Private Reinforcement Learning with PAC and Regret Guarantees

Giuseppe Vietri* Borja Balle† Akshay Krishnamurthy‡ Zhiwei Steven Wu§

Abstract

Motivated by high-stakes decision-making domains like personalized medicine where user information is inherently sensitive, we design privacy preserving exploration policies for episodic reinforcement learning (RL). We first provide a meaningful privacy formulation using the notion of joint differential privacy (JDP)–a strong variant of differential privacy for settings where each user receives their own sets of output (e.g., policy recommendations). We then develop a private optimism-based learning algorithm that simultaneously achieves strong PAC and regret bounds, and enjoys a JDP guarantee. Our algorithm only pays for a moderate privacy cost on exploration: in comparison to the non-private bounds, the privacy parameter only appears in lower-order terms. Finally, we present lower bounds on sample complexity and regret for reinforcement learning subject to JDP.

1 Introduction

Privacy-preserving machine learning is critical to the deployment of data-driven solutions in applications involving sensitive data. Differential privacy (DP) [DMNS06] is a de-facto standard for designing algorithms with strong privacy guarantees for individual data. Large-scale industrial deployments – e.g. by Apple [Tea17], Google [EPK14] and the US Census Bureau [Abo18] – and general purpose DP tools for machine learning [ACP19] and data analysis [HBAL19, WZL+19] exemplify that existing methods are well-suited for simple data analysis tasks (e.g. averages, histograms, frequent items) and batch learning problems where the training data is available beforehand. While these techniques cover a large number of applications in the central and (non-interactive) local models, they are often insufficient to tackle machine learning applications involving other threat models. This includes federated learning problems [KMA+19, LSTS19] where devices cooperate to learn a joint model while preserving their individual privacy, and, more generally, interactive learning in the spirit of the reinforcement learning (RL) framework [SB18].

In this paper we contribute to the study of reinforcement learning from the lens of differential privacy. We consider sequential decision-making tasks where users interact with an agent for the duration of a fixed-length episode. At each time-step the current user reveals a state to the agent, which responds with an appropriate action and receives a reward generated by the user. Like in standard RL, the goal of the agent is to learn a policy that maximizes the rewards provided by the users. However, our focus is on situations where the states and rewards that users provide to the agent might contain sensitive information. While users might be ready to reveal such information to an agent in order to receive a service, we assume they want to prevent third parties from making unintended inferences about their personal data. This includes external parties who might have access to the policy learned by the agent, as well as malicious users who can probe the agent's behavior to trigger actions informed by its interactions with previous users. For example,

*Department of Computer Science and Engineering, University of Minnesota. Supported by the GAANN fellowship from the U.S. Department of Education. Email: vietr002@umn.edu

†Now at DeepMind. Email: borja.balle@gmail.com

‡Microsoft Research, New York, NY. Email: akshaykr@microsoft.com

§Department of Computer Science and Engineering, University of Minnesota. Email: zstevenwu@cmu.edu

[PWZ+19] recently showed how RL policies can be probed to reveal information about the environment where the agent was trained.

The question we ask in this paper is: how should the learnings an agent can extract from an episode be balanced against the potential information leakages arising from the behaviors of the agent that are informed by such learnings? We answer the question by making two contributions to the analysis of the privacy-utility trade-off in reinforcement learning: (1) we provide the first privacy-preserving RL algorithm with formal accuracy guarantees, and (2) we provide lower bounds on the regret and number of sub-optimal episodes for any differentially private RL algorithm. To measure the privacy provided by episodic RL algorithms we introduce a notion of episodic joint differential privacy (JDP) under continuous observation, a variant of joint differential privacy [KPRU14] that captures the potential information leakages discussed above.

Overview of our results. We study reinforcement learning in a fixed-horizon episodic Markov decision process with S states, A actions, and episodes of length H. We first provide a meaningful privacy formulation for this general learning problem with a strong relaxation of differential privacy: joint differential privacy (JDP) under continual observation, controlled by a privacy parameter $\varepsilon \geq 0$ (larger ε means less privacy). Under this formulation, we give the first known RL sample complexity and regret upper and lower bounds with formal privacy guarantees. First, we present a new algorithm, PUCB, which satisfies ε-JDP in addition to two utility guarantees: it finds an α-optimal policy with a sample complexity of

$$\tilde{O}\left(\frac{SAH^4}{\alpha^2} + \frac{S^2AH^4}{\varepsilon\alpha}\right) ,$$

and achieves a regret rate of

$$\tilde{O}\left(H^2\sqrt{SAT} + \frac{SAH^3 + S^2AH^3}{\varepsilon}\right)$$

over T episodes. In both of these bounds, the first terms $\frac{SAH^4}{\alpha^2}$ and $H^2\sqrt{SAT}$ are the non-private sample complexity and regret rates, respectively. The privacy parameter ε only affects the lower order terms – for sufficiently small approximation α and sufficiently large T, the "cost" of privacy becomes negligible.

We also provide new lower bounds for ε-JDP reinforcement learning. Specifically, by incorporating ideas from existing lower bounds for private learning into constructions of hard MDPs, we prove a sample complexity bound of

$$\tilde{\Omega}\left(\frac{SAH^2}{\alpha^2} + \frac{SAH}{\varepsilon\alpha}\right)$$

and a regret bound of

$$\tilde{\Omega}\left(\sqrt{HSAT} + \frac{SAH}{\varepsilon}\right) .$$

As expected, these lower bounds match our upper bounds in the dominant term (ignoring H and polylogarithmic factors). We also see that necessarily the utility cost for privacy grows linearly with the state space size, although this does not match our upper bounds. Closing this gap is an important direction for future work.

1.1 Related Work

Most previous works on differentially private interactive learning with partial feedback concentrate on bandit-type problems, including on-line learning with bandit feedback [TS13, AS17], multi-armed bandits [MT15, TD16, TD17, TD18], and linear contextual bandits [NR18, SS18]. These works generally differ on the assumed reward models under which utility is measured (e.g. stochastic, oblivious adversarial, adaptive adversarial) and the concrete privacy definition being used (e.g. privacy when observing individual actions or sequences of actions, and privacy of reward or reward and observation in the contextual setting). [BDT19] provides a comprehensive account of different privacy definitions used in the bandit literature.

Much less work has addressed DP for general RL. For policy evaluation in the batch case, [BGP16] propose regularized least-squares algorithms with output perturbation and bound the excess risk due to the privacy constraints. For the control problem with private rewards and public states, [WH19] give a differentially private Q-learning algorithm with function approximation.

On the RL side, as we are initiating the study of RL with differential privacy, we focus on the well-studied tabular setting. While a number of algorithms with utility guarantees and lower bound constructions are known for this setting [Kak03, AOM17, DLB17], we are not aware of any work addressing the privacy issues that are fundamental in high-stakes applications.

2 Preliminaries

2.1 Markov Decision Processes

A fixed-horizon *Markov decision process* (MDP) with time-dependent dynamics can be formalized as a tuple $M = (\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, p_0, H)$. $\mathcal{S}$ is the state space with cardinality S. $\mathcal{A}$ is the action space with cardinality A. $\mathcal{R}(s_h, a_h, h)$ is the reward distribution on the interval $[0,1]$ with mean $r(s_h, a_h, h)$. $\mathcal{P}$ is the transition kernel, given time step h, action a_h and, state s_h the next state is sampled from $s_{t+1} \sim \mathcal{P}(.|s_h, a_h, h)$. Let p_0 be the initial state distribution at the start of each episode, and H be the number of time steps in an episode.

In our setting, an agent interacts with an MDP by following a (deterministic) policy $\pi \in \Pi$, which maps states s and timestamps h to actions, i.e., $\pi(s,h) \in \mathcal{A}$. The *value function* in time step $h \in [H]$ for a policy π is defined as:

$$\begin{aligned} V_h^\pi(s) &= \mathbb{E}\left[\sum_{i=h}^{H} r(s_i, a_i, i) \middle| s_h = s, \pi\right] \\ &= r(s, \pi(s,h), h) + \sum_{s' \in \mathcal{S}} V_{h+1}^\pi(s') \mathcal{P}(s'|s, \pi(s,h), h) \ . \end{aligned}$$

The *expected total reward* for policy π during an entire episode is:

$$\rho^\pi = \mathbb{E}\left[\sum_{i=1}^{H} r(s_i, a_i, i) \middle| \pi\right] = p_0^\top V_1^\pi \ .$$

The *optimal value function* is given by $V_h^*(s) = \max_{\pi \in \Pi} V_h^\pi(s)$. Any policy π such that $V_h^\pi(s) = V_h^*(s)$ for all $s \in \mathcal{S}$ and $h \in [H]$ is called optimal. It achieves the optimal expected total reward $\rho^* = \max_{\pi \in \Pi} \rho^\pi$.

The goal of an RL agent is to learn a near-optimal policy after interacting with an MDP for a finite number of episodes T. During each episode $t \in [T]$ the agent follows a policy π_t informed by previous interactions, and after the last episode it outputs a final policy π.

Definition 1. An agent is (α, β)-*probably approximately correct* (PAC) with sample complexity $f(S, A, H, \frac{1}{\alpha}, \log(\frac{1}{\beta}))$, if with probability at least $1 - \beta$ it follows an α-optimal policy π such that $\rho^* - \rho^\pi \leq \alpha$ except for at most $f(S, A, H, \frac{1}{\alpha}, \log(\frac{1}{\beta}))$ episodes.

Definition 2. The (expected cumulative) *regret* of an agent after T episodes is given by

$$\mathrm{Regret}(T) = \sum_{t=1}^{T} (\rho^* - \rho^{\pi_t}) \ ,$$

where $\pi_1, \ldots \pi_T$ are the policies followed by the agent on each episode.

2.2 Privacy in RL

In some RL application domains such as personalized medical treatments, the sequence of states and rewards received by a reinforcement learning agent may contain sensitive information. For example, individual users may interact with an RL agent for the duration of an episode and reveal sensitive information in order to obtain a service from the agent. This information affects the final policy produced by the agent, as well as the actions taken by the agent in any subsequent interaction. Our goal is to prevent damaging inferences about a user's sensitive information in the context of the interactive protocol in algorithm 1 summarizing the interactions between an RL agent $\mathcal{M}$ and T distinct users.

Algorithm 1: Episodic RL Protocol

```
Input: Agent M and users u_1,...,u_T
for t ∈ [T] do
    for h ∈ [H] do
        u_t sends state s_h^(t) to M
        M sends action a_h^(t) to u_t
        u_t sends reward r_h^(t) to M
    end
end
M releases policy π
```

Throughout the execution of this protocol the agent observes a collection of T state-reward trajectories of length H. Each user u_t gets to observe the actions chosen by the agent during the t-th episode, as well as the final policy π. To preserve the privacy of individual users we enforce a (joint) differential privacy criterion: upon changing one of the users in the protocol, the information observed by the other $T-1$ participants will not change substantially. This criterion must hold even if the $T-1$ participants collude adversarially, by e.g., crafting their states and rewards to induce the agent to reveal information about the remaining user.

Formally, we write $U = (u_1, \ldots, u_T)$ to denote a sequence of T users participating in the RL protocol. Technically speaking a user can be identified with a tree of depth H encoding the state and reward responses they would give to all the A^H possible sequences of actions the agent can choose. During the protocol the agent only gets to observe the information along a single root-to-leaf path in each user's tree. For any $t \in [T]$, we write $\mathcal{M}_{-t}(U)$ to denote all the outputs excluding the output for episode t during the interaction between $\mathcal{M}$ and U. This captures all the outputs which might leak information about the t-th user in interactions after the t-th episode, as well as all the outputs from earlier episodes where other users could be submitting information to the agent adversarially to condition its interaction with the t-th users.

We also say that two user sequences U and U' are t-neighbors if they only differ in their t-th user.

Definition 3. A randomized RL agent $\mathcal{M}$ is *ε-jointly differentially private under continual observation* (JDP) if for all $t \in [T]$, all t-neighboring user sequences U, U', and all events $E \subseteq \mathcal{A}^{H \times [T-1]} \times \Pi$ we have

$$\Pr[\mathcal{M}_{-t}(U) \in E] \leq e^{\varepsilon} \Pr[\mathcal{M}_{-t}(U') \in E] \quad .$$

This definition extends to the RL setting the one used in [SS18] for designing privacy-preserving algorithms for linear contextual bandits. The key distinctions is that in our definition each user interacts with the agent for H time-steps (in bandit problems one usually has $H = 1$), and we also allow the agent to release the learned policy at the end of the learning process.

Another distinction is that our definition holds for all past and future outputs. In contrast, the definition of JDP in [SS18] only captures future episodes; hence, it only protects against collusion from future users.

To demonstrate that our definition gives a stronger privacy protection, we use a simple example. Consider an online process that takes as input a stream of binary bits $u = (u_1, \ldots, u_T)$, where $u_t \in \{0,1\}$ is the data of user t, and on each round t the mechanism outputs the partial sum $m_t(u) = \sum_{i=1}^{t} u_i$. Then the following trivial mechanism satisfies JDP for m's future outputs (as in the JDP definition of [SS18]): First, sample once from the Laplace mechanism $\xi \sim \mathrm{Lap}(\varepsilon)$ before the rounds begin, and on each round output $\tilde{m}_t(u) = m_t(u)+\xi$. Note that the view of any future user $t' > t$ is $\tilde{m}_{t'}(u)$. Now let u be a binary stream with user t bit on and let w be identical to u but with user t bit off. Then, by the differential-privacy guarantee of the Laplace mechanism, a user $t' > t$ cannot distinguish between $\tilde{m}_{t'}(u)$ and $\tilde{m}_{t'}(w)$. Furthermore, any coalition of future users cannot provide more information about user t. Therefore this simple mechanism satisfies the JDP definition from [SS18].

However the simple counting mechanism with one round of Laplace noise does not satisfy JDP for past and future outputs as in our JDP definition 3. To see why, suppose that user $t-1$ and user $t+1$ collude in the following way: For input u, the view of user $t-1$ is $\tilde{m}_{t-1}(u)$ and the view of user $t+1$ is $\tilde{m}_{t+1}(u)$. They also know their own data u_{t-1}, u_{t+1}. Then they can recover the data of the t-th user as follows

$$\tilde{m}_{t+1}(u) - u_{t+1} - \tilde{m}_{t-1}(u) = m_{t+1}(u) + \xi - u_{t+1} - m_{t-1}(u) - \xi = \sum_{i=1}^{t+1} u_i - u_{t+1} - \sum_{i=1}^{t-1} u_i = u_t$$

Remark. *1. would the algorithm leak more info for the returning user? yes, but we could bound using group privacy. 2. would other users be affected? no, because JDP prevents arbitrary collusion*

2.3 Counting Mechanism

The algorithm we describe in the next section maintains a set of counters to keep track of events that occur when interacting with the MDP. We denote by $\widehat{n}_t(s,a,h)$ the count of visits to state tuple (s,a,h) right before episode t, where $a \in \mathcal{A}$ is the action taken on state $s \in \mathcal{S}$ and time-step $h \in [H]$. Likewise $\widehat{m}_t(s,a,s',h)$ is the count of going from state s to s' after taking actions a before episode t. Finally, we have the counter $\widehat{r}_t(s,a,h)$ for the total reward received by taking action a on state s and time h before episode t. Then, on episode t, the counters are sufficient to create an estimate of the MDP dynamics to construct a policy for episode t. The challenge is that the counters depend on the sequence of states and actions, which is considered sensitive data in this work. Therefore the algorithm must release the counts in a privacy-preserving way, and we do this the private counters proposed by [CSS11] and [DNPR10].

