Improved regret bounds in tabular RL under local differential privacy

Hugo Richard
Joint team Fairplay, Criteo AI Labs, ENSAE CREST, INRIA
France
research.hugo.richard@gmail.com

Abstract

We study regret minimization for time-homogeneous episodic tabular reinforcement learning with S states and A actions under non-interactive ϵ-local differential privacy (LDP). In each episode $k \in [K]$, the decision-maker sends a policy to a user, the user executes the policy for H steps, and sends an ϵ-locally differentially private message back to the decision-maker. When $H = 1$, the problem reduces to an LDP contextual bandit with S contexts, A arms and horizon K. We prove that the minimax regret is $\min\left(K, \frac{\sqrt{\max(S,\exp(\epsilon)+1)SAK(\exp(\epsilon)+1)}}{\exp(\epsilon)-1}\right)$, up to constants and a $\sqrt{\log(A)}$ factor. The upper bound is achieved by an adaptation of EXP3 and a unary encoding mechanism. When $H > 1$, and for small ϵ, we show a lower bound in $\Omega(HS\frac{\sqrt{AK}}{\epsilon})$ and an upper bound in $\tilde{O}(H^2S^2\frac{\sqrt{AK}}{\epsilon})$. The upper bound is achieved by an algorithm based on UCBVI and a unary encoding mechanism. For small ϵ, these bounds improve the previously known lower bound by a factor $\sqrt{S}$ and the previously known upper bound by a factor of $H\sqrt{A}$.

1 Introduction

Reinforcement learning (RL) is the problem of learning to make decisions through interaction. Typical applications of RL include recommender systems, online advertising, healthcare, routing or robotics Kober et al. [2013], Afsar et al. [2023], Cai et al. [2017], Yu et al. [2023], Nazari et al. [2018]. The entity running the RL algorithm is called the *decision-maker*, which could be a dating application, an ad displaying platform or a scientist running a medical experiment. In the classical RL framework, an interaction proceeds as follows. The decision-maker observes the current *state* which could be the profile of a user, the context of an ad or the health status of the subject and decides on an *action* to take based on the observed state which could be a partner recommendation, the display of an ad or the choice of treatment to give to a subject. Conditionally on the state and action, the decision-maker receives a random numerical *reward* which could be a like of the recommended person, a click on the displayed ad or a numerical measure of the subject's health. The next state is a random function of the current state and action. The goal of reinforcement learning is to optimize the *policy* used by the decision-maker to choose actions from previous observations so that the sum of rewards gathered over time is maximized.

In classical RL, the focus is on the decision-maker and the constraints of the *user*, which could be a client of a dating application, a viewer of the ad, or a subject of the medical experiment are not taken into account. A primary concern of users in the above applications is privacy. A user may not want to publicly disclose their activity on a dating application, their browsing history, or their health condition. While the decision-maker may be trustworthy, it can be hacked or leak data by accident. For instance, the dating application Ashley Madison was hacked in 2015 exposing the data of 32 million users Jones [2016]. Another example is the UK Biobank data containing health and

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

genetic information of 500 000 subjects that was partially uploaded on Github by accident Rocher and Morley [2026].

While the UK Biobank data is de-identified, meaning the personal information such as the name or the address of the subjects is removed, journalists from the Guardian were able to re-identify some subjects Devlin and Burgis [2026]. The fact that removing personal information is not sufficient to protect the privacy of users is well-documented Sweeney [2000], Narayanan and Shmatikov [2008], Rocher et al. [2019] and justifies the use of more rigorous privacy frameworks such as local differentially privacy (LDP) Duchi et al. [2013].

The above considerations motivate the study of reinforcement learning under LDP. We consider a tabular RL problem with S states and A possible actions, where the decision-maker learns an optimal policy by interacting with K different users sequentially. The *transition kernel*—the probabilities of moving from one state to another when an action is taken—and reward distributions are fixed (but unknown) and the same across all users. Each user interaction corresponds to one *episode*, where the initial state varies from users to users. When interacting with a user the decision-maker sends a policy to the user who executes it for H steps and applies a randomized *mechanism* to privatize its data before sending the privatized data back to the decision-maker. The mechanism is chosen by the decision-maker but must satisfy ϵ-LDP constraints which guarantee a limit on the amount of information that the decision-maker - or hackers - can learn about the user.

Our contributions We first study the case $H = 1$. We prove that the minimax regret rate is given by

$$\min\left(K, \frac{\sqrt{\max(S, \exp(\epsilon)+1)SAK(\exp(\epsilon)+1)}}{\exp(\epsilon)-1}\right)$$

up to constants and a $\sqrt{\log(A)}$ factor. This result is a consequence of Theorem 1 showing the lower bound and Theorem 2 providing an algorithm and a mechanism whose regret matches the lower bound up to constants and a $\sqrt{\log(A)}$ factor.

Second, we study the case $H > 1$. A reduction from the bandit case implies a new lower bound shown in Corollary 1 improving upon previous work Garcelon et al. [2021] uniformly across all parameters. The improvement is particularly significant when $\epsilon \leq 1$, gaining, in this case, a factor $\sqrt{S}$ over the previously known lower bound. Then, we present Theorem 3 which provides an algorithm and a mechanism whose regret, for $\epsilon \leq 1$, is in $\tilde{O}(\frac{H^2S^2\sqrt{AK}}{\epsilon})$ improving, in this high privacy regime, over the previous upper bound [Garcelon et al., 2021, Theorem 5] by a factor of $H\sqrt{A}$. We emphasize that, for $\epsilon \leq 1$, a multiplicative gap of order HS remains between this upper bound and the lower bound of Corollary 1, which is in $\Omega(\frac{HS\sqrt{AK}}{\epsilon})$. Closing this gap is left as an open question (see Section 6).

2 Related Work

Regret minimization for non-private time-homogeneous episodic tabular RL is well studied. A lower bound for the minimax regret of order $\Omega(H\sqrt{SAK})$ is obtained by Azar et al. [2017] thanks to a reduction to contextual bandits. Many works Azar et al. [2017], Dann et al. [2019], Zanette and Brunskill [2019], Zhang et al. [2021] study regret minimization in RL and various algorithms have been proposed. The best upper bound up to date is given by a model-based algorithm that achieves a regret in $\tilde{O}(H\sqrt{SAK} + HS^2A)$ up to log factors Zhang et al. [2021]. This result shows that tabular RL is essentially as hard as contextual bandits. Whether a model-free algorithm can achieve a leading rate of $H\sqrt{SAK}$ is an active area of research Jin et al. [2018], Zhang et al. [2020].

The combination of online decision-making and local differential privacy is well studied in the bandits case. In classical multi-armed bandits, the optimal regret under ϵ-LDP is in $\Theta(\max(\frac{\sqrt{AK}}{\epsilon}, \sqrt{AK}))$ Ren et al. [2020]. We recover this rate up to constants and log factor by setting $S = 1$ in Theorem 1 and Theorem 2.

The works most related to ours are those studying RL under LDP constraints Garcelon et al. [2021], Qiao and Wang [2023], Chowdhury and Zhou [2022b]. The work of Garcelon et al. [2021] studies the same time-homogeneous setting as us, while Qiao and Wang [2023] and Chowdhury and Zhou [2022b] study the time-varying case where rewards and transition matrices change over time. Since

their bounds are for this more general setting, they are loose if applied directly. Our main point of comparison is therefore Garcelon et al. [2021]. We improve the lower bound obtained in Garcelon et al. [2021] and give an upper bound improving that of Garcelon et al. [2021] in the high privacy regime $\epsilon \leq H\sqrt{A}$.

Although tabular contextual bandit under LDP has never been studied specifically, some generalizations were considered in the literature. Contextual linear bandits under LDP have been studied in Zheng et al. [2020]. After embedding the reward using one-hot features and rescaling so that the unit-norm assumption on the parameter is satisfied, the guarantee of Zheng et al. gives in the tabular setting a rate in $\tilde{O}\big(\frac{(SA)^{5/4}K^{3/4}}{\epsilon}\big)$, which is suboptimal. If contexts are chosen i.i.d., then, under some assumptions, an $O(K^{1/2})$ rate can be recovered Han et al. [2021], Pavlovic et al. [2025]. The setting we study differs in that there are no restrictions on how contexts are chosen. Another generalization is contextual nonparametric bandits under LDP, which have been studied in Ma et al. [2025], where the authors obtain sublinear regret under regularity assumptions (Assumptions 2.2 and 2.3) and a margin assumption (Assumption 2.4). These assumptions do not necessarily hold in the setting we study. Lastly, a related problem is LDP contextual-bandit pricing Zhao et al. [2026], Chen et al. [2023], where the price is the action and the revenue is the reward. However, none of the results of Zhao et al. [2026] or Chen et al. [2023] apply in the setting we consider because of incompatible assumptions. Notably, in Zhao et al., the reward has a specific parametric form derived from a GLM assumption, and in Chen et al., the reward is assumed to be strongly concave. In contrast, none of these restrictions are imposed in the setting we study.

Besides LDP, other notions of privacy have been studied in bandits and RL. In *global differential privacy*, the decision-maker is trusted but the sequence of action it takes must not reveal too much information about the users' reward Azize and Basu [2022], Azize et al. [2026]. This setting is only relevant for multi-armed bandits since in contextual bandits and RL, a linear regret is unavoidable Shariff and Sheffet [2018]. This motivates the notion of *joint differential privacy* Shariff and Sheffet [2018], Vietri et al. [2020] which is weaker. It guarantees that the actions taken by the decision-maker at time t do not reveal too much information about the users' information up to time $t' < t$. Intermediate notions based on shuffling have also been studied for linear contextual bandits Chowdhury and Zhou [2022a], Garcelon et al. [2022].

3 Episodic Finite-Horizon Tabular RL under LDP

Notations For an integer n, we use $[n]$ to denote the set $\{1, 2, \dots, n\}$. For a finite set $\mathcal{X}$, we use $\Delta(\mathcal{X})$ to denote the set of functions $\nu : \mathcal{X} \to [0, 1]$ such that $\sum_{x\in\mathcal{X}} \nu(x) = 1$ and denote by $\nu(\cdot)$ the corresponding distribution over $\mathcal{X}$. We use $f(x) = \Omega(g(x))$ to denote that there exists a constant $C > 0$ such that $f(x) \geq Cg(x)$, $f(x) = O(g(x))$ to denote that there exists a constant $C > 0$ such that $f(x) \leq Cg(x)$ and $f(x) = \Theta(g(x))$ if $f(x) = \Omega(g(x))$ and $f(x) = O(g(x))$. We use $\tilde{O}(\cdot)$, $\tilde{\Omega}(\cdot)$ and $\tilde{\Theta}(\cdot)$ to hide constants and logarithmic factors. We denote $\mathcal{B}(p)$ the Bernoulli distribution with parameter p. $D_{\text{KL}}(\nu, \nu') = \mathbb{E}_\nu\left[\log \frac{d\nu}{d\nu'}\right]$ and $D_{\text{TV}}(\nu, \nu') = \sup_A |\nu(A) - \nu'(A)|$ denote the Kullback-Leibler divergence and total variation distance between distributions ν and ν'.

Non-private interaction, value functions and regret A Markov Decision Process (MDP) is defined by the tuple $I = (\mathcal{S} = [S], \mathcal{A} = [A], P, r, (\rho^k)_{k=1}^K)$, where $\mathcal{S}$ is the set of S states, $\mathcal{A} = [A]$ the set of A actions, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the *transition kernel*, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the expected *reward function* and $\rho^k \in \Delta(\mathcal{S})$ is the distribution of the starting state at episode k for each $k \in [K]$. We call $\mathfrak{M}(S, A, K)$, or simply $\mathfrak{M}$ when the context is clear, the set of MDP with S states, A actions and K episodes. The above description corresponds to the *tabular* case where the number of states, actions and episodes are finite.

In the non-private setting, the decision-maker interacts directly with the MDP. The interaction consists of K episodes each consisting of H steps. At episode k, the decision maker selects a policy $\pi^k = (\pi_h^k : \mathcal{S} \to \Delta(\mathcal{A}))_{h\in[H]}$ and samples a starting state S_1^k from ρ^k. Then, at step $h \in [H]$, the decision maker observes $S_h^k \in \mathcal{S}$, samples an action $A_h^k \in \mathcal{A}$ from $\pi_h^k(S_h^k, \cdot)$, obtains a reward $R_h^k \in [0, 1]$ and transitions to $S_{h+1}^k \in \mathcal{S}$. The reward R_h^k and the next state S_{h+1}^k are sampled independently conditionally on (S_h^k, A_h^k), with $\mathbb{E}[R_h^k \mid S_h^k, A_h^k] = r(S_h^k, A_h^k)$ and $S_{h+1}^k \mid S_h^k, A_h^k \sim P(S_h^k, A_h^k, \cdot)$. The *trajectory* of episode k is denoted by $\tau^k = (S_1^k, A_1^k, R_1^k, \dots, S_H^k, A_H^k, R_H^k, S_{H+1}^k) \in \mathbb{T}$ where

$\mathbb{T} = (\mathcal{S} \times \mathcal{A} \times [0,1])^H \times \mathcal{S}$ is the set of possible trajectories. The *value* of a policy π at step h on MDP I at state s is defined as

$$V_h^{I,\pi}(s) = \mathbb{E}\left[\sum_{i=h}^{H} R_i^k \mid S_h^k = s\right].$$

where the expectation is taken over the trajectory τ^k. We often omit the superscript I when it is clear from the context. The *optimal value function* is defined as $V_h^*(s) = \max_\pi V_h^{I,\pi}(s)$. Lastly, note that the policies we use depend only on the current state and step. It can be shown that using history-dependent policies instead does not increase the optimal value function [Puterman, 2014, Theorem 4.4.2]. We call the process by which the decision maker updates policies from episode to episode the (non-private) *algorithm* $\mathfrak{A} = (\mathfrak{A}_k : \mathbb{T}^{k-1} \to (\mathcal{S} \to \Delta(\mathcal{A}))^H)_{k\in[K]}$ and call $\mathbb{A}^{\text{public}}$ the set of non-private algorithms. where $\mathfrak{A}_k$ maps the past $k-1$ trajectories $(\tau^1, \dots, \tau^{k-1})$ to the policy π^k.

The performance of an algorithm $\mathfrak{A}$ on an MDP I is computed by the *regret* defined as

$$\mathcal{R}^{\text{public}}(\mathfrak{A}, I) = \mathbb{E}\left[\sum_{k=1}^{K}\left(V_1^{I,*}(S_1^k) - V_1^{I,\pi^k}(S_1^k)\right)\right], \quad \pi^k \sim \mathfrak{A}_k(\tau^1, \dots, \tau^{k-1}),$$

and the difficulty of the problem is quantified by the *minimax regret* defined as

$$\mathfrak{R}^{\text{public}} = \inf_{\mathfrak{A}\in\mathbb{A}^{\text{public}}} \sup_{I\in\mathfrak{M}} \mathcal{R}(\mathfrak{A}, I).$$

Private interaction and local differential privacy In the setting we study, the user does not allow the decision maker to access its true data. The algorithm will therefore not have access to the trajectories but only to a version privatized by a mechanism $\mathcal{M} : \mathbb{T} \to \Delta(\mathcal{Y})$ satisfying the following ϵ-LDP guarantee: for any two trajectories τ, τ' and any measurable subset $Y \subset \mathcal{Y}$, we must have

$$\mathbb{P}(\mathcal{M}(\tau) \in Y) \leq e^\epsilon \mathbb{P}(\mathcal{M}(\tau') \in Y).$$

We call $\mathbb{M}_\epsilon$ the set of such ϵ-LDP mechanisms.

