_L\}\times[\![1,A]\!]} R(K,I) = R(K,I^\star)$.

$$R(K,I^\star) \geq \frac{(H-d)\sqrt{KLA}}{64\min\left\{(\exp(\varepsilon)-1),\frac{1}{2}\right\}} \tag{62}$$

Thus we have that there exists an MDP such that its frequentist regret is $\Omega\left(\frac{H\sqrt{SAK}}{\min\{1,\exp(\varepsilon)-1\}}\right)$.

C Regret Upper Bound (Proof of Theorem 2)

In this section, we prove Thm 2.

Theorem. *For any privacy mechanism $\mathcal{M}$ satisfying Def. 1 and Def. 2 with $\varepsilon_0 > 0$, $\delta_0 \geq 0$ and bounds $c_{k,1}(\varepsilon_0, \delta_0, .)$, $c_{k,2}(\varepsilon_0, \delta_0, .)$, $c_{k,3}(\varepsilon_0, \delta_0, .)$ and $c_{k,4}(\varepsilon_0, \delta_0, .)$, for any $\delta > 0$ the regret of* LDP-OBI *is bounded with probability at least $1 - \delta$ by:*

$$R(\text{LDP-OBI}, K) \leq \tilde{\mathcal{O}}\Bigg(\max\Bigg\{ HS\sqrt{AT}, SAH^2 c_{K,3}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 K^2}\right), H^2S^2Ac_{K,4}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 K^2}\right), \\ SAHc_{K,2}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 K^2}\right), SAHc_{K,1}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 K^2}\right)\Bigg\}\Bigg) \tag{63}$$

In addition, the combinaison of $\mathcal{M}$ and LDP-OBI *is $(\varepsilon_0, \delta_0)$-LDP.*

Good Event: Before proceeding the proof of the regret we define a good event under which all concentration inequalities holds with probability at least $1 - \delta$. First, we define the event that all inequalities from Def. 2 holds. Let:

$$L_{1,k} = \bigcap_{s,a} \left\{ \left| \widetilde{R}_k(s,a) - R_k(s,a) \right| \leq c_{k,1}(\varepsilon_0, \delta_0, 3\delta/2k^2\pi^2) \right\}$$
$$L_{2,k} = \bigcap_{s,a} \left\{ \left| \widetilde{N}^r_k(s,a) - N^r_k(s,a) \right| \leq c_{k,2}(\varepsilon_0, \delta_0, 3\delta/2k^2\pi^2) \right\}$$
$$L_{3,k} = \bigcap_{s,a} \left\{ \left| \sum_{s'} N^p_k(s,a,s') - \sum_{s‘} \widetilde{N}^p_k(s,a,s') \right| \leq c_{k,3}(\varepsilon_0, \delta_0, 3\delta/2k^2\pi^2) \right\}$$
$$L_{4,k} = \bigcap_{s,a,s'} \left\{ \left| N^p_k(s,a,s') - \widetilde{N}^p_k(s,a,s') \right| \leq c_{k,4}(\varepsilon_0, \delta_0, 3\delta/2k^2\pi^2) \right\}$$

then thanks to Def. 2 we have :

$$\mathbb{P}\left(\bigcup_{k=1}^{+\infty} L^c_{1,k} \cup L^c_{2,k} \cup L^c_{3,k} \cup L^c_{4,k} \right) \leq \sum_{k=1}^{+\infty} \frac{3\delta}{\pi^2 k^2} = \frac{\delta}{4} \tag{64}$$

In addition, for all $k \in \mathbb{N}^\star$, we can define $\overline{r}_k(s,a) = R_k(s,a)/N^r_k(s,a)$ and $\overline{p}_k = N^p_k(s,a,s')/\sum_{s'} N^p_k(s,a,s')$ as the empirical reward and transition probability computed with the non-private counters. Note that in this case $N_k(s,a) := N^r_k(s,a) = \sum_{s'} N^p_k(s,a,s')$. We also define $\overline{\beta}^r_k(\delta, s, a) = \sqrt{\frac{2\ln(1/\delta)}{N_k(s,a)}}$ and $\overline{\beta}^p_k(\delta, s, a) = \sqrt{\frac{14S\log(1/\delta)}{N_k(s,a)}}$. as the size of the confidence intervals using Hoeffding and Weissman inequalities. Thus, we get:

$$\mathbb{P}\left(\bigcup_{k=1}^{+\infty} \bigcup_{s,a} |\overline{r}_k(s,a) - r(s,a)| \geq \overline{\beta}^r_k(3\delta/4\pi^2 SAHk^3, s, a) \right) \tag{65}$$
$$\leq \sum_{k=1}^{+\infty} \sum_{s,a} \mathbb{P}\left(|\overline{r}_k(s,a) - r(s,a)| \geq \sqrt{\frac{2\ln(4\pi^3 SAHk^3/3\delta)}{N_k(s,a)}} \right) \tag{66}$$
$$\leq \sum_{k=1}^{+\infty} \sum_{s,a} \sum_{n=0}^{kH} \mathbb{P}\left(|\overline{r}_k(s,a) - r(s,a)| \geq \sqrt{\frac{2\ln(4\pi^2 SAHk^3/3\delta)}{n}} \right) \leq \sum_{k=1}^{+\infty} \sum_{s,a} \sum_{n=0}^{kH} \frac{3\delta}{4\pi^2 SHAk^3} \leq \frac{\delta}{8} \tag{67}$$

The same result holds for the transition dynamics, that is to say:

$$\mathbb{P}\left(\bigcup_{k=1}^{+\infty} \bigcup_{s,a} ||\overline{p}_k(\cdot|s,a) - p(\cdot|s,a)||_1 \geq \overline{\beta}^p_k(3\delta/4\pi^2 SAHk^3, s, a) \right) \leq \frac{\delta}{8} \tag{68}$$

Thus we can define the good event $\mathcal{G}_k$ by:

$$\mathcal{G}_k = \bigcap_{l=1}^{k-1} \bigcap_{i=1}^{4} L_{i,l} \cap \bigcap_{s,a} \left\{ |\overline{r}_l(s,a) - r(s,a)| \le \overline{\beta}_l^r(3\delta/(4\pi^2 SAHl^3), s, a) \right\} \tag{69}$$

$$\cap \left\{ ||\overline{p}_k(\cdot|s,a) - p(\cdot|s,a)||_1 \le \overline{\beta}_k^p(3\delta/(4\pi^2 SAHl^3), s, a) \right\} \tag{70}$$

Then $\mathbb{P}\left(\bigcap_{k=1}^{+\infty} \mathcal{G}_k\right) \ge 1 - \delta/2$ and $\mathcal{G}_k \subset \sigma(\mathcal{H}_k)$ (i.e., the history before episode k).

Optimism: For each episode k, the value function $V_{k,1}$ computed by LDP-OBI is optimistic, that is to say: $V_{k,h}(s) \ge V_h^\star(s)$ for any h and state s. We sum up this with the following lemma:

Lemma 1. *For any episode $k \in [\![1,k]\!]$, the value function $V_{k,1}$ computed by running Alg. 2 is such that with probability $1-\delta$:*

$$\forall s \in \mathcal{S}, h \in [\![1,H]\!] \qquad V_{k,h}(s) \ge V_h^\star(s) \tag{71}$$

Proof. Fix an episode k then we proceed by backward induction conditioned on the event $\mathcal{G}_k$:

- For $h = H$, we have for any state s and action a:

$$V_{k,H}(s) \ge Q_{k,H}(s,a) \ge \widetilde{r}_k(s,a) + \beta_k^r(s,a) \ge r(s,a) \text{ thanks to Eq. (9)} \tag{72}$$

- For $h < H$ when the property is true for $h+1$, we get for any state-action (s,a):

$$\begin{aligned} V_{k,h}(s) \ge Q_{k,h}(s,a) &= \widetilde{r}_k(s,a) + \beta_k^r(s,a) + \widetilde{p}_k(\cdot|s,a)^\intercal V_{k,h+1} + H\beta_k^p(s,a) && (73) \\ &\ge r(s,a) + p(\cdot|s,a)^\intercal V_{k,h+1} \ge Q_h^\star(s,a) && (74) \end{aligned}$$

  where we used the fact that $\|(\widetilde{p}_k(\cdot|s,a) - p(\cdot|s,a))^\intercal V_{k,h+1}\| \le \|\widehat{p}_k(\cdot|s,a) - p(\cdot|s,a)\|_1 \|V_{k,h+1}\|_\infty \le H\beta_k^p(s,a)$ and the inductive hypothesis.