A private counter mechanism takes as input a stream $\sigma = (\sigma_1 \ldots, \sigma_T) \in [0,1]^T$ and on any round t releases and approximation of the prefix count $c(\sigma)(t) = \sum_{i=1}^{t} \sigma_i$. In this work we will denote PC as the binary mechanism of [CSS11] and [DNPR10] with parameters ε and T. This mechanism produces a monotonically increasing count and satisfies the following accuracy guarantee: Let $\mathcal{M} := \mathrm{PC}(T, \varepsilon)$ be a private counter and $c(\sigma)(t)$ be the true count on episode t, then given a stream σ, with probability at least $1-\beta$, simultaneously for all $1 \le t \le T$, we have

$$|\mathcal{M}(\sigma)(t) - c(\sigma)(t)| \le \frac{4}{\varepsilon} \ln(1/\beta) \log(T)^{5/2} \ .$$

While the stated bound above holds for a single ε-DP counter, our algorithm needs to maintain more than S^2AH many counters. A naive allocation of the privacy budget across all these counters will require noise with scale polynomially with S, A, and H. However, we will leverage the fact that the total change across all counters a user can have scales with the length of the episode H, which allows us to add a much smaller amount of noise that scales linearly in H.

3 The PUCB Algorithm

In this section, we introduce the Private Upper Confidence Bound algorithm (PUCB), a JDP algorithm with both PAC and regret guarantees. The pseudo-code for PUCB is in algorithm 2. At a high level, the algorithm is a private version of the UBEV algorithm [DLB17]. UBEV keeps track of three types of statistics about the history, including (a) the average empirical reward for taking action a in state s at time h, denoted $\widehat{r}_t(s,a,h)$, (b) the number of times the agent has taken action a in state s at time h, denoted $\widehat{n}_t(s,a,h)$, and (c) the number

Algorithm 2: Private Upper Confidence Bound (PUCB)

Parameters: Privacy parameter ε, target failure probability β
Input: Maximum number of episodes T, horizon H, state space $\mathcal{S}$, action space $\mathcal{A}$
$\varepsilon' := \varepsilon/(3H)$
for $s, a, s', h \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$ **do**
  Initialize private counters: $\widetilde{r}(s,a,h), \widetilde{n}(s,a,h), \widetilde{m}(s,a,s',h) := \mathrm{PC}(T, \varepsilon', \beta)$
end
for $t \leftarrow 1$ ***to*** T **do**
  Private planning: $\widetilde{Q}_t^+ := \mathtt{PrivQ}(\widetilde{r}, \widetilde{n}, \widetilde{m}, \varepsilon)$
  for $h \leftarrow 1$ ***to*** H **do**
    Let s denote the state during step h and episode t
    Execute $a := \arg\max_{a'} \widetilde{Q}_t^+(s, a', h)$
    Observe $r \sim \mathcal{R}(s,a,h)$ and $s' \sim \mathcal{P}(.|s,a,h)$
    Feed r to $\widetilde{r}(s,a,h)$
    Feed 1 to $\widetilde{n}(s,a,h)$ and $\widetilde{m}(s,a,s',h)$ and 0 to all other counters $\widetilde{n}(\cdot,\cdot,h)$ and $\widetilde{m}(\cdot,\cdot,\cdot,h)$
  end
end

of times the agent has taken action a in state s at time h and transitioned to s', denoted $\widehat{m}_t(s,a,s',h)$. In each episode t, UBEV uses these statistics to compute a policy via dynamic programming, executes the policy, and updates the statistics with the observed trajectory. [DLB17] compute the policy using an optimistic strategy and establish both PAC and regret guarantees for this algorithm.

Of course, as the policy depends on the statistics from the previous episodes, UBEV as is does not satisfy JDP. On the other hand, the policy executed only depends on the previous episodes *only* through the statistics $\widehat{r}_t, \widehat{n}_t, \widehat{m}_t$. If we maintain and use private versions of these statistics, and we set the privacy level appropriately, we can ensure JDP.

To do so PUCB initializes one private counter mechanism for each $\widehat{r}_t, \widehat{n}_t, \widehat{m}_t$ ($2SAH + S^2AH$ counters in total). At episode t, we compute the policy using optimism as in UBEV, but we use only the private counts $\widetilde{r}_t, \widetilde{n}_t, \widetilde{m}_t$ released from the counter mechanisms. We require that each set of counters is $(\varepsilon/3)$ JDP, and so with

$$E_\varepsilon = \frac{3}{\varepsilon} H \log\left(\frac{2SAH + S^2AH}{\beta'}\right) \log(T)^{5/2},$$

we can ensure that with probability at least $1-\beta$:

$$\forall t \in [T] : |\widetilde{n}_t(s,a,h) - \widehat{n}_t(s,a,h)| < E_\varepsilon \quad ,$$

where $\widehat{n}_t, \widetilde{n}_t$ are the count and release at the beginning of the t^{th} episode. The guarantee is uniform in (s,a,h) and also holds simultaneously for $\widetilde{r}$ and $\widetilde{m}$.

To compute the policy, we define a bonus function $\widetilde{\mathrm{conf}}(s,a,h)$ for each (s,a,h) tuple, which can be decomposed into two parts $\widetilde{\phi}_t(s,a,h)$ and $\widetilde{\psi}_t(s,a,h)$, where

$$\widetilde{\phi}_t(s,a,h) = \sqrt{\frac{2\ln(T/\beta')}{\max(\widetilde{n}_t(s,a,h) - E_\varepsilon, 1)}} \quad ,$$

$$\widetilde{\psi}_t(s,a,h) = (1 + SH)\left(\frac{3E_\varepsilon}{\widetilde{n}_t(s,a,h)} + \frac{2E_\varepsilon^2}{\widetilde{n}_t(s,a,h)^2}\right) \quad .$$

The term $\widetilde{\phi}_t(\cdot)$ roughly corresponds to the sampling error, while $\widetilde{\psi}_t(\cdot)$ corresponds to errors introduced by the private counters. Using this bonus function, we use dynamic programming to compute an optimistic private Q-function in Algorithm 3. The algorithm here is a standard batch Q-learning update, with $\widetilde{\mathrm{conf}}(\cdot)$

Algorithm 3: $\texttt{PrivQ}(\widetilde{r}, \widetilde{n}, \widetilde{m}, \varepsilon, \beta)$

Input: Private counters $\widetilde{r}, \widetilde{n}, \widetilde{m}$, privacy parameter ε, target failure probability β

$E_\varepsilon := \frac{3}{\varepsilon} H \log\left(\frac{2SAH+S^2AH}{\beta'}\right) \log(T)^{5/2}$

$\widetilde{V}_{H+1}(s) := 0 \quad \forall s \in \mathcal{S}$

for $h \leftarrow H$ ***to*** 1 **do**

  for $s, a \in \mathcal{S} \times \mathcal{A}$ **do**

    if $\widetilde{n}_t(s,a,h) \geq 2E_\varepsilon$ **then**

      $\widetilde{\mathrm{conf}}_t(s,a,h) := (H+1)\widetilde{\phi}_t(s,a,h) + \widetilde{\psi}_t(s,a,h)$

    else

      $\widetilde{\mathrm{conf}}_t(s,a,h) := H$

    end

    $\widetilde{Q}_t(s,a,h) := \frac{1}{\widetilde{n}_t(s,a,h)} \left(\widetilde{r}_t(s,a,h) + \sum_{s' \in \mathcal{S}} \widetilde{V}_{h+1}(s') \widetilde{m}_t(s,a,s',h)\right)$

    $\widetilde{Q}^+_t(s,a,h) := \min\left\{H, \widetilde{Q}_t(s,a,h) + \widetilde{\mathrm{conf}}_t(s,a,h)\right\}$

  end

  $\widetilde{V}_h(s) := \max_a \widetilde{Q}^+_t(s,a,h) \quad \forall s \in \mathcal{S}$

end

Output: $\widetilde{Q}^+_t$

serving as an optimism bonus. The resulting Q-function, called $\widetilde{Q}^+$, encodes a greedy policy, which we use for the t^{th} episode.

4 Privacy Analysis of PUCB

We show that releasing the sequence of actions by algorithm PUCB satisfies JDP with respect to any user on an episode changing his data. Formally,

Theorem 1. *Algorithm (2) PUCB is ε-JDP.*

To prove theorem 1, we use the *billboard lemma* due to [HHR+16] which says that an algorithm is JDP if the output sent to each user is a function of the user's private data and a common signal computed with standard differential privacy. We state the formal lemma:

Lemma 2 (Billboard lemma [HHR+16])**.** *Suppose $\mathcal{M} : U \to \mathcal{R}$ is ε-differentially private. Consider any set of functions $f_i : U_i \times \mathcal{R} \to \mathcal{R}'$ where U_i is the portion of the database containing the i's user data. The composition $\{f_i(\Pi_i U, \mathcal{M}(U))\}$ is ε-joint differentially private, where $\Pi_i : U \to U_i$ is the projection to i's data.*

Let $U_{<t}$ denote the data of all users before episode t and u_t denote the data of the user during episode t. Algorithm PUCBkeeps track of all events on users $U_{<t}$ in a differentially-private way with private counters $\widetilde{r}_t, \widetilde{n}_t, \widetilde{m}_t$. These counters are given to the procedure $\texttt{PrivQ}$ which computes a Q-function $\widetilde{Q}^+_t$, and induces the policy $\pi_t(s,h) := \max_a \widetilde{Q}^+_t(s,a,h)$ to be use by the agent during episode t. Then the output during episode t is generated the policy π_t and the private data of the user u_t according to the protocol 1, the output on a single episode is: $\left(\pi_t\left(s_1^{(t)}, 1\right), \ldots, \pi_t\left(s_H^{(t)}, H\right)\right)$. By the billboard lemma 2, the composition of the output of all T episodes, and the final policy $\left(\left\{\left(\pi_t(s_1^{(t)}, 1), \ldots, \pi_t(s_H^{(t)}, H)\right)\right\}_{t \in [T]}, \pi_T\right)$ satisfies ε-JDP if the policies $\{\pi_t\}_{t \in [T]}$ are computed with a ε-DP mechanism.

Then it only remains to show that the noisy counts satisfy ε-DP. First, consider the counters for the number of visited states. The algorithm PUCB runs SAH parallel private counters, one for each state tuple (s,a,h). Each counter is instantiated with a $\varepsilon/(3H)$-differentially private mechanism which takes an input an event stream $\widehat{n}(s,a,h) = \{0,1\}^T$ where the ith bit is set to 1 if a user visited the state tuple (s,a,h) during episode i and 0 otherwise. Hence each stream $\widehat{n}(s,a,h)$ is the data for a private counter. The next claim says that the total ℓ_1 sensitivity over all streams is bounded by H:

Claim 1. *Let U, U' be two t-neighboring user sequences, in the sense that they are only different in the data for episode t. For each $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, let $\widehat{n}(s,a,h)$ be the event stream corresponding to user sequence U and $\widehat{n}'(s,a,h)$ be the event stream corresponding to U'. Then the total ℓ_1 distance of all stream is given by the following claim:*

$$\sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \|\widehat{n}(s,a,h) - \widehat{n}'(s,a,h)\|_1 \leq H$$

Proof. The proof follows from the fact that on any episode t a user visits at most H states. □

Finally we use a result from [HHR+16, Lemma 34] which states that the composition of the SAH $(\varepsilon/3H)$-DP counters for $\widehat{n}(\cdot)$ satisfy $(\varepsilon/3)$-DP as long as the ℓ_1 sensitivity of the counters is H as shown in claim 1. We can apply the same analysis to show that the counters corresponding to the empirical reward $\widehat{r}(\cdot)$ and the transitions $\widehat{m}(\cdot)$ are both also $\varepsilon/3$-differentially private. Putting it all together releasing the noisy counters is ε-differentially private.

5 PAC and Regret Analysis of PUCB

Now that we have established PUCB is JDP, we turn to utility guarantees. We establish two forms of utility guarantee namely a PAC sample complexity bound, and a regret bound. In both cases, comparing to UBEV, we show that the price for JDP is quite mild. In both bounds the privacy parameter interacts quite favorably with the "error parameter."

We first state the PAC guarantee.

Theorem 3 (PAC guarantee for PUCB). *Let T be the maximum number of episodes and ε the JDP parameter. Then for any $\alpha \in (0,H]$ and $\beta \in (0,1)$, algorithm PUCB with parameters (ε, β) follows a policy that with probability at least $1-\beta$ is α-optimal on all but*

$$O\left(\left(\frac{SAH^4}{\alpha^2} + \frac{S^2AH^4}{\varepsilon\alpha}\right)\mathrm{polylog}\left(T,S,A,H,\tfrac{1}{\alpha},\tfrac{1}{\beta},\tfrac{1}{\varepsilon}\right)\right)$$

episodes.

The theorem states that if we run PUCB for many episodes, it will act near-optimally in a large fraction of them. The number of episodes where the algorithm acts suboptimally scales polynomially with all the relevant parameters. In particular, notice that in terms of the utility parameter α, the bound scales as $1/\alpha^2$. In fact the first term here matches the guarantee for the non-private algorithm UBEV up to polylogarithmic factors. On the other hand, the privacy parameter ε appears only in the term scaling as $1/\alpha$. In the common case where α is relatively small, this term is typically of a lower order, and so the price for privacy here is relatively low.

Analogous to the PAC bound, we also have a regret guarantee.

Theorem 4 (Regret bound for PUCB). *With probability at least $1-\beta$, the regret of PUCB up to episode T is at most*

$$O\left(\left(H^2\sqrt{SAT} + \frac{SAH^3 + S^2AH^3}{\varepsilon}\right)\mathrm{polylog}\left(T,S,A,H,\tfrac{1}{\beta},\tfrac{1}{\varepsilon}\right)\right) .$$

A similar remark to the PAC bound applies here: the privacy parameter only appears in the polylog(T) terms, while the leading order term scales as $\sqrt{T}$. In this guarantee it is clear that as T gets large, the utility price for privacy is essentially negligible.

We also remark that both bounds have "lower order" terms that scale with S^2. This is quite common for tabular reinforcement algorithms [DLB17, AOM17]. We find it quite interesting to observe that the privacy parameter ε interacts with this term, but not with the so-called "leading" term in these guarantees.

Proof Sketch. The proofs for both results parallel the arguments in [DLB17] for the analysis of UBEV. The main differences arises from the fact that we have adjusted the confidence interval $\widetilde{\text{conf}}$ to account for the noise in the releases of $\widetilde{r}, \widetilde{n}, \widetilde{m}$. In [DLB17] the bonus is crucially used to establish optimism, and the final guarantees are related to the over-estimation incurred by these bonuses. We focus on these two steps in this sketch, with a full proof deferred to the appendix.