The interaction then proceeds as follows. The decision maker first selects a privacy mechanism $\mathcal{M}$. Then, at episode k, the decision maker selects a policy π^k and sends it to user k, who executes the policy π^k for H steps and observes the resulting trajectory τ^k which is then privatized by the mechanism $\mathcal{M}$ to produce a privatized statistic $M^k = \mathcal{M}(\tau^k)$ sent back to the decision maker.

The process by which the decision maker updates policies from episode to episode is now characterized by a *private algorithm* $\mathfrak{A} = (\mathfrak{A}_k : \mathcal{Y}^{k-1} \to (\mathcal{S} \to \Delta(\mathcal{A}))^H)_{k\in[K]}$ and the set of private algorithms is denoted by $\mathbb{A}^{\text{private}}$. Here $\mathcal{Y}^{k-1}$ is the Cartesian-product space containing all possible sequences of $k-1$ private messages: at episode k, the realized input to the algorithm is the history $(M^1, \dots, M^{k-1}) \in \mathcal{Y}^{k-1}$.

Furthermore, the performance of the decision maker is now determined by the choice of mechanism $\mathcal{M}$ and the private algorithm $\mathfrak{A}$. We still measure the performance by the regret given by

$$\mathcal{R}(\mathfrak{A}, I, \mathcal{M}) = \mathbb{E}\left[\sum_{k=1}^{K}\left(V_1^{I,*}(S_1^k) - V_1^{I,\pi^k}(S_1^k)\right)\right], \quad \begin{array}{r} M_1, \dots, M_K \sim \mathcal{M}(\tau^1), \dots, \mathcal{M}(\tau^K) \\ \pi^k \sim \mathfrak{A}_k((M_i)_{i\in[k-1]}) \end{array}$$

and the minimax regret is defined as

$$\mathfrak{R} = \inf_{\substack{\mathfrak{A}\in\mathbb{A}^{\text{private}} \\ \mathcal{M}\in\mathbb{M}_\epsilon}} \sup_{I\in\mathfrak{M}} \mathcal{R}(\mathfrak{A}, I, \mathcal{M}).$$

In the following sections, we present upper and lower bounds on $\mathfrak{R}$.

4 Private contextual bandits

The current lower bounds available in the literature in either the non-private Azar et al. [2017] or private Garcelon et al. [2021] case use a reduction to contextual bandits. It is therefore natural to first study this simpler case. We therefore set in this section $H = 1$. In this case, the transition kernel P plays no role and since A and S are fixed, an instance of contextual bandits is characterized by the reward function r and the initial distributions $\{\rho^k\}_{k=1}^K$.

4.1 Lower bound

In Theorem 1, we show a lower bound on $\mathfrak{R}$.

Theorem 1. *Assume $A \geq 2$, for any ϵ-LDP mechanism $\mathcal{M}$ and any algorithm $\mathfrak{A}$ that the decision maker can choose, there exists an instance I characterized by a reward function r and initial distributions $\{\rho^k\}_{k=1}^K$ such that the regret is lower bounded as*

$$\mathcal{R}(\mathfrak{A}, \mathcal{M}, I) \geq \frac{1}{32} \min\left(K, \frac{\sqrt{\max(S, \exp(\epsilon)+1)SAK(\exp(\epsilon)+1)}}{\exp(\epsilon)-1}\right)$$

Proof sketch (see Appendix A for the full proof). Fix an algorithm $\mathfrak{A}$, a mechanism $\mathcal{M}$ and a gap $\Delta > 0$. For each $\phi \in \mathcal{A}^{\mathcal{S}}$ and $\sigma \in \mathcal{S}^K$, define the hard instance $I_{\phi,\sigma}$ by

$$r_{\phi,\sigma}(s,a) = \frac{1}{2} + \Delta \mathbb{1}\{a = \phi(s)\}, \qquad \rho^k_{\phi,\sigma}(s) = \mathbb{1}\{s = \sigma_k\}.$$

Let $I_{\alpha,c}$ be sampled uniformly from this family and let I^0_s be the same random instance with all rewards at state s replaced by $1/2$. A standard change-of-measure argument lower bounds the expected regret by the gap times the probability of not identifying the best action, and upper bounds this identification probability through $D_{\mathrm{KL}}((M_1, \ldots, M_K)_{s,0} \| (M_1, \ldots, M_K)_{s,j})$.

The improvement comes from controlling this KL term using both LDP contractions of [Asoodeh and Zhang, 2024, Theorems 1 and 2], namely

$$D_{\mathrm{KL}}(M_1 \| M_2) \leq \frac{(\exp(\epsilon)-1)^2}{(\exp(\epsilon)+1)^2} D_{\mathrm{KL}}(\tau_1 \| \tau_2),$$
$$D_{\mathrm{KL}}(M_1 \| M_2) \leq \frac{(\exp(\epsilon)-1)^2}{\exp(\epsilon)} D_{\mathrm{TV}}(\tau_1, \tau_2)^2.$$

Combining these inequalities with the chain rule yields the two regimes in the bound. Optimizing Δ and using the trivial upper bound $\mathcal{R} \leq K$ concludes. □

When $\epsilon \leq 1$, the lower bound of Theorem 1 is in $\Omega(\frac{S\sqrt{AK}}{\epsilon})$. Compared to the non-private rate of $\tilde{\Theta}(\sqrt{SAK})$ obtained by Azar et al. [2017], there is a multiplicative cost of $\frac{\sqrt{S}}{\epsilon}$ due to privacy. When $\epsilon \leq 1$, the rate of private bandits with SA arms is in $\Theta(\sqrt{SAK}/\epsilon)$, the results of Theorem 1 shows that private contextual bandit with S context and A arms is strictly harder as a factor of order $\sqrt{S}$ is gained.

Intuitively the difficulty in private bandits with S context and A arms is that since context must remain private, when a privatized reward is received it is not clear from which bandit it came. The computation of the TV distance in the proof of Theorem 1 in Appendix A shows that a factor $\frac{1}{S}$ appears become of the randomness of the context. This ultimately leads to the lower bound in $\frac{S\sqrt{AK}}{\epsilon}$ in the high privacy regime $\epsilon \leq 1$.

Remark 1 (Sequentially interactive mechanisms)**.** *In the setting of Section 3, the same mechanism $\mathcal{M}$ privatizes every trajectory. Theorem 1 holds more generally for* sequentially interactive *ϵ-LDP mechanisms (Definition 1 in Appendix A), for which the ϵ-LDP channel used at episode k may depend on the previous messages $M^1, \ldots, M^{k-1}$. The proof only uses that, given the past messages, the same ϵ-LDP channel is applied under the two instances being compared, whether or not this channel depends on the history. The proof in Appendix A is written for this more general class.*

4.2 Upper bound

In this subsection, we show that the lower bound of Theorem 1 is essentially tight by showing that there exists an algorithm and a mechanism achieving a regret upper bound matching the lower bound up to constants and a multiplicative $\sqrt{\log(A)}$ factor.

The algorithm we use is based on Exp3 with forced exploration Auer et al. [2002] and the mechanism we use is based on the optimized unary encoding mechanism Wang et al. [2017]. We name the procedure Exp3-OUE and give an implementation in Algorithm 1.

We first describe the optimized unary encoding mechanism (OUE). Given $v \in \{0,1\}^d$ with $\|v\|_1 \leq 1$, the OUE mechanism outputs a vector $\text{OUE}(v) \in \{0,1\}^d$ where each coordinate is generated independently as follows:

$$\text{OUE} : v \in \{w \in \{0,1\}^d : \|w\|_1 \leq 1\} \to \text{OUE}(v) \in \{0,1\}^d \quad \text{such that}$$
$$\forall i \in [d], \text{OUE}(v)_i \mid v_i \sim \begin{cases} \mathcal{B}(\frac{1}{2}), \text{if } v_i = 1 \\ \mathcal{B}\left(\frac{1}{\exp(\epsilon)+1}\right), \text{if } v_i = 0 \end{cases} . \tag{1}$$

As shown in Wang et al. [2017], OUE is ϵ-LDP. This mechanism is optimal for histogram estimation under LDP Canonne and Gentle [2024]. It is therefore a natural candidate for our problem.

We now describe Exp3-OUE informally. At episode k, the user receives a deterministic policy π^k from the decision-maker and executes it, which yields a trajectory $\tau^k = (S^k, A^k, R^k)$. Then, the user binarizes the reward by sampling $B^k \sim \mathcal{B}(R^k)$ and constructs a vector V^k of dimension S where $V^k(s) = \mathbb{1}\{S^k = s, B^k = 1\}$. This vector is then privatized by OUE yielding a vector $M^k = \text{OUE}(V^k)$ sent back to the decision-maker. The decision maker then builds an importance-weighted estimator $\hat{V}^k(s,a)$ of the reward for each state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$ and uses it to update the weights of Exp3 as $w_{k+1} = w_k \exp(\eta \hat{V}^k)$ where η is a learning rate parameter and w_k is known from previous episodes (we set $w_1 = (1, \dots, 1)$). Following the Exp3 algorithm, these weights are used to compute the new policy as in Equation (2). For every state $s \in \mathcal{S}$, we sample α_s from $\Pi^{k+1}(\cdot, s)$ and define the deterministic policy π^{k+1} by $\pi^{k+1}(s) = \alpha_s$. The policy π^{k+1} is then sent to the user of episode $k+1$ and the process continues.

Algorithm 1: Exp3 with optimized unary encoding for private contextual bandits (Exp3-OUE)

Input: ϵ, K, $\mathcal{A}$, $\mathcal{S}$, η, γ

▷ Decision maker (public operations)

Initialize $w_1(s,a) = 1$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$

Define the distributions $\Pi^1(\cdot, s)$ as $\Pi^1(a,s) = (1-\gamma)\frac{w_1(s,a)}{\sum_{a' \in \mathcal{A}} w_1(s,a')} + \frac{\gamma}{A}$ for each $s, a \in \mathcal{S} \times \mathcal{A}$

Set $\pi^1(s) = \alpha_s$ where $\alpha_s \sim \Pi^1(\cdot, s)$ and send π^1 to user 1.

for $k = 1$ **to** K **do**

▷ User k (private operations)

Observe S^k, select $A^k = \pi^k(S^k)$, observe R^k, and sample $B^k \sim \mathcal{B}(R^k)$

Construct $V^k(s) = \mathbb{1}\{S^k = s\} B^k$ for all $s \in \mathcal{S}$

Construct $M^k \in \{0,1\}^{\mathcal{S}} = \text{OUE}(V^k)$ via the OUE mechanism Equation (1)

Send M^k to the decision maker

▷ Decision maker (public operations)

For each $s \in \mathcal{S}$ and $a \in \mathcal{A}$, construct the estimator

$$\hat{V}^k(s,a) = M^k(s) \frac{\mathbb{1}\{a = \pi^k(s)\}}{\Pi^k(a,s)}$$

Update the weights as for each $s, a \in \mathcal{S} \times \mathcal{A}$,

$$w_{k+1}(s,a) = w_k(s,a) \exp(\eta \hat{V}^k(s,a)).$$

Define $\Pi^{k+1}(\cdot, s)$ as for each $s, a \in \mathcal{S} \times \mathcal{A}$,

$$\Pi^{k+1}(a,s) = (1-\gamma)\frac{w_{k+1}(s,a)}{\sum_{a' \in \mathcal{A}} w_{k+1}(s,a')} + \frac{\gamma}{A}. \tag{2}$$

Set $\pi^{k+1}(s) = \alpha_s$ where $\alpha_s \sim \Pi^{k+1}(\cdot, s)$ for each $s \in \mathcal{S}$ and send π^{k+1} to user $k+1$.

In essence, Exp3-OUE is Exp3 applied to privatized input, but two implementation choices make a significant difference compared to a direct Laplace-based LDP-OBI Garcelon et al. [2021] adapted to the bandit case. First, in Exp3-OUE, the decision-maker sends a deterministic realization π^k of Π^k rather than Π^k directly, so the user privatizes a vector indexed by states rather than by state-action

pairs, improving the dependence on A. Second, the user binarizes the reward and applies OUE, which improves the dependence on ϵ. The estimator $\hat{V}^k(s,a)$ also gives the right bias-variance trade-off; debiasing it would increase the regret.

In Theorem 2, we precisely characterize the regret of Exp3-OUE and show that it matches the lower bound of Theorem 1 up to constants and a multiplicative $\sqrt{\log(A)}$ factor.

Theorem 2. *Assume $A \geq 2$, for any instance I characterized by a reward function r and initial distributions $\{\rho^k\}_{k=1}^K$, the regret of Algorithm 1 run with $\gamma = A\eta$ and*

$$\eta = \min\Big(\frac{1}{2A}, \sqrt{\frac{S\ln(A)}{2(e-1)KA\big(\frac{1}{2} - \frac{1}{\exp(\epsilon)+1} + \frac{S}{\exp(\epsilon)+1}\big)}}\Big)$$

is upper bounded as

$$\mathcal{R}(\mathfrak{A}, \mathcal{M}, I) \leq 16 \min\left(K, \frac{\sqrt{\max(S, \exp(\epsilon)+1)SAK(\exp(\epsilon)+1)\log(A)}}{\exp(\epsilon)-1}\right)$$

Proof sketch (see Appendix B for the full proof). The proof follows the standard Exp3 potential argument [Auer et al., 2002, Theorem 3.1]. Let

$$\Phi_k = \sum_{s\in\mathcal{S}} \log\Big(\sum_{a\in\mathcal{A}} w_k(s,a)\Big).$$

Since $\hat{V}^k(s,a) \in [0, A/\gamma]$ and $\gamma = A\eta$, we have $\eta\hat{V}^k(s,a) \in [0,1]$. Expanding $\Phi_{k+1} - \Phi_k$, summing over k, using $\Phi_1 = S\log(A)$ and comparing with the best fixed action $a^*(s)$ gives, after taking expectations,

$$\mathbb{E}\Big[\sum_{k=1}^K r(S^k, a^*(S^k)) - \sum_{k=1}^K r(S^k, A^k)\Big] \leq \frac{S\ln(A)}{\eta(\frac{1}{2} - \frac{1}{\exp(\epsilon)+1})} + \frac{2(e-1)KA\eta(\frac{1}{2} - \frac{1}{\exp(\epsilon)+1} + \frac{S}{\exp(\epsilon)+1})}{\frac{1}{2} - \frac{1}{\exp(\epsilon)+1}}.$$

Plugging in the chosen learning rate and using that regret is at most K yields the theorem. □

As a consequence of Theorem 1 and Theorem 2, we have that the lower bound in Theorem 1 is tight up to log-factor. Therefore, the minimax regret $\mathfrak{R}$ for private contextual bandits satisfies

$$\mathfrak{R} = \tilde{\Theta}\left(\min\left(K, \frac{\sqrt{\max(S, \exp(\epsilon)+1)SAK(\exp(\epsilon)+1)}}{\exp(\epsilon)-1}\right)\right).$$

5 Private Tabular Reinforcement Learning

In this section, we attempt to extend the lower bound of Theorem 1 and the upper bound of Theorem 2 to the case of finite-horizon MDPs. We first focus on lower bound and see in Corollary 1 that the usual reduction from reinforcement learning to contextual bandits applies in the private setting as well. This yield a lower bound for private tabular reinforcement learning that improves over previous work Garcelon et al. [2021]. We then establish an upper bound in Theorem 3, adapting the ideas used to prove Theorem 2 to the case of finite-horizon MDPs. In the high privacy regime $\epsilon \leq 1$, this upper bound improves over known upper bounds for private tabular reinforcement learning Garcelon et al. [2021], Qiao and Wang [2023], Chowdhury and Zhou [2022b].

5.1 Lower bound

We first show the following corollary of Theorem 1.