□

Regret Decomposition: We are now ready to analyze the regret of LDP-OBI. Consider an episode k, then, conditioned on $\mathcal{G}_k$:

$$\begin{aligned} V_1^\star(s_{k,1}) - V_1^{\pi_k} s_{k,1} &\le V_{k,1}(s_{k,1}) - V_1^{\pi_k}(s_{k,1}) && (75) \\ &\le \widetilde{r}_k(s_{k,1},a_{k,1}) + \beta_k^r(s_{k,1},a_{k,1}) - r(s_{k,1},a_{k,1}) + \widetilde{p}_k(\cdot|s,a)^\intercal V_{k,2} - p(\cdot|s,a)^\intercal V_2^{\pi_k} + H\beta_k^p(s_{k,1},a_{k,1}) && (76) \\ &\le \underbrace{(p(\cdot|s,a)^\intercal(V_{k,2} - V_2^{\pi_k}) - (V_{k,2}(s_{k,2}) - V_2^{\pi_k}(s_{k,2}))}_{:=\eta_{k,1}} + V_{k,2}(s_{k,2}) - V_2^{\pi_k}(s_{k,2}) + 2H\beta_k^p(s_{k,1},a_{k,1}) && (77) \\ &\quad + 2\beta_k^r(s_{k,1},a_{k,1}) \\ &= \sum_{h=1}^{H-1} \eta_{k,h} + 2\sum_{h=1}^{H} \beta_k^r(s_{k,h},a_{k,h}) + H\beta_k^p(s_{k,h},a_{k,h}) && (78) \end{aligned}$$

Then, observe that $(\eta_{k,h})_{k,h}$ is a Martingale Difference Sequence with respect to the history before episode k and thanks to Azuma-Hoeffding inequality we have that with probability at least $1-\delta/2$, $\sum_{k=1}^K \sum_{h=1}^{H-1} \eta_{k,h} \le 2H\sqrt{KH\ln(2/\delta)}$. Therefore, we have with probability at least $1-\delta$:

$$R(\text{LDP-OBI}, K) \le 2\sum_{k=1}^{K}\sum_{h=1}^{H} \beta_k^r(s_{k,h},a_{k,h}) + H\beta_k^p(s_{k,h},a_{k,h}) + \underbrace{2H\sqrt{T\ln(2/\delta)}}_{\text{MDS error term}} \tag{79}$$

Let $\nu_k(s,a) = \sum_{h=1}^{H} \mathbb{1}_{\{s_{k,h}=s, a_{k,h}=a\}}$. Then summing over the reward bonus and using the fact that $\alpha > 1$, we get:

$$\begin{aligned}\sum_{k=1}^{K}\sum_{h=1}^{H} \beta_k^r(s_{k,h}, a_{k,h}) = \sum_{s,a,k} \frac{\nu_k(s,a) L_{k,r}}{\sqrt{\widetilde{N}_k^r(s,a) + \alpha c_{k,2}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 k^2}\right)}} &+ \sum_{s,a,k} \frac{\nu_k(s,a)(\alpha+1) c_{k,2}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 k^2}\right)}{\alpha c_{k,2}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 k^2}\right) + \widetilde{N}_k^r(s,a)} \\ &+ \sum_{s,a,k} \frac{\nu_k(s,a) c_{k,1}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 k^2}\right)}{\alpha c_{k,2}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 k^2}\right) + \widetilde{N}_k^r(s,a)}\end{aligned} \tag{80}$$

where $L_{k,r} = \sqrt{2 \ln\left(\frac{4\pi^2 SAHk^3}{3\delta}\right)}$. Then,

$$\begin{aligned}(80) \leq \sum_{s,a,k} \frac{\nu_k(s,a) L_{k,r}}{\sqrt{N_k(s,a) + (\alpha-1) c_{k,2}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 k^2}\right)}} &+ \frac{\nu_k(s,a)(\alpha+1) c_{k,2}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 k^2}\right)}{(\alpha-1) c_{k,2}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 k^2}\right) + N_k(s,a)} \qquad (81) \\ &+ \sum_{s,a,k} \frac{\nu_k(s,a) c_{k,1}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 k^2}\right)}{(\alpha-1) c_{k,2}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 k^2}\right) + N_k(s,a)} \\ \leq \sum_{s,a,k} \frac{\nu_k(s,a) L_{K,r}}{\sqrt{N_k(s,a)}} + \left((\alpha+1) c_{K,2}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 K^2}\right) + c_{K,1}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 K^2}\right)\right) &\sum_{k,s,a} \frac{\nu_k(s,a)}{N_k(s,a)} \qquad (82) \\ \leq 2\left((\alpha+1) c_{K,2}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 K^2}\right) + c_{K,1}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 K^2}\right)\right) &SA(\ln(2TSA) + H) \qquad (83) \\ &+ \sqrt{6\ln\left(14SAT/\delta\right)}\left(\sqrt{2SAT} + HSA\right)\end{aligned}$$

where the last inequality comes from Lem. 19 in (Jaksch et al., 2010). For the sum of the bonus on the transition dynamics we have that:

$$\begin{aligned}\sum_{k=1}^{K}\sum_{h=1}^{H} H\beta_k^p(s_{k,h}, a_{k,h}) = \sum_{s,a,k} \frac{H\nu_k(s,a) L_{k,p}}{\sqrt{\widetilde{N}_k^p(s,a) + \alpha c_{k,3}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 k^2}\right)}} &+ \sum_{s,a,k} \frac{HS\nu_k(s,a) c_{k,4}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 k^2}\right)}{\alpha c_{k,3}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 k^2}\right) + \widetilde{N}_k^p(s,a)} \\ &+ \sum_{s,a,k} \frac{H\nu_k(s,a)(\alpha+1) c_{k,3}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 k^2}\right)}{\alpha c_{k,3}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 k^2}\right) + \widetilde{N}_k^p(s,a)}\end{aligned} \tag{84}$$

where $L_{k,p} = \sqrt{14S \ln\left(\frac{4\pi^2 SAHk^3}{3\delta}\right)}$. Then,

$$\begin{aligned}(84) \leq \sum_{s,a,k} \frac{H\nu_k(s,a) L_{k,p}}{\sqrt{N_k(s,a) + (\alpha-1) c_{k,3}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 k^2}\right)}} &+ \sum_{s,a,k} \frac{H\nu_k(s,a)(\alpha+1) c_{k,3}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 k^2}\right)}{(\alpha-1) c_{k,3}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 k^2}\right) + N_k(s,a)} \qquad (85) \\ &+ \sum_{k,s,a} \frac{HSc_{k,4}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 k^2}\right)}{(\alpha-1) c_{k,3}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 k^2}\right) + N_k(s,a)} \\ \leq \sum_{s,a,k} \frac{H\nu_k(s,a) L_{K,p}}{\sqrt{N_k(s,a)}} + \left((\alpha+1) c_{K,3}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 K^2}\right) + Sc_{K,4}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 K^2}\right)\right) &\sum_{k,s,a} \frac{H\nu_k(s,a)}{N_k(s,a)} \qquad (86) \\ \leq 2SAH\left((\alpha+1) c_{K,3}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 K^2}\right) + Sc_{K,4}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2 K^2}\right)\right) &(\ln(2TSA) + H) \qquad (87) \\ &+ H\sqrt{46S\ln\left(14SAT/\delta\right)}\left(\sqrt{2SAT} + HSA\right)\end{aligned}$$

where the last inequality comes from (Jaksch et al., 2010, Lem. 19) and (Fruit et al., 2020, Lem. 8). Hence putting everything together, we get that with probability $1-\delta$:

$$\begin{aligned}R(\text{LDP-OBI}, K) \leq 2SAH &\left((\alpha+1)c_{K,3}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2K^2}\right) + Sc_{K,4}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2K^2}\right)\right)(\ln(2TSA)+H)\\ &+H\sqrt{46S\ln(14SAT/\delta)}(\sqrt{2SAT}+HSA) + \sqrt{6\ln(14SAT/\delta)}(\sqrt{2SAT}+HSA)\\ &+2\left((\alpha+1)c_{K,2}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2K^2}\right) + c_{K,1}\left(\varepsilon_0, \delta_0, \frac{3\delta}{2\pi^2K^2}\right)\right)SA(\ln(2TSA)+H) + 2H\sqrt{T\ln(2/\delta)}\end{aligned}$$

In addition, because LDP-OBI has only access to the privatized data, that is to say it only uses the output of $\mathcal{M}(\{(s_{k,h}, a_{k,h}, r_{k,h})_{h\leq H}\})$ for each episode k, the LDP constraint is satsified as long as the privacy mechanism $\mathcal{M}$ satisfies Def. 1.

Note: the proof of this regret upper-bound relies on concentration inequalities more generally used in the average reward regret minimization setting. That is to say, we directly study the error between the estimated model and the true model, i.e., $|\widetilde{r}_k - r|$ and $||\widetilde{p}_k(. \mid s,a) - p(. \mid s,a)||_1$ for each s, a. In the non-private setting, it is possible to get a more refined regret using more precise concentration inequalities, mainly Bernstein inequality and other tools introduced in (Azar et al., 2017). However, in the private setting, using such results only leads to a gain in lower order terms and terms independent of ε while the technical derivations are much more intricate.

D Posterior Sampling for Local Differential Privacy

The Posterior Sampling for Reinforcement Learning algorithm (PSRL, Osband et al., 2013) is a Thompson Sampling based algorithm for Reinforcement Learning. It works by maintaining a Bayesian posterior distribution over MDPs. We focus on a particular instantiation of PSRL where for each state-action pair (s, a) we have an independent Gaussian prior for the reward distribution and a Dirichlet prior for the transition dynamics. With those priors, the posterior distributions are Normal-Gamma and Dirichlet distributions.