First we verify optimism. Fix episode t and state tuple (s, a, h), and let us abbreviate the latter simply by x. Assume that $\widetilde{V}_{h+1}$ is private and optimistic in the sense that $\widetilde{V}_{h+1}(s) \geq V^*_{h+1}(s)$, for all $s \in \mathcal{S}$. First define the empirical Q-value

$$\widehat{Q}_t(\mathrm{x}) = \frac{\widehat{r}_t(\mathrm{x}) + \sum_{s' \in \mathcal{S}} \widetilde{V}_{h+1}(s') \widehat{m}_t(\mathrm{x}, s')}{\widehat{n}_t(\mathrm{x})} \ .$$

The optimistic Q-function, which is similar to the one used by [DLB17], is given by

$$\widehat{Q}^+_t(\mathrm{x}) = \widehat{Q}_t(\mathrm{x}) + (H+1)\widehat{\phi}_t(\mathrm{x}) \ ,$$

where $\widehat{\phi}_t(\mathrm{x}) = \sqrt{\frac{T/\beta'}{\widehat{n}_t(\mathrm{x})}}$. A standard concentration argument shows that $\widehat{Q}^+_t \geq Q^\star$, assuming that $\widetilde{V}_{h+1} \geq V^\star_{h+1}$.

Of course, both $\widehat{Q}_t$ and $\widehat{Q}^+_t$ involve the non-private counters $\widehat{r}, \widehat{n}, \widehat{m}$, so they are *not* available to our algorithm. Instead, we construct a surrogate for the empirical Q-value using the private releases:

$$\widetilde{Q}_t(\mathrm{x}) = \frac{\widetilde{r}_t(\mathrm{x}) + \sum_{s' \in \mathcal{S}} \widetilde{V}_{h+1}(s') \widetilde{m}_t(\mathrm{x}, s')}{\widetilde{n}_t(\mathrm{x})} \ .$$

Our analysis involves relating $\widetilde{Q}_t$ which the algorithm has access to, with $\widehat{Q}_t$ which is non-private. To do this, note that by the guarantee for the counting mechanism, we have

$$\widehat{Q}_t(\mathrm{x}) \leq \frac{\widetilde{r}_t(\mathrm{x}) + E_\varepsilon + \sum_{s' \in \mathcal{S}} \widetilde{V}_{h+1}(s')(\widetilde{m}_t(\mathrm{x}, s') + E_\varepsilon)}{\widetilde{n}_t(\mathrm{x}) - E_\varepsilon} \ . \tag{1}$$

Next, we use the following elementary fact.

Claim 2. *Let $y \in \mathbb{R}$ be any positive real number. Then for all $x \in \mathbb{R}$ with $x \geq 2y$ it holds that $\frac{1}{x-y} \leq \frac{1}{x} + \frac{2y}{x^2}$.*

If $\widetilde{n}_t(\mathrm{x}) \geq 2E_\varepsilon$, then we can apply claim 2 to equation (1), along with the facts that $\widetilde{V}_{h+1}(s') \leq H$ and $\widetilde{r}_t(\mathrm{x}) \leq \widetilde{n}_t(\mathrm{x}) + 2E_\varepsilon \leq 2\widetilde{n}_t(\mathrm{x})$, to upper bound $\widehat{Q}_t$ by $\widetilde{Q}_t$. This gives:

$$\begin{aligned} \widehat{Q}_t(\mathrm{x}) &\leq \widetilde{Q}_t(\mathrm{x}) + \left(\frac{1}{\widetilde{n}_t(\mathrm{x})} + \frac{2E_\varepsilon}{\widetilde{n}_t(\mathrm{x})^2} \right) \cdot (1 + SH)E_\varepsilon \\ &= \widetilde{Q}_t(\mathrm{x}) + \widetilde{\psi}_t(\mathrm{x}) \ . \end{aligned}$$

Therefore, we see that $\widetilde{Q}_t(\mathrm{x}) + \widetilde{\psi}_t(\mathrm{x})$ dominates $\widehat{Q}_t(\mathrm{x})$. Accordingly, if we inflate by $\widetilde{\phi}_t(\mathrm{x})$ – which is clearly an upper bound on $\widehat{\phi}_t(\mathrm{x})$ – we account for the statistical fluctuations and can verify optimism. In the event that $\widetilde{n}_t(\mathrm{x}) \leq 2E_\varepsilon$, we simply upper bound $Q^* \leq H$.

For the over-estimation, the bonus we have added is $\widetilde{\phi}_t(\mathrm{x}) + \widetilde{\psi}_t(\mathrm{x})$, which is closely related to the original bonus $\widehat{\phi}_t(\mathrm{x})$. The essential property for our bonus is that it is not significantly larger than the original one $\widehat{\phi}_t(\mathrm{x})$. Indeed, $\widehat{\phi}_t(\mathrm{x})$ scales as $1/\sqrt{\widetilde{n}_t(\mathrm{x})}$ while $\widetilde{\psi}_t(\mathrm{x})$ scales roughly as $E_\varepsilon/\widetilde{n}_t(\mathrm{x}) + E_\varepsilon^2/\widetilde{n}_t(\mathrm{x})^2$, which is lower order in the dependence on $\widetilde{n}_t(\mathrm{x})$. Similarly, the other sources of error here only have lower order effects on the over-estimation.

In detail, there are three sources of error. First, $\widetilde{\phi}_t(\mathrm{x})$ is within a constant factor of $\widehat{\phi}_t(\mathrm{x})$ since we are focusing on rounds where $\widetilde{n}_t(\mathrm{x}) \geq 2E_\varepsilon$. Second, as the policy suboptimality is related to the bonuses on the states and actions we are likely to visit, we cannot have many rounds where $\widetilde{n}_t(\mathrm{x}) \leq 2E_\varepsilon$, since all of the private counters are increasing. A similar argument applies for $\widetilde{\psi}_t(\mathrm{x})$: we can ignore states that we visit infrequently, and the private counters $\widetilde{n}_t(\mathrm{x})$ for states that we visit frequently increase rapidly enough to introduce minimal additional error. Importantly, in the latter two arguments, we have terms of the form $E_\varepsilon/\widetilde{n}_t(\mathrm{x})$, while $\widehat{\phi}_t(\mathrm{x})$ itself scales as $\sqrt{1/\widetilde{n}_t(\mathrm{x})}$, which dominates in terms of the accuracy parameter α or the number of episodes T. As such we obtain PAC and regret guarantees where the privacy parameter ε does not appear in the dominant terms.

6 Lower Bounds

In this section we prove the following lower bounds on the sample complexity and regret for any PAC RL agent providing joint differential privacy.

Theorem 5 (PAC Lower Bound)**.** *Let $\mathcal{M}$ be an RL agent satisfying ε-JDP. Suppose that $\mathcal{M}$ is (α,β)-PAC for some $\beta \in (0,1/8)$. Then, there exists a fixed-horizon episodic MDP where the number of episodes until the algorithm's policy is α-optimal with probability at least $1-\beta$ satisfies*

$$\mathbb{E}[n_{\mathcal{M}}] \geq \Omega\left(\frac{SAH^2}{\alpha^2} + \frac{SAH}{\alpha\varepsilon}\ln\left(\frac{1}{\beta}\right)\right) .$$

Theorem 6 (Private Regret Lower Bound)**.** *For any ε JDP-algorithm $\mathcal{M}$ there exist an MDP M with S states A actions over H time steps per episode such that for any initial state $s \in \mathcal{S}$ the expected regret of $\mathcal{M}$ after T steps is*

$$\mathbb{E}[\mathrm{Regret}(T)] = \Omega\left(\sqrt{HSAT} + \frac{SAH\log(T)}{\varepsilon}\right)$$

for any $T \geq S^{1.1}$.

Here we present the proof steps for the sample complexity lower bound in Theorem 5. The proof for the regret lower bound in Theorem 6 follows from a similar argument and is deferred to the appendix.

To obtain Theorem 5, we go through two intermediate lower bounds: one for private best-arm identification in multi-armed bandits problems (Lemma 8), and one for private RL in a relaxed scenario where the initial state of each episode is considered public information (Lemma 10). At first glance our arguments look similar to other techniques that provide lower bounds for RL in the non-private setting by leveraging lower bounds for bandits problems, e.g. [SLL09, DB15]. However, getting this strategy to work in the private case is significantly more challenging because one needs to ensure the notions of privacy used in each of the lower bounds are compatible with each other. Since this is the main challenge to prove Theorem 5, we focus our presentation on the aspects that make the private lower bound argument different from the non-private one, and defer the rest of details to the appendix.

6.1 Lower Bound for Best-Arm Identification

The first step is a lower bound for best-arm identification for differentially private multi-armed bandits algorithms. This considers mechanisms $\mathcal{M}$ interacting with users via the MAB protocol described in algorithm 4, where we assume arms $a^{(t)}$ come from some finite space $\mathcal{A}$ and rewards are binary, $r^{(t)} \in \{0,1\}$. Recall that T denotes the total number of users. Our lower bound applies to mechanisms for this protocol that satisfy standard DP in the sense that the adversary has access to all the outputs $\mathcal{M}(U) = (a^{(1)},\ldots,a^{(T)},\hat{a})$ produced by the mechanism.

Definition 4. A MAB mechanism $\mathcal{M}$ is ε-DP if for any neighboring user sequences U and U' differing in a single user, and all events $E \subseteq \mathcal{A}^{T+1}$ we have

$$\Pr[\mathcal{M}(U) \in E] \leq e^{\varepsilon}\Pr[\mathcal{M}(U') \in E] .$$

To measure the utility of a mechanism for performing *best-arm identification* in MABs we consider a stochastic setting with independent arms. In this setting each arm $a \in \mathcal{A}$ produces rewards following a Bernoulli distribution with expectation $\bar{P}_a$ and the goal is to identify high probability an optimal arm a^* with expected reward $\bar{P}_{a^*} = \max_{a \in \mathcal{A}} \bar{P}_a$. A problem instance can be identified with the vector of expected rewards $\bar{P} = (\bar{P}_a)_{a \in \mathcal{A}}$.

Algorithm 4: MAB Protocol for Best-Arm Identification

Input: Agent $\mathcal{M}$ and users $u_1, \ldots, u_T$
for $t \in [T]$ **do**
  $\mathcal{M}$ sends arm $a^{(t)}$ to u_t
  u_t sends reward $r^{(t)}$ to $\mathcal{M}$
end
$\mathcal{M}$ releases arm $\hat{a}$

The lower bound result relies on the following adaptation of the coupling lemma from [KV17, Lemma 6.2].

Lemma 7. *Fix any arm $a \in [k]$. Now consider any pair of MAB instances $\mu, \nu \in [0,1]^k$ both with k arms and time horizon T, such that $\|\mu_a - \nu_a\|_{tv} < \alpha$ and $\|\mu_{a'} - \nu_{a'}\|_{tv} = 0$ for all $a' \neq a$. Let $R \sim B(\mu)^T$ and $Q \sim B(\nu)^T$ be the sequence of T rounds of rewards sampled under μ and ν respectively, and let $\mathcal{M}$ be any ε-DP multi-armed bandit algorithm. Then, for any event E such that under event E arm a is pulled less than t times,*

$$Pr_{\mathcal{M},R}[E] \leq e^{6\varepsilon t \alpha} Pr_{\mathcal{M},Q}[E]$$

Lemma 8 (Private MAB Lower Bound)**.** *Let $\mathcal{M}$ be a MAB best-arm identification algorithm satisfying ε-DP that succeeds with probability at least $1-\beta$, for some $\beta \in (0, 1/4)$. For any MAB instance $\bar{P}$ and any α-suboptimal arm a with $\alpha > 0$ (i.e. $\bar{P}_a = \bar{P}_{a^*} - \alpha$), the number of times that $\mathcal{M}$ pulls arm a during the protocol satisfies*

$$\mathbb{E}[n_a] > \frac{1}{24\varepsilon\alpha} \ln\left(\frac{1}{4\beta}\right) .$$

Proof. Let a^* be the optimal arm under $\bar{P}$ and a an α-suboptimal arm. We construct an alternative MAB instance $\bar{Q}$ by exchanging the rewards of a and a^*: $\bar{Q}_a = \bar{P}_{a^*}$, $\bar{Q}_{a^*} = \bar{P}_a$, and the rest of rewards are identical on both instances. Note that now a^* is α-suboptimal under $\bar{Q}$.

Let $t_a = \frac{1}{24\varepsilon\alpha} \ln\left(\frac{1-2\beta}{2\beta}\right)$ and n_a is the number of times the policy $\mathcal{M}$ pulls arm a. We suppose that $\mathbb{E}_{\bar{P}}[n_a] \leq t_a$ and derive a contradiction.

Define A to be the event that arm n_a is pulled less than $4t_a$ times, that is $A := \{n_a \leq 4t_a\}$. From Markov's inequality we have

$$\begin{aligned} t_a \geq \mathbb{E}_{\bar{P}}[n_a] &\geq 4t_a \Pr_{\bar{P}}[n_a > 4t_a] && (2) \\ &= 4t_a (1 - \Pr_{\bar{P}}[n_a \leq 4t_a]) \ , && (3) \end{aligned}$$

where the first inequality (2) comes from the assumption that $\mathbb{E}[n_a] \leq t_a$. From (3) above it follows that $\Pr_{\bar{P}}[A] \geq 3/4$. We also let B be the event that arm a^* is selected. Since arm a^* is optimal under $\bar{P}$, our assumption on $\mathcal{M}$ implies $\Pr_{\bar{P}}[B] \geq 1 - \beta$.