Corollary 1. *Assume $S \geq 4$ and $A \geq 2$ and $H \geq 2$. For any ϵ-LDP mechanism $\mathcal{M}$ and any algorithm $\mathfrak{A}$ that the decision maker can choose, there exists an instance I of finite-horizon MDP such that the regret is lower bounded as*

$$\mathcal{R}(\mathfrak{A}, \mathcal{M}, I) \geq \frac{H}{128} \min\left(K, \frac{\sqrt{\max(S, \exp(\epsilon)+1)SAK(\exp(\epsilon)+1)}}{\exp(\epsilon)-1}\right)$$

Proof sketch (see Appendix C for the full proof). Consider a finite-horizon MDP with $\mathcal{S} = [S], \mathcal{A} = [A]$. ρ^k has support over $[S-2]$. The other 2 states are labelled "good" and "bad". Both good and bad states are absorbing and yield a reward of 1 and 0 respectively. For any $s \in [S-2]$, any actions lead to the good state with some probability and to the bad state with the remaining probability with the reward being 0 in both cases. This MDP is identical to a contextual bandit problem with $S-2$ contexts and A actions except that rewards are scaled by $H-1$. Then, choose P and ρ such that the hard instance of Theorem 1 is obtained and the lower bound is proved. □

The lower bound in Garcelon et al. [2021] is given (up to constants) by $H\frac{\sqrt{SAK}}{\min(\exp(\epsilon)-1,1)}$ a rate that is lower than that of Corollary 1 for any value of H, S, A, K, ϵ. More precisely, the two bounds coincide when ϵ is on the order of $\log S$ or larger, and when $\epsilon \leq \log S$ the improvement is a multiplicative factor of order $\sqrt{S/\exp(\epsilon)}$.

The reason why the lower bound in Theorem 1 dominates is because in Garcelon et al. [2021], authors construct a tree in which one starts at the root and moves deterministically depending on the actions played, the number of leaves being of order S, and in which one special leaf-action pair is slightly more rewarding than the others. Their construction therefore reduces the problem to a private bandit instance with SA arms. In contrast, our lower bound reduces the problem to a private contextual bandit with S context and A arms which is a more difficult problem.

5.2 Upper bound

In this sub-section, we show an upper bound for the minimax regret improving over known upper bounds in the high privacy regime.It is obtained using a variant of UCBVI as the algorithm and the optimized unary encoding as the ϵ-LDP mechanism.

We name the procedure UCBVI-OUE, which we now present informally. A precise implementation is available in Algorithm 2. At episode k, the user receives a deterministic policy π^k and an index J^k from the decision-maker. The user executes the policy π^k, which yields a trajectory $\tau^k = (S_1^k, A_1^k, R_1^k, \dots, S_H^k, A_H^k, R_H^k)$. Then, the user binarizes the reward at step J^k by $B_{J^k}^k \sim \mathcal{B}(R_{J^k}^k)$. The user then constructs a vector $I^k = (\mathbb{1}\left\{S_{J^k}^k = s, B_{J^k}^k = b, S_{J^k+1}^k = s'\right\})_{s\in\mathcal{S}, b\in\{0,1\}, s'\in\mathcal{S}}$ of dimension $2S^2$ and privatizes it using the OUE mechanism to get $M^k = \text{OUE}(I^k)$. M^k is then sent back to the decision-maker who uses it to build for each $s \in \mathcal{S}$ and $a \in \mathcal{A}$, confidence sets $\mathcal{C}^k(s,a)$ which contain $P(s,a,\cdot)$ and $r(s,a)$ with high probability. Then, the decision maker computes optimistic Q-value estimates $\tilde{Q}^{k,(+)}$ using the confidence sets by backward induction as in Equation (6). Then the new policy is computed as $\pi_h^{k+1}(s) = \text{argmax}_{a\in\mathcal{A}} \tilde{Q}_h^{k,(+)}(s,a)$ for each $s \in \mathcal{S}$ and $h \in [H]$ and J^{k+1} is sampled uniformly at random from $[H]$. π^{k+1} and J^{k+1} are then sent to the user of episode $k+1$ and the process continues.

In addition to the two ideas already used in the case of contextual bandits, which are sending a deterministic policy instead of a randomized one and binarizing the reward to use the OUE mechanism, UCBVI-OUE also uses the idea of sending the indicator of only a triplet $S_{J^k}^k, B_{J^k}^k, S_{J^k+1}^k$ from the trajectory chosen uniformly at random as opposed to sending the indicator of the whole trajectory. This reduces the amount of noise that needs to be added to ensure LDP and ultimately leads to an estimator that has lower variance.

We now present Theorem 3, which gives an upper bound on the regret of Algorithm 2.

Theorem 3. *Assume $S \geq 4$, $A \geq 2$ and $H \geq 2$. For any instance I of finite-horizon MDP, the regret of Algorithm 2 is upper bounded up to constants and log-factors (see the exact expression in Appendix D) as*

$$\mathcal{R}(\mathfrak{A}, \mathcal{M}, I) \leq \tilde{O}\left(H^2S^2\sqrt{\Upsilon_\epsilon^2 AK} + H^2S^2A\Upsilon_\epsilon\right), \quad \Upsilon_\epsilon = \frac{\exp(\epsilon)+1}{\exp(\epsilon)-1}$$

Proof sketch (see Appendix D for the full proof). Let

$$N^k(s,a,b,s') = \sum_{i\in[k]}\sum_{h=1}^{H} \mathbb{1}\left\{S_h^i = s, B_h^i = b, S_{h+1}^i = s', \pi_h^i(s) = a\right\}$$

Algorithm 2: UCBVI with optimized unary encoding for private tabular RL (UCBVI-OUE)

Input: ϵ, K, $\mathcal{A}$, $\mathcal{S}$, H, $\Upsilon_\epsilon = \frac{\exp(\epsilon)+1}{\exp(\epsilon)-1}$, $q_\epsilon = \frac{1}{\exp(\epsilon)+1}$.

▷ Decision maker (public operations)

Initialize $\tilde{Q}_h^0(s,a) = H$ for all $s, a, h \in \mathcal{S} \times \mathcal{A} \times [H]$

Send $\pi_h^1(s) = \operatorname{argmax}_{a\in\mathcal{A}} \tilde{Q}_h^0(s,a)$ to user 1 for each $s \in \mathcal{S}$ and $h \in [H]$.

Sample J^1 uniformly at random from $[H]$ and send J^1 to user 1.

for $k = 1$ **to** K **do**

▷ User k (private operations)

Play according to policy π^k and observe $(S_h^k, A_h^k, R_h^k, S_{h+1}^k)_{h\in[H]}$

Sample $B_h^k \sim \mathcal{B}(R_h^k)$ for each $h \in [H]$

Construct $I^k(s,b,s') = \mathbb{1}\left\{S_{J^k}^k = s, B_{J^k}^k = b, S_{J^k+1}^k = s'\right\}$ for all $s, s' \in \mathcal{S}$, $b \in \{0,1\}$.

Construct $M^k \in \{0,1\}^{\mathcal{S}} = \text{OUE}(I^k)$ via the OUE mechanism Equation (1)

Send M^k to the decision maker

▷ Decision maker (public operations)

For each $s, b, s' \in \mathcal{S} \times \{0,1\} \times \mathcal{S}$, define $\hat{I}^k(s,b,s') = 2\Upsilon_\epsilon(M^k(s,b,s') - q_\epsilon)$

For each $s, a, b, s' \in \mathcal{S} \times \mathcal{A} \times \{0,1\} \times \mathcal{S}$, compute $\beta(s,a)$ as in Equation (7) and define

$$\hat{N}^k(s,a,b,s') = H \sum_{i\in[k]} \hat{I}^i(s,b,s')\mathbb{1}\left\{\pi_{J^i}^i(s) = a\right\} \tag{3}$$

$$\tilde{N}^{k,(-)}(s,a) = \max(1, \sum_{b\in\{0,1\}, s'\in\mathcal{S}} (\hat{N}^k(s,a,b,s') - \beta_k(s,a))) \tag{4}$$

Then, build confidence sets

$$\mathcal{C}^k(s,a) = \Bigg\{\nu \in \Delta(\mathcal{S}), \rho \in [0,1] : |\tilde{N}^{k,(-)}(s,a)\rho - \sum_{s'\in\mathcal{S}} \hat{N}^k(s,a,1,s')| \le 7S\beta_k(s,a)$$
$$\|\tilde{N}^{k,(-)}(s,a)\nu(\cdot) - \sum_{b\in\{0,1\}} \hat{N}^k(s,a,b,\cdot)\|_1 \le 8S\beta_k(s,a)\Bigg\} \tag{5}$$

Set $\tilde{Q}_{H+1}^{k,(+)}(s,a) = 0$ for each $s, a \in \mathcal{S} \times \mathcal{A}$ and then compute

$$\tilde{Q}_h^{k,(+)}(s,a) = \max_{\nu,\rho\in\mathcal{C}^k(s,a)} \left(\rho + \sum_{s'\in\mathcal{S}} \nu(s') \max_{a'\in\mathcal{A}} \tilde{Q}_{h+1}^{k,(+)}(s',a')\right) \tag{6}$$

Send $\pi_h^{k+1}(s) = \operatorname{argmax}_{a\in\mathcal{A}} \tilde{Q}_h^{k,(+)}(s,a)$ to user $k+1$ for each $s \in \mathcal{S}$ and $h \in [H]$.

Sample J^{k+1} uniformly at random from $[H]$ and send J^{k+1} to user $k+1$.

and let $\hat{N}^k$ be the private estimator in Equation (4). Freedman's inequality Hsu et al. [2015] implies (Lemma 1) that, with probability at least $1-\delta$, for all s, a, b, s', k,

$$|\hat{N}^k(s,a,b,s') - N^k(s,a,b,s')| \leq \beta_k(s,a),$$

$$\beta_k(s,a) = \sqrt{20H\Upsilon_\epsilon^2 \sum_{h\in[H],i\in[k]} \mathbb{1}\{a=\pi_h^i(s)\} \log\frac{48K^2S^2A}{\delta}} + 4H\Upsilon_\epsilon \log\frac{48K^2S^2A}{\delta}. \quad (7)$$

Furthermore, Lemma 3 shows $\mathcal{C}^k(s,a)$ Equation (5) contains with high probability the true reward and transition probabilities for all s, a, k. By Lemma 2, under these events, the optimistic Q-values then satisfy the Bellman-type bounds

$$0 \leq \tilde{Q}_h^{k,(+)}(s,a) - \left(r(s,a) + \sum_{s'\in\mathcal{S}} P(s' \mid s,a) \max_{a'\in\mathcal{A}} \tilde{Q}_{h+1}^{k,(+)}(s',a')\right) \leq b_k(s,a,h)$$

$$\tilde{Q}_h^{k,+}(s,a) \leq H-h+1$$

where $b_k(s,a,h) = \min(H-h+1, 18S\beta_K(s,a)H/\max(1, N^k(s,a) - 4S\beta_K(s,a)))$. As shown by Lemma 4, optimism reduces regret to bounding $\sum_{k,h} b_{k-1}(S_h^k, A_h^k)$; the counting argument in Appendix D gives the stated rate. □

The upper bound shown in Theorem 3 is clearly sub-optimal in the low privacy regime. As $\epsilon \to \infty$, the upper bound becomes $\tilde{O}(H^2S^2\sqrt{AK})$ which is much worse than the optimal non-private regret rate in $\Theta(H\sqrt{SAK})$. Furthermore, the private term in the upper bound of Garcelon et al. [2021] is given by $H^3S^2A\sqrt{K}/\epsilon$ which is better than the bound of Theorem 3 as soon as $\epsilon \geq H\sqrt{A}$. However, when $\epsilon \leq H\sqrt{A}-2$, the bound of Theorem 3 dominates and when $\epsilon \leq 1$, the gain is of order $H\sqrt{A}$. The gain of the $\sqrt{A}$ factor is linked to the fact that we send a deterministic version of the policy and that we privatize a randomly chosen triplet from the trajectory. Indeed, privatizing a quantity such as $\sum_{h\in[H]} \mathbb{1}\{S_h^k = s, A_h^k = a, S_{h+1}^k = s'\}$ by adding Laplace noise as done in Garcelon et al. [2021] gives an estimator of variance H^2/ϵ^2. Instead, privatizing $\mathbb{1}\left\{S_{J^k}^k = s, A_{J^k}^k = a, S_{J^k+1}^k = s'\right\}$ can be done with a noise of scale $1/\epsilon^2$ and the variance of the resulting estimator is on the order of $H/\epsilon^2 \sum_{h=1}^H \mathbb{1}\{\pi_h^k(s) = a\}$ which can be much smaller for actions rarely selected. The gain of the H factor is due to a tighter analysis. In Garcelon et al. [2021], the summation of the bonuses introduces an additional factor on top of the one due to the Bellman updates and the one due to the privacy-noise variance (see their Equation (50)). In contrast, our careful handling of the counts $\sum_{k,h} b_{k-1}(S_h^k, A_h^k)$ in the proof of Theorem 3 allows us not to pay this additional factor.

6 Conclusion and future work

We improved regret bounds for tabular RL under local differential privacy. For contextual bandits, the minimax regret is characterized up to constants and a $\sqrt{\log(A)}$ factor Theorems 1 and 2. For finite-horizon RL, our reduction Corollary 1 gives a lower bound that is H times the contextual bandit lower bound. The improvement over previous results increases as ϵ decreases, yielding a $\sqrt{S}$ improvement when $\epsilon \leq 1$. Our UCBVI-OUE upper bound Theorem 3 improves over previous results when $\epsilon \leq \max(H\sqrt{A}-2, 1)$, which includes the high privacy regime $\epsilon \leq 1$, where a $H\sqrt{A}$ improvement is obtained. We note that when $\epsilon \to \infty$ the upper bound of Theorem 3 becomes $\tilde{O}(H^2S^2\sqrt{AK})$ which is much worse than the optimal non-private regret rate in $\Theta(H\sqrt{SAK})$. In fact, when $\epsilon \geq H\sqrt{A}$, the approach of Garcelon et al. [2021] is preferable.

The main open question is whether the lower bound of Corollary 1 is tight. In non-private problems, contextual bandits and tabular RL differ by a factor H, and it is natural to conjecture that this remains true under LDP. How to close the remaining HS gap in the high-privacy regime is then the most pressing question. We conjecture that closing the S-factor gap requires model-free approaches, as model-based methods struggle with the high dimensionality of the transition kernel. However, existing analyses of model-free algorithms produce suboptimal rates even without privacy constraints. These challenging questions remain open for future investigation.

References

M. M. Afsar, T. Crump, and B. Far. Reinforcement Learning based Recommender Systems: A Survey. *ACM Computing Surveys*, 55(7):1–38, July 2023. ISSN 0360-0300, 1557-7341. doi: 10.1145/3543846.

S. Asoodeh and H. Zhang. Contraction of Locally Differentially Private Mechanisms. *IEEE Journal on Selected Areas in Information Theory*, 5:385–395, 2024. ISSN 2641-8770. doi: 10.1109/JSAIT.2024.3397305.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The Nonstochastic Multiarmed Bandit Problem. *SIAM Journal on Computing*, 32(1):48–77, Jan. 2002. ISSN 0097-5397, 1095-7111. doi: 10.1137/S0097539701398375.

M. G. Azar, I. Osband, and R. Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

A. Azize and D. Basu. When privacy meets partial information: A refined analysis of differentially private bandits. *Advances in Neural Information Processing Systems*, 35:32199–32210, 2022.

A. Azize, M. Jourdan, A. A. Marjani, and D. Basu. Differentially Private Best-Arm Identification, Apr. 2026.

H. Cai, K. Ren, W. Zhang, K. Malialis, J. Wang, Y. Yu, and D. Guo. Real-Time Bidding by Reinforcement Learning in Display Advertising. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 661–670, Cambridge United Kingdom, Feb. 2017. ACM. ISBN 978-1-4503-4675-7. doi: 10.1145/3018661.3018702.