At the beginning of episode k and for a given pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, let $\alpha_k(s, a) \in (\mathbb{R}_+^\star)^S$ be such that the posterior distribution over the transition dynamics is $\mathrm{Dir}(\alpha_k(s, a))$. In addition, let's note $\mu_k(s, a) \in \mathbb{R}$, $\lambda_k(s, a) \in \mathbb{R}_+^\star$, $\nu_k(s, a) \in \mathbb{R}_+^\star$ and $\beta_k(s, a) \in \mathbb{R}_+^\star$ the parameters of the Normal-Gamma posterior distributions. Using standard results from Bayesian Learning we have that:

$$\forall s' \in \mathcal{S} \qquad \alpha_k(s, a) = \alpha_0(s, a) + N_k(s, a, s') \tag{88}$$

$$\lambda_k(s, a) = \lambda_0(s, a) + N_k(s, a) \tag{89}$$

$$\nu_k(s, a) = \nu_0(s, a) + \frac{N_k(s, a)}{2} \tag{90}$$

$$\mu_k(s, a) = \frac{\lambda_0(s, a)\mu_0(s, a) + N_k(s, a)\hat{R}_k(s, a)}{\lambda_0(s, a) + N_k(s, a)} \tag{91}$$

$$\beta_k(s, a) = \beta_0(s, a) + \frac{1}{2}\widehat{\mathrm{Var}}(R(s, a)) + \frac{N_k(s, a)\lambda_0(s, a)}{2(\lambda_0(s, a) + N_k(s, a))}\left(\hat{R}_k(s, a) - \mu_0(s, a)\right)^2 \tag{92}$$

where $\alpha_0, \mu_0, \lambda_0, \nu_0, \beta_0$ are prior parameters provided at the beginning of the algorithm. We denote by $N_k(s, a)$, the number of visits to the state-action pair (s, a), $N_k(s, a, s')$ the number visits to (s, a, s'), $\hat{R}_k(s, a)$ the average reward observed for (s, a) and $\widehat{\mathrm{Var}}(R(s, a))$ the empirical variance for (s, a).

At each episode k, PSRL samples an MDP from the posterior distributions, then computes the optimal policy and executes it in the true MDP. Osband et al. (2013) showed that the *Bayesian* regret of this algorithm is bounded by $\tilde{O}\left(HS\sqrt{AT}\right)$.

Locally Differentially Private Posterior Sampling for Reinforcement Learning: We now discuss how to adapt PSRL to ensure it is locally differentially private. Def. 1 states that LDP is ensured at the collection time of trajectories therefore it is enough for us to design a LDP posterior sampling algorithm which takes as input the trajectories outputted by a mechanism similar to Alg. 3. Here, we use the LDP mechanism to pertub the statistics used to define the parameters of the posterior distribution in PSRL. More precisely, we replace the aggregate counts in Eqs. 88-92 by noisy counts provided by an LDP mechanism. In order to do this, we need to modify the initial values of those parameters to guarantee they are non-negative.

In this appendix, we assume that the privacy-preserving mechanism $\mathcal{M}$ is such that for a given trajectory X, $\mathcal{M}(X) = (\widetilde{R}_X, \widetilde{R}_{2,X}, \widetilde{N}_X^r, \widetilde{N}_X^p)$ where $\widetilde{R}_X, \widetilde{R}_{2,X}, \widetilde{N}_X^r$ and $\widetilde{N}_X^p$ are noisy version of the following aggregate statistics:

$$R_X(s, a) = \sum_{h=1}^{H} r_h \mathbb{1}_{\{s_h = s, a_h = a\}}, \qquad R_{2,X}(s, a) = \sum_{h=1}^{H} r_h^2 \mathbb{1}_{\{s_h = s, a_h = a\}}$$

$$N_X^r(s, a) = \sum_{h=1}^{H} \mathbb{1}_{\{s_h = s, a_h = a\}}, \qquad N_X^p(s, a, s') = \sum_{h=1}^{H-1} \mathbb{1}_{\{s_h = s, a_h = a, s_{h+1} = s'\}}$$

In particular, $\widetilde{R}_X, \widetilde{N}_X^r$ and $\widetilde{N}_X^p$ are defined as for the optimistic algorithm in Section 3.1 and $\widetilde{R}_{2,X}$ is a privatized version of $R_{2,X}(s, a) = \sum_{h=1}^{H} r_h^2 \mathbb{1}_{\{s_h = s, a_h = a\}}$ for a trajectory X.

Algorithm 4: LDP-PSRL

```
Input: Initial values: $\alpha_0, \mu_0, \lambda_0, \nu_0$ and $\beta_0$
for episodes $k = 1, \dots, K$ do
    Draw empirical MDP, $\theta_k$ from the posterior and compute $\pi_k$ as the optimal policy for MDP $\theta_k$
    User $u_k$ executes policy $\pi_k$, collect trajectory $X_k = \{(s_{k,h}, a_{k,h}, r_{k,h}) \mid h \leq H\}$
    Update noisy counts with $(\widetilde{R}_{X_k}(s,a), \widetilde{R}_{X_k,2}(s,a), \widetilde{N}^r_{X_k}(s,a), \widetilde{N}^p_{X_k}(s,a))$ and posterior distribution
```

The posterior updates we use in LDP-PSRL are then:

$$\forall s' \in \mathcal{S} \qquad \widetilde{\alpha}_k(s,a) = \alpha_0(s,a) + \widetilde{N}^p_k(s,a,s') \tag{93}$$

$$\widetilde{\mu}_k(s,a) = \frac{\lambda_0(s,a)\mu_0(s,a) + \widetilde{R}_k(s,a)}{\lambda_0(s,a) + \widetilde{N}^r_k(s,a)} \tag{94}$$

$$\widetilde{\lambda}_k(s,a) = \lambda_0(s,a) + \widetilde{N}^r_k(s,a) \tag{95}$$

$$\widetilde{\nu}_k(s,a) = \tilde{\alpha}_0(s,a) + \frac{\widetilde{N}^r_k(s,a)}{2} \tag{96}$$

$$\widetilde{\beta}_k(s,a) = \beta_0(s,a) + \frac{\lambda_0(s,a)\widetilde{N}^r_k(s,a)\mu_0^2(s,a) - \widetilde{R}^2_k(s,a)}{2(\lambda_0(s,a) + \widetilde{N}^r_k(s,a))} + \frac{1}{2}\sum_{l \leq k-1} \widetilde{R}_{2,l} - \frac{\mu_0(s,a)\widetilde{R}_k(s,a)}{\lambda_0(s,a) + \widetilde{N}^r_k(s,a)} \tag{97}$$

In the following, we choose the Laplace mechanism as our privacy-preserving mechanism for LDP-PSRL. That is to say for each trajectory X, we add independent Laplace variables to $(R_X(s,a), R_{X,2}(s,a), N^r_X(s,a), N^p_X(s,a))$ with parameter $8H/\varepsilon$. Following the same argument outlined in the proof of Thm. 3, we can show that this privacy-preserving mechanism is $(\varepsilon, 0)$-LDP.

To ensure positivity, by concentration of Laplace variables we set the initial values of the parameters of the posterior distributions to:

$$\alpha_0(s,a,s') = \max\{\sqrt{KS}, \ln(6S^2A/\delta)\}\frac{\sqrt{8\ln(6S^2A/\delta)}}{\varepsilon_0} \tag{98}$$

$$\mu_0(s,a) = 0 \tag{99}$$

$$\lambda_0(s,a) = \max\{\sqrt{K}, \ln(6SA/\delta)\}\frac{\sqrt{8\ln(6SA/\delta)}}{\varepsilon_0} \tag{100}$$

$$\nu_0(s,a) = \max\{\sqrt{K}, \ln(6SA/\delta)\}\frac{\sqrt{8\ln(6SA/\delta)}}{\varepsilon_0} \tag{101}$$

$$\beta_0(s,a) = 5\max\{\sqrt{K}, \ln(6SA/\delta)\}\frac{\sqrt{8\ln(6SA/\delta)}}{\varepsilon_0} \tag{102}$$

where K is the total number of episodes.

The pseudocode of LDP-PSRL is reported in Alg. 4. While we have shown that this algorithm is ε-LDP and empirically outperforms optimistic approaches, we leave the regret analysis to future work.

E Other Privacy Preserving Mechanism:

We have shown in App. A.1 that the Laplace mechanism, Alg. 3, satisfies Def. 2. However it is not the only mechanism to do so. In this appendix we present the Gaussian and Bernoulli mechanism and show that these also satisfy Def. 2.

E.1 Gaussian Mechanism:

The Gaussian mechanism is a fundamental mechanism in the differential privacy literature (see e.g., Dwork and Roth, 2014). However, contrary to the Laplace mechanism the Gaussian mechanism can only guarantees (ε, δ)-LDP for $\delta > 0$. The mechanism is based on the same idea as the Laplace mechanism, that is to say it adds Gaussian noise to the result of a given computation on the input data. This noise is centered and the standard deviation $\sigma(\varepsilon, \delta)$ is $\frac{cH}{\epsilon_0}$.