Now let E be the event that both A and B occur, that is $E = A \cap B$. We combine the lower bound of $\Pr_{\bar{P}}[A]$ and $\Pr_{\bar{P}}[B]$ to get a lower bound for $\Pr_{\bar{P}}[E]$. First we show that $\Pr_{\bar{P}}[B|A] \geq 3/4 - \beta$:

$$\begin{aligned} 1 - \beta &\leq \Pr_{\bar{P}}[B|A] \Pr_{\bar{P}}[A] + \Pr_{\bar{P}}[B|A^c] \Pr_{\bar{P}}[A^c] \\ &\leq \Pr_{\bar{P}}[B|A] + \Pr_{\bar{P}}[A^c] \leq \Pr_{\bar{P}}[B|A] + 1/4 \ . \end{aligned}$$

By replacing in the lower bounds for $\Pr_{\bar{P}}[A]$ and $\Pr_{\bar{P}}[B|A]$ we obtain:

$$\Pr_{\bar{P}}[E] = \Pr_{\bar{P}}[A]\Pr_{\bar{P}}[B|A] \geq \frac{3}{4}\left(\frac{3}{4} - \beta\right) \ .$$

On instance $\bar{Q}$ arm a^* is suboptimal, hence we have that $\Pr_{\bar{Q}}[E] \leq \beta$. Now we apply the group privacy property (Lemma 7) where the number of observations is $4t_a$ and $t_a = \frac{1}{24\varepsilon\alpha}\ln\left(\frac{1/2-\beta}{\beta}\right)$ to obtain

$$\begin{aligned}\frac{3}{4}\left(\frac{3}{4} - \beta\right) &\leq \Pr_{\bar{P}}[E] \leq e^{6\varepsilon\alpha 4t_a}\Pr_{\bar{Q}}[E] \\ &\leq e^{6\varepsilon\alpha 4t_a}\beta = \frac{1}{2} - \beta \ . \end{aligned} \tag{4}$$

But $\frac{3}{4}\left(\frac{3}{4} - \beta\right) > \frac{1}{2} - \beta$ for $\beta \in (0, 1/4)$, therefore (4) is a contradiction. □

6.2 Lower Bound for RL with Public Initial State

To leverage the lower bound for private best-arm identification in the RL setting we first consider a simpler setting where the initial state of each episode is public information. This means that we consider agents $\mathcal{M}$ interacting with a variant of the protocol in Algorithm 1 where each user t releases their first state $s_1^{(t)}$ in addition to sending it to the agent. We model this scenario by considering agents whose inputs (U, S_1) include the sequence of initial states $S_1 = (s_1^{(1)}, \ldots, s_1^{(T)})$, and define the privacy requirements in terms of a different notion of neighboring inputs: two sequences of inputs (U, S_1) and (U', S_1') are t-neighboring if $u_{t'} = u'_{t'}$ for all $t \neq t'$ and $S_1 = S_1'$. That is, we do not expect to provide privacy in the case where the user that changes between U and U' also changes their initial state, since in this case making the initial state public already provides evidence that the user changed. Note, however, that u_t and u'_t can provide different rewards for actions taken by the agent on state $s_1^{(t)}$.

Definition 5. A randomized RL agent $\mathcal{M}$ is ε-JDP under continual observation in the *public initial state* setting if for all $t \in [T]$, all t-neighboring user-state sequences (U, S_1), (U', S_1'), and all events $E \subseteq A^{H\times[T-1]} \times \Pi$ we have

$$\Pr[\mathcal{M}_{-t}(U, S_1) \in E] \leq e^{\varepsilon}\Pr[\mathcal{M}_{-t}(U', S_1') \in E] \ .$$

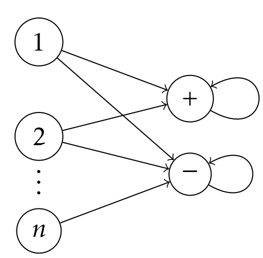

Figure 1: Class of hard MDP instances used in the lower bound.

We obtain a lower bound on the sample complexity of PAC RL agents that satisfy JDP in the public initial state setting by constructing a class of hard MDPs shown in Figure 2. An MDP in this class has state space $\mathcal{S} := [n] \cup \{+, -\}$ and action space $\mathcal{A} := \{0, \ldots, m\}$. On each episode, the agent starts on one of the initial states $\{1, \ldots, n\}$ chosen uniformly at random. On each of the initial states the agent has $m+1$ possible actions and transitions can only take it to one of two possible absorbing states $\{+, -\}$. Lastly, if the current state is either one of $\{+, -\}$ then the only possible transition is a self loop, hence the agent will in that state until the end of the episode. We assume in these absorbing states the agent can only take a fixed action. Every action which transitions to state + provides reward 1 while actions transitioning to state − provide reward 0. In particular, in each episode the agent either receives reward H or 0.

Such an MDP can be seen as consisting of n parallel MAB problems. Each MAB problem determines the transition probabilities between the initial state $s \in \{1,\dots,n\}$ and the absorbing states $\{+,-\}$. We index the possible MAB problems in each initial state by their optimal arm, which is always one of $\{0,\dots,m\}$. We write $I_s \in \{0,\dots,m\}$ to denote the MAB instance in initial state s, and define the transition probabilities such that $\Pr[+|s,0] = 1/2 + \alpha'/2$ and $\Pr[+|s,a'] = 1/2$ for $a' \neq I_s$ for all I_s, and for $I_s \neq 0$ we also have $\Pr[+|s,I_s] = 1/2 + \alpha'$. Here α' is a free parameter to be determined later. We succinctly represent an MDP in the class by identifying the optimal action (i.e. arm) in each initial state: $I := (I_1,\dots,I_n)$.

To show that our MAB lower bounds imply lower bounds for an RL agent interacting with MDPs in this class we prove that collecting the first action taken by the agent in all episodes t with a fixed initial state $s_1^{(t)} = s \in [n]$ simulates the execution of an ε-DP MAB algorithm.

Let $\mathcal{M}$ be an RL agent and (U, S_1) a user-state input sequence with initial states from some set $\mathcal{S}_1$. Let $\mathcal{M}(U,S_1) = (\vec{a}^{(1)},\dots,\vec{a}^{(T)},\pi) \in \mathcal{A}^{H\times T} \times \Pi$ be the collection of all outputs produced by the agent on inputs U and S_1. For every $s \in \mathcal{S}_1$ we write $\mathcal{M}_{1,s}(U,S_1)$ to denote the restriction of the previous trace to contain just the first action from all episodes starting with s together with the action predicted by the policy at states s:

$$\mathcal{M}_{1,s}(U,S_1) := \left(a_1^{(t_{s,1})},\dots,a_1^{(t_{s,T_s})},\pi(s)\right) ,$$

where T_s is the number of occurrences of s in S_1 and $t_{s,1},\dots,t_{s,T_s}$ are the indices of these occurrences. Furthermore, given $s \in \mathcal{S}_1$ we write $U_s = (u_{t_{s,1}},\dots,u_{t_{s,T_s}})$ to denote the set of users whose initial state equals s.

Lemma 9. *Let (U,S_1) be a user-state input sequence with initial states from some set $\mathcal{S}_1$. Suppose $\mathcal{M}$ is an RL agent that satisfies ε-JDP in the public initial state setting. Then, for any $s \in \mathcal{S}_1$ the trace $\mathcal{M}_{1,s}(U,S_1)$ is the output of an ε-DP MAB mechanism on input U_s.*

Using Lemmas 8 and 9 and a reduction from RL lower bounds to bandits lower bounds yields the second term in the following result. The first terms follows directly from the non-private lower bound in [DB15].

Lemma 10. *Let $\mathcal{M}$ be an RL agent satisfying ε-JDP in the public initial state setting. Suppose that $\mathcal{M}$ is (α,β)-PAC for some $\beta \in (0,1/8)$. Then, there exists a fixed-horizon episodic MDP where the number of episodes until the algorithm's policy is α-optimal with probability at least $1-\beta$ satisfies*

$$\mathbb{E}[n_{\mathcal{M}}] \geq \Omega\left(\frac{SAH^2}{\alpha^2} + \frac{SAH}{\alpha\varepsilon}\ln\left(\frac{1}{\beta}\right)\right) .$$

Finally, Theorem 5 follows from Lemma 10 by observing that any RL agent $\mathcal{M}$ satisfying ε-JDP also satisfies ε-JDP in the public state setting (see lemma 11 and see appendix for proof).

Lemma 11. *Any RL agent $\mathcal{M}$ satisfying ε-JDP also satisfies ε-JDP in the public state setting.*

7 Conclusion

In this paper, we initiate the study of differentially private algorithms for reinforcement learning. On the conceptual level, we formalize the privacy desiderata via the notion of joint differential privacy, where the algorithm cannot strongly base future decisions off sensitive information from previous interactions. Under this formalism, we provide a JDP algorithm and establish both PAC and regret utility guarantees for episodic tabular MDPs. Our results show that the utility cost for privacy is asymptotically negligible in the large accuracy regime. We also establish the first lower bounds for reinforcement learning with JDP.

A natural direction for future work is to close the gap between our upper and lower bounds. A similar gap remains open for tabular RL *without* privacy considerations, but the setting is more difficult with privacy, so it may be easier to establish a lower bound here. We look forward to pursuing this direction, and hope that progress will yield new insights into the non-private setting.

Beyond the tabular setup considered in this paper, we believe that designing RL algorithms providing state and reward privacy in non-tabular settings is a promising direction for future work with considerable potential for real-world applications.

8 Acknowledgements

Giuseppe Vietri has been supported by the GAANN fellowship from the U.S. Department of Education. We want to thank Matthew Joseph, whose comments improved our definition of joint-differential-private.

References

[Abo18] John M Abowd. The us census bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2867–2867, 2018.

[ACP19] Galen Andrew, Steve Chien, and Nicolas Papernot. Tensorflow privacy. `https://github.com/tensorflow/privacy`, 2019.

[AOM17] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org, 2017.

[AS17] Naman Agarwal and Karan Singh. The price of differential privacy for online learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 32–40. JMLR. org, 2017.

[BDT19] Debabrota Basu, Christos Dimitrakakis, and Aristide Tossou. Differential privacy for multi-armed bandits: What is it and what is its cost? *arXiv preprint arXiv:1905.12298*, 2019.

[BGP16] Borja Balle, Maziar Gomrokchi, and Doina Precup. Differentially private policy evaluation. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 2130–2138, 2016.

[CSS11] T-H Hubert Chan, Elaine Shi, and Dawn Song. Private and continual release of statistics. *ACM Transactions on Information and System Security (TISSEC)*, 14(3):26, 2011.

[DB15] Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2818–2826, 2015.

[DLB17] Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5713–5723, 2017.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[DNPR10] Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N Rothblum. Differential privacy under continual observation. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 715–724. ACM, 2010.

[EPK14] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067, 2014.

[HBAL19] Naoise Holohan, Stefano Braghin, Pól Mac Aonghusa, and Killian Levacher. Diffprivlib: The IBM differential privacy library. *CoRR*, abs/1907.02444, 2019.

[HHR+16] Justin Hsu, Zhiyi Huang, Aaron Roth, Tim Roughgarden, and Zhiwei Steven Wu. Private matchings and allocations. *SIAM Journal on Computing*, 45(6):1953–1984, 2016.

[JOA10] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

[Kak03] Kakade. On the sample complexity of reinforcement learning. *Diss. University of London*, 2003.

[KMA+19] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning, 2019.

[KPRU14] Michael J. Kearns, Mallesh M. Pai, Aaron Roth, and Jonathan Ullman. Mechanism design in large games: incentives and privacy. In *Innovations in Theoretical Computer Science, ITCS'14, Princeton, NJ, USA, January 12-14, 2014*, pages 403–410, 2014.

[KV17] Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals. *arXiv preprint arXiv:1711.03908*, 2017.

[LSTS19] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions, 2019.

[MT15] Nikita Mishra and Abhradeep Thakurta. (nearly) optimal differentially private stochastic multi-arm bandits. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 592–601, 2015.

[NR18] Seth Neel and Aaron Roth. Mitigating bias in adaptive data gathering via differential privacy. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 3717–3726, 2018.

[PWZ+19] Xinlei Pan, Weiyao Wang, Xiaoshuai Zhang, Bo Li, Jinfeng Yi, and Dawn Song. How you act tells a lot: Privacy-leaking attack on deep reinforcement learning. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019*, pages 368–376, 2019.

[SB18] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[SLL09] Alexander L Strehl, Lihong Li, and Michael L Littman. Reinforcement learning in finite mdps: Pac analysis. *Journal of Machine Learning Research*, 10(Nov):2413–2444, 2009.

[SS18] Roshan Shariff and Or Sheffet. Differentially private contextual linear bandits. In *Advances in Neural Information Processing Systems*, pages 4296–4306, 2018.

[TD16] Aristide CY Tossou and Christos Dimitrakakis. Algorithms for differentially private multi-armed bandits. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[TD17] Aristide Charles Yedia Tossou and Christos Dimitrakakis. Achieving privacy in the adversarial multi-armed bandit. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[TD18] Aristide CY Tossou and Christos Dimitrakakis. On the differential privacy of thompson sampling with gaussian prior. *arXiv preprint arXiv:1806.09192*, 2018.

[Tea17] Apple Differential Privacy Team. Learning with privacy at scale. `https://machinelearning.apple.com/2017/12/06/learning-with-privacy-at-scale.html`, 2017.

[TS13] Abhradeep Guha Thakurta and Adam Smith. (nearly) optimal algorithms for private online learning in full-information and bandit settings. In *Advances in Neural Information Processing Systems*, pages 2733–2741, 2013.

[WH19] Baoxiang Wang and Nidhi Hegde. Privacy-preserving q-learning with functional noise in continuous spaces. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 11323–11333. 2019.

[WZL+19] Royce J Wilson, Celia Yuxin Zhang, William Lam, Damien Desfontaines, Daniel Simmons-Marengo, and Bryant Gipson. Differentially private sql with bounded user contribution, 2019.

A Private Counters

We use the binary mechanism of [CSS11] and [DNPR10] to keep track of important events in a differentially private way. The error of the counter is given by the following theorem:

Algorithm 5: Binary Meachanism $\mathcal{B}$

Input: Time upper bound T, privacy parameter ε, stream $\sigma \in \{0,1\}^T$
$\varepsilon' \leftarrow \varepsilon / \log T$
for $t \leftarrow 1$ ***to*** T **do**
  Express t in binary form: $t = \sum_j \mathrm{Bin}_j(j) \cdot 2^j$
  Let $i := \min\{j : \mathrm{Bin}_j(j) \neq 0\}$
  $\alpha_i \leftarrow \sum_{j<i} \alpha_j + \sigma(t)$
  for $j \longleftarrow 0$ ***to*** $i-1$ **do**
    $\alpha_j \leftarrow 0, \hat{\alpha}_j \leftarrow 0$
  end
  $\hat{\alpha}_i \leftarrow \alpha_i + \mathrm{Lap}\left(\frac{1}{\varepsilon'}\right)$
  Output at time t $\mathcal{B}(t) \leftarrow \sum_{j:\mathrm{Bin}_j(T)=1} \hat{\alpha}_j$
end

Theorem 12 (Theorem 4.1 in [DNPR10])**.** *The counter algorithm 5 run with parameters T, ε, β, yields a T-bounded counter with ε-differential privacy, such that with probability at least $1-\beta$ the error for all prefixes $1 \leq t \leq T$ is at most $\frac{4}{\varepsilon} \log(1/\beta) \log^{2.5}(T)$.*

B PAC and Regret Analysis of algorithm PUCB

In this section we provide the complete PAC and Regret analysis of algorithm PUCB corresponding to theorem 3 and 4 respectively. We begin by analyzing the PAC sample complexity.

B.1 PAC guarantee for PUCB. Proof of theorem 3

We restate the PAC guarantee.

Theorem (PAC guarantee for PUCB. Theorem 3)**.** *Let T be the maximum number of episodes and ε the JDP parameter. Then for any $\alpha \in (0, H]$ and $\beta \in (0,1)$, algorithm PUCB with parameters (ε, β) follows a policy that with probability at least $1-\beta$ is α-optimal on all but*

$$O\left(\left(\frac{SAH^4}{\alpha^2} + \frac{S^2AH^4}{\varepsilon\alpha}\right)\mathrm{polylog}\left(T, S, A, H, \tfrac{1}{\alpha}, \tfrac{1}{\beta}, \tfrac{1}{\varepsilon}\right)\right)$$

episodes.