C. L. Canonne and A. Gentle. Locally Private Histograms in All Privacy Regimes, Sept. 2024.

X. Chen, S. Miao, and Y. Wang. Differential privacy in personalized pricing with nonparametric demand models. *Operations Research*, 71(2):581–602, 2023.

S. R. Chowdhury and X. Zhou. Shuffle private linear contextual bandits. *arXiv preprint arXiv:2202.05567*, 2022a.

S. R. Chowdhury and X. Zhou. Differentially private regret minimization in episodic markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6375–6383, 2022b.

C. Dann, L. Li, W. Wei, and E. Brunskill. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pages 1507–1516. PMLR, 2019.

H. Devlin and T. Burgis. Confidential Health Records from {UK Biobank} Project Exposed Online. 2026.

J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438. IEEE, 2013.

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

E. Garcelon, K. Chaudhuri, V. Perchet, and M. Pirotta. Privacy amplification via shuffling for linear contextual bandits. In *International Conference on Algorithmic Learning Theory*, pages 381–407. PMLR, 2022.

Y. Han, Z. Liang, Y. Wang, and J. Zhang. Generalized linear bandits with local differential privacy. *Advances in neural information processing systems*, 34:26511–26522, 2021.

D. J. Hsu, A. Kontorovich, and C. Szepesvari. Mixing Time Estimation in Reversible Markov Chains from a Single Sample Path. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan. Is Q-learning Provably Efficient?, July 2018.

S. Jones. Having an affair may shorten your life: The ashley Madison suicides. *Ga. St. UL Rev.*, 33: 455, 2016.

J. Kober, J. A. Bagnell, and J. Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, Sept. 2013. ISSN 0278-3649, 1741-3176. doi: 10.1177/0278364913495721.

T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Y. Ma, F. Jiang, Z. Zhao, H. Yang, and Y. Yu. Locally private nonparametric contextual multi-armed bandits. *arXiv preprint arXiv:2503.08098*, 2025.

A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (Sp 2008)*, pages 111–125. IEEE, 2008.

M. Nazari, A. Oroojlooy, L. Snyder, and M. Takác. Reinforcement learning for solving the vehicle routing problem. *Advances in neural information processing systems*, 31, 2018.

N. Pavlovic, S. Salgia, and Q. Zhao. Differential privacy in kernelized contextual bandits via random projections. *arXiv preprint arXiv:2507.13639*, 2025.

M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

D. Qiao and Y.-X. Wang. Near-optimal differentially private reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 9914–9940. PMLR, 2023.

W. Ren, X. Zhou, J. Liu, and N. B. Shroff. Multi-Armed Bandits with Local Differential Privacy, July 2020.

L. Rocher and J. Morley. Protecting health data at UK Biobank, 2026.

L. Rocher, J. M. Hendrickx, and Y.-A. De Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature communications*, 10(1):3069, 2019.

R. Shariff and O. Sheffet. Differentially private contextual linear bandits. *Advances in Neural Information Processing Systems*, 31, 2018.

L. Sweeney. Simple demographics often identify people uniquely. Technical Report Data Privacy Working Paper No. 3, Carnegie Mellon University, Pittsburgh, PA, 2000. URL `https://dataprivacylab.org`.

G. Vietri, B. Balle, A. Krishnamurthy, and S. Wu. Private reinforcement learning with pac and regret guarantees. In *International Conference on Machine Learning*, pages 9754–9764. PMLR, 2020.

T. Wang, J. Blocki, N. Li, and S. Jha. Locally differentially private protocols for frequency estimation. In *26th USENIX Security Symposium (USENIX Security 17)*, pages 729–745, 2017.

A. C.-C. Yao. Probabilistic computations: Toward a unified measure of complexity. In *18th Annual Symposium on Foundations of Computer Science (Sfcs 1977)*, pages 222–227. IEEE, 1977.

C. Yu, J. Liu, S. Nemati, and G. Yin. Reinforcement Learning in Healthcare: A Survey. *ACM Computing Surveys*, 55(1):1–36, Jan. 2023. ISSN 0360-0300, 1557-7341. doi: 10.1145/3477600.

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

Z. Zhang, Y. Zhou, and X. Ji. Almost Optimal Model-Free Reinforcement Learning via Reference-Advantage Decomposition. https://arxiv.org/abs/2004.10019v2, Apr. 2020.

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

Z. Zhao, F. Jiang, and Y. Yu. Contextual dynamic pricing: Algorithms, optimality, and local differential privacy constraints. *Journal of the American Statistical Association*, pages 1–16, 2026.

K. Zheng, T. Cai, W. Huang, Z. Li, and L. Wang. Locally differentially private (contextual) bandits learning. *Advances in Neural Information Processing Systems*, 33:12300–12310, 2020.

A Proof of Theorem 1

We prove Theorem 1 for the following class of mechanisms, which contains the mechanisms of Section 3.

Definition 1 (Sequentially interactive ϵ-LDP mechanism)**.** *Following Duchi et al. [2013], a* sequentially interactive *mechanism is a sequence of channels $\mathcal{M} = (Q^k)_{k\in[K]}$ where, for each trajectory $\tau \in \mathbb{T}$ and each sequence of past messages $y \in \mathcal{Y}^{k-1}$, $Q^k(\cdot \mid \tau, y)$ is a probability distribution over $\mathcal{Y}$. At episode k, user k sends $M^k \sim Q^k(\cdot \mid \tau^k, (M^1, \dots, M^{k-1}))$, drawn independently of everything else given τ^k and $(M^1, \dots, M^{k-1})$. The mechanism is ϵ-LDP if, for every $k \in [K]$ and every $y \in \mathcal{Y}^{k-1}$, the channel $Q^k(\cdot \mid \cdot, y)$ is ϵ-LDP: for any two trajectories τ, τ' and any measurable subset $Y \subset \mathcal{Y}$,*

$$Q^k(Y \mid \tau, y) \le e^\epsilon Q^k(Y \mid \tau', y).$$

The (non-interactive) mechanisms of Section 3 are the special case where $Q^k(\cdot \mid \tau, y)$ is the distribution of $\mathcal{M}(\tau)$ for every k and y. Since the policy π^k of a deterministic algorithm is a function of $(M^1, \dots, M^{k-1})$, the channel used at episode k may in particular depend on π^k. The regret is defined as in Section 3, with M^k drawn as above.

Proof. Fix $\Delta > 0$, we will choose its value later. Given $\phi \in \mathcal{A}^\mathcal{S}$ and $\sigma \in \mathcal{S}^K$, we define an instance $I_{\phi,\sigma} = (r_{\phi,\sigma}, \rho_{\phi,\sigma})$ by

$$r_{\phi,\sigma}(s,a) = \frac{1}{2} + \Delta \mathbb{1}\{a = \phi(s)\}, \quad \rho^k_{\phi,\sigma}(s) = \mathbb{1}\{s = \sigma_k\}$$

Then consider the random instance $I = (r_I, \rho_I) = I_{\alpha,c}$ where $\alpha(s)$ is chosen uniformly at random from $\mathcal{A}$ for each $s \in \mathcal{S}$ and c_k is chosen uniformly at random from $\mathcal{S}$ for each $k \in [K]$. In other words, I is uniformly distributed over the set of instances $\{I_{\phi,\sigma} : \phi \in \mathcal{A}^\mathcal{S}, \sigma \in \mathcal{S}^K\}$. Lastly, consider instance $I^0_s = (r^0_s, \rho^0_s)$ defined as I except that $r^0_s(s,a) = \frac{1}{2}$ for each $a \in \mathcal{A}$ and $r^0_s(s',a) = r_I(s',a)$ for each $s' \neq s$ and $a \in \mathcal{A}$.

Fix a deterministic algorithm $\mathfrak{A}$ and a sequentially interactive ϵ-LDP mechanism $\mathcal{M} = (Q^k)_{k\in[K]}$ (Definition 1). We will show a lower bound on the expected regret of $(\mathfrak{A}, \mathcal{M})$ on the random instance I and then conclude the proof by application of Yao's minimax principle Yao [1977].

The expected regret can be expressed as

$$
\begin{aligned}
&\mathbb{E}[\mathcal{R}(\mathfrak{A}, \mathcal{M}, I)] \\
&= \mathbb{E}\big[\sum_{k=1}^{K} \Big(V^*(S^k) - V^{\pi^k}(S^k)\Big)\big] \\
&= \sum_{s\in\mathcal{S}} \mathbb{E}\big[\sum_{k=1}^{K} \mathbb{1}\big\{S^k = s\big\} \Big(V^*(s) - V^{\pi^k}(s)\Big)\big] \\
&= \sum_{s\in\mathcal{S}} \mathbb{E}\big[\sum_{k=1}^{K} \mathbb{1}\big\{S^k = s\big\}\, \Delta \mathbb{1}\big\{A^k \neq \alpha(s)\big\}\big] \\
&= \Delta\Big(K - \sum_{s\in\mathcal{S}} \mathbb{E}\Big[\sum_{k=1}^{K} \mathbb{1}\big\{S^k = s, A^k = \alpha(s)\big\}\Big]\Big) \\
&= \Delta\Big(K - \sum_{s\in\mathcal{S}}\sum_{j\in\mathcal{A}} \mathbb{E}\Big[\sum_{k=1}^{K} \mathbb{1}\big\{A^k = j\big\} \mid S^k = s, \alpha(s) = j\Big] \frac{1}{SA}\Big) \\
&= \Delta\Big(K - \sum_{s\in\mathcal{S}}\sum_{j\in\mathcal{A}} \mathbb{E}\Big[\underbrace{\sum_{k=1}^{K} \mathbb{E}\Big[\mathbb{1}\big\{A^k = j\big\} \mid S^k = s, (M^i)_{i\in[k-1]}\Big]}_{N_{s,j}} \mid \alpha(s) = j\Big] \frac{1}{SA}\Big). \qquad (8)
\end{aligned}
$$

Call $\nu_{s,j}$ and $\nu_{s,0}$ the distributions of the sequence of privatized messages $(M^i)_{i\in[K]}$ on instance I conditioned on $\alpha(s) = j$ and on instance I_s^0 respectively, and denote $\mathbb{E}_{s,j}$ and $\mathbb{E}_{s,0}$ the corresponding expectations.

Since $N_{s,j}$ is bounded by K and is a function of the sequence of privatized messages $(M^i)_{i\in[K]}$, from Auer et al. [2002] (see also [Lattimore and Szepesvári, 2020, Exercise 15.2]), we have

$$
\mathbb{E}_{s,j}\Big[N_{s,j}\Big] \leq \mathbb{E}_{s,0}\Big[N_{s,j}\Big] + K\sqrt{\frac{1}{2} D_{\mathrm{KL}}(\nu_{s,0}\|\nu_{s,j})}. \qquad (9)
$$

Let us now detail the KL divergence $D_{\mathrm{KL}}(\nu_{s,0}\|\nu_{s,j})$. To simplify notation, we write KL between random variables to mean the KL between the distributions they are sampled from. Furthermore, for a random variable X that depends on the instance we consider, we write $D_{\mathrm{KL}}(X \mid \alpha(s) = 0 \| X \mid \alpha(s) = j)$ to mean the KL divergence between the distribution of X when the instance is sampled from I_s^0 and the distribution of X when the instance is sampled from $I \mid \alpha(s) = j$.

For any $\ell \in [K]$, let $H^\ell = (M^i)_{i=1}^{\ell}$, we have

$$
\begin{aligned}
&D_{\mathrm{KL}}((M^i)_{i\in[K]} \mid \alpha(s) = 0 \| (M^i)_{i\in[K]} \mid \alpha(s) = j) \\
&= \sum_{\ell=0}^{K-1} \mathbb{E}_{s,0}\Big[D_{\mathrm{KL}}(M^{\ell+1} \mid \alpha(s) = 0, H^\ell \| M^{\ell+1} \mid \alpha(s) = j, H^\ell)\Big]. \qquad (10)
\end{aligned}
$$

where the last equality is a direct consequence of the chain rule for KL divergence. We now detail the KL divergence term $D_{\mathrm{KL}}(M^{\ell+1} \mid \alpha(s) = 0, H^\ell \| M^{\ell+1} \mid \alpha(s) = j, H^\ell)$. Let $P_0^{\ell+1}$ and $P_j^{\ell+1}$ denote the distributions of $(S^{\ell+1}, A^{\ell+1}, R^{\ell+1}) \mid \alpha(s) = 0, H^\ell$ and of $(S^{\ell+1}, A^{\ell+1}, R^{\ell+1}) \mid \alpha(s) = j, H^\ell$. Conditionally on H^ℓ, the message $M^{\ell+1}$ is obtained by passing $(S^{\ell+1}, A^{\ell+1}, R^{\ell+1})$ through the channel $Q^{\ell+1}(\cdot \mid \cdot, H^\ell)$. Hence, the distributions of $M^{\ell+1} \mid \alpha(s) = 0, H^\ell$ and of $M^{\ell+1} \mid \alpha(s) = j, H^\ell$ are $P_0^{\ell+1}Q^{\ell+1}$ and $P_j^{\ell+1}Q^{\ell+1}$, where $PQ^{\ell+1}$ denotes the distribution of the output of $Q^{\ell+1}(\cdot \mid \cdot, H^\ell)$ when its input is drawn from P. The same channel appears in both distributions, and it is ϵ-LDP by Definition 1. Therefore, [Asoodeh and Zhang, 2024, Theorem 2]

gives

$$D_{\text{KL}}(M^{\ell+1} \mid \alpha(s)=0, H^\ell \| M^{\ell+1} \mid \alpha(s)=j, H^\ell) \\ \leq 4\frac{(\exp(\epsilon)-1)^2}{\exp(\epsilon)} D_{\text{TV}}(S^{\ell+1}, A^{\ell+1}, R^{\ell+1} \mid \alpha(s)=0, H^\ell \| S^{\ell+1}, A^{\ell+1}, R^{\ell+1} \mid \alpha(s)=j, H^\ell)^2 \tag{11}$$

We now upper bound the TV divergence term. Notice that, the distribution of $S^{\ell+1}$ is independent of α and H^ℓ, and the distribution of $A^{\ell+1}$ given $S^{\ell+1}$ and H^ℓ is also independent of α since $\pi^{\ell+1}$ is a function of H^ℓ. Furthermore, given $S^{\ell+1}$ and $A^{\ell+1}$, the reward $R^{\ell+1}$ is independent of H^ℓ. Since both instances share the optimal actions $\phi(s')$ at the states $s' \neq s$, its distribution is the same under both instances unless $S^{\ell+1}=s$ and $A^{\ell+1}=j$, in which case it is $\mathcal{B}(\frac{1}{2})$ under I_s^0 and $\mathcal{B}(\frac{1}{2}+\Delta)$ under $I \mid \alpha(s)=j$. Therefore, we have

$$\begin{aligned} &D_{\text{TV}}(S^{\ell+1}, A^{\ell+1}, R^{\ell+1} \mid \alpha(s)=0, H^\ell \| S^{\ell+1}, A^{\ell+1}, R^{\ell+1} \mid \alpha(s)=j, H^\ell) \\ &= \frac{1}{2} \sum_{x\in\mathcal{S}, a\in\mathcal{A}, r\in\{0,1\}} \Big| \mathbb{P}(S^{\ell+1}=x, A^{\ell+1}=a, R^{\ell+1}=r \mid \alpha(s)=0, H^\ell) \\ &- \mathbb{P}(S^{\ell+1}=x, A^{\ell+1}=a, R^{\ell+1}=r \mid \alpha(s)=j, H^\ell) \Big| \\ &= \frac{1}{2} \sum_{x\in\mathcal{S}, a\in\mathcal{A}, r\in\{0,1\}} \mathbb{P}(S^{\ell+1}=x) \mathbb{P}(A^{\ell+1}=a \mid S^{\ell+1}=x, H^\ell) \\ &\Big| \mathbb{P}(R^{\ell+1}=r \mid S^{\ell+1}=x, A^{\ell+1}=a, \alpha(s)=0, H^\ell) \\ &- \mathbb{P}(R^{\ell+1}=r \mid S^{\ell+1}=x, A^{\ell+1}=a, \alpha(s)=j, H^\ell) \Big| \\ &= \mathbb{P}(S^{\ell+1}=s) \mathbb{P}(A^{\ell+1}=j \mid S^{\ell+1}=s, H^\ell) D_{\text{TV}}(\mathcal{B}(\frac{1}{2}), \mathcal{B}(\frac{1}{2}+\Delta)) \\ &= \frac{\Delta}{S} \mathbb{P}(A^{\ell+1}=j \mid S^{\ell+1}=s, H^\ell) \end{aligned} \tag{12}$$