Algorithm 5: Gaussian mechanism for LDP

```
Input: Trajectory: $X = \{(s_h, a_h, r_h) \mid h \leq H\}$, Privacy Parameter: $\varepsilon_0, c$,
Draw $(Y_{i,X}(s,a))_{(s,a)\in\mathcal{S}\times\mathcal{A}, i\leq 2}$ i.i.d $\mathcal{N}\left(0,\sigma^2\right)$ and $(Z_X(s,a,s'))_{(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}}$ i.i.d $\mathcal{N}\left(0,\sigma^2\right)$ and independent from
  $Y_{i,X}$ for $i \in \{1,2\}$ with $\sigma = cH/\varepsilon_0$
for $(s,a) \in \mathcal{S}\times\mathcal{A}$ do
    $\widetilde{R}_X(s,a) = \sum_{h=1}^{H} r_h \mathbb{1}_{\{s_h=s, a_h=a\}} + Y_{1,X}(s,a)$
    $\widetilde{N}^r_X(s,a) = \sum_{h=1}^{H} \mathbb{1}_{\{s_h=s, a_h=a\}} + Y_{2,X}(s,a)$
    for $s' \in \mathcal{S}$ do
        $\widetilde{N}^p_X(s,a,s') = \sum_{h=1}^{H-1} \mathbb{1}_{\{s_h=s, a_h=a, s_{h+1}=s'\}} + Z_X(s,a,s')$
Output: $(\widetilde{R}_X, \widetilde{N}^r_X, \widetilde{N}^p_X) \in \mathbb{R}^{S\times A}\times\mathbb{R}^{S\times A}\times\mathbb{R}^{S\times A\times S}$
```

In the following, we show that the Gaussian mechanism almost satisfies Def. 2. The Gaussian mechanism can not guarantee $(\varepsilon_0, 0)$-LDP for any $\varepsilon_0 > 0$, however we show that it satisfies the other necessary conditions, including (ε_0, δ)-LDP for any $\delta > 0$. First, the mechanism guarantees Local Differential Privacy for high enough noise.

Proposition 4. *For any $1 \geq \varepsilon_0 > 0$ and $\delta_0 > 0$ and parameter $c > 4\ln\left(\frac{24}{\delta_0}\right)$, the Gaussian mechanism, Alg. 5, is $(\varepsilon_0, \delta_0)$-LDP.*

Proof of Prop. 4: The proof is based on the proof presented in (Dwork and Roth, 2014). Similarly to the proof of Prop. 2 let's consider two trajectories $X = \{(s_h, a_h, r_h) \mid h \leq H\}$ and $X' = \{(s'_h, a'_h, r'_h) \mid h \leq H\}$ and also denote the output of the private randomizer $\mathcal{M}$ by $\mathcal{M}(X) = (\widetilde{R}_X, \widetilde{N}^r_X, \widetilde{N}^p_X)$ and $\mathcal{M}(X') = (\widetilde{R}_{X'}, \widetilde{N}^r_{X'}, \widetilde{N}^p_{X'})$.

For a given vector $r \in \mathbb{R}^{S\times A}$,

$$\frac{\mathbb{P}\left(\forall(s,a), \widetilde{R}_X(s,a) = r_{s,a} \mid X\right)}{\mathbb{P}\left(\forall(s,a), \widetilde{R}_{X'}(s,a) = r_{s,a} \mid X'\right)} = \prod_{s,a} \frac{\mathbb{P}\left(Y_{1,X}(s,a) = \sum_{h=1}^{H} r_h \mathbb{1}_{\{s_h=s,a_h=a\}} - r_{s,a} \mid X\right)}{\mathbb{P}\left(Y_{1,X'}(s,a) = \sum_{h=1}^{H} r'_h \mathbb{1}_{\{s'_h=s,a'_h=a\}} - r_{s,a} \mid X'\right)} \tag{103}$$

since the Gaussian distribution is symmetric. Then,

$$\begin{aligned}\prod_{s,a} & \frac{\mathbb{P}\left(Y_{1,X}(s,a) = \sum_{h=1}^{H} r_h \mathbb{1}_{\{s_h=s,a_h=a\}} - r_{s,a} \mid X\right)}{\mathbb{P}\left(Y_{1,X'}(s,a) = \sum_{h=1}^{H} r'_h \mathbb{1}_{\{s'_h=s,a'_h=a\}} - r_{s,a} \mid X'\right)} \\ &= \prod_{s,a} \exp\left(\frac{\left(\sum_{h=1}^{H} r_h \mathbb{1}_{\{s_h=s,a_h=a\}} - r_{s,a}\right)^2 - \left(\sum_{h=1}^{H} r'_h \mathbb{1}_{\{s'_h=s,a'_h=a\}} - r_{s,a}\right)^2}{2\sigma^2}\right)\end{aligned} \tag{104}$$

But, developping the squared term, we get:

$$\begin{aligned}\left(\sum_{h=1}^{H} r_h \mathbb{1}_{\{s_h=s,a_h=a\}} - r_{s,a}\right)^2 &= \left(\sum_{h=1}^{H} r_h \mathbb{1}_{\{s_h=s,a_h=a\}} - \sum_{h=1}^{H} r'_h \mathbb{1}_{\{s'_h=s,a'_h=a\}} + \sum_{h=1}^{H} r'_h \mathbb{1}_{\{s'_h=s,a'_h=a\}} - r_{s,a}\right)^2 \\ &= \left(\sum_{h=1}^{H} r_h \mathbb{1}_{\{s_h=s,a_h=a\}} - \sum_{h=1}^{H} r'_h \mathbb{1}_{\{s'_h=s,a'_h=a\}}\right)^2 + \left(\sum_{h=1}^{H} r'_h \mathbb{1}_{\{s'_h=s,a'_h=a\}} - r_{s,a}\right)^2 \\ &\quad + 2\left(\sum_{h=1}^{H} r_h \mathbb{1}_{\{s_h=s,a_h=a\}} - \sum_{h=1}^{H} r'_h \mathbb{1}_{\{s'_h=s,a'_h=a\}}\right)\left(\sum_{h=1}^{H} r'_h \mathbb{1}_{\{s'_h=s,a'_h=a\}} - r_{s,a}\right)\end{aligned}$$

Hence developping the squared term we get:

$$\begin{aligned}(104) = \prod_{s,a} \exp\Bigg(\frac{1}{2\sigma^2}\Bigg(\left(\sum_{h=1}^{H} r_h \mathbb{1}_{\{s_h=s,a_h=a\}} - \sum_{h=1}^{H} r'_h \mathbb{1}_{\{s'_h=s,a'_h=a\}}\right)^2 \\ - 2\left(\sum_{h=1}^{H} r_h \mathbb{1}_{\{s_h=s,a_h=a\}} - r'_h \mathbb{1}_{\{s'_h=s,a'_h=a\}}\right)\left(\sum_{h=1}^{H} r'_h \mathbb{1}_{\{s'_h=s,a'_h=a\}} - r_{s,a}\right)\Bigg)\Bigg).\end{aligned} \tag{105}$$

But, $\sum_{s,a}\left(\sum_{h=1}^{H} r_h \mathbb{1}_{\{s_h=s,a_h=a\}} - \sum_{h=1}^{H} r'_h \mathbb{1}_{\{s'_h=s,a'_h=a\}}\right)^2 \leq 2H^2$ because for each step h, $r_h \in [0,1]$. By the same reasonning, we have $\sum_{s,a}\left|\left(\sum_{h=1}^{H} r_h \mathbb{1}_{\{s_h=s,a_h=a\}} - r'_h \mathbb{1}_{\{s'_h=s,a'_h=a\}}\right)\sum_{h=1}^{H} r'_h \mathbb{1}_{\{s'_h=s,a'_h=a\}}\right| \leq H^2$. Therefore, we have:

$$\begin{aligned}(104) &\leq \exp\left(\frac{1}{2\sigma^2}\left(2\sum_{s,a}\left(\sum_{h=1}^{H} r_h \mathbb{1}_{\{s_h=s,a_h=a\}} - r'_h \mathbb{1}_{\{s'_h=s,a'_h=a\}}\right) r_{s,a} + 3H^2\right)\right) \\ &\leq \exp\left(\frac{1}{2\sigma^2}\left(2\sqrt{2}H\sqrt{\sum_{s,a} r_{s,a}^2} + 3H^2\right)\right)\end{aligned} \tag{106}$$

Therefore if $||r||_2 \leq \frac{\sigma^2\varepsilon_0}{3\sqrt{2}H} - \frac{3H}{2\sqrt{2}}$, Eq. (106) is bounded by $\exp(\varepsilon_0/3)$. To finish, let's partition $\mathbb{R}^{S\times A}$ in two subspaces $R_1 = \left\{x \in \mathbb{R}^{S\times A} \mid ||x||_2 \leq \frac{c^2H}{3\sqrt{2}\varepsilon_0} - \frac{3H}{2\sqrt{2}}\right\}$ and $R_2 = \left\{x \in \mathbb{R}^{S\times A} \mid ||x||_2 > \frac{c^2H}{3\sqrt{2}\varepsilon_0} - \frac{3H}{2\sqrt{2}}\right\}$ where we used the fact that $\sigma = cH/\varepsilon_0$ with c a constant to be chosen later. Then for $c^2 \geq 4\ln\left(\frac{3}{\delta_1}\right)$, for δ_1 to be chosen later, $\mathbb{P}(Y_{1,X} \in R_2) \leq \delta_1$ and $\mathbb{P}(Y_{1,X'} \in R_2) \leq \delta_1$. Thus for Eq. (103):

$$\begin{aligned}\mathbb{P}\left(\forall(s,a), \widetilde{R}_X(s,a) = r_{s,a} \mid X\right) &= \mathbb{P}\left(\forall(s,a), \widetilde{R}_X(s,a) = r_{s,a} \mid X\right)\mathbb{1}_{\{r-(\sum_{h=1}^{H} r_h \mathbb{1}_{\{s_h=s,a_h=a))\}})_{s,a}\in R_1\}} && (107)\\ &\quad + \mathbb{P}\left(\forall(s,a), \widetilde{R}_X(s,a) = r_{s,a} \mid X\right)\mathbb{1}_{\{r-(\sum_{h=1}^{H} r_h \mathbb{1}_{\{s_h=s,a_h=a\}})_{s,a}\in R_2\}} \\ &\leq \exp(\varepsilon_0/3)\mathbb{P}\left(\forall(s,a), \widetilde{R}_{X'}(s,a) = r_{s,a} \mid X'\right)\mathbb{1}_{\{r-(\sum_{h=1}^{H} r_h \mathbb{1}_{\{s_h=s,a_h=a))\}})_{s,a}\in R_1\}} && (108)\\ &\quad + \mathbb{P}(Y_{1,X} \in R_2) \\ &\leq \exp(\varepsilon_0/3)\mathbb{P}\left(\forall(s,a), \widetilde{R}_{X'}(s,a) = r_{s,a} \mid X'\right) + \delta_1 && (109)\end{aligned}$$