The term $\frac{S^2AH^4}{\varepsilon\alpha}$ in theorem 3 is the extra sample complexity due to the constraint of differential privacy. Importantly, as we will show in section 6, the privacy term matches the lower bound in ε and α. Although it remains an open problem whether the dependence on S^2 in the lower order term is necessary for privacy.

Proof. of theorem 3. We use a similar approach as in [DLB17] which uses the concept of nice episodes but we modify their definition of nice episodes to account for the noise the algorithm adds in order to preserve privacy. Denote by $[T]$ the set of all episodes where T is the maximum number of episodes and $\mathrm{x} := (s, a, h)$.

The term $w_t(\mathrm{x})$ gives the probability of visiting state tuple x after following policy π_t during episode t. Let the set of nice episodes $N \subset [T]$ be defined as

$$N := \left\{ t : w_t(\mathrm{x}) < w_{min} \text{ or } \frac{1}{4} \sum_{i<t} w_i(\mathrm{x}) \geq \ln \frac{SAH}{\beta'} + 2E_\varepsilon \right\}$$

The number of suboptimal episodes is bounded by the number of suboptimal nice episodes plus the number of non-nice episodes. In this section we demonstrate how to bound each individually.

Optimality gap decomposition Fix some episode t and let π_t be the policy produced by algorithm 2. The optimality gap for episode t is denoted by $\Delta_t := V_1^\star - V_1^{\pi_t}$. In section B.4 we show how to construct the optimistic Q-value function $\widetilde{Q}_t^+$ used by algorithm PUCB. And in section B.5 we use the optimism of $\widetilde{Q}_t^+$ to decompose the optimality gap as follows:

$$\Delta_t \leq \sum_{\mathrm{x} \in \mathcal{S} \times \mathcal{A} \times [H]} w_t(\mathrm{x}) \widetilde{\mathrm{conf}}_t(\mathrm{x}) \tag{5}$$

We define a set $L_t := \{\mathrm{x} : w_t(\mathrm{x}) < w_{min}\}$ where $w_{min} := \frac{\alpha}{3SH^2}$. The set L_t contains all state tuples with low probability of being visited during episode t by following policy π_t. We can now decompose equation (5) further

$$\Delta_t \leq \sum_{\mathrm{x} \in L_t} w_t(\mathrm{x}) \widetilde{\mathrm{conf}}_t(\mathrm{x}) + \sum_{\mathrm{x} \notin L_t} w_t(\mathrm{x}) \widetilde{\mathrm{conf}}_t(\mathrm{x})$$

We choose w_{min} such that $\sum_{\mathrm{x} \in L_t} w_t(\mathrm{x}) \widetilde{\mathrm{conf}}_t(\mathrm{x}) \leq \frac{\alpha}{3}$. Hence the gap is upper bounded by:

$$\Delta_t \leq \frac{\alpha}{3} + \sum_{\mathrm{x} \notin L_t} w_t(\mathrm{x}) \widetilde{\mathrm{conf}}_t(\mathrm{x})$$

Now we only need to bound the number of episodes where the term $\sum_{\mathrm{x} \notin L_t} w_t(\mathrm{x}) \widetilde{\mathrm{conf}}_t(\mathrm{x})$ is greater than $2\alpha/3$.

Bounding suboptimal episodes First we bound the number of suboptimal nice episodes. Note that from algorithm PUCB we have $\widetilde{\mathrm{conf}}_t(\mathrm{x}) = (H+1)\widetilde{\phi}_t(\mathrm{x}) + \widetilde{\psi}_t(\mathrm{x})$ if $\widetilde{n}_t(\mathrm{x}) \geq 2E_\varepsilon$ otherwise $\widetilde{\mathrm{conf}}_t(\mathrm{x}) = H$. However, we use a properties of nice episodes (from lemma 19) which says that if $\mathrm{x} \notin L_t$ and t is a nice episode then $\widetilde{n}_t(\mathrm{x}) \geq 2E_\varepsilon$. Therefore, if t is a nice episode, we can replace every $\widetilde{\mathrm{conf}}_t(\mathrm{x})$ term and the gap can be upper bounded by

$$\Delta_t \leq \frac{\alpha}{3} + \Delta_{1,t} + \Delta_{2,t}$$

where

$$\Delta_{1,t} = \sum_{\mathrm{x} \notin L_t} w_t(\mathrm{x})(H+1)\widetilde{\phi}_t(\mathrm{x}) \qquad \text{and} \qquad \Delta_{2,t} = \sum_{\mathrm{x} \notin L_t} w_t(\mathrm{x}) \widetilde{\psi}_t(\mathrm{x})$$

Now we bound the number of nice episodes where term $\Delta_{2,t}$ is greater than $\alpha/3$. Recal that $\widetilde{\psi}_t(\mathrm{x}) = (1+SH)\left(\frac{3E_\varepsilon}{\widetilde{n}_t(\mathrm{x})} + \frac{2E_\varepsilon^2}{\widetilde{n}_t(\mathrm{x})^2}\right)$. We use lemma 19 again that says that w.p at leat $1-\beta$, if t is a nice episode and if $\mathrm{x} \notin L_t$ then we have $\widetilde{n}_t(\mathrm{x}) \geq 2E_\varepsilon$. Thus, the following is true: $\frac{E_\varepsilon^2}{\widetilde{n}_t(\mathrm{x})^2} < \frac{E_\varepsilon}{\widetilde{n}_t(\mathrm{x})}$ on nice episode t. We can upper bound the gap $\Delta_{2,t}$ with

$$\Delta_{2,t} \leq \sum_{\mathrm{x} := (s,a,h) \notin L_t} w_t(\mathrm{x})(1+SH)\frac{10E_\varepsilon}{\widetilde{n}_t(\mathrm{x})} \tag{6}$$

In section B.6 we show how to bound the number of nice episodes where the term from right side of inequality (6) is bigger than $\alpha/3$. That is we use lemma 21 from section B.6 with $r = 1$ to show that the number of nice episodes $t \in N$ where $\Delta_{2,t} > \frac{\alpha}{3}$ is at most

$$\frac{240 E_\varepsilon SAH}{\alpha} \text{polylog}(T, S, A, H, \tfrac{1}{\varepsilon}, \tfrac{1}{\beta}, \tfrac{1}{\alpha}) \tag{7}$$

For gap $\Delta_{1,t}$ we have the upper bound

$$\Delta_{1,t} \leq \sum_{x \notin L_t} w_t(\mathrm{x})(H+1)\sqrt{\frac{2\ln(T/\beta')}{\max(\widetilde{n}_t(\mathrm{x}) - E_\varepsilon, 1)}} \tag{8}$$

We use lemma 21 from section B.6 again with $r = 1/2$ to show that the right side of equation 8 is greater than $\alpha/3$ on at most

$$\frac{18SAH^4}{\alpha^2} \text{polylog}(T, S, A, H, \tfrac{1}{\beta}, \tfrac{1}{\alpha}) \tag{9}$$

nice episodes. Finally, lemma 20 from B.6 says that the set of non-nice episodes is at most

$$\frac{120S^2AH^4}{\alpha\varepsilon} \text{polylog}(T, S, A, H, \tfrac{1}{\beta}) \tag{10}$$

Combining equations 10, 9, and 7 gives the most number of α-suboptimal episodes

$$O\left(\left(\frac{S^2AH^4}{\varepsilon\alpha} + \frac{SAH^3}{\alpha^2}\right) \text{polylog}(T, S, A, H, \tfrac{1}{\beta}, \tfrac{1}{\alpha})\right)$$

conmpleting the proof.

□

B.2 Regret bound for PUCB. Proof of theorem 4

In this section we layout the proof the regret bound from theorem 4. We reuse some tools from the previous PAC analysis and also use similar techniques as in [AOM17]. As in the PAC analysis the key to getting the right dependence on α and ε lies in the decomposition of the confidence bounds. We restate the theorem below and the provide the proof.

Theorem (Regret bound for PUCB. Theorem 4.). *With probability at least $1 - \beta$, the regret of PUCB up to episode T is at most*

$$O\left(\left(H^2\sqrt{SAT} + \frac{SAH^3 + S^2AH^3}{\varepsilon}\right) \text{polylog}\left(T, S, A, H, \tfrac{1}{\beta}, \tfrac{1}{\varepsilon}\right)\right) .$$

Proof. of theorem 4 Denote by $[T]$ the set of all episodes where T is the maximum number of episodes. Let $\mathrm{x} := (s, a, h)$. The term $w_t(\mathrm{x})$ gives the probability of visiting state tuple x after following policy π_t during episode t.

Let $\Delta_t := V_1^\star - V_1^{\pi_t}$ be the optimality gap for episode t given that the learner plays policy π_t, then the expected regret of the learner at episode $T \in [T]$ is given by

$$\text{Regret}(T) = \sum_{t=1}^{T} \Delta_t \tag{11}$$

Optimality gap decomposition In section B.4 we show how to construct the optimistic Q-value function $\widetilde{Q}_t^+$ used by algorithm PUCB. And in section B.5 we use the optimism of $\widetilde{Q}_t^+$ to decompose the optimality gap as follows:

$$\Delta_t \le \sum_{h=1}^{H} \mathbb{E}_{s\sim\pi_t(\cdot,h)} \widetilde{\mathrm{conf}}_t(s,\pi_t(s,h),h) \tag{12}$$

Therefore the regret is bounded by

$$\mathrm{Regret}(T) \le \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{s\sim\pi_t(\cdot,h)} \left[\widetilde{\mathrm{conf}}_t(s,\pi_t(s,h),h)\right] \tag{13}$$

For brevity let $\mathrm{x}_{t,h} := (s_{t,h}, \pi_t(s_{t,h}), h)$ where $s_{t,h}$ is the state visited by the agent during episode t and time h. Then we can bound the regret by

$$\mathrm{Regret}(T) \le \sum_{t=1}^{T}\sum_{h=1}^{H} \left(\mathbb{E}_{s\sim\pi_t(\cdot,h)} \left[\widetilde{\mathrm{conf}}_t(s,\pi_t(s,h),h)\right] - \widetilde{\mathrm{conf}}_t(\mathrm{x}_{t,h})\right) + \sum_{t=1}^{T}\sum_{h=1}^{H} \widetilde{\mathrm{conf}}_t(\mathrm{x}_{t,h}) \tag{14}$$

Next step to get the regret bound is to bound each term from equation (14) individually.

Bounding martingale sequence $\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{s\sim\pi_t(\cdot,h)} \left[\widetilde{\mathrm{conf}}_t(s,\pi_t(s,h),h)\right] - \widetilde{\mathrm{conf}}_t(\mathbf{x}_{t,h})$: The first of equation (14) is sequence of random variables. Azuma's concentration bound says that $\Pr[X_n - X_0 > b] < \exp\left(\frac{-2b^2}{\sum_{i=1}^{n} c_i^2}\right)$ for martingale sequence (X_i) such that $|X_i - X_{i-1}| < c_i$. Let

$$X_t = \sum_{i=1}^{t}\sum_{h=1}^{H} \left(\mathbb{E}_{s\sim\pi_i(\cdot,h)} \left[\widetilde{\mathrm{conf}}_i(s,\pi_i(s,h),h)\right] - \widetilde{\mathrm{conf}}_i(\mathrm{x}_{i,h})\right)$$

be a sequence of T random variables where each X_t depends on the realizations of the previous $X_1,\ldots,X_{t-1}$. Then it follows by the boundness of $\widetilde{\mathrm{conf}}(\cdot)$ that $|X_t - X_{t-1}| \le H^2$ and that each random variable X_t in the sequence has mean zero. Hence $X_1,\ldots,X_T$ is a martingale sequence and we can apply Azuma's inequality to get that on round T with probability at least $1-\beta'$ we have

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \left(\mathbb{E}_{s\sim\pi_t(\cdot,h)} \left[\widetilde{\mathrm{conf}}_t(s,\pi_t(s,h),h)\right] - \widetilde{\mathrm{conf}}_t(\mathrm{x}_{t,h})\right) \le H^2\sqrt{\tfrac{1}{2}T\log(1/\beta')} \tag{15}$$

The last step is to set fail probability to $\beta' = \frac{\beta}{T}$ and apply union bound over T rounds

Bounding Exploration bonus term $\sum_{t=1}^{T}\sum_{h=1}^{H} \widetilde{\mathrm{conf}}_t(\mathbf{x}_{t,h})$: Next, we focus on bounding the second term from equation (14). As seen before $(x_{t,1},\ldots,x_{t,H})$ the a sequence of state-tuples corresponding to the trajectory observed by the agent during episode t. Let N be the set of episodes with each state-tuple in the trajectory visited at least $3E_\varepsilon$ many times, that is,

$$N := \left\{t \in [T] : \quad \forall_{h\in[H]}, \widehat{n}_t(\mathrm{x}_{t,h}) \ge 3E_\varepsilon\right\}$$

Then we can decompose the second term from equation (14) using N as follows:

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \widetilde{\mathrm{conf}}_t(\mathrm{x}_{t,h}) = \sum_{t\in N}\sum_{h=1}^{H} \widetilde{\mathrm{conf}}_t(\mathrm{x}_{t,h}) + \sum_{t\notin N}\sum_{h=1}^{H} \widetilde{\mathrm{conf}}_t(\mathrm{x}_{t,h}) \tag{16}$$

There is only a finite number of episodes with $t \notin N$. To bound the maximum cardinality of the set $\{t : t \notin N\}$, consider the smallest visitation count in the trajectory of episode t, let's denote it by $m_t = \min_h \widehat{n}_t(\mathrm{x}_{t,h})$.