Plugin the above upper bound into (11) we get

$$D_{\text{KL}}(M^{\ell+1} \mid \alpha(s)=0, H^\ell \| M^{\ell+1} \mid \alpha(s)=j, H^\ell) \\ \leq \frac{4(\exp(\epsilon)-1)^2}{\exp(\epsilon)} \Delta^2 \frac{1}{S^2} \mathbb{P}(A^{\ell+1}=j \mid S^{\ell+1}=s, H^\ell)^2 \tag{13}$$

Then, plugin the above upper bound into the chain rule decomposition of KL divergence Equation (10) we get

$$D_{\text{KL}}((M^i)_{i\in[K]} \mid \alpha(s)=0 \| (M^i)_{i\in[K]} \mid \alpha(s)=j) \\ \leq 4\frac{(\exp(\epsilon)-1)^2}{\exp(\epsilon)} \Delta^2 \frac{1}{S^2} \sum_{\ell=0}^{K-1} \mathbb{E}_{s,0}\left[\mathbb{P}(A^{\ell+1}=j \mid S^{\ell+1}=s, H^\ell)^2 \right].$$

Lastly, plugin the above upper bound into the KL inequality (9) we get

$$\mathbb{E}_{s,j}\left[N_{s,j}\right] \leq \mathbb{E}_{s,0}\left[N_{s,j}\right] \\ + K\sqrt{2\frac{(\exp(\epsilon)-1)^2}{\exp(\epsilon)} \Delta^2 \frac{1}{S^2} \sum_{\ell=0}^{K-1} \mathbb{E}_{s,0}\left[\mathbb{P}(A^{\ell+1}=j \mid S^{\ell+1}=s, H^\ell)^2 \right]}.$$

Summing over $j \in \mathcal{A}$, using Cauchy-Schwarz inequality and the fact that $\sum_{j\in\mathcal{A}} \mathbb{P}(A^{\ell+1} = j \mid S^{\ell+1} = s, H^\ell)^2 \leq 1$, we get

$$\sum_{j\in\mathcal{A}} \mathbb{E}_{s,j}\left[N_{s,j}\right] \leq \sum_{j\in\mathcal{A}} \mathbb{E}_{s,0}\left[N_{s,j}\right] + K\sqrt{2A\sum_{\ell=0}^{K-1}\frac{(\exp(\epsilon)-1)^2}{\exp(\epsilon)}\Delta^2\frac{1}{S^2}}$$
$$= K + K\sqrt{2AK\frac{(\exp(\epsilon)-1)^2}{\exp(\epsilon)}\Delta^2\frac{1}{S^2}}.$$

Then, we set $\Delta = \min(\frac{1}{2}, \frac{\sqrt{A}}{4\sqrt{2K\frac{(\exp(\epsilon)-1)^2}{\exp(\epsilon)}\frac{1}{S^2}}})$, the above bound becomes

$$\sum_{j\in\mathcal{A}} \mathbb{E}_{s,j}\left[N_{s,j}\right] \leq K + \frac{1}{4}AK$$

Finally, plugin this upper bound into the regret decomposition (8) we get

$$\begin{aligned}
\mathbb{E}[\mathcal{R}(\mathfrak{A}, \mathcal{M}, I)] &\geq \Delta\left(K - \sum_{s\in\mathcal{S}}\sum_{j\in\mathcal{A}} \mathbb{E}_{s,j}\left[N_{s,j}\right]\frac{1}{SA}\right) \\
&\geq \Delta\left(K - S\frac{K + \frac{1}{4}AK}{SA}\right) \\
&\geq \Delta\left(\frac{3}{4}K - K/A\right) \\
&\geq \Delta\left(\frac{3}{4}K - K/2\right) \\
&\geq \Delta\left(\frac{1}{4}K\right) \\
&\geq \min(\frac{K}{8}, \frac{S\sqrt{AK\exp(\epsilon)}}{16\sqrt{2}(\exp(\epsilon)-1)}) \\
&\geq \frac{1}{32}\min(K, \frac{S\sqrt{AK(\exp(\epsilon)+1)}}{(\exp(\epsilon)-1)})
\end{aligned}$$

Then we re-start the proof again and develop the same argument but replacing the bound used in Equation (11) by the KL contraction of [Asoodeh and Zhang, 2024, Theorem 1], applied to the same ϵ-LDP channel $Q^{\ell+1}(\cdot \mid \cdot, H^\ell)$:

$$\begin{aligned}
&D_{\mathrm{KL}}(M^{\ell+1} \mid \alpha(s) = 0, H^\ell \| M^{\ell+1} \mid \alpha(s) = j, H^\ell) \\
&\leq \frac{(\exp(\epsilon)-1)^2}{(\exp(\epsilon)+1)^2} D_{\mathrm{KL}}(S^{\ell+1}, A^{\ell+1}, R^{\ell+1} \mid \alpha(s) = 0, H^\ell \| S^{\ell+1}, A^{\ell+1}, R^{\ell+1} \mid \alpha(s) = j, H^\ell) \\
&= \frac{(\exp(\epsilon)-1)^2}{(\exp(\epsilon)+1)^2} \\
&\mathbb{E}[D_{\mathrm{KL}}(R^{\ell+1} \mid \alpha(s) = 0, H^\ell, S^{\ell+1}, A^{\ell+1} \| R^{\ell+1} \mid \alpha(s) = j, H^\ell, S^{\ell+1}, A^{\ell+1}) \mid \alpha(s) = 0, H^\ell] \\
&= \frac{(\exp(\epsilon)-1)^2}{(\exp(\epsilon)+1)^2} \mathbb{P}(S^{\ell+1} = s, A^{\ell+1} = j \mid H^\ell, \alpha(s) = 0) D_{\mathrm{KL}}(\mathcal{B}(\frac{1}{2}) \| \mathcal{B}(\frac{1}{2} + \Delta)) \\
&\leq 4\frac{(\exp(\epsilon)-1)^2}{(\exp(\epsilon)+1)^2}\Delta^2\frac{1}{S}\mathbb{P}(A^{\ell+1} = j \mid H^\ell, S^{\ell+1} = s)
\end{aligned}$$

where the last inequality is a consequence of the fact that $D_{\mathrm{KL}}(\mathcal{B}(\frac{1}{2}) \| \mathcal{B}(\frac{1}{2} + \Delta)) \leq 4\Delta^2$ for $\Delta \leq \frac{1}{4}$. The bound we obtain is analogous to the one given in Equation (13).

Proceeding as before by plugin the above upper bound in the chain rule decomposition of KL divergence Equation (10), then into the KL inequality Equation (9), summing over j using the identity

$\sum_{j\in\mathcal{A}} \mathbb{P}(A^{\ell+1}=j \mid S^{\ell+1}=s, H^\ell)=1$ and applying Cauchy-Schwarz, we get

$$\sum_{j\in\mathcal{A}} \mathbb{E}_{s,j}\left[N_{s,j}\right] \leq \sum_{j\in\mathcal{A}} \mathbb{E}_{s,0}\left[N_{s,j}\right] + K\sqrt{2A\sum_{\ell=0}^{K-1}\frac{(\exp(\epsilon)-1)^2}{(\exp(\epsilon)+1)^2}\Delta^2\frac{1}{S}}$$
$$= K + K\sqrt{2AK\frac{(\exp(\epsilon)-1)^2}{(\exp(\epsilon)+1)^2}\Delta^2\frac{1}{S}}.$$

Taking $\Delta = \min(\frac{1}{4}, \frac{\sqrt{A}}{4\sqrt{2K\frac{(\exp(\epsilon)-1)^2}{(\exp(\epsilon)+1)^2}\frac{1}{S}}})$, we get

$$\sum_{j\in\mathcal{A}} \mathbb{E}_{s,j}\left[N_{s,j}\right] \leq K + \frac{1}{4}AK$$

Then the proof follows the same steps and yield the bound

$$\mathbb{E}[\mathcal{R}(\mathfrak{A},\mathcal{M},I)] \geq \min(\frac{K}{16}, \frac{\sqrt{SAK}(\exp(\epsilon)+1)}{16\sqrt{2}(\exp(\epsilon)-1)})$$
$$\geq \frac{1}{32}\min(K, \frac{\sqrt{SAK}(\exp(\epsilon)+1)}{(\exp(\epsilon)-1)})$$

Putting the two bounds together we get

$$\mathbb{E}[\mathcal{R}(\mathfrak{A},\mathcal{M},I)] \geq \frac{1}{32}\min(K, \frac{\sqrt{\max(S,\exp(\epsilon)+1)SAK(\exp(\epsilon)+1)}}{\exp(\epsilon)-1})$$

□

B Proof of Theorem 2

Proof. The proof closely follows the standard analysis of Exp3 [Auer et al., 2002, Theorem 3.1].

Form the potential

$$\Phi_k = \sum_{s\in\mathcal{S}} \ln\left(\sum_{a\in\mathcal{A}} w_k(s,a)\right).$$

We have

$$\Phi_{k+1}-\Phi_k = \sum_{s\in\mathcal{S}} \ln\left(\frac{\sum_{a\in\mathcal{A}} w_{k+1}(s,a)}{\sum_{a\in\mathcal{A}} w_k(s,a)}\right)$$
$$= \sum_{s\in\mathcal{S}} \ln\left(\sum_{a\in\mathcal{A}} \frac{w_k(s,a)}{\sum_{a'\in\mathcal{A}} w_k(s,a')}\exp(\eta\hat{V}^k(s,a))\right)$$

Then notice that $\hat{V}^k(s,a) \in \left[0, \frac{A}{\gamma}\right]$. We choose η such that $\gamma = A\eta \leq \frac{1}{2}$ so that $\eta\hat{V}^k(s,a) \in [0,1]$. We can then can apply successively the inequalities $\exp(x) \leq 1+x+(e-2)x^2$ for $x\in[0,1]$ and $\ln(1+x)\leq x$ for $x\in\mathbb{R}$ to get

$$\begin{aligned}
&\Phi_{k+1}-\Phi_k \\
&\leq \sum_{s\in\mathcal{S}} \ln\left(1+\sum_{a\in\mathcal{A}}\frac{w_k(s,a)}{\sum_{a'\in\mathcal{A}} w_k(s,a')}\eta\hat{V}^k(s,a) + \sum_{a\in\mathcal{A}}\frac{w_k(s,a)}{\sum_{a'\in\mathcal{A}} w_k(s,a')}(e-2)\eta^2\hat{V}^k(s,a)^2\right) \\
&\leq \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\frac{w_k(s,a)}{\sum_{a'\in\mathcal{A}} w_k(s,a')}\eta\hat{V}^k(s,a) + \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\frac{w_k(s,a)}{\sum_{a'\in\mathcal{A}} w_k(s,a')}(e-2)\eta^2\hat{V}^k(s,a)^2 \\
&\leq \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\frac{\Pi^k(s,a)}{1-\gamma}\eta\hat{V}^k(s,a) + \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\frac{\Pi^k(s,a)}{1-\gamma}(e-2)\eta^2\hat{V}^k(s,a)^2
\end{aligned}$$

Then, summing over k, we get

$$\begin{aligned}&\Phi_{K+1} - \Phi_1\\ &\leq \sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\frac{\Pi^k(s,a)}{1-\gamma}\eta\hat{V}^k(s,a) + \sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\frac{\Pi^k(s,a)}{1-\gamma}(e-2)\eta^2\hat{V}^k(s,a)^2\end{aligned}$$

We have $\Phi_1 = \sum_{s\in\mathcal{S}} \ln\left(\sum_{a\in\mathcal{A}} w_1(s,a)\right) = S\ln(A)$ and

$$\begin{aligned}\Phi_{K+1} &= \sum_{s\in\mathcal{S}} \ln\left(\sum_{a\in\mathcal{A}} w_{K+1}(s,a)\right)\\ &= \sum_{s\in\mathcal{S}} \ln\left(\sum_{a\in\mathcal{A}} \exp\left(\eta\sum_{k=1}^{K}\hat{V}^k(s,a)\right)\right)\\ &\geq \sum_{s\in\mathcal{S}} \ln\left(\exp\left(\eta\sum_{k=1}^{K}\hat{V}^k(s,a^*(s))\right)\right)\\ &= \sum_{s\in\mathcal{S}} \eta\sum_{k=1}^{K}\hat{V}^k(s,a^*(s))\end{aligned}$$

where $a^*(s) = \mathrm{argmax}_{a\in\mathcal{A}}\, r(s,a)$.

We therefore get the following identity

$$\begin{aligned}&\sum_{s\in\mathcal{S}}\sum_{k=1}^{K}\hat{V}^k(s,a^*(s))\\ &\leq \frac{S\ln(A)}{\eta} + \sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\frac{\Pi^k(s,a)}{1-\gamma}\hat{V}^k(s,a) + \sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\frac{\Pi^k(s,a)}{1-\gamma}(e-2)\eta\hat{V}^k(s,a)^2\\ &= \sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\Pi^k(s,a)\hat{V}^k(s,a)\\ &+ \frac{S\ln(A)}{\eta} + \frac{\gamma}{1-\gamma}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\Pi^k(s,a)\hat{V}^k(s,a) + \sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\frac{\Pi^k(s,a)}{1-\gamma}(e-2)\eta\hat{V}^k(s,a)^2\end{aligned} \tag{14}$$

The term $\sum_{a\in\mathcal{A}}\Pi^k(s,a)\hat{V}^k(s,a)^2$ can be upper bounded as

$$\begin{aligned}&\sum_{a\in\mathcal{A}}\Pi^k(s,a)\hat{V}^k(s,a)^2\\ &= \sum_{a\in\mathcal{A}}\Pi^k(s,a)M^k(s)\frac{\mathbb{1}\{a=\pi^k(s)\}}{\Pi^k(a,s)}\hat{V}^k(s,a)\\ &\leq \sum_{a\in\mathcal{A}}\hat{V}^k(s,a)\end{aligned}$$

Then, we compute $\mathbb{E}[\hat{V}^k(s,a)]$.

$$\begin{aligned}\mathbb{E}[\hat{V}^k(s,a)] &= \mathbb{E}[M^k(s)\frac{\mathbb{1}\{a=\pi^k(s)\}}{\Pi^k(a,s)}]\\ &= \mathbb{E}[\mathbb{E}[M^k(s)\mid V^k(s)]\frac{\mathbb{1}\{a=\pi^k(s)\}}{\Pi^k(a,s)}]\end{aligned}$$

The term $\mathbb{E}[M^k(s) \mid V^k(s)]$ is given by

$$\begin{aligned}\mathbb{E}[M^k(s) \mid V^k(s)] &= V^k(s)\frac{1}{2} + (1 - V^k(s))\frac{1}{\exp(\epsilon)+1} \\ &= V^k(s)(\frac{1}{2} - \frac{1}{\exp(\epsilon)+1}) + \frac{1}{\exp(\epsilon)+1} \\ &= \mathbb{1}\{S^k = s\}\, B^k(\frac{1}{2} - \frac{1}{\exp(\epsilon)+1}) + \frac{1}{\exp(\epsilon)+1}\end{aligned}$$

which yield

$$\begin{aligned}&\mathbb{E}[\hat{V}^k(s,a)] \\ &= \mathbb{E}[\frac{\mathbb{1}\{S^k = s, A^k = a\}}{\Pi^k(a,s)} B^k(\frac{1}{2} - \frac{1}{\exp(\epsilon)+1})] + \frac{1}{\exp(\epsilon)+1} \\ &= \mathbb{E}[\mathbb{1}\{S^k = s\}]r(s,a)(\frac{1}{2} - \frac{1}{\exp(\epsilon)+1}) + \frac{1}{\exp(\epsilon)+1}\end{aligned}$$

Taking expectations in Equation (14), we get

$$\begin{aligned}&(\frac{1}{2} - \frac{1}{\exp(\epsilon)+1})\mathbb{E}[\sum_{k=1}^{K} r(S^k, a^*(S^k)) - \sum_{k=1}^{K} r(S^k, A^k)] \\ &\leq \frac{S\ln(A)}{\eta} + \frac{\gamma}{1-\gamma}\sum_{k=1}^{K}(\frac{1}{2} - \frac{1}{\exp(\epsilon)+1} + \frac{S}{\exp(\epsilon)+1}) \\ &+ \sum_{k=1}^{K}\frac{A}{1-\gamma}(e-2)\eta(\frac{1}{2} - \frac{1}{\exp(\epsilon)+1} + \frac{S}{\exp(\epsilon)+1}) \\ &\leq \frac{S\ln(A)}{\eta} + 2(e-1)KA\eta(\frac{1}{2} - \frac{1}{\exp(\epsilon)+1} + \frac{S}{\exp(\epsilon)+1})\end{aligned}$$

where we used $\gamma = A\eta$ and $\gamma \leq \frac{1}{2}$.