We get the same results for $\widetilde{N}^r$ and $\widetilde{N}^p$. Then, because $(Y_{i,X}(s,a))_{i\leq 2,(s,a)\in\mathcal{S}\times\mathcal{A}}$, $(Z_X(s,a,s'))_{(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}}$ are independent, see Alg. 5 it holds that:

$$\mathbb{P}\left(\widetilde{R}_X = r, \widetilde{N}^r_X = n, \widetilde{N}^p_X = n' \mid X\right) = \mathbb{P}\left(\widetilde{R}_X = r \mid X\right)\mathbb{P}\left(\widetilde{N}^r_X = n \mid X\right)\mathbb{P}\left(\widetilde{N}^p_X = n' \mid X\right)$$

and so,

$$\begin{aligned}\mathbb{P}\Big(\mathcal{M}(X) = (r,n,n') \mid X\Big) &= \mathbb{P}\left(\widetilde{R}_X = r, \widetilde{N}^r_X = n, \widetilde{N}^p_X = n' \mid X\right) && (110)\\ &= \mathbb{P}\left(\widetilde{R}_X = r \mid X\right)\mathbb{P}\left(\widetilde{N}^r_X = n \mid X\right)\mathbb{P}\left(\widetilde{N}^p_X = n' \mid X\right) && (111)\end{aligned}$$

Then for any two trajectories X and X', we have:

$$\mathbb{P}\left(\widetilde{R}_X = r \mid X\right)\mathbb{P}\left(\widetilde{N}^r_X = n \mid X\right)\mathbb{P}\left(\widetilde{N}^p_X = n' \mid X\right) \leq \left(\exp(\varepsilon_0/3)\mathbb{P}\left(\widetilde{R}_{X'} = r \mid X'\right) + \delta_1\right) \tag{112}$$

$$\times\left(\exp(\varepsilon_0/3)\mathbb{P}\left(\widetilde{N}^r_{X'} = n \mid X'\right) + \delta_1\right) \tag{113}$$

$$\times\left(\exp(\varepsilon_0/3)\mathbb{P}\left(\widetilde{N}^p_{X'} = n' \mid X'\right) + \delta_1\right) \tag{114}$$

$$\leq \exp(\varepsilon_0)\mathbb{P}\left(\widetilde{R}_{X'} = r \mid X'\right)\mathbb{P}\left(\widetilde{N}^r_{X'} = n \mid X'\right)\mathbb{P}\left(\widetilde{N}^p_{X'} = n' \mid X'\right) + 2\delta_1\exp\left(2\varepsilon_0/3\right) \tag{115}$$

$$+2\delta_1^2\exp\left(\varepsilon_0/3\right) + \delta_1^3 \tag{116}$$

Thus by choosing $\delta_1 = \delta_0/8$, it holds that $2\delta_1\exp\left(2\varepsilon_0/3\right) + 2\delta_1^2\exp\left(\varepsilon_0/3\right) + \delta_1^3 \leq \delta_0$ for $\varepsilon_0 \leq 1$, and so we can conclude that the Gaussian mechanism is $(\varepsilon_0, \delta_0)$-LDP. □

In addition, the precision of the Gaussian mechanism is of the same order as the Laplace mechanism, that is to say:

Proposition 5. *The Gaussian mechanism, Alg. 5, with parameter $\varepsilon_0 > 0$ and $c^2 \geq 4\ln\left(\frac{24}{\delta_0}\right)$ for any $\delta_0 > 0$ satisfies Def. 2 for any $\delta > 0$ and $k \in \mathbb{N}^\star$ with:*

$$c_{k,1}(\varepsilon_0,\delta_0,\delta) = c_{k,2}(\varepsilon_0,\delta_0,\delta) = c_{k,4}(\varepsilon_0,\delta_0,\delta) = \max\left\{\frac{cH}{\varepsilon_0}\sqrt{(k-1)\ln\left(\frac{6SA}{\delta}\right)}, 1\right\}$$

$$c_{k,3}(\varepsilon_0,\delta_0,\delta) = \max\left\{\frac{cH}{\varepsilon_0}\sqrt{(k-1)S\ln\left(\frac{6SA}{\delta}\right)}, 1\right\}$$

This result shows that using the Gaussian mechanism rather than the Laplace mechanism would not lead to improved regret rate as the utilities $c_{k,1}, c_{k,2}, c_{k,3}, c_{k,4}$ have the same depency of S, A, H, ε_0 and k . Moreover, the Gaussian mechanism only guarantees LDP for $\delta > 0$ whereas using the Laplace mechanism ensures that we can guarantee LDP for $\delta = 0$ as well.

Proof of Prop. 5: Following the same steps as in the proof of Prop 2, we have that at the beginning of episode k with probability at least $1 - \frac{\delta}{3SA}$:

$$\left|\widetilde{R}_k(s,a) - R_k(s,a)\right| = \left|\sum_{l<k}(\widetilde{R}_{X_l}(s,a) - R_{X_l}(s,a))\right| \tag{117}$$

$$= \left|\sum_{l<k}\left(Y_{1,X_l}(s,a) + \sum_{h=1}^{H} r_h \mathbb{1}_{\{s_{l,h}=s, a_{l,h}=a\}}\right) - \sum_{l<k}\sum_{h=1}^{H} r_h \mathbb{1}_{\{s_{l,h}=s, a_{l,h}=a\}}\right| \tag{118}$$

$$= \left|\sum_{l=1}^{k-1} Y_{1,X_l}(s,a)\right| \leq \sigma\sqrt{2(k-1)\ln\left(\frac{6SA}{\delta}\right)} \tag{119}$$

thanks to Chernoff bounds. The same result follows for $\widetilde{N}^r$ and $\widetilde{N}^p$. Therefore, the Gaussian mechanism satisfies Def. 2 with $c_{k,1}(\varepsilon_0,\delta_0,\delta) = c_{k,2}(\varepsilon_0,\delta_0,\delta) = c_{k,4}(\varepsilon_0,\delta_0,\delta)$ with:

$$c_{k,1}(\varepsilon_0,\delta_0,\delta) = \max\left\{\frac{cH}{\varepsilon_0}\sqrt{(k-1)\ln\left(\frac{6SA}{\delta}\right)}, 1\right\} \tag{120}$$

with $c > 0$ and:

$$c_{k,3}(\varepsilon_0,\delta_0,\delta) = \max\left\{\frac{cH}{\varepsilon_0}\sqrt{(k-1)S\ln\left(\frac{6SA}{\delta}\right)}, 1\right\} \tag{121}$$

where $c_{k,3}(\varepsilon_0,\delta_0,\delta)$ is defined as $\left|\sum_{s'} N^p_k(s,a,s') - \sum_{s'} \widetilde{N}^p_k(s,a,s')\right| \leq c_{k,3}(\varepsilon_0,\delta_0,\delta)$. □

E.2 Bernoulli Mechanism:

The second alternative mechanism we consider is the Bernoulli mechanism. In general, this mechanism is used for discrete data like indicator functions $(\mathbb{1}_{\{s_h=s,a_h=a\}})_{h,s,a}$. We therefore use this mechanism to privatize the number of visits of a state-action pair and state-action-next-state tuple for each trajectory. With the assumption that reward are supported in $[0,1]$, we can also use this mechanism for privatizing the cumulative reward of a given trajectory. Contrary to previous mechanisms, the output of the Bernoulli mechanism is three vectors, two of size $H \times S \times A$, and the last one of size $(H-1) \times S \times A \times S$. We slightly modify the requirements of Def. 2 by changing the size of the output of the privacy preserving mechanism. We summarize the mechanism in Alg. 6.