By a pigeon-hole argument, after SA episodes m_t must increase by at least one. It follows that after $3E_\varepsilon SA$ episodes we have that $m_t \geq 3E_\varepsilon$. Hence $|\{t : t \notin N\}| \leq 3E_\varepsilon SA$. Therefore, we can bound the second term of equation (16), $\sum_{t \notin N} \sum_{h=1}^{H} \widetilde{\mathrm{conf}}_t(\mathrm{x}_{t,h})$, by

$$\sum_{t \notin N} \sum_{h=1}^{H} \widetilde{\mathrm{conf}}_t(\mathrm{x}_{t,h}) \leq 3E_\varepsilon SAH \widetilde{\mathrm{conf}}_t(\mathrm{x}_{t,h}) \leq 3E_\varepsilon SAH^2 = \tfrac{3}{\varepsilon} SAH^3 \log\Big(\tfrac{2SAH+S^2AH}{\beta'}\Big) \log(T)^{5/2} \tag{17}$$

The first inequality of (17) follows from $|\{t : t \notin N\}| \leq 3E_\varepsilon SA$, and the second inequality from $\widetilde{\mathrm{conf}}_t(\mathrm{x}_{t,h}) \leq H$. We get the last equality by setting $E_\varepsilon = \frac{3}{\varepsilon} H \log\Big(\frac{2SAH+S^2AH}{\beta'}\Big) \log(T)^{5/2}$. Next note that if $t \in N$ then visited states $\mathrm{x}_{t,h}$ on episode t have been seen at least $3E_\varepsilon$ many times, i.e, $\widehat{n}_t(\mathrm{x}_{t,h}) \geq 3E_\varepsilon$. Then on the high probability event $|\widetilde{n}_t(\mathrm{x}_{t,h}) - \widehat{n}_t(\mathrm{x}_{t,h})| \leq E_\varepsilon$ it follows that $\widetilde{n}_t(\mathrm{x}_{t,h}) \geq 2E_\varepsilon$ hence, from algorithm 3, we have that $\widetilde{\mathrm{conf}}_t(\mathrm{x}_{t,h}) = \widetilde{\psi}_t(\mathrm{x}_{t,h}) + (H+1)\widetilde{\phi}_t(\mathrm{x}_{t,h})$. Now we can decompose the first term of 16 as

$$\sum_{t \in N} \sum_{h=1}^{H} \widetilde{\mathrm{conf}}_t(\mathrm{x}_{t,h}) = (H+1) \sum_{t \in N} \sum_{h=1}^{H} \widetilde{\phi}_t(\mathrm{x}_{t,h}) + \sum_{t \in N} \sum_{h=1}^{H} \widetilde{\psi}_t(\mathrm{x}_{t,h}) \tag{18}$$

We now bound the first term of equation (18). We will use a pigeon-hole argument for the next step which goes as follows. If for all state tuples $\mathrm{x}_{t,h}$ in the trajectory of the agent we have $\widehat{n}_t(\mathrm{x}_{t,h}) \geq 1$ then

$$\sum_{t=1}^{T} \sum_{h=1}^{H} \frac{1}{\widehat{n}_t(\mathrm{x}_{t,h})} \leq \sum_{(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]} \sum_{m=1}^{\widehat{n}_t(s,a,h)} \tfrac{1}{m} \leq SAH \ln(T)$$

The last inequality follows from the fact that $\widehat{n}_t(\cdot) \leq T$ and the bound $\sum_{m=1}^{T} = \frac{1}{1} + \frac{1}{2} + \ldots + \frac{1}{T} \leq \ln(T) + 1$. Recall that

$$\widetilde{\phi}_t(\mathrm{x}_{t,h}) := \sqrt{\frac{2\ln(T/\beta')}{\max(\widetilde{n}_t(\mathrm{x}_{t,h}) - E_\varepsilon, 1)}}$$

and let $L = 2\ln(T/\beta')$. We are ready to bound the first term of equation (18)

$$\begin{aligned}
(H+1) \sum_{t \in N} \sum_{h=1}^{H} \widetilde{\phi}_t(\mathrm{x}_{t,h}) &\leq (H+1)\sqrt{L} \sum_{t \in N} \sum_{h=1}^{H} \sqrt{\frac{1}{\widetilde{n}_t(\mathrm{x}_{t,h}) - E_\varepsilon}} \\
&\leq (H+1)\sqrt{L} \sum_{t \in N} \sum_{h=1}^{H} \sqrt{\frac{1}{\widehat{n}_t(\mathrm{x}_{t,h}) - 2E_\varepsilon}} && (19) \\
&\leq (H+1)\sqrt{L} \sqrt{\sum_{t \in N} \sum_{h=1}^{H} 1} \sqrt{\sum_{t \in N} \sum_{h=1}^{H} \frac{1}{\widehat{n}_t(\mathrm{x}_{t,h}) - 2E_\varepsilon}} && (20) \\
&\leq (H+1)\sqrt{L}\sqrt{TH} \sqrt{\sum_{\mathrm{x} \in \mathcal{S} \times \mathcal{A} \times [H]} \sum_{m=E_\varepsilon}^{\widehat{n}_t(\mathrm{x}) - 2E_\varepsilon} \frac{1}{m}} && (21) \\
&\leq (H+1)\sqrt{L}\sqrt{TH}\sqrt{SAH \ln(T)} && (22) \\
&\leq H^2 \sqrt{SAT} \mathrm{polylog}(T, S, A, H, 1/\beta) && (23)
\end{aligned}$$

For the inequality (19) we use fact that on the good event we have $\widetilde{n}_t(\mathrm{x}_{t,h}) \geq \widehat{n}_t(\mathrm{x}_{t,h}) - E_\varepsilon$. For inequality (20) we use Cauchy-Schwarz inequality. Then we use the pigeon-hole principle for inequality (21). Finally for inequality (22) we use the bound $\frac{1}{1} + \frac{1}{2} + \ldots + \frac{1}{T} \leq \ln(T) + 1$

Next we bound the second term of equation (18). Recall that

$$\widetilde{\psi}_t(\mathrm{x}_{t,h}) := (1+SH)\left(\frac{3E_\varepsilon}{\widetilde{n}_t(\mathrm{x}_{t,h})} + \frac{2E_\varepsilon^2}{\widetilde{n}_t(\mathrm{x}_{t,h})^2}\right)$$

Since we are considering episodes in $t \in N$ for each $\mathrm{x}_{t,h}$ we have that $\widehat{n}(\mathrm{x}_{t,h}) \geq 3E_\varepsilon$ and thus $\widetilde{n}_t(\mathrm{x}_{t,h}) \geq 2E_\varepsilon$ on the high-probability good event. Then it follows that we can upper bound $\frac{E_\varepsilon^2}{\widetilde{n}_t(\mathrm{x}_{t,k})^2}$ by $\frac{E_\varepsilon}{\widetilde{n}_t(\mathrm{x}_{t,k})}$. We end up with

$$\begin{aligned}
\sum_{t\in N}\sum_{h=1}^{H}\widetilde{\psi}_t(\mathrm{x}_{t,h}) &\leq (1+SH)5E_\varepsilon \sum_{t\in N}\sum_{h=1}^{H}\frac{1}{\widetilde{n}_t(\mathrm{x}_{t,h})} \\
&\leq (1+SH)5E_\varepsilon \sum_{t\in N}\sum_{h=1}^{H}\frac{1}{\widehat{n}_t(\mathrm{x}_{t,h}) - E_\varepsilon} && (24)\\
&\leq (1+SH)5E_\varepsilon \sum_{\mathrm{x}\in\mathcal{S}\times\mathcal{A}\times[H]}\ \sum_{m=2E_\varepsilon}^{\widehat{n}_t(\mathrm{x})-E_\varepsilon}\frac{1}{m} && (25)\\
&\leq 5S^2AH^2E_\varepsilon\log(T) \\
&\leq \frac{15}{\varepsilon}S^2AH^3\log(\tfrac{1}{\beta'})\log(T)^{7/2}\mathrm{polylog}(S,A,H) && (26)
\end{aligned}$$

Putting together inequalities (15), (17) , (23), (26) we have

$$\begin{aligned}
\mathrm{Regret}(T) \leq H^2\sqrt{\tfrac{1}{2}T\log(T/\beta)} + H^2\sqrt{SAT}\mathrm{polylog}(T,S,A,H,1/\beta) \\
+\frac{15}{\varepsilon}S^2AH^3\log(\tfrac{1}{\beta'})\log(T)^{7/2}\mathrm{polylog}(S,A,H) \\
+\tfrac{3}{\varepsilon}SAH^3\log\left(\tfrac{2SAH+S^2AH}{\beta'}\right)\log(T)^{5/2}
\end{aligned}$$

□

B.3 Error bounds

The proof of theorem 3 relies on the confidence bounds on the empirical estimates of the sufficient statistics $\widehat{n}_t(s,a,h)$, $\widehat{r}_t(s,a,h)$ and $\widehat{m}_t(s,a,h)$ as well as the error upper bound on the private estimates $\widetilde{n}_t(s,a,h)$, $\widetilde{r}_t(s,a,h)$ $\widetilde{m}_t(s,a,h)$.

We will combine the confidence bounds from definition 6 to construct a confidence interval $\widetilde{\mathrm{conf}}(s,a,h)$ for the private Q-values The main challenge lie in getting the correct sample complexity dependence on the target accuracy α and the privacy parameter ε.

Definition 6. The fail event

$$F = \bigcup_{t=1}^{T}\left[F_t^N \cup F_t^R \cup F_t^V \cup F_t^{PR} \cup F_t^{PN} \cup F_t^{PM}\right]$$

where T is the maximum number of episodes is defined as

$$F_t^N = \left\{\exists s,a,h : \widehat{n}_t(s,a,h) < \frac{1}{2}\sum_{i<t} w_{i,h}(s,a) - \ln\frac{SAH}{\beta'}\right\}$$

$$F_t^R = \left\{\exists s,a,t : \left|\frac{\widehat{r}_t(s,a,h)}{\widehat{n}_t(s,a,h)} - r(s,a,h)\right| \geq \widehat{\phi}_t(s,a,h)\right\}$$

$$F_t^V = \left\{\exists s,a,h : |(\widehat{\mathcal{P}}_t(s,a,h) - \mathcal{P}(s,a,h))^\top V_{h+1}^* | \geq H\widehat{\phi}_t(s,a,h)\right\}$$
$$F_t^{PR} = \{\exists s,a,h : |\widetilde{r}_t(s,a,h) - \widehat{r}_t(s,a,h)| \geq E_\varepsilon\}$$
$$F_t^{PN} = \{\exists s,a,h : |\widetilde{n}_t(s,a,h) - \widehat{n}_t(s,a,h)| \geq E_\varepsilon\}$$
$$F_t^{PM} = \{\exists s,a,s',h : |\widetilde{m}_t(s,a,s',h) - \widehat{m}_t(s,a,s',h)| \geq E_\varepsilon\}$$

where $\widehat{\phi}_t(s,a,h) = \sqrt{\frac{T/\beta'}{\widehat{n}_t(s,a,h)}}$ and $E_\varepsilon = \frac{3}{\varepsilon}H\log\left(\frac{2SAH+S^2AH}{\beta'}\right)\log(T)^{5/2}$.

Lemma 13. *The fail event F from definition 6 occurs with probability at most β.*

Proof. From [DLB17, Corollary E.4] we have that $\Pr\left[\bigcup_{t=1}^T F_t^N\right] \leq \beta'$. Using the standard Chernoff Bound inequality and union bound over T rounds we have that $\Pr\left[\bigcup_{t=1}^T F_t^R\right] \leq \beta'$ and $\Pr\left[\bigcup_{t=1}^T F_t^V\right] \leq \beta'$.

There is a total of $SAH+S^2AH$ private counters. From the error bound (theorem 12) of the private counter from algorithm 5 with fail probability set to $\frac{\beta'}{SAH+S^2AH}$ and applying union bound over all $SAH+S^2AH$ counters we have

$$\Pr\left[\bigcup_{t=1}^T \left(F_t^{PR} \cup F_t^{PN} \cup F_t^{PM}\right)\right] \leq \beta'$$

Finally setting $\beta' = \frac{\beta}{4}$ and applying another union bound over 4 fail events we obtain that $\Pr[F] \leq \beta$. □

B.4 Q-optimism

The proof will follow the principle of optimism under uncertainty as in previous work [DLB17]. In order to obtain the right sample complexity dependence on α and ε we must construct a confidence bound that disentangles the sampling error from the empirical estimates and the error from the private counters.

We will construct a confidence bound for any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and any episode $t \in [T]$. To reduce notation clutter we will use $\mathrm{x} := (s,a,h)$ in place of (s,a,h). Our objective in this section is to use the private counts to construct an optimistic and private Q-function $\widetilde{Q}_t^+$. We say that a Q-function $\widehat{Q}^+$ is optimistic with respect to the Q-function induced from the optimal policy Q^* if for all x we have $\widehat{Q}_t(\mathrm{x}) \geq Q^*(\mathrm{x})$ with high probability. To that end we first construct the optimistic but non-private $\widehat{Q}_t^+$ using the non-private counters. Let us first define the empirical Q-value estimate $\widehat{Q}_t$ as

$$\widehat{Q}_t(\mathrm{x}) = \frac{\widehat{r}_t(\mathrm{x}) + \sum_{s' \in \mathcal{S}} \widetilde{V}_{h+1}(s')\widehat{m}_t(\mathrm{x},s')}{\widehat{n}_t(\mathrm{x})}$$

and the optimistic Q-function on episode t is by

$$\widehat{Q}_t^+(\mathrm{x}) = \widehat{Q}_t(\mathrm{x}) + (H+1)\widehat{\phi}_t(\mathrm{x}) \tag{27}$$

where $\widehat{\phi}_t(\mathrm{x}) = \sqrt{\frac{T/\beta'}{\widehat{n}_t(\mathrm{x})}}$.

To show that $\widehat{Q}_t^+(\mathrm{x})$ is optimistic with respect to $Q^*(\mathrm{x})$ it suffices to show that $\widetilde{V}_h$ is optimistic with respect to V^* which follows from induction on $h \in [H]$ and from the standard concentration bounds. These two requirements are formilized in lemma 16 and 17 below.

To construct a private Q-function we must use the private counts. We denote the private empirical Q-value estimate on episode t by

$$\widetilde{Q}_t(\mathrm{x}) = \frac{\widetilde{r}_t(\mathrm{x}) + \sum_{s' \in \mathcal{S}} \widetilde{V}_{h+1}(s')\widetilde{m}_t(\mathrm{x},s')}{\widetilde{n}_t(\mathrm{x})}$$

The optimistic private Q-function is defined as

$$\widetilde{Q}_t^+(\mathrm{x}) = \widetilde{Q}_t(\mathrm{x}) + (H+1)\widetilde{\phi}_t(\mathrm{x}) + \widetilde{\psi}_t(\mathrm{x}) \tag{28}$$

Theorem 14. *On the good event F^c, the Q-function $\widetilde{Q}_t^+$ from equation 28 is optimistic. That is, for any $t \in [T]$ we have $\widetilde{Q}_t^+(x) \geq \widehat{Q}_t^+(x)$ for all tuples $x = (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$.*

Proof. First use lemma 15 to show that the private optimismtic $\widetilde{Q}^+$ is optimistic with respect to $\widehat{Q}^+$ from equation 27. From lemma 17 we have that the Q-function $\widehat{Q}^+$ is optimistic. Putting it all together, on the good event F^c, for any round t and all state tuples $\mathrm{x} = (s,a,h)$ we have

$$\widetilde{Q}_t^+(\mathrm{x}) \geq \widehat{Q}_t^+(\mathrm{x}) \geq Q^*(\mathrm{x})$$

completing the proof. □

Lemma 15. *On the good event F^c, for any $t \in [T]$ we have $\widetilde{Q}_t^+(x) \geq \widehat{Q}_t^+(x)$ for all tuples $x = (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$.*

Proof. We attempt to construct a confidence bound for $\widetilde{Q}_t$. By using the error bound E_ε of the private counters, we can upper bound $\widehat{Q}_t$ in terms of the private counters as follows

$$\widehat{Q}_t(\mathrm{x}) \leq \frac{\widetilde{r}_t(\mathrm{x}) + E_\varepsilon + \sum_{s' \in \mathcal{S}} \widetilde{V}_{h+1}(s')(\widetilde{m}_t(\mathrm{x},s') + E_\varepsilon)}{\widetilde{n}_t(\mathrm{x}) - E_\varepsilon} \tag{29}$$

The following claim will help us recover $\widetilde{Q}_t$ from 29.