So that

$$\begin{aligned}&\mathbb{E}[\sum_{k=1}^{K} r(S^k, a^*(S^k)) - \sum_{k=1}^{K} r(S^k, A^k)] \\ &\leq \frac{S\ln(A)}{\eta(\frac{1}{2} - \frac{1}{\exp(\epsilon)+1})} + \frac{2(e-1)KA\eta(\frac{1}{2} - \frac{1}{\exp(\epsilon)+1} + \frac{S}{\exp(\epsilon)+1})}{(\frac{1}{2} - \frac{1}{\exp(\epsilon)+1})}\end{aligned}$$

Choosing $\eta = \min(\frac{1}{2A}, \sqrt{\frac{S\ln(A)}{2(e-1)KA\left(\frac{1}{2} - \frac{1}{\exp(\epsilon)+1} + \frac{S}{\exp(\epsilon)+1}\right)}})$ yields in case $\eta < \frac{1}{2A}$,

$$\begin{aligned}&\mathbb{E}[\sum_{k=1}^{K} r(S^k, a^*(S^k)) - \sum_{k=1}^{K} r(S^k, A^k)] \\ &= 2\frac{\sqrt{2(e-1)KAS\ln(A)(\frac{1}{2} - \frac{1}{\exp(\epsilon)+1} + \frac{S}{\exp(\epsilon)+1})}}{(\frac{1}{2} - \frac{1}{\exp(\epsilon)+1})} \\ &= 2\frac{\sqrt{2(e-1)KAS\ln(A)\big(2(\exp(\epsilon)+1)^2 + 4(S-1)(\exp(\epsilon)+1)\big)}}{\exp(\epsilon)-1} \\ &\leq 8\frac{\exp(\epsilon)+1}{\exp(\epsilon)-1}\sqrt{SKA\ln(A)} + 8\frac{\sqrt{\exp(\epsilon)+1}}{\exp(\epsilon)-1}S\sqrt{KA\ln(A)} \qquad (15)\end{aligned}$$

where we used $\frac{1}{2} - \frac{1}{\exp(\epsilon)+1} = \frac{\exp(\epsilon)-1}{2(\exp(\epsilon)+1)}$.

If $\eta = \frac{1}{2A}$, then we have

$$\begin{aligned}
\frac{1}{2A} &\leq \sqrt{\frac{S\ln(A)}{2(e-1)KA\big(\frac{1}{2} - \frac{1}{\exp(\epsilon)+1} + \frac{S}{\exp(\epsilon)+1}\big)}} \\
&\implies K \leq \frac{2AS\ln(A)}{(e-1)\big(\frac{1}{2} - \frac{1}{\exp(\epsilon)+1} + \frac{S}{\exp(\epsilon)+1}\big)} \\
&\implies K \leq \frac{4AS\ln(A)(\exp(\epsilon)+1)}{(e-1)\big(\exp(\epsilon) + 2(S-1)\big)} \\
&\implies K \leq \frac{4AS\ln(A)(\exp(\epsilon)+1)}{\big(\exp(\epsilon) - 1\big)}
\end{aligned}$$

Finally, notice that the regret is always upper bounded by K and we can then write

$$\begin{aligned}
K &= \sqrt{K}\sqrt{K} \\
&\leq \sqrt{K}\sqrt{\frac{4AS\ln(A)(\exp(\epsilon)+1)}{\big(\exp(\epsilon) - 1\big)}} \\
&\leq 8\frac{\exp(\epsilon)+1}{\exp(\epsilon)-1}\sqrt{SKA\ln(A)}
\end{aligned}$$

therefore, the bound Equation (15) holds in all cases.

Finally, using the identity $x + y \leq 2\max(x, y)$, we get

$$\begin{aligned}
&\mathbb{E}[\sum_{k=1}^{K} r(S^k, a^*(S^k)) - \sum_{k=1}^{K} r(S^k, A^k)] \\
&\leq 16\max\left(\frac{\exp(\epsilon)+1}{\exp(\epsilon)-1}\sqrt{SKA\ln(A)}, \frac{\sqrt{\exp(\epsilon)+1}}{\exp(\epsilon)-1}S\sqrt{KA\ln(A)}\right) \\
&\leq 16\min(K, \frac{\sqrt{\max(S, \exp(\epsilon)+1)SAK(\exp(\epsilon)+1)\ln(A)}}{\exp(\epsilon)-1})
\end{aligned}$$

□

C Proof of Corollary 1

Proof. Fix an algorithm $\mathfrak{A}$ and an ϵ-LDP mechanism $\mathcal{M}$ and consider the instance $\tilde{I} = (\tilde{\mathcal{S}}, \tilde{\mathcal{A}}, \tilde{r}, \{\tilde{\rho}^k\}_{k=1}^K)$ of Theorem 1 that achieves high regret in the private contextual bandit setting with $S-2$ states and A arms. In instance $\tilde{I}$, rewards are sampled from a Bernoulli distribution with mean $\tilde{r}(s,a)$ for each $s \in \tilde{\mathcal{S}}$ and $a \in \tilde{\mathcal{A}}$

We construct an instance I' with S states and A actions of finite-horizon MDP $(\mathcal{S}, \mathcal{A}, P, r, \{\rho^k\}_{k=1}^K)$ as follows. The action space is $\mathcal{A} = \tilde{\mathcal{A}}$. The state space is $\mathcal{S} = \tilde{\mathcal{S}} \cup \{\text{good}, \text{bad}\}$ having two additional states. All states have reward zero, except the good state which has reward one for every action. The good state and the bad state are absorbing. The initial distributions are defined as $\rho^k(\text{good}) = \rho^k(\text{bad}) = 0$ and $\rho^k(s) = \tilde{\rho}^k(s)$ for each $s \in \tilde{\mathcal{S}}$. Then, we choose the transition probabilities as

$$P(\text{good} \mid s, a) = \tilde{r}(s,a), \quad P(\text{bad} \mid s, a) = 1 - \tilde{r}(s,a)$$

for each $s \in \tilde{\mathcal{S}}$ and $a \in \mathcal{A}$. Fix $s \in \mathcal{S} \setminus \{\text{good}, \text{bad}\}$, $a \in \mathcal{A}$, $k \in [K]$, fix any policy π and sample a trajectory $(S_h^k, A_h^k)_{h=1}^H$ such that $S_1^k = s$ and $A_1^k = a$. We have

$$\begin{aligned}\sum_{h\in[H]} r'(S_h^k, A_h^k) &= r'(s,a) + \sum_{h=2}^{H} r'(S_h^k, A_h^k) \\ &= r'(s,a) + (H-1)\mathbb{1}\left\{S_2^k = \text{good}\right\} \\ &= (H-1)\mathbb{1}\left\{S_2^k = \text{good}\right\}.\end{aligned}$$

Notice that $\mathbb{1}\left\{S_2^k = \text{good}\right\}$ is a Bernoulli random variable with mean $\tilde{r}(s,a)$. Therefore, instance I is equivalent to that of $\tilde{I}$ but with rewards scaled by $H-1$. Furthermore, the bandit mechanism induced by $\mathcal{M}$ is ϵ-LDP private since each bandit observation is mapped to a trajectory and then privatized. Therefore, the lower bound of Theorem 1 applies with a multiplicative factor of $H-1$ and we have

$$\mathcal{R}(\mathfrak{A}, \mathcal{M}, I) \geq \frac{H-1}{32} \min\left(K, \frac{\sqrt{\max(S-2, \exp(\epsilon)+1)(S-2)AK(\exp(\epsilon)+1)}}{\exp(\epsilon)-1}\right)$$

Then since $H \geq 2$ and $S \geq 4$, we have $H - 1 \geq \frac{H}{2}$ and $S - 2 \geq \frac{S}{2}$ which concludes the proof. □

D Proof of Theorem 3

Proof. Define for each $s, a \in \mathcal{S} \times \mathcal{A}$, $b \in \{0,1\}$ and $s' \in \mathcal{S}$,

$$N^k(s,a,b,s') = \sum_{i\in[k]} \sum_{h\in[H]} \mathbb{1}\left\{S_h^i = s, A_h^i = a, B_h^i = b, S_{h+1}^i = s'\right\}$$

The following Lemma 1 quantifies the deviation of $\hat{N}^k(s,a,b,s')$ from $N^k(s,a,b,s')$.

Lemma 1. *Let $\mathcal{E}_1$ be the event that for each $s \in \mathcal{S}, a \in \mathcal{A}, b \in \{0,1\}, s' \in \mathcal{S}, k \in [K]$ and $\delta \leq \frac{1}{e}$, the following holds*

$$\begin{aligned}&|\hat{N}^k(s,a,b,s') - N^k(s,a,b,s')| \\ &\leq \sqrt{20H\Upsilon_\epsilon^2 \sum_{h\in[H], i\in[k]} \mathbb{1}\left\{a = \pi_h^i(s)\right\} \log(\frac{8KS^2A\log(16K)}{\delta})} + 4H\Upsilon_\epsilon \log(\frac{8KS^2A\log(16K)}{\delta}) \\ &\leq \underbrace{\sqrt{20H\Upsilon_\epsilon^2 \sum_{h\in[H], i\in[k]} \mathbb{1}\left\{a = \pi_h^i(s)\right\} \log(\frac{48K^2S^2A}{\delta})} + 4H\Upsilon_\epsilon \log(\frac{48K^2S^2A}{\delta})}_{:=\beta_k(s,a)}\end{aligned}$$

Then $\mathbb{P}(\mathcal{E}_1) \geq 1 - \delta$.

The proof of Lemma 1 relies on Freedman's inequality for martingales. We use the following version from Hsu et al. [2015][Corollary 1, with $c = 2$, $x = \log(\frac{1}{\delta})$].

Theorem 4 (Freedman's inequality Hsu et al. [2015])**.** *Let $(X_i)_{i\in[K]}$ adapted to a filtration $(\mathcal{F}_i)_{i\in[K]}$ such that $\mathbb{E}[X_i \mid \mathcal{F}_{i-1}] = 0$ and $|X_i| \leq b$ for each $i \in [K]$ and $b \in \mathbb{R}_+$.*

Define $S_k = \sum_{i=1}^k X_i$ and $V_k = \sum_{i=1}^k \mathbb{E}[X_i^2 \mid \mathcal{F}_{i-1}]$ for each $k \in [K]$. Then,

$$\mathbb{P}\left(\exists k \in [K] : S_k \geq \sqrt{4V_k \log(\frac{1}{\delta})} + \frac{4}{3}b\log(\frac{1}{\delta})\right) \leq (1 + \max(1, \lceil \log(2K/\log(\frac{1}{\delta}))\rceil))\delta$$

from which we derive the following corollary.

Corollary 2. *Under the same assumptions as Theorem 4, if $\frac{1}{\delta} \geq e$, we have,*

$$\mathbb{P}\left(\exists k \in [K] : S_k \geq \sqrt{4V_k \log(\frac{\log(16K)}{\delta})} + \frac{4}{3} b \log(\frac{\log(16K)}{\delta})\right) \leq \delta$$

Proof of Lemma 1. Define the martingale difference sequence $(X_k)_{k\in[K]}$ where

$$X_k = H\hat{I}^k(s,b,s')\mathbb{1}\{\pi^k_{J^k}(s) = a\} - \sum_{h\in[H]} \mathbb{1}\{S^k_h = s, A^k_h = a, B^k_h = b, S^k_{h+1} = s'\}.$$

Define a filtration $(\mathcal{F}_k)_{k\in[K]}$ adapted to $(X_k)_{k\in[K]}$ by

$$\mathcal{F}_k = \sigma((J^i, S^i_j, A^i_j, B^i_j, S^i_{j+1}, M^i)_{j\in[H],i\in[k]}).$$

It holds that $\mathbb{E}[X_k \mid \mathcal{F}_{k-1}] = 0$ for each $k \in [K]$ and since $|\hat{I}^k| \leq 2\Upsilon_\epsilon$ for each $k \in [K]$, we have that $|X_k| \leq 3H\Upsilon_\epsilon$ for each $k \in [K]$. Moreover, we have

$$\begin{aligned}
&\mathbb{E}[X_k^2 \mid \mathcal{F}_{k-1}] \\
&= \mathbb{E}\left[\left(H\hat{I}^k(s,b,s')\mathbb{1}\{\pi^k_{J^k}(s) = a\} - \sum_{h\in[H]} \mathbb{1}\{S^k_h = s, A^k_h = a, B^k_h = b, S^k_{h+1} = s'\}\right)^2 \mid \mathcal{F}_{k-1}\right] \\
&\leq 5H\Upsilon_\epsilon^2 \sum_{h\in[H]} \mathbb{1}\{a = \pi^k_h(s)\}
\end{aligned}$$

Then Freedman's inequality applied to $(X_k)_k$ and $(-X_k)_k$ along with a union bound over $s, a, b, s' \in \mathcal{S} \times \mathcal{A} \times \{0,1\} \times \mathcal{S} \times [K]$ and over the two resulting inequalities gives the proof. □

Then define Q^*_h for each $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ by $Q^*_{H+1}(s,a) = 0$ and

$$Q^*_h(s,a) = r(s,a) + \sum_{s'\in\mathcal{S}} P(s' \mid s,a) \max_{a'\in\mathcal{A}} Q^*_{h+1}(s',a')$$

Then, we show in Lemma 2 the following optimism property for $\tilde{Q}^{k,(+)}$.