Algorithm 6: Bernoulli mechanism for LDP

```
Input: Trajectory: $X = \{(s_h, a_h, r_h) \mid h \leq H\}$, Privacy Parameter: $\varepsilon_0$
for $(s,a) \in \mathcal{S} \times \mathcal{A}$ do
    for $h = 1, \dots, H$ do
        Sample $Y_{1,X}(h,s,a) \sim \mathrm{Ber}\left(\frac{\exp(\varepsilon_0)-1}{\exp(\varepsilon_0)+1} r_h \mathbb{1}_{\{s_h=s,a_h=a\}} + \frac{1}{\exp(\varepsilon_0)+1}\right)$
        $\widetilde{R}_X(h,s,a) = \frac{\exp(\varepsilon_0)+1}{\exp(\varepsilon_0)-1}\left(Y_{1,X}(h,s,a) - \frac{1}{\exp(\varepsilon_0)+1}\right)$
        Sample $\widetilde{n}^r_X(h,s,a) \sim \mathrm{Ber}\left(\frac{\exp(\varepsilon_0)-1}{\exp(\varepsilon_0)+1} \mathbb{1}_{\{s_h=s,a_h=a\}} + \frac{1}{\exp(\varepsilon_0)+1}\right)$
        $\widetilde{N}^r_X(h,s,a) = \frac{\exp(\varepsilon_0)+1}{\exp(\varepsilon_0)-1}\left(\widetilde{n}^r_X(h,s,a) - \frac{1}{\exp(\varepsilon_0)+1}\right)$
        if $h < H-1$ then
            for $s' \in \mathcal{S}$ do
                Sample $\widetilde{n}^p_X(h,s,a,s') \sim \mathrm{Ber}\left(\frac{\exp(\varepsilon_0)-1}{\exp(\varepsilon_0)+1} \mathbb{1}_{\{s_h=s,a_h=a,s_{h+1}=s'\}} + \frac{1}{\exp(\varepsilon_0)+1}\right)$
                $\widetilde{N}^p_X(h,s,a,s') = \frac{\exp(\varepsilon_0)+1}{\exp(\varepsilon_0)-1}\left(\widetilde{n}^p_X(h,s,a,s') - \frac{1}{\exp(\varepsilon_0)+1}\right)$
Output: $(\widetilde{R}_X, \widetilde{N}^r_X, \widetilde{N}^p_X) \in \left\{\frac{-1}{\exp(\varepsilon_0)-1}, \frac{\exp(\varepsilon_0)}{\exp(\varepsilon_0)-1}\right\}^{HSA} \times \left\{\frac{-1}{\exp(\varepsilon_0)-1}, \frac{\exp(\varepsilon_0)}{\exp(\varepsilon_0)-1}\right\}^{HSA} \times \left\{\frac{-1}{\exp(\varepsilon_0)-1}, \frac{\exp(\varepsilon_0)}{\exp(\varepsilon_0)-1}\right\}^{(H-1)SAS}$
```

Just as for the Gaussian mechanism, we show that Alg. 6 satisfies Def. 2. We begin by showing that this mechanism satisfies $(\varepsilon_0, 0)$-LDP for any $\varepsilon_0 > 0$.

Proposition 6. *For any $\varepsilon > 0$, the Bernoulli mechanism, Alg. 6, with parameter $\varepsilon_0 = \varepsilon/6H$ is $(\varepsilon, 0)$-LDP.*

Proof of Prop. 6: Just as in the proof of Prop. 4 and Prop. 2, let's consider two trajectories $X = \{(s_h, a_h, r_h) \mid h \leq H\}$ and $X' = \{(s'_h, a'_h, r'_h) \mid h \leq H\}$ and also denote the output of the private randomizer $\mathcal{M}$ by $\mathcal{M}(X) = (\widetilde{R}_X, \widetilde{N}^r_X, \widetilde{N}^p_X)$ and $\mathcal{M}(X') = (\widetilde{R}_{X'}, \widetilde{N}^r_{X'}, \widetilde{N}^p_{X'})$.

For a given $r \in \left\{\frac{-1}{\exp(\varepsilon_0)-1}, \frac{\exp(\varepsilon_0)}{\exp(\varepsilon_0)-1}\right\}^{HSA}$, we have that:

$$\begin{aligned}\frac{\mathbb{P}\left(\forall(h,s,a), \widetilde{R}_X(h,s,a) = r_{h,s,a} \mid X\right)}{\mathbb{P}\left(\forall(h,s,a), \widetilde{R}_{X'}(h,s,a) = r_{h,s,a} \mid X'\right)} &= \prod_{h,s,a} \left(\frac{\frac{\exp(\varepsilon_0)-1}{\exp(\varepsilon_0)+1} r_h \mathbb{1}_{\{s_h=s,a_h=a\}} + \frac{1}{\exp(\varepsilon_0)+1}}{\frac{\exp(\varepsilon_0)-1}{\exp(\varepsilon_0)+1} r'_h \mathbb{1}_{\{s'_h=s,a'_h=a\}} + \frac{1}{\exp(\varepsilon_0)+1}}\right)^{y^r_{h,s,a}} \times \\ &\times \left(\frac{1 - \left(\frac{\exp(\varepsilon_0)-1}{\exp(\varepsilon_0)+1} r_h \mathbb{1}_{\{s_h=s,a_h=a\}} + \frac{1}{\exp(\varepsilon_0)+1}\right)}{1 - \left(\frac{\exp(\varepsilon_0)-1}{\exp(\varepsilon_0)+1} r'_h \mathbb{1}_{\{s'_h=s,a'_h=a\}} + \frac{1}{\exp(\varepsilon_0)+1}\right)}\right)^{1-y^r_{h,s,a}}\end{aligned} \tag{122}$$

where for every $(h,s,a) \in H \times \mathcal{S} \times \mathcal{A}$, we define $y^r_{h,s,a} = \frac{\exp(\varepsilon_0)-1}{\exp(\varepsilon_0)+1} r + \frac{1}{\exp(\varepsilon_0)+1}$ belongs to $\{0,1\}$ because $r \in \left\{\frac{-1}{\exp(\varepsilon_0)-1}, \frac{\exp(\varepsilon_0)}{\exp(\varepsilon_0)-1}\right\}^{HSA}$. Eq. (122) can be rewritten as:

$$(122) = \prod_{h,s,a} \left(\frac{(\exp(\varepsilon_0)-1) r_h \mathbb{1}_{\{s_h=s,a_h=a\}} + 1}{(\exp(\varepsilon_0)-1) r'_h \mathbb{1}_{\{s'_h=s,a'_h=a\}} + 1}\right)^{y^r_{h,s,a}} \left(\frac{\exp(\varepsilon_0) - (\exp(\varepsilon_0)-1) r_h \mathbb{1}_{\{s_h=s,a_h=a\}}}{\exp(\varepsilon_0) - (\exp(\varepsilon_0)-1) r'_h \mathbb{1}_{\{s'_h=s,a'_h=a\}}}\right)^{1-y^r_{h,s,a}} \tag{123}$$

Then for a given (h, s, a), because $r_h \in [0, 1]$ we have:

$$\frac{(\exp(\varepsilon_0) - 1)r_h \mathbb{1}_{\{s_h = s, a_h = a\}} + 1}{(\exp(\varepsilon_0) - 1)r'_h \mathbb{1}_{\{s'_h = s, a'_h = a\}} + 1} \leq \begin{cases} \exp(\varepsilon_0) & \text{if } \mathbb{1}_{\{s_h = s, a_h = a\}} = \mathbb{1}_{\{s'_h = s, a'_h = a\}} = 1 \\ 1 & \text{if } \mathbb{1}_{\{s_h = s, a_h = a\}} = \mathbb{1}_{\{s'_h = s, a'_h = a\}} = 0 \\ \exp(\varepsilon_0) & \text{if } \mathbb{1}_{\{s_h = s, a_h = a\}} = 1 \text{ and } \mathbb{1}_{\{s'_h = s, a'_h = a\}} = 0 \\ 1 & \text{if } \mathbb{1}_{\{s_h = s, a_h = a\}} = 0 \text{ and } \mathbb{1}_{\{s'_h = s, a'_h = a\}} = 1 \end{cases} \tag{124}$$

$$\frac{\exp(\varepsilon_0) - (\exp(\varepsilon_0) - 1)r_h \mathbb{1}_{\{s_h = s, a_h = a\}}}{\exp(\varepsilon_0) - (\exp(\varepsilon_0) - 1)r'_h \mathbb{1}_{\{s'_h = s, a'_h = a\}}} \leq \begin{cases} \exp(\varepsilon_0) & \text{if } \mathbb{1}_{\{s_h = s, a_h = a\}} = \mathbb{1}_{\{s'_h = s, a'_h = a\}} = 1 \\ 1 & \text{if } \mathbb{1}_{\{s_h = s, a_h = a\}} = \mathbb{1}_{\{s'_h = s, a'_h = a\}} = 0 \\ 1 & \text{if } \mathbb{1}_{\{s_h = s, a_h = a\}} = 1 \text{ and } \mathbb{1}_{\{s'_h = s, a'_h = a\}} = 0 \\ \exp(\varepsilon_0) & \text{if } \mathbb{1}_{\{s_h = s, a_h = a\}} = 0 \text{ and } \mathbb{1}_{\{s'_h = s, a'_h = a\}} = 1 \end{cases} \tag{125}$$

Therefore, we can simplify each term in (123) by:

$$\frac{(\exp(\varepsilon_0) - 1)r_h \mathbb{1}_{\{s_h = s, a_h = a\}} + 1}{(\exp(\varepsilon_0) - 1)r'_h \mathbb{1}_{\{s'_h = s, a'_h = a\}} + 1} \leq \exp\left(\varepsilon_0 \left(\mathbb{1}_{\{s_h = s, a_h = a\}} + \mathbb{1}_{\{s'_h = s, a'_h = a\}}\right)\right)$$
$$\frac{\exp(\varepsilon_0) - (\exp(\varepsilon_0) - 1)r_h \mathbb{1}_{\{s_h = s, a_h = a\}}}{\exp(\varepsilon_0) - (\exp(\varepsilon_0) - 1)r'_h \mathbb{1}_{\{s'_h = s, a'_h = a\}}} \leq \exp\left(\varepsilon_0 \left(\mathbb{1}_{\{s_h = s, a_h = a\}} + \mathbb{1}_{\{s'_h = s, a'_h = a\}}\right)\right)$$