Claim. *2 Let $y \in \mathbb{R}$ be any positive real number. Then for all $x \in \mathbb{R}$ with $x \geq 2y$ it holds that $\frac{1}{x-y} \leq \frac{1}{x} + \frac{2y}{x^2}$*

To get $\widetilde{Q}_t$ back from inequality 29 we apply claim 2 which allows us write $\frac{1}{\widetilde{n}_t(\mathrm{x}) - E_\varepsilon} \leq \frac{1}{\widetilde{n}_t(\mathrm{x})} + \frac{2E_\varepsilon}{\widetilde{n}_t(\mathrm{x})}$ when $\widetilde{n}_t(\mathrm{x}) \geq 2E_\varepsilon$. Then

$$\begin{aligned}
\widehat{Q}_t(\mathrm{x}) &\leq \left(\frac{1}{\widetilde{n}_t(\mathrm{x})} + \frac{2E_\varepsilon}{\widetilde{n}_t(\mathrm{x})^2}\right)\left(\widetilde{r}_t(\mathrm{x}) + E_\varepsilon + \sum_{s' \in \mathcal{S}} \widetilde{V}_{h+1}(s')(\widetilde{m}_t(\mathrm{x},s') + E_\varepsilon)\right) \\
&\leq \frac{\widetilde{r}_t(\mathrm{x}) + \sum_{s' \in \mathcal{S}} \widetilde{V}_{h+1}(s')\widetilde{m}_t(\mathrm{x},s')}{\widetilde{n}_t(\mathrm{x})} + 3\frac{E_\varepsilon + SHE_\varepsilon}{\widetilde{n}_t(\mathrm{x})} + 2\frac{E_\varepsilon^2 + SHE_\varepsilon^2}{\widetilde{n}_t(\mathrm{x})^2} \\
&= \widetilde{Q}_t(\mathrm{x}) + \widetilde{\psi}_t(\mathrm{x})
\end{aligned}$$

So far we have the following bound when $\widetilde{n}_t(x) \geq 2E_\varepsilon$

$$Q^*(\mathrm{x}) \leq \widehat{Q}_t^+(\mathrm{x}) \leq \widetilde{Q}_t(\mathrm{x}) + \widetilde{\psi}_t(\mathrm{x}) + (H+1)\widehat{\phi}_t(\mathrm{x})$$

In the case when $\widetilde{n}_t(x) < 2E_\varepsilon$ we can simly upper bound Q^* by H. The last step is to replace the term $\widehat{\phi}_t(\mathrm{x})$ which is not a private object. We again use the error bound E_ε from the private counters to write

$$\widehat{\phi}_t(\mathrm{x}) = \sqrt{\frac{T/\beta'}{\widehat{n}_t(\mathrm{x})}} \leq \sqrt{\frac{2\ln(T/\beta')}{\max(\widetilde{n}_t(\mathrm{x}) - E_\varepsilon, 1)}} = \widetilde{\phi}_t(\mathrm{x})$$

Finally we can write an expression for the optimistic private Q-function as

$$\widetilde{Q}^+(\mathrm{x}) := \begin{cases} \widetilde{Q}(\mathrm{x}) + \widetilde{\psi}_t(\mathrm{x}) + (H+1)\widetilde{\phi}_t(\mathrm{x}) & \text{if } \widetilde{n}_t(\mathrm{x}) \geq 2E_\varepsilon \\ H & \text{otherwise} \end{cases}$$

Therefore we have by construction that $\widetilde{Q}^+(\mathrm{x}) \geq \widehat{Q}^+(\mathrm{x})$. □

Lemma 16 (Value function optimism)**.** *On the event F^c, the value function from algorithm 3 is optimismtic. i.e. for all $s \in \mathcal{S}$ and all $h \in [H]$ we have $\widetilde{V}_h(s) \geq V_h^*(s)$*

Lemma 17. *On the good event F^c, $\widehat{Q}_t^+(x)$ is optimistic with respect to $Q^*(x)$*

Proof. Let the empirical mean reward on round t be $\bar{r}_t(\mathrm{x}) = \frac{\widehat{r}_t(\mathrm{X})}{\widehat{n}_t(\mathrm{X})}$ and let $\widehat{P}_t\widetilde{V}_{h+1} = \frac{\sum_{s' \in \mathcal{S}} \widetilde{V}_{h+1}(s')(\widehat{m}_t(\mathrm{X}, s')}{\widehat{n}_t(\mathrm{X})}$. Now we can write $\widehat{Q}_t^+$ as

$$\widehat{Q}_t^+(\mathrm{x}) = \bar{r}_t(\mathrm{x}) + \widehat{P}_t\widetilde{V}_{h+1} + (H+1)\widehat{\phi}_t(\mathrm{x})$$

On the event F^c we have that $\frac{\widehat{r}_t(\mathrm{X})}{\widehat{n}_t(\mathrm{X})} - r(\mathrm{x}) \leq \widehat{\phi}_t(\mathrm{x})$ and $(\widehat{P}_t - P)V_{h+1}^* \leq H\widehat{\phi}_t(\mathrm{x})$. Furthermore, from the value function optimism lemma (16) we have that for all $s \in \mathcal{S}$, and all $h \in [H]$, $\widetilde{V}_h(s) \geq V_h^*(s)$. Putting it all together:

$$\begin{aligned}
\widehat{Q}_t^+(\mathrm{x}) - Q^*(\mathrm{x}) &= \bar{r}_t(\mathrm{x}) - r(\mathrm{x}) + \widehat{P}_t\widetilde{V}_{h+1} - PV_{h+1}^* + (H+1)\widehat{\phi}_t(\mathrm{x}) \\
&\geq \bar{r}(\mathrm{x}) - r(\mathrm{x}) + (\widehat{P} - P)V_{h+1}^* + \widehat{\phi}_t(\mathrm{x}) + H\widehat{\phi}_t(\mathrm{x}) \\
&\geq 0
\end{aligned}$$

Completing the proof. □

Proof of lemma 16

Proof. The proof proceeds by induction. Fixing any state $s \in \mathcal{S}$, we must show that $\widetilde{V}_h(s) \geq V_h^*(s)$ for all $h \in [H]$. For the base case, note that $\widetilde{V}_{H+1}(s) = V_h^*(s) = 0$. Now assume that for any $h \leq H$, $\widetilde{V}_{h+1}(s) \geq V_{h+1}^*(s)$. Then we must show that $\widetilde{V}_h(s) \geq V_h^*(s)$. First write out the equation for $V_h^*(s)$:

$$V_h^*(s) = \max_{a^* \in \mathcal{A}} \Big(r(s, a^*, h) + P(s, a^*, h)V_{h+1}^* \Big) \tag{30}$$

where $r(s,a,h)$ is the true mean reward and $P(s,a,h)$ the true transition function for state tuple (s,a,h). Let a^* be the action corresponding to equation (30). Next we write out the equation for $\widetilde{V}_h(s)$

$$\begin{aligned}
\widetilde{V}_h(s) &= \max_{a \in \mathcal{A}} \frac{\widetilde{r}_t(s,a,h) + \sum_{s'} \widetilde{V}_{h+1}\widetilde{m}_t(s,a,s',h)}{\widetilde{n}_t(s,a,h)} + \widetilde{\mathrm{conf}}_t(s,a,h) \\
&\geq \frac{\widetilde{r}_t(s,a^*,h) + \sum_{s'} \widetilde{V}_{h+1}\widetilde{m}_t(s,a^*,s',h)}{\widetilde{n}_t(s,a^*,h)} + \widetilde{\mathrm{conf}}_t(s,a^*,h) \\
&= \widetilde{Q}_t^+(s,a^*,h)
\end{aligned}$$

Next we use lemma 15 which says that on event F^c, $\widetilde{Q}_t^+(s,a^*,h) \geq \widehat{Q}_t^+(s,a^*,h)$ to get a lower bound for $\widetilde{V}_h$:

$$\widetilde{V}_h(s) \geq \frac{\widehat{r}_t(s,a^*,h) + \sum_{s'\mathcal{S}} \widetilde{V}_h(s')\widehat{m}_t(s,a^*,s',h)}{\widehat{n}_t(s,a^*,h)} + (H+1)\widehat{\phi}_t(s,a^*,h)$$

Letting

$$\frac{\sum_{s'\mathcal{S}} \widetilde{V}_h(s')\widehat{m}_t(s,a^*,s',h)}{\widehat{n}_t(s,a^*,h)} = \widehat{P}(s,a^*,h)\widetilde{V}_{h+1}$$

and applying the inductive step (i.e. $\forall_s \widetilde{V}_{h+1}(s) \geq V_{h+1}^*(s)$) we get

$$\widetilde{V}_h(s) \geq \frac{\widehat{r}_t(s,a^*,h)}{\widehat{n}_t(s,a^*,h)} + \widehat{P}(s,a^*,h)V_{h+1}^* + (H+1)\widehat{\phi}_t(s,a^*,h)$$

Next we use the concentration bound from definition 6

$$\widehat{P}(s,a^*,h)V^*_{h+1} \geq P(s,a^*,h)V^*_{h+1} - H\widehat{\phi}_t(s,a^*,h)$$

and then we use the definition of $V^*_h(s)$

$$P(s,a^*,h)V^*_{h+1} = -r(s,a^*,h) + V^*_h(s)$$

to get

$$\begin{aligned}\widetilde{V}_h(s) &\geq \frac{\widetilde{r}(s,a^*,h)}{\widetilde{n}(s,a^*,h)} + P(s,a^*,h)V^*_{h+1} + \widehat{\phi}(s,a^*,h) \\ &= \frac{\widetilde{r}(s,a^*,h)}{\widetilde{n}(s,a^*,h)} - r(s,a^*,h) + \widehat{\phi}(s,a^*,h) + V^*_h(s)\end{aligned}$$

Next note that on event F^c, we have

$$\frac{\widetilde{r}(s,a^*,h)}{\widetilde{n}(s,a^*,h)} - r(s,a^*,h) + \widehat{\phi}(s,a^*,h) > 0$$

Hence it follows that $\widetilde{V}_h(s) \geq V^*_h(s)$, completing the proof. □

B.5 Optimality gap

The next step is to decompose the optimality gap $\Delta_t := V_1^\star - V_1^{\pi_t}$ for episode t. The following lemma states that we can upper bound Δ_t by the weighted sum of confidence terms:

Lemma 18. *Let π_t be the policy played by algorithm 2 during episode t. Let $w_t(s,a,h)$ be the probability of visiting state tuple (s,a,h) during episode t. Then the optimality gap is bounded by*

$$V_1^\star - V_1^{\pi_t} \leq \sum_{h=1}^{H} \mathbb{E}_{s\sim\pi_t(\cdot,h)}\widetilde{\text{conf}}_t(s,\pi_t(s,h),h) = \sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} w_t(s,a,h)\widetilde{\text{conf}}_t(s,a,h) \tag{31}$$

Proof. On episode t, let $\widetilde{Q}^+$ be the private and optimismtic Q-fuction from algorithm 3. Given that the learner is following deterministic policy $\pi_t : \mathcal{S}\times[H] \to \mathcal{A}$, then for within episode time-step h we use the short hand notation $r_h(s) := \frac{\widetilde{r}_t(s,\pi_t(s,h),h)}{\widetilde{n}_t(s,\pi_t(s,h),h)}$ to denote the private mean reward estimate on state s and $p_h(s,s') := \frac{\widetilde{m}_t(s,\pi_t(s,h),s',h)}{\widetilde{n}_t(s,\pi_t(s,h),h)}$ to denote the transition probability from state s to state s'. Let $s_1,\ldots,s_H$ be random variables, where each s_h represents the state visited during time-step h after following policy π_t. Next let $\mathbb{E}_{s_1,\ldots,s_H\sim\pi_t}$ denote the expectation only over the randomness of the states $s_1,\ldots,s_H$ after following the deterministic policy π_t on the MDP. For brevity let us use $\mathbb{E}_{\pi_t}$ instead of $\mathbb{E}_{s_1,\ldots,s_H\sim\pi_t}$. Then, from the definition of $\widetilde{Q}^+_t$ in algorithm 3 we have

$$\mathbb{E}_{\pi_t}\widetilde{Q}^+_t(s_1,\pi_t(s_1,1),1) = \mathbb{E}_{\pi_t} r_1(s_1) + \mathbb{E}_{\pi_t} p(s_1,s_2)\widetilde{V}_2(s_2) + \mathbb{E}_{\pi_t}\widetilde{\text{conf}}_t(s_1,\pi_t(s_1,1),1)$$

Since $p(s_1,s_2) \leq 1$, if we set $\widetilde{V}_2(s_2) = \widetilde{Q}^+_t(s_2,\pi_t(s_2,2),2)$ it follows that

$$\mathbb{E}_{\pi_t}\widetilde{Q}^+_t(s_1,\pi_t(s_1,1),1) - \mathbb{E}_{\pi_t} r_1(s_1) \leq \mathbb{E}_{\pi_t}\widetilde{Q}^+_t(s_2,\pi_t(s_2),2) + \mathbb{E}_{\pi_t}\widetilde{\text{conf}}_t(s_1,\pi_t(s_1,h),1) \tag{32}$$

Then the optimality gap on episode t is

$$V_1^\star - V_1^{\pi_t} = V^\star - \sum_{h=1}^{H}\mathbb{E}_{\pi_t} r_h(s_h)$$

$$= \mathbb{E}_{\pi_t} Q^\star(s_1, \pi^\star(s_1), 1) - \sum_{h=1}^{H} \mathbb{E}_{\pi_t} r_h(s_h) \tag{33}$$

(Optimism) (34)

$$\le \mathbb{E}_{\pi_t} \widetilde{Q}_t^+(s_1, \pi^*(s_1), 1) - \sum_{h=1}^{H} \mathbb{E}_{\pi_t} r_h(s_h) \tag{35}$$

(Greedy of policy π_t) (36)

$$\le \mathbb{E}_{\pi_t} \widetilde{Q}_t^+(s_1, \pi_t(s_1, 1), 1) - \sum_{h=1}^{H} \mathbb{E}_{\pi_t} r_h(s_h) \tag{37}$$

$$\le \mathbb{E}_{\pi_t} \widetilde{Q}_t^+(s_1, \pi_t(s_1, 1), 1) - \mathbb{E}_{\pi_t} r_1(s_1) - \sum_{h=2}^{H} \mathbb{E}_{\pi_t} r_h(s_h)$$

(Apply inequality 32) (38)

$$\le \mathbb{E}_{\pi_t} \widetilde{\mathrm{conf}}_t(s_1, \pi_t(s_1, 1), 1) + \mathbb{E}_{\pi_t} \widetilde{Q}_t^+(s_2, \pi_t(s_2, 2), 2) - \sum_{h=2}^{H} \mathbb{E}_{\pi_t} r_h(s_h) \tag{39}$$

The equality 33 is from the definition of V_1^* and $V_1^{\pi_t}$. Then inequality 34 follows from the optimism of $\widetilde{Q}_t^+$. The inequality 36 follows from π_t following a greedy strategy. And finally, inequality 38 is from applying inequality 32.
Since $\widetilde{Q}_t^+(s, a, H+1) = 0$ for all state-action tuples (s, a) then applying equation (32) $H - 1$ more times, we get

$$V_1^\star - V_1^{\pi_t} \le \sum_{h=1}^{H} \mathbb{E}_{s_h \sim \pi_t} \widetilde{\mathrm{conf}}_t(s_h, \pi_t(s_h, h), h) \tag{40}$$

□

B.6 Nice Episodes

The goal in this section is to bound the number of suboptimal episodes. We use a similar approach as in [DLB17] which uses the concept of nice episodes but we modify their definition of nice episodes to to account for the noise the algorithm adds in order to preserve privacy. We formally define nice episodes in definition 7. The rest of the proof proceeds by bounding the number of episodes that are not nice and bounding the number of nice suboptimal episodes.