Lemma 2. *With probability at least $1 - 3\delta$, for each $k \in [K]$, $s \in \mathcal{S}, h \in [H]$ and $a \in \mathcal{A}$, we have*

$$0 \leq \tilde{Q}^{k,(+)}_h(s,a) - \left(r(s,a) + \sum_{s'\in\mathcal{S}} P(s' \mid s,a) \max_{a'\in\mathcal{A}} \tilde{Q}^{k,(+)}_{h+1}(s',a')\right) \leq b_k(s,a,h) \tag{16}$$

$$\tilde{Q}^{k,+}_h(s,a) \leq H - h + 1 \tag{17}$$

where

$$b_k(s,a,h) = \min(H - h + 1, \frac{18S\beta_K(s,a)H}{\max(1, N^k(s,a) - 4S\beta_K(s,a))}).$$

Proof of Lemma 2. Equation (17) follows easily by induction. The case $H + 1$ is satisfied since $\tilde{Q}^{k,(+)}_{H+1}(s,a) = 0$ then assuming $\tilde{Q}^{k,(+)}_{h+1}(s,a) \leq H - h$, we have since $\rho \leq 1$ and $\sum_{s'\in\mathcal{S}} \nu(s') = 1$ that $\tilde{Q}^{k,(+)}_h(s,a) \leq H - h + 1$.

For Equation (16), we first need to show the following Lemma 3 stating that P and r belong to the confidence sets.

Lemma 3. *Call $\mathcal{E}_2$ the event that for each $s \in \mathcal{S}$, $a \in \mathcal{A}, k \in [K]$, the following holds*

$$P(s,a,\cdot), r(s,a) \in \mathcal{C}^k(s,a)$$

Then $\mathbb{P}(\mathcal{E}_2 \cap \mathcal{E}_1) \geq 1 - 3\delta$.

Proof of Lemma 3. First if $\tilde{N}^{k,(-)}(s,a) = 1$ and $\mathcal{E}_1$ holds, we have for each $s, a, k \in \mathcal{S} \times \mathcal{A} \times [K]$ that

$$\begin{aligned}
\sum_{s' \in \mathcal{S}, b \in \{0,1\}} N^k(s,a,b,s') &\leq \sum_{s' \in \mathcal{S}, b \in \{0,1\}} (\hat{N}^k(s,a,b,s') + \beta_k(s,a)) && \text{(By } \mathcal{E}_1\text{)} \\
&\leq \sum_{s' \in \mathcal{S}, b \in \{0,1\}} (\hat{N}^k(s,a,b,s') + \beta_k(s,a)) \\
&= \sum_{s' \in \mathcal{S}, b \in \{0,1\}} \hat{N}^k(s,a,b,s') + 2S\beta_k(s,a) \\
&\leq \tilde{N}^{k,(-)}(s,a) + 4S\beta_k(s,a) \\
&= 1 + 4S\beta_k(s,a)
\end{aligned}$$

and thus

$$\begin{aligned}
&|\tilde{N}^{k,(-)}(s,a) r(s,a) - \sum_{s' \in \mathcal{S}} \hat{N}^k(s,a,1,s')| \\
&= |r(s,a) - \sum_{s' \in \mathcal{S}} \hat{N}^k(s,a,1,s')| \\
&\leq 1 + |\sum_{s' \in \mathcal{S}} \hat{N}^k(s,a,1,s')| \\
&\leq 1 + \sum_{s' \in \mathcal{S}} |\hat{N}^k(s,a,1,s')| \\
&\leq 1 + \sum_{s' \in \mathcal{S}} N^k(s,a,1,s') + \beta_k(s,a)) && \text{(By } \mathcal{E}_1 \text{ and } \beta_k(s,a) > 0\text{)} \\
&\leq 2 + 5S\beta_k(s,a) \\
&\leq 7S\beta_k(s,a)
\end{aligned}$$

and

$$\begin{aligned}
&\|\tilde{N}^{k,(-)}(s,a) P(s,a,\cdot) - \sum_{b \in \{0,1\}} \hat{N}^k(s,a,b,\cdot)\|_1 \\
&= \|P(s,a,\cdot) - \sum_{b \in \{0,1\}} \hat{N}^k(s,a,b,\cdot)\|_1 \\
&\leq 1 + \sum_{b \in \{0,1\}, s' \in \mathcal{S}} |\hat{N}^k(s,a,b,s')| \\
&\leq 1 + \sum_{b \in \{0,1\}, s' \in \mathcal{S}} (N^k(s,a,b,s') + \beta_k(s,a)) && \text{(By } \mathcal{E}_1\text{)} \\
&\leq 2 + 6S\beta_k(s,a) \\
&\leq 8S\beta_k(s,a)
\end{aligned}$$

which implies $\mathcal{E}_2$.

We now study the case $\tilde{N}^{k,(-)}(s,a) > 1$. Define the martingale difference sequence $(X_k)_{k \in [K]}$ where

$$X_k = H\hat{I}^k(s,b,s') \mathbb{1}\{\pi^k_{J^k}(s) = a\} - \sum_{h \in [H]} \mathbb{1}\{S^k_h = s, A^k_h = a, B^k_h = b\} P(s' \mid s,a).$$

Then, denoting

$$N^k(s,a,b) = \sum_{i \in [k]} \sum_{h \in [H]} \mathbb{1}\{S^i_h = s, A^i_h = a, B^i_h = b\}$$

and applying the same steps as in the proof of Lemma 1, we show that with probability at least $1 - \delta$, for each $s \in \mathcal{S}, a \in \mathcal{A}, b \in \{0,1\}, s' \in \mathcal{S}, k \in [K]$, it holds that

$$|N^k(s,a,b)P(s,a,s') - \hat{N}^k(s,a,b,s')| \leq \beta_k(s,a).$$

Equivalently, if we denote $\mathcal{E}_3$ the event that for each $s, a, b, s', k \in \mathcal{S} \times \mathcal{A} \times \{0,1\} \times \mathcal{S} \times [K]$,

$$|N^k(s,a,b)P(s,a,s') - \hat{N}^k(s,a,b,s')| \leq \beta_k(s,a),$$

then $\mathbb{P}(\mathcal{E}_3) \geq 1 - \delta$.

Consider $\mathcal{E}_1$ and $\mathcal{E}_3$ holds and $\tilde{N}^{k,(-)}(s,a) > 1$ meaning

$$\tilde{N}^{k,(-)}(s,a) = \sum_{b \in \{0,1\}, s' \in \mathcal{S}} (\hat{N}^k(s,a,b,s') - \beta_k(s,a)).$$

Then, we have that for each $s, a, s', k \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [K]$,

$$\begin{aligned}
&\sum_{b \in \{0,1\}} \hat{N}^k(s,a,b,s') \\
&\geq \sum_{b \in \{0,1\}} \big(N^k(s,a,b)P(s,a,s') - \beta_k(s,a)\big) && \text{(By } \mathcal{E}_3\text{)} \\
&\geq \sum_{b \in \{0,1\}} \left(\sum_{y \in \mathcal{S}} \big(\hat{N}^k(s,a,b,y) - \beta_k(s,a)\big)P(s,a,s') - \beta_k(s,a) \right) && \text{(By } \mathcal{E}_1\text{)} \\
&= \tilde{N}^{k,(-)}(s,a)P(s,a,s') - 2\beta_k(s,a) && \text{(By } \tilde{N}^{k,(-)}(s,a) > 1\text{)}
\end{aligned}$$

Similarly, we can write for each $s, a, s', k \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [K]$,

$$\begin{aligned}
&\sum_{b \in \{0,1\}} \hat{N}^k(s,a,b,s') \\
&\leq \sum_{b \in \{0,1\}} \big(N^k(s,a,b)P(s,a,s') + \beta_k(s,a)\big) && \text{(By } \mathcal{E}_3\text{)} \\
&\leq \sum_{b \in \{0,1\}} \left(\sum_{y \in \mathcal{S}} \big(\hat{N}^k(s,a,b,y) + \beta_k(s,a)\big)P(s,a,s') + \beta_k(s,a) \right) && \text{(By } \mathcal{E}_1\text{)} \\
&= \big(\tilde{N}^{k,(-)}(s,a) + 2 \sum_{b \in \{0,1\}, y \in \mathcal{S}} \beta_k(s,a)\big)P(s,a,s') + 2\beta_k(s,a) && \text{(By } \tilde{N}^{k,(-)}(s,a) > 1\text{)} \\
&\leq \tilde{N}^{k,(-)}(s,a)P(s,a,s') + 4S\beta_k(s,a)P(s,a,s') + 2\beta_k(s,a)
\end{aligned}$$

so we get

$$|\tilde{N}^{k,(-)}(s,a)P(s,a,s') - \sum_{b \in \{0,1\}} \hat{N}^k(s,a,b,s')| \leq 4S\beta_k(s,a)P(s,a,s') + 2\beta_k(s,a)$$

which implies

$$\|\tilde{N}^{k,(-)}(s,a)P(s,a,\cdot) - \sum_{b \in \{0,1\}} \hat{N}^k(s,a,b,\cdot)\|_1 \leq 6S\beta_k(s,a)$$

meaning $P(s,a,\cdot)$ satisfies the corresponding constraint of the confidence set $\mathcal{C}^k(s,a)$.

For the other constraint, we consider the martingale difference sequence $(X_k)_{k \in [K]}$ where

$$X_k = H\hat{I}^k(s,1,s')\mathbb{1}\{\pi^k_{j^k}(s) = a\} - \sum_{h \in [H]} \mathbb{1}\{S^k_h = s, A^k_h = a, S^k_{h+1} = s'\}\, r(s,a).$$

and apply the same steps as in the proof of Lemma 1 to show that with probability at least $1 - \delta$, for each $s \in \mathcal{S}, a \in \mathcal{A}, k \in [K], s' \in \mathcal{S}$, it holds that

$$|N^k(s,a,s')r(s,a) - \hat{N}^k(s,a,1,s')| \leq \beta_k(s,a).$$

Equivalently, if we denote $\mathcal{E}_4$ the event that for each $s, a, s', k \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [K]$,

$$|N^k(s,a,s')r(s,a) - \hat{N}^k(s,a,1,s')| \leq \beta_k(s,a),$$

then $\mathbb{P}(\mathcal{E}_4) \geq 1-\delta$. Consider $\mathcal{E}_1$ and $\mathcal{E}_4$ hold and $\tilde{N}^{k,(-)}(s,a) > 1$. We have that for each $s, a, s', k \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [K]$,

$$
\begin{aligned}
&\sum_{s' \in \mathcal{S}} \hat{N}^k(s,a,1,s') \\
&\geq \sum_{s' \in \mathcal{S}} \big(N^k(s,a,s')r(s,a) - \beta_k(s,a)\big) && \text{(By } \mathcal{E}_4\text{)} \\
&\geq \sum_{s' \in \mathcal{S}} \Bigg(\big(\sum_{b \in \{0,1\}} \hat{N}^k(s,a,b,s') - 2\beta_k(s,a)\big) r(s,a) - \beta_k(s,a) \Bigg) && \text{(By } \mathcal{E}_1\text{)} \\
&\geq \tilde{N}^{k,(-)}(s,a) r(s,a) - S\beta_k(s,a) && \text{(By } \tilde{N}^{k,(-)}(s,a) > 1\text{)}
\end{aligned}
$$

and

$$
\begin{aligned}
&\sum_{s' \in \mathcal{S}} \hat{N}^k(s,a,1,s') \\
&\leq \sum_{s' \in \mathcal{S}} \big(N^k(s,a,s')r(s,a) + \beta_k(s,a)\big) && \text{(By } \mathcal{E}_4\text{)} \\
&\leq \sum_{s' \in \mathcal{S}} \Bigg(\sum_{b \in \{0,1\}} \big(\hat{N}^k(s,a,b,s') + \beta_k(s,a)\big) r(s,a) + \beta_k(s,a) \Bigg) && \text{(By } \mathcal{E}_1\text{)} \\
&\leq (\tilde{N}^{k,(-)}(s,a) + 4S\beta_k(s,a)) r(s,a) + S\beta_k(s,a) && \text{(By } \tilde{N}^{k,(-)}(s,a) > 1\text{)} \\
&\leq \tilde{N}^{k,(-)}(s,a) r(s,a) + 5S\beta_k(s,a)
\end{aligned}
$$

So it holds that $\mathcal{E}_2 \supseteq \mathcal{E}_1 \cap \mathcal{E}_3 \cap \mathcal{E}_4$ and therefore $\mathbb{P}(\mathcal{E}_2 \cap \mathcal{E}_1) \geq 1 - 3\delta$. □

In the rest of the proof, we assume $\mathcal{E}_2 \cap \mathcal{E}_1$ holds. First, we show the left-hand side of Equation (16). We have that for each $s, a \in \mathcal{S} \times \mathcal{A}$ and $h \in [H]$,

$$
\begin{aligned}
&\tilde{Q}_h^{k,(+)}(s,a) - \big(r(s,a) + \sum_{s' \in \mathcal{S}} P(s' \mid s,a) \max_{a' \in \mathcal{A}} \tilde{Q}_{h+1}^{k,(+)}(s',a')\big) \\
&= \max_{\nu, \rho \in \mathcal{C}^k(s,a)} \Bigg(\rho + \sum_{s' \in \mathcal{S}} \nu(s') \max_{a' \in \mathcal{A}} \tilde{Q}_{h+1}^{k,(+)}(s',a') \Bigg) \\
&- \Bigg(r(s,a) + \sum_{s' \in \mathcal{S}} P(s' \mid s,a) \max_{a' \in \mathcal{A}} \tilde{Q}_{h+1}^{k,(+)}(s',a') \Bigg) \\
&\geq 0 && \text{(By } \mathcal{E}_2\text{)}
\end{aligned}
$$

Then, we show the right-hand side of Equation (16). By $\mathcal{E}_2$ and the definition of $\mathcal{C}^k(s,a)$, it holds that for any $\nu, \rho \in \mathcal{C}^k(s,a)$,

$$
\begin{aligned}
&|\tilde{N}^{k,(-)}(s,a)(\rho - r(s,a))| \\
&\leq |\tilde{N}^{k,(-)}(s,a)\rho - \sum_{s' \in \mathcal{S}} \hat{N}^k(s,a,1,s')| + |\sum_{s' \in \mathcal{S}} \hat{N}^k(s,a,1,s') - \tilde{N}^{k,(-)}(s,a) r(s,a)| \\
&\leq 14S\beta_K(s,a)
\end{aligned}
$$

Due to $\mathcal{E}_1$, $\tilde{N}^{k,(-)}(s,a) \geq \max(1, N^k(s,a) - 4S\beta_K(s,a))$ so that we have

$$
|\rho - r(s,a)| \leq \frac{14S\beta_K(s,a)}{\max(1, N^k(s,a) - 4S\beta_K(s,a))}
$$

Similarly for ν, we have

$$
\begin{aligned}
&\|\tilde{N}^{k,(-)}(s,a)(\nu - P(s,a,\cdot))\|_1 \\
&\leq \|\tilde{N}^{k,(-)}(s,a)\nu - \sum_{b \in \{0,1\}} \hat{N}^k(s,a,b,\cdot)\|_1 + \|\sum_{b \in \{0,1\}} \hat{N}^k(s,a,b,\cdot) - \tilde{N}^{k,(-)}(s,a) P(s,a,\cdot)\|_1 \\
&\leq 16S\beta_K(s,a)
\end{aligned}
$$

and due to $\mathcal{E}_1$, $\tilde{N}^{k,(-)}(s,a) \geq \max(1, N^k(s,a) - 4S\beta_K(s,a))$ so that we have

$$\|\nu - P(s,a,\cdot)\|_1 \leq \frac{16S\beta_K(s,a)}{\max(1, N^k(s,a) - 4S\beta_K(s,a))}$$