Hence, using the two inequalities above:

$$\begin{aligned} (123) &\leq \prod_{h,s,a} \exp\left(y^r_{h,s,a}\varepsilon_0 \left(\mathbb{1}_{\{s_h = s, a_h = a\}} + \mathbb{1}_{\{s'_h = s, a'_h = a\}}\right) + (1 - y^r_{h,s,a})\varepsilon_0 \left(\mathbb{1}_{\{s'_h = s, a'_h = a\}} + \mathbb{1}_{\{s_h = s, a_h = a\}}\right)\right) \\ &= \prod_{h,s,a} \exp\left(\varepsilon_0 \left(\mathbb{1}_{\{s_h = s, a_h = a\}} + \mathbb{1}_{\{s'_h = s, a'_h = a\}}\right)\right) \\ &= \exp\left(2\varepsilon_0 H\right) \end{aligned}$$

In addition, let's consider $m \in \left\{\frac{-1}{e^{\varepsilon_0} - 1}, \frac{e^{\varepsilon_0}}{e^{\varepsilon_0} - 1}\right\}^{H \times S \times A}$ and $y = \frac{\exp(\varepsilon_0) - 1}{\exp(\varepsilon_0) + 1} m + \frac{1}{\exp(\varepsilon_0) + 1} \in \{0, 1\}$, we then have that:

$$\begin{aligned} \frac{\mathbb{P}\left(\forall(h, s, a), \widetilde{N}^r_X(h, s, a) = m_{h,s,a} \mid X\right)}{\mathbb{P}\left(\forall(h, s, a), \widetilde{N}^r_{X'}(h, s, a) = m_{h,s,a} \mid X'\right)} = \prod_{h,s,a} &\left(\frac{\frac{\exp(\varepsilon_0) - 1}{\exp(\varepsilon_0) + 1} \mathbb{1}_{\{s_h = s, a_h = a\}} + \frac{1}{\exp(\varepsilon_0) + 1}}{\frac{\exp(\varepsilon_0) - 1}{\exp(\varepsilon_0) + 1} \mathbb{1}_{\{s'_h = s, a'_h = a\}} + \frac{1}{\exp(\varepsilon_0) + 1}}\right)^{y_{h,s,a}} \times \\ &\times \left(\frac{1 - \left(\frac{\exp(\varepsilon_0) - 1}{\exp(\varepsilon_0) + 1} \mathbb{1}_{\{s_h = s, a_h = a\}} + \frac{1}{\exp(\varepsilon_0) + 1}\right)}{1 - \left(\frac{\exp(\varepsilon_0) - 1}{\exp(\varepsilon_0) + 1} \mathbb{1}_{\{s'_h = s, a'_h = a\}} + \frac{1}{\exp(\varepsilon_0) + 1}\right)}\right)^{1 - y_{h,s,a}} \end{aligned} \tag{126}$$

Which can be rewritten as:

$$\begin{aligned} \frac{\mathbb{P}\left(\forall(h, s, a), \widetilde{N}^r_X(h, s, a) = m_{h,s,a} \mid X\right)}{\mathbb{P}\left(\forall(h, s, a), \widetilde{N}^r_{X'}(h, s, a) = m_{h,s,a} \mid X'\right)} = \prod_{h,s,a} &\left(\frac{(\exp(\varepsilon_0) - 1)\mathbb{1}_{\{s_h = s, a_h = a\}} + 1}{(\exp(\varepsilon_0) - 1)\mathbb{1}_{\{s'_h = s, a'_h = a\}} + 1}\right)^{y_{h,s,a}} \times \\ &\times \left(\frac{\exp(\varepsilon_0) - (\exp(\varepsilon_0) - 1)\mathbb{1}_{\{s_h = s, a_h = a\}}}{\exp(\varepsilon_0) - (\exp(\varepsilon_0) - 1)\mathbb{1}_{\{s'_h = s, a'_h = a\}}}\right)^{1 - y_{h,s,a}} \end{aligned} \tag{127}$$

Thus for a given (h, s, a):

$$\frac{(\exp(\varepsilon_0) - 1)\mathbb{1}_{\{s_h = s, a_h = a\}} + 1}{(\exp(\varepsilon_0) - 1)\mathbb{1}_{\{s'_h = s, a'_h = a\}} + 1} = \begin{cases} 1 & \text{if } \mathbb{1}_{\{s_h = s, a_h = a\}} = \mathbb{1}_{\{s'_h = s, a'_h = a\}} \\ \exp(\varepsilon_0) & \text{if } \mathbb{1}_{\{s_h = s, a_h = a\}} = 1 \text{ and } \mathbb{1}_{\{s'_h = s, a'_h = a\}} = 0 \\ \exp(-\varepsilon_0) & \text{if } \mathbb{1}_{\{s_h = s, a_h = a\}} = 0 \text{ and } \mathbb{1}_{\{s'_h = s, a'_h = a\}} = 1 \end{cases} \tag{128}$$

$$\frac{\exp(\varepsilon_0) - (\exp(\varepsilon_0) - 1)\mathbb{1}_{\{s_h = s, a_h = a\}}}{\exp(\varepsilon_0) - (\exp(\varepsilon_0) - 1)\mathbb{1}_{\{s'_h = s, a'_h = a\}}} = \begin{cases} 1 & \text{if } \mathbb{1}_{\{s_h = s, a_h = a\}} = \mathbb{1}_{\{s'_h = s, a'_h = a\}} \\ \exp(-\varepsilon_0) & \text{if } \mathbb{1}_{\{s_h = s, a_h = a\}} = 1 \text{ and } \mathbb{1}_{\{s'_h = s, a'_h = a\}} = 0 \\ \exp(\varepsilon_0) & \text{if } \mathbb{1}_{\{s_h = s, a_h = a\}} = 0 \text{ and } \mathbb{1}_{\{s'_h = s, a'_h = a\}} = 1 \end{cases} \tag{129}$$

Therefore, here again we can simplify each term in (127) by:

$$\frac{(\exp(\varepsilon_0)-1)\mathbb{1}_{\{s_h=s,a_h=a\}}+1}{(\exp(\varepsilon_0)-1)\mathbb{1}_{\{s'_h=s,a'_h=a\}}+1} \leq \exp\left(\varepsilon_0\left(\mathbb{1}_{\{s_h=s,a_h=a\}}-\mathbb{1}_{\{s'_h=s,a'_h=a\}}\right)\right)$$

$$\frac{\exp(\varepsilon_0)-(\exp(\varepsilon_0)-1)\mathbb{1}_{\{s_h=s,a_h=a\}}}{\exp(\varepsilon_0)-(\exp(\varepsilon_0)-1)\mathbb{1}_{\{s'_h=s,a'_h=a\}}} \leq \exp\left(\varepsilon_0\left(\mathbb{1}_{\{s_h=s,a_h=a\}}-\mathbb{1}_{\{s'_h=s,a'_h=a\}}\right)\right)$$

Therefore:

$$\begin{aligned}(127) &= \prod_{h,s,a}\exp\left(y_{h,s,a}\varepsilon_0\left(\mathbb{1}_{\{s_h=s,a_h=a\}}-\mathbb{1}_{\{s'_h=s,a'_h=a\}}\right)+(1-y_{h,s,a})\varepsilon_0\left(\mathbb{1}_{\{s'_h=s,a'_h=a\}}-\mathbb{1}_{\{s_h=s,a_h=a\}}\right)\right)\\ &= \prod_{h,s,a}\exp\left((2y_{h,s,a}-1)\varepsilon_0\left(\mathbb{1}_{\{s_h=s,a_h=a\}}-\mathbb{1}_{\{s'_h=s,a'_h=a\}}\right)\right)\\ &\leq \exp\left(2\varepsilon_0 H\right)\end{aligned}$$

Using the same reasonning we have that for any $m' \in \left\{-\frac{1}{\exp(\varepsilon_0)-1}, \frac{\exp(\varepsilon_0)}{\exp(\varepsilon_0)-1}\right\}^{(H-1)\times S\times A\times S}$:

$$\frac{\mathbb{P}\left(\forall(h,s,a,s'), \widetilde{N}^p_X(h,s,a,s') = m'_{h,s,a,s'} \mid X\right)}{\mathbb{P}\left(\forall(h,s,a,s'), \widetilde{N}^p_{X'}(h,s,a,s') = m'_{h,s,a,s'} \mid X'\right)} \leq \exp(2\varepsilon_0 H) \tag{130}$$

We conclude the proof the same way as the proof of Prop. 3. □

In addition, the precision $c_{k,1}$, $c_{k,2}$, $c_{k,3}$ and $c_{k,4}$ of the Bernoulli mechanism are still of order $\sqrt{k}$ just as the Gaussian and Laplace mechanisms. From the below proposition, we see that although the dependency on S is worst than with the Laplace or Gaussian mechanisms, the dependence on ε_0 is better for small ε_0. Indeed, we have an additional factor S for $c_{k,3}$ compared to the other mechanisms but those terms scale with $1/(\exp(\varepsilon_0)-1)$ instead of the worse dependency $1/\varepsilon$.

Proposition 7. *The Bernoulli mechanism, Alg. 6, with parameter $\varepsilon_0 > 0$ satisfies Def. 2 for any $\delta > 0$ and $k \in \mathbb{N}^\star$ with:*

$$\begin{aligned}c_{k,1}(\varepsilon_0,\delta) = c_{k,2}(\varepsilon_0,\delta) &= \max\left\{1, \frac{2\exp(\varepsilon_0)-1}{\exp(\varepsilon_0)-1}\sqrt{\frac{(k-1)H}{2}\ln\left(\frac{4SA}{\delta}\right)}\right\}\\ c_{k,3}(\varepsilon_0,\delta) &= \max\left\{1, \frac{S(2\exp(\varepsilon_0)-1)}{\exp(\varepsilon_0)-1}\sqrt{\frac{(k-1)H}{2}\ln\left(\frac{4SA}{\delta}\right)}\right\}\\ c_{k,4}(\varepsilon_0,\delta) &= \max\left\{1, \frac{2\exp(\varepsilon_0)-1}{\exp(\varepsilon_0)-1}\sqrt{\frac{(k-1)H}{2}\ln\left(\frac{4S^2A}{\delta}\right)}\right\}\end{aligned}$$

Proof of Prop. 7: Let's consider a given state-action-next state tuple, (s,a,s'), then when summing over h:

$$\left|\sum_{h=1}^{H}\widetilde{N}^r_k(h,s,a) - \sum_{l<k}\sum_{h=1}^{H}\mathbb{1}_{\{s_{l,h}=s,a_{l,h}=a\}}\right| = \left|\sum_{h=1}^{H}\sum_{l<k}\widetilde{N}^r_{X_l}(h,s,a) - \mathbb{1}_{\{s_{l,h}=s,a_{l,h}=a\}}\right| \tag{131}$$

We now construct a filtration $(\mathcal{F}_{k,h})_{k,h}$ such that $(\widetilde{N}^r_{X_l}(h,s,a) - \mathbb{1}_{\{s_{l,h}=s,a_{l,h}=a\}})_{l,h}$ is a Martingale Difference Sequence. For an episode k and step h, define $\mathcal{F}_{k,h} = \sigma(\{(s_{l,j},a_{l,j},r_{l,j})_{j\leq H}, \mathcal{M}((s_{l,j},a_{l,j},r_{l,j})_{j\leq H})\} \mid l < k\} \cup \{(s_{k,j},a_{k,j},r_{k,j})_{j\leq h}\})$ to be the filtration that contains the history before episode k. Then $\mathbb{1}_{\{s_{k,h}=s,a_{k,h}=a\}}$ is $\mathcal{F}_{k,h}$-measurable and thus we have:

$$\begin{aligned}\mathbb{E}\left(\widetilde{N}^r_{X_k}(h,s,a) - \mathbb{1}_{\{s_{k,h}=s,a_{k,h}=a\}} \mid \mathcal{F}_{k,h}\right) = \frac{\exp(\varepsilon_0)+1}{\exp(\varepsilon_0)-1}&\left(\mathbb{E}\left(\widetilde{n}_{X_k}(h,s,a) \mid \mathcal{F}_{k,h}\right) - \frac{1}{\exp(\varepsilon_0)+1}\right)\\ &-\mathbb{1}_{\{s_{k,h}=s,a_{k,h}=a\}} = 0\end{aligned}$$

where $\tilde{n}_{X_k}(h,s,a)$ is a Bernoulli random variable generated by Alg. 6 for each step h, state s, action a and trajectory X_k. And $\left|\widetilde{N}^r_{X_k}(h,s,a) - \mathbb{1}_{\{s_{k,h}=s,a_{k,h}=a\}}\right| \leq \frac{2\exp(\varepsilon_0)-1}{\exp(\varepsilon_0)-1}$. Then thanks to Azuma-Hoeffding inequality we have that with probability at least $1-\delta/(4SA)$:

$$\left|\sum_{h=1}^{H}\widetilde{N}^r_k(h,s,a) - \sum_{l<k}\sum_{h=1}^{H}\mathbb{1}_{\{s_{l,h}=s,a_{l,h}=a\}}\right| \leq \frac{2\exp(\varepsilon_0)-1}{\exp(\varepsilon_0)-1}\sqrt{\frac{(k-1)H}{2}\ln\left(\frac{4SA}{\delta}\right)} \tag{132}$$

With the same reasonning, we have with probability at least $1-\delta/4S^2A$:

$$\left|\sum_{h=1}^{H}\widetilde{N}^p_k(h,s,a,s') - \sum_{l<k}\sum_{h=1}^{H-1}\mathbb{1}_{\{s_{l,h}=s,a_{l,h}=a,s_{l,h+1}=s'\}}\right| \leq \frac{2\exp(\varepsilon_0)-1}{\exp(\varepsilon_0)-1}\sqrt{\frac{(k-1)H}{2}\ln\left(\frac{4S^2A}{\delta}\right)} \tag{133}$$

Also, we have:

$$\left|\sum_{h=1}^{H}\widetilde{R}^r_k(h,s,a) - \sum_{l<k}\sum_{h=1}^{H}r_h\mathbb{1}_{\{s_{l,h}=s,a_{l,h}=a\}}\right| \leq \frac{2\exp(\varepsilon_0)-1}{\exp(\varepsilon_0)-1}\sqrt{\frac{(k-1)H}{2}\ln\left(\frac{4SA}{\delta}\right)} \tag{134}$$

with $\widetilde{R}^r_k(h,s,a) = \sum_{l<k}\widetilde{R}_{X_l}$. Finally, with probability at least $1-\delta/4SA$:

$$\left|\sum_{h=1}^{H}\sum_{s'}\widetilde{N}^p_k(h,s,a,s') - \sum_{s'}\sum_{l<k}\sum_{h=1}^{H-1}\mathbb{1}_{\{s_{l,h}=s,a_{l,h}=a,s_{l,h+1}=s'\}}\right| \leq \frac{S(2e^{\varepsilon_0}-1)}{e^{\varepsilon_0}-1}\sqrt{\frac{(k-1)H}{2}\ln\left(\frac{4SA}{\delta}\right)} \tag{135}$$

Compared to bounds, we derived for previous mechanisms there is an additional factor $\sqrt{S}$. This comes from using a triangular inequality instead of using concntration inequalities like in previous mechanisms. Then thanks to a union bound over the state-action pair and the state-action-next state tuple we have that the Bernoulli mechanism satisfies Def. 2 with:

$$c_{k,1}(\varepsilon_0,\delta) = c_{k,2}(\varepsilon_0,\delta) = \max\left\{1, \frac{2\exp(\varepsilon_0)-1}{\exp(\varepsilon_0)-1}\sqrt{\frac{(k-1)H}{2}\ln\left(\frac{4SA}{\delta}\right)}\right\} \tag{136}$$

$$c_{k,3}(\varepsilon_0,\delta) = \max\left\{1, \frac{S(2\exp(\varepsilon_0)-1)}{\exp(\varepsilon_0)-1}\sqrt{\frac{(k-1)H}{2}\ln\left(\frac{4SA}{\delta}\right)}\right\}, \tag{137}$$

$$c_{k,4}(\varepsilon_0,\delta) = \max\left\{1, \frac{2\exp(\varepsilon_0)-1}{\exp(\varepsilon_0)-1}\sqrt{\frac{(k-1)H}{2}\ln\left(\frac{4S^2A}{\delta}\right)}\right\} \tag{138}$$

□

F Additional Experiment

In this appendix, we explore a second experiment, in which we use the same the RandomMDP environment with the same parameters as in Sec. 6 in order to investigate the effect of differential privacy on the learning process. For this, we run the UCB-VI algorithm for $K = 10^3$ episodes and collect the aggregate noisy statistics, $(\widetilde{R}_K(s,a))_{(s,a)\in\mathcal{S}\times\mathcal{A}}$, $(\widetilde{N}^r_K(s,a))_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ and $(\widetilde{N}^p_K(s,a,s'))_{(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}}$ that have been generated by using the Laplace mechanism for each episode. We collect those statistics, 10^3 times. We compare the histogram of those noisy statistics to that of the noiseless statistics used by UCB-VI in Fig. 3. This demonstrates that, as expected, there is much more variation in the statistics provided by the private mechanism. In Fig. 4, we applied the Laplace mechanism to two different random trajectories, X and X'. We can see that, after applying the Laplace mechanism, the two distinct trajectories become almost indistinguishable. These two figures combined demonstrate the difficulty of learning from locally differentially private data.

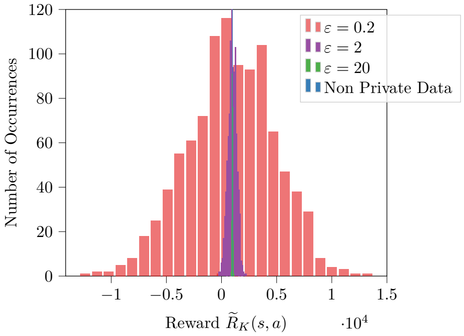

Figure 3: Aggregate reward for privatized data with $\varepsilon \in \{0.2, 2, 20\}$ and non-privatized data for state 0 and action 1

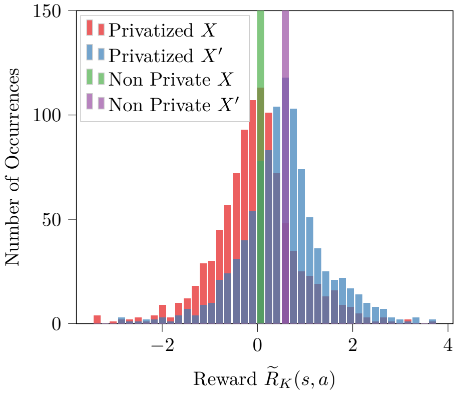

Figure 4: Privatized cumulative reward over an episode for a given state-action pair and two different trajectories X and X' with $\varepsilon = 20$ for state 0 and action 1