Then, we conclude the proof by writing for each $s, a \in \mathcal{S} \times \mathcal{A}$ and $h \in [H]$,

$$\begin{aligned}
&\tilde{Q}_h^{k,(+)}(s,a) - \big(r(s,a) + \sum_{s' \in \mathcal{S}} P(s' \mid s,a) \max_{a' \in \mathcal{A}} \tilde{Q}_{h+1}^{k,(+)}(s',a')\big) \\
&= \max_{\nu, \rho \in \mathcal{C}^k(s,a)} \left(\rho + \sum_{s' \in \mathcal{S}} \nu(s') \max_{a' \in \mathcal{A}} \tilde{Q}_{h+1}^{k,(+)}(s',a')\right) \\
&- \left(r(s,a) + \sum_{s' \in \mathcal{S}} P(s' \mid s,a) \max_{a' \in \mathcal{A}} \tilde{Q}_{h+1}^{k,(+)}(s',a')\right) \\
&\leq \max_{\nu, \rho \in \mathcal{C}^k(s,a)} \left(r(s,a) + \frac{14S\beta_K(s,a)}{\max(1, N^k(s,a) - 4S\beta_K(s,a))}\right. \\
&\left. + \sum_{s' \in \mathcal{S}} \nu(s') \max_{a' \in \mathcal{A}} \tilde{Q}_{h+1}^{k,(+)}(s',a')\right) \\
&- \left(r(s,a) + \sum_{s' \in \mathcal{S}} P(s' \mid s,a) \max_{a' \in \mathcal{A}} \tilde{Q}_{h+1}^{k,(+)}(s',a')\right) \\
&= \frac{14S\beta_K(s,a)}{\max(1, N^k(s,a) - 4S\beta_K(s,a))} + \max_{\nu \in \mathcal{C}^k(s,a)} \sum_{s' \in \mathcal{S}} (\nu(s') - P(s' \mid s,a)) \max_{a' \in \mathcal{A}} \tilde{Q}_{h+1}^{k,(+)}(s',a') \\
&\leq \frac{14S\beta_K(s,a)}{\max(1, N^k(s,a) - 4S\beta_K(s,a))} + \max_{\nu \in \mathcal{C}^k(s,a)} \|\nu(\cdot) - P(s,a,\cdot)\|_1 (H-h) \\
&\leq \frac{14S\beta_K(s,a)}{\max(1, N^k(s,a) - 4S\beta_K(s,a))} + \frac{16S\beta_K(s,a)(H-h)}{\max(1, N^k(s,a) - 4S\beta_K(s,a))} \\
&\leq \frac{18S\beta_K(s,a)H}{\max(1, N^k(s,a) - 4S\beta_K(s,a))}
\end{aligned}$$

Furthermore, since $\tilde{Q}_h^{k,(+)}(s,a) \leq H - h + 1$ and

$$\tilde{Q}_h^{k,(+)}(s,a) - \big(r(s,a) + \sum_{s' \in \mathcal{S}} P(s' \mid s,a) \max_{a' \in \mathcal{A}} \tilde{Q}_{h+1}^{k,(+)}(s',a')\big) \geq 0,$$

we have

$$\tilde{Q}_h^{k,(+)}(s,a) - \big(r(s,a) + \sum_{s' \in \mathcal{S}} P(s' \mid s,a) \max_{a' \in \mathcal{A}} \tilde{Q}_{h+1}^{k,(+)}(s',a')\big) \leq H - h + 1$$

thus concluding the proof. □

Then, we show the optimism of $\tilde{V}^{k,(+)}(s) = \max_{a \in \mathcal{A}} \tilde{Q}_h^{k,(+)}(s,a)$ for each $s, h, k \in \mathcal{S} \times [H] \times [K] \cup \{0\}$. Define $V_h^*(s) = \max_{a \in \mathcal{A}} Q_h^*(s,a)$ for each $s \in \mathcal{S}$ and $h \in [H]$ where $Q_h^*(s,a) = r(s,a) + \sum_{s' \in \mathcal{S}} P(s' \mid s,a) \max_{a' \in \mathcal{A}} Q_{h+1}^*(s',a')$ for each $s \in \mathcal{S}, a \in \mathcal{A}$ and $h \in [H]$ with the convention that $Q_{H+1}^*(s,a) = 0$ for each $s \in \mathcal{S}, a \in \mathcal{A}$. We show the following result.

Lemma 4. *Under $\mathcal{E}_2 \cap \mathcal{E}_1$, for each $s \in \mathcal{S}, h \in [H], k \in [K] \cup \{0\}$, we have*

$$\tilde{V}_h^{k,(+)}(s) \geq V_h^*(s).$$

Proof of Lemma 4. The result for $k = 0$ comes from $V_h^* \leq H$. For $k > 0$, assume $\mathcal{E}_2 \cap \mathcal{E}_1$ holds. We show the result by induction. The case $H + 1$ is satisfied since $\tilde{V}_{H+1}^{k,(+)}(s) = 0$ and $V_{H+1}^*(s) = 0$ for

each $s \in \mathcal{S}$. Then, assuming $\tilde{V}_{h+1}^{k,(+)}(s) \geq V_{h+1}^*(s)$ for each $s \in \mathcal{S}$, we have that for each $s \in \mathcal{S}$,

$$
\begin{aligned}
&\tilde{V}_h^{k,(+)}(s) - V_h^*(s) \\
&= \max_{a \in \mathcal{A}} \tilde{Q}_h^{k,(+)}(s,a) - \max_{a \in \mathcal{A}} Q_h^*(s,a) \\
&\geq \max_{a \in \mathcal{A}} r(s,a) + \sum_{s' \in \mathcal{S}} P(s' \mid s,a) \tilde{V}_{h+1}^{k,(+)}(s') \\
&- \max_{a \in \mathcal{A}} \big(r(s,a) + \sum_{s' \in \mathcal{S}} P(s' \mid s,a) V_{h+1}^*(s') \big) \qquad \text{(By } \mathcal{E}_2 \text{ and } \mathcal{E}_1 \text{ and Lemma 2)} \\
&\geq \max_{a \in \mathcal{A}} r(s,a) + \sum_{s' \in \mathcal{S}} P(s' \mid s,a) V_{h+1}^*(s') \\
&- \max_{a \in \mathcal{A}} \big(r(s,a) + \sum_{s' \in \mathcal{S}} P(s' \mid s,a) V_{h+1}^*(s') \big) \qquad \text{(By } \mathcal{E}_2 \text{ and } \mathcal{E}_1 \text{ and induction hypothesis)} \\
&\geq 0
\end{aligned}
$$

□

Finally, we show the regret bound. Assume $\mathcal{E}_2 \cap \mathcal{E}_1$ holds. We will first show by induction that for $k \in [K]$ and $h \in [H]$,

$$
\begin{aligned}
&\mathbb{E}[\tilde{V}_h^{k-1,(+)}(S_h^k) - V_h^{\pi^k}(S_h^k) \mid \mathcal{F}_{k-1}, S_h^k] \\
&\leq \mathbb{E}[\sum_{i=h}^{H} b_{k-1}(S_i^k, \pi_i^k(S_i^k)) \mid \mathcal{F}_{k-1}, S_h^k]
\end{aligned} \tag{18}
$$

The above result is true for $h = H+1$ since $\tilde{V}_{H+1}^{k-1,(+)}(s) = 0$ and $V_{H+1}^{\pi^k}(s) = 0$ for each $s \in \mathcal{S}$. Then, assuming the result holds for $h+1$, we have that for each $k \in [K]$ that

$$
\begin{aligned}
&\mathbb{E}[\tilde{V}_h^{k-1,(+)}(S_h^k) - V_h^{\pi^k}(S_h^k) \mid \mathcal{F}_{k-1}, S_h^k] \\
&= \mathbb{E}[\max_{a \in \mathcal{A}} \tilde{Q}_h^{k-1,(+)}(S_h^k, a) - Q_h^{\pi^k}(S_h^k, \pi_h^k(S_h^k)) \mid \mathcal{F}_{k-1}, S_h^k] \\
&= \mathbb{E}[\tilde{Q}_h^{k-1,(+)}(S_h^k, \pi_h^k(S_h^k)) - Q_h^{\pi^k}(S_h^k, \pi_h^k(S_h^k)) \mid \mathcal{F}_{k-1}, S_h^k] && \text{(By the definition of } \pi^k\text{)} \\
&\leq \mathbb{E}[\tilde{V}_{h+1}^{k-1,(+)}(S_{h+1}^k) + b_{k-1}(S_h^k, \pi_h^k(S_h^k)) - V_{h+1}^{\pi^k}(S_{h+1}^k) \mid \mathcal{F}_{k-1}, S_h^k] && \text{(By } \mathcal{E}_2 \text{ and } \mathcal{E}_1 \text{ and Lemma 2)} \\
&\leq \mathbb{E}[b_{k-1}(S_h^k, \pi_h^k(S_h^k)) + \mathbb{E}[\sum_{i=h+1}^{H} b_{k-1}(S_i^k, \pi_i^k(S_i^k)) \mid \mathcal{F}_{k-1}, S_{h+1}^k] \mid \mathcal{F}_{k-1}, S_h^k] \\
&= \mathbb{E}[\sum_{i=h}^{H} b_{k-1}(S_i^k, \pi_i^k(S_i^k)) \mid \mathcal{F}_{k-1}, S_h^k]
\end{aligned}
$$

Then, the regret bound can be derived as follows. Given $s \in \mathcal{S}$, we have

$$
\begin{aligned}
&\sum_{k=1}^{K} \big(V_1^*(s) - V_1^{\pi^k}(s) \big) \\
&\leq \sum_{k=1}^{K} \big(\tilde{V}_1^{k-1,(+)}(s) - V_1^{\pi^k}(s) \big) && \text{(By Lemma 4)} \\
&\leq \sum_{k=1}^{K} \mathbb{E}[\sum_{h=1}^{H} b_{k-1}(S_h^k, A_h^k) \mathbb{1}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \mid S_1^k = s] + 3H^2 K\delta && \text{(By Equation (18))}
\end{aligned}
$$

We set $\delta = \frac{1}{3K}$ and then focus on $\sum_{k=1}^{K} \sum_{h=1}^{H} b_{k-1}(S_h^k, A_h^k)$.

$$\sum_{k=1}^{K}\sum_{h=1}^{H} b_{k-1}(S_h^k, A_h^k)$$
$$= \sum_{k=1}^{K}\sum_{h=1}^{H} \min\big(H-h+1, \frac{18S\beta_K(S_h^k, \pi_h^k(S_h^k))H}{\max(1, N^{k-1}(S_h^k, \pi_h^k(S_h^k)) - 4S\beta_K(S_h^k, \pi_h^k(S_h^k)))}\big)$$

Define

$$N_h^k(s,a) = N^{k-1}(s,a) + \sum_{i=1}^{h-1} \mathbb{1}\big\{S_i^k = s, A_i^k = a\big\}$$

for each $s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]$ and $k \in [K]$. It holds that $N^{k-1}(s,a) \geq N_h^k(s,a) - H$ so we can write

$$\sum_{k=1}^{K}\sum_{h=1}^{H} b_{k-1}(S_h^k, A_h^k)$$
$$\leq \sum_{k=1}^{K}\sum_{h=1}^{H} \min\big(H-h+1, \frac{18S\beta_K(S_h^k, \pi_h^k(S_h^k))H}{\max(1, N_h^k(S_h^k, \pi_h^k(S_h^k)) - 4S\beta_K(S_h^k, \pi_h^k(S_h^k)) - H)}\big)$$
$$= \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{n\in[KH]} \mathbb{1}\big\{S_h^k = s, A_h^k = a, N_h^k(s,a) = n\big\}$$
$$\min\big(H, \frac{18S\beta_K(s,a)H}{\max(1, n - 4S\beta_K(s,a) - H)}\big) \tag{19}$$

Then, notice that given $s, a, n \in \mathcal{S}\times\mathcal{A}\times[KH]$, $\mathbb{1}\big\{S_h^k = s, A_h^k = a, N_h^k(s,a) = n\big\}$ is zero except for maybe a single couple k, h we call $k(s,a,n), h(s,a,n)$, so we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H} b_{k-1}(S_h^k, A_h^k)$$
$$\leq \sum_{s,a,n\in\mathcal{S}\times\mathcal{A}\times[KH]} \mathbb{1}\Big\{S_{h(s,a,n)}^{k(s,a,n)} = s, A_{h(s,a,n)}^{k(s,a,n)} = a, N_{h(s,a,n)}^{k(s,a,n)}(s,a) = n\Big\}$$
$$\min(H, \frac{18S\beta_K(s,a)H}{\max(1, n - 4S\beta_K(s,a) - H)})$$
$$\leq \sum_{s,a,n\in\mathcal{S}\times\mathcal{A}\times[KH]} \min(H, \frac{18S\beta_K(s,a)H}{\max(1, n - 4S\beta_K(s,a) - H)})$$
$$\leq 4HS\sum_{s\in\mathcal{S},a\in\mathcal{A}}\beta_K(s,a) + H^2SA + 18S\sum_{s,a\in\mathcal{S}\times\mathcal{A}}\beta_K(s,a)H\log(KH)$$
$$\leq 22HS\log(3KH)\sum_{s,a\in\mathcal{S}\times\mathcal{A}}\beta_K(s,a) + H^2SA$$

and then we can bound $\sum_{s,a\in\mathcal{S}\times\mathcal{A}}\beta_K(s,a)$ as follows.

$$\sum_{s\in\mathcal{S},a\in\mathcal{A}}\beta_K(s,a)$$
$$\leq \sum_{s\in\mathcal{S},a\in\mathcal{A}}\sqrt{20H\Upsilon_\epsilon^2\sum_{h\in[H],i\in[K]}\mathbb{1}\big\{a=\pi_h^i(s)\big\}\log(144K^3S^2A)}$$
$$+ 4HSA\Upsilon_\epsilon\log(148K^3S^2A)$$
$$\leq S\sqrt{20\Upsilon_\epsilon^2AH^2K\log(148K^3S^2A)} + 4HSA\Upsilon_\epsilon\log(148K^3S^2A)$$

The final regret bound is therefore given by

$$\sum_{k\in[K]} \left(V_1^*(s) - V_1^{\pi^k}(s)\right)$$
$$\leq 22H^2S^2\sqrt{20\Upsilon_\epsilon^2 AK\log(148K^3S^2A)}\log(3KH) + 88H^2S^2A\Upsilon_\epsilon\log(148K^3S^2A)\log(3KH)$$
$$+ H^2SA + H^2$$
$$\leq 22H^2S^2\sqrt{20\Upsilon_\epsilon^2 AK\log(148K^3S^2A)}\log(3KH) + 92H^2S^2A\Upsilon_\epsilon\log(148K^3S^2A)\log(3KH)$$

□

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the minimax regret bounds claimed in the paper. All claims are supported by formal theorems (Theorem 1, Theorem 2, Corollary 1, Theorem 3) with complete proofs, and the scope (time-homogeneous episodic tabular RL under non-interactive ϵ-LDP) is precisely defined.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The conclusion section discusses the remaining gap of HS between lower and upper bounds for the high-privacy regime ($\epsilon \leq 1$), the sub-optimality of the upper bound in the low-privacy regime, and the conjecture that the S^2 dependency is unimprovable for model-based algorithms.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theorems and corollaries are numbered, their assumptions are explicitly stated, and complete proofs are provided in the appendix (Appendices A, B and D). Proof sketches in the main body provide intuition for the full appendix proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: The paper is purely theoretical and contains no experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: The paper is purely theoretical and contains no experiments requiring code or data.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper is purely theoretical and contains no experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper is purely theoretical and contains no experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper is purely theoretical and contains no experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is foundational theoretical research on privacy-preserving reinforcement learning. The work contributes to understanding fundamental limits of private learning, with no direct path to negative applications.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper releases no models, datasets, or other assets that pose a risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use any external code, datasets, or models. All prior works are properly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper introduces no new datasets, models, or code assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: ChatGPT 5.5 was used to brainstorm proof strategies, verifying proposed algorithms and proofs, proofreading and polishing the manuscript. The authors verified the correctness and originality of all content.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.