Recal that $\widetilde{n}_t(s, a, h)$ represents the private count of the number of times state triplet s, a, h has been visited right before episode t. And E_ε is the error of the ε-differentially private counter, that is, on any episode t

$$|\widetilde{n}_t(s, a, h) - \widehat{n}_t(s, a, h)| < E_\varepsilon$$

where $\widehat{n}_t(s, a, h)$ is the true count.

Definition 7 (Nice Episodes. Similar to definition 2 in [DLB17])**.** Let $w(s, a, h)$ be the probability of visiting state s and taking action a during episode t and time-step h after following policy π_t. An episode t is nice if and only if for all $s, \in \mathcal{S}$, $a, \in \mathcal{A}$ and $h \in [H]$ the following two conditions hold:

$$w_t(s, a, h) \le w_{min} \quad \vee \quad \frac{1}{4} \sum_{i<t} w_i(s, a, h) \ge \ln \frac{SAH}{\beta'} + 2E_\varepsilon$$

Lemma 19. *If an episode t is nice, then on F^c for all s,a,h the following statement holds*

$$w_t(s,a,h) < w_{min} \quad \vee \quad \widetilde{n}_t(s,a,h) > \frac{1}{4}\sum_{i<t} w_i(s,a,h) + E_\varepsilon$$

Plus, it follows that if $w_t(s,a,h) > w_{min}$ then $\widetilde{n}_t(s,a,h) > 2E_\varepsilon$.

Proof. Since we consider the event F^c it holds for all s,a,h triplets

$$\widehat{n}_t(s,a,h) > \frac{1}{2}\sum_{i<t} w_i(s,a,h) - \ln(\frac{SAH}{\delta'})$$

and

$$\widetilde{n}_t(s,a,h) > \widehat{n}_t(s,a,h) - E_\varepsilon > \frac{1}{2}\sum_{i<t} w_i(s,a,h) - \ln(\frac{SAH}{\delta'}) - E_\varepsilon > \frac{1}{4}\sum_{i<t} w_i(s,a,h)$$

□

Lemma 20 (Non-nice Episodes. [DLB17])**.** *On the good event F^c, the number of episodes that are not nice is at most*

$$\frac{120S^2AH^4}{\alpha\varepsilon}\text{polylog}(S,A,H,1/\beta)$$

Lemma 21 (Nice Episodes rate. [DLB17])**.** *Let $r \geq 1$ and fix $C > 0$ which can depend polynomially on the relevant quantities and let $D \geq 1$ which can depend poly-logarithmically on the relevant quantities. Finally let $\alpha' > 0$ be the target accuracy and let $\widetilde{n}_t(s,a,h)$ be the private estimate count with error E_ε. Then*

$$\sum_{(s,a,h)\notin L_t} w_h(s,a,h)\left(\frac{C(\log(T)+D)}{\widetilde{n}_t(s,a,h)-E_\varepsilon}\right)^{1/r} \leq \alpha'$$

on all but at most

$$\frac{8CSAH^r}{(\alpha')^r}\text{polylog}(T,S,A,H,1/\varepsilon,1/\beta',1/\alpha')$$

nice episodes.

Proof. The proof follows mostly from the argument in [DLB17, Lemma E.3]. Let $x := (s,a,h)$ denote a state tuple. Define the gap in episode t by

$$\Delta_t = \sum_{\mathrm{x}\notin L_t} w_t(\mathrm{x})\left(\frac{C(\log(T)+D)}{\widetilde{n}_t(\mathrm{x})-E_\varepsilon}\right)^{\frac{1}{r}} = \sum_{\mathrm{x}\notin L_t} w_t(\mathrm{x})^{1-\frac{1}{r}}\left(w_t(\mathrm{x})\frac{C(\log(T)+D)}{\widetilde{n}_t(\mathrm{x})-E_\varepsilon}\right)^{\frac{1}{r}}$$

Using Hölder's inequality

$$\Delta_t \leq \left(\sum_{\mathrm{x}\notin L_t} CH^{r-1} w_t(\mathrm{x})\frac{C(\log(T)+D)}{\widetilde{n}_t(\mathrm{x})-E_\varepsilon}\right)^{\frac{1}{r}}$$

Now we the properties of nice episodes from lemma 19 and the fact that $\sum_{i<t} w_t(\mathrm{x}) \geq 4\ln(SAH/\delta') \geq 2$. Then on the good event F^c we have the following bound

$$\widetilde{n}_t(\mathrm{x}) \geq \frac{1}{4}\sum_{i<t} w_i(\mathrm{x}) + E_\varepsilon \geq \frac{1}{8}\sum_{i\leq t} w_i(\mathrm{x}) + E_\varepsilon$$

The function $\frac{\ln(T)+D}{x}$ is monotonically decreasing for $x \geq 0$. Then we bound

$$\begin{aligned}\Delta_t^r &\leq \sum_{\mathrm{x} \notin L_t} CH^{r-1} w_t(\mathrm{x}) \frac{C(\log(T)+D)}{\widetilde{n}_t(\mathrm{x}) - E_\varepsilon} \\ &\leq 8CH^{r-1} \sum_{\mathrm{x} \notin L_t} w_t(\mathrm{x}) \frac{(\log(T)+D)}{\sum_{i \leq t} w_i(\mathrm{x})}\end{aligned}$$

Let the set of nice episodes be $N = \left\{t : w_t(\mathrm{x}) < w_{min} \text{ or } \frac{1}{4}\sum_{i<t} w_i(\mathrm{x}) \geq \ln \frac{SAH}{\beta'} + 2E_\varepsilon\right\}$ and define a set $M = \{t : \Delta_t > \alpha'\} \cap N$ to be the set of suboptimal nice epidoes. We know that $|M| \leq T$. Finally we can bound the total number of suboptimal nice episodes by

$$\begin{aligned}\sum_{t \in M} \Delta_t^r &\leq \sum_{t \in M} 8CH^{r-1}(\log(T)+D) \sum_{\mathrm{x} \notin L_t} \frac{w_t(\mathrm{x})}{\sum_{i \leq t} w_i(\mathrm{x})} \\ &\leq 8CH^{r-1}(\log(T)+D) \sum_{\mathrm{x} \notin L_t} \sum_{t \in K} \frac{w_t(\mathrm{x})}{\sum_{i \leq t} w_i(\mathrm{x}) \mathbb{1}(w_i(\mathrm{x}) > w_{\min})}\end{aligned}$$

For every $x = (s, a, h)$ consider the sequence $w_i(\mathrm{x}) \in [w_{\min}, 1]$ with $i \in I = \{w_i(\mathrm{x}) \geq w_{\min}\}$ and apply lemma 22 to get

$$\sum_{t \in M} \frac{w_t(\mathrm{x})}{\sum_{i \leq t} w_i(\mathrm{x}) \mathbb{1}(w_i(\mathrm{x}) > w_{\min})} \leq \ln\left(\frac{Te}{w_{\min}}\right)$$

Therefore we have

$$\sum_{t \in M} \Delta_t^r \leq 8CSAH^r(\log(T)+D) \ln\left(\frac{Te}{w_{\min}}\right)$$

□

Since each episode has to contribute at least $(\alpha')^r$ to this bound we have

$$|M| \leq \frac{8CSAH^r(\log(T)+D)\ln\left(\frac{Te}{w_{\min}}\right)}{(\alpha')^2}$$

Completing the proof.

Lemma 22. *[DLB17, Lemma E.5] Let a_i be a sequence taking values in $[a_{min}, 1]$ with $a_{min} > 0$ and $m > 0$, then* $\sum_{k=1}^{m} \frac{a_i}{\sum_{i=1}^{k} a_i} \leq \ln(\frac{me}{a_{min}})$

C PAC and Regret Lower Bound Proofs

C.1 PAC Lower Bound. Proof of theorem 5

In this section we provide the analysis of PAC lower bound from section 6. Below is the proof of theorem 5.

Theorem (PAC Lower Bound. Theorem 5)**.** *Let $\mathcal{M}$ be an RL agent satisfying ε-JDP. Suppose that $\mathcal{M}$ is (α, β)-PAC for some $\beta \in (0, 1/8)$. Then, there exists a fixed-horizon episodic MDP where the number of episodes until the algorithm's policy is α-optimal with probability at least $1-\beta$ satisfies*

$$\mathbb{E}[n_{\mathcal{M}}] \geq \Omega\left(\frac{SAH^2}{\alpha^2} + \frac{SAH}{\alpha\varepsilon}\ln\left(\frac{1}{\beta}\right)\right) .$$

Proof of Theorem 5. The proof follows five main steps: 1) We consider the easier case of JDP with public-initial-state setting, (see definition 5) and a class of hard-MDPs (fig. 2). 2) In lemma 8, we give a sample complexity lower bound of differentially-private best-arm-identification for MAB. 3) In lemma 9, we show that learning the MDP with JDP in the public initial state setting is the same as learning S best-arm-identification MAB instances with differential privacy (DP). 4) We use lemma 8 and lemma 9 to get lemma 10, which gives a lower bound for any RL agent with JDP in the public-initial-state setting. 5) Finally, lemma 11 shows that given that class of hard MDPs, any agent satisfying JDP in the public-initial-state setting also satisfies JDP. Therefore the lower bound in lemma 10 applies, and that concludes the proof of theorem 5. □

C.2 Proofs from Section 6.1

The lower bound result relies on the following adaptation of the coupling lemma from [KV17, Lemma 6.2].

Lemma 23 ([KV17]). *For every pair of distributions $\mathbb{D}_{\theta_0}$ and $\mathbb{D}_{\theta_1}$, every (ε,δ)-differentially private mechanism $M(x_1,\ldots,x_n)$, if $\mathbb{M}_{\theta_0}$ and $\mathbb{M}_{\theta_1}$ are two induced marginal distributions on the output of M evaluated on input dataset $X_1,\ldots,X_n$ sampled i.i.d from $\mathbb{D}_{\theta_0}$ and $\mathbb{D}_{\theta_1}$ respectively, $\varepsilon' = 6\varepsilon n\|\mathbb{D}_{\theta_0} - \mathbb{D}_{\theta_1}\|_{tv}$ and $\delta' = 4\delta e^{\varepsilon'}\|\mathbb{D}_{\theta_0} - \mathbb{D}_{\theta_1}\|_{tv}$, then, for every event E,*

$$\mathbb{M}_{\theta_0}(E) \le e^{\varepsilon'}\mathbb{M}_{\theta_1}(E) + \delta'$$

Lemma (Lemma 7.). *Fix any arm $a \in [k]$. Now consider any pair of MAB instances $\mu,\nu \in [0,1]^k$ both with k arms and time horizon T, such that $\|\mu_a - \nu_a\|_{tv} < \alpha$ and $\|\mu_{a'} - \nu_{a'}\|_{tv} = 0$ for all $a' \neq a$. Let $R \sim B(\mu)^T$ and $Q \sim B(\nu)^T$ be the sequence of T rounds of rewards sampled under μ and ν respectively, and let $\mathcal{M}$ be any ε-DP multi-armed bandit algorithm. Then, for any event E such that under event E arm a is pulled less than t times,*

$$Pr_{\mathcal{M},R}[E] \le e^{6\varepsilon t\alpha} Pr_{\mathcal{M},Q}[E]$$

Proof. We can think of algorithm $M(R)$ as taking as input a tape of t pre-generated rewards for arm a, denote this tape as $R = (r_1,\ldots r_t)$. If $M(R)$ is executed with input tape R, then when $M(R)$ pulls arm a for the j^{th} time the j^{th} entry R_j is revealed and removed from R. If $M(R)$ runs out of the tape R then the reward is drawn from the real distribution of arm a (i.e. P_a). Lastly, if $M(R)$ pulls some arm $a' \neq a$ then the reward is drawn from the real distribution $P_{a'}$.

Note that if R is sampled from the real distribution of arm a i.e $R \sim P_a^t$, then M and $M(R)$ are equivalent. That is, for any event E,

$$\Pr_{M,P}[E] = \Pr_{M(R),P}[E]$$

Under this construction, the event that $M(R)$ pulls arm a less than t times, is the same as the event that $M(R)$ consumes less than t entries of the tape R. By the assumption of the event E under consideration, if $M(R)$ consumes at least t entries of the tape then we can say that event E fails to happen. Therefore, in order to evaluate the event E we only need to initialize $M(R)$ with tapes of size t. Furthermore, we treat the input tape R as the data of M and we claim that $M(R)$ is (ε,δ)-differentially private on R.

Now we apply lemma (23) to bound the probability of E under $M(R_p)$ and $M(R_q)$, where R_p and R_q are the input tapes each generated with t i.i.d samples from distribution P_a and Q_a respectively.

$$\Pr_{M(R_p),P}[E] \le e^{6\varepsilon t\Delta_a}\Pr_{M(R_q),Q}[E]$$

This implies $\Pr_{M,P}[E] \le e^{6\varepsilon t\Delta_a}\Pr_{M,Q}[E]$. □

C.3 Proofs from Section 6.2

Lemma (Lemma 9.). *Let (U,S_1) be a user-state input sequence with initial states from some set $\mathcal{S}_1$. Suppose $\mathcal{M}$ is an RL agent that satisfies ε-JDP in the public initial state setting. Then, for any $s \in \mathcal{S}_1$ the trace $\mathcal{M}_{1,s}(U,S_1)$ is the output of an ε-DP MAB mechanism on input U_s.*

Proof of Lemma 9. Fix (U, S_1), s and U_s as in the statement. Recall that T_s is the number of times state s is in S_1. Observe that $\mathcal{M}_{1,s}(U, S_1)$ has the output type expected from a MAB mechanism on input U_s. Fix an event $E \subseteq \mathcal{A}^{T_s+1}$ on the first action from all episodes starting with s together with the action predicted by the policy at state s. For any $\bar{a} = (\bar{a}^{(1)}, \ldots, \bar{a}^{(T_s)}, \hat{a}) \in E$ we define the event $E_{\bar{a}} \subseteq \mathcal{A}^{H\times T} \times \Pi$ by

$$E_{\bar{a}} = \{(a_h^t)_{h\in[H],t\in[T]}, \pi | a_1^{(t_1)} = \bar{a}^{(1)}, \ldots, a_1^{(t_{T_s})} = \bar{a}^{(T_s)}, \pi(s) = \hat{a}\} \ .$$

where $a_1^{t_i}$ is the first action in the ith episode where state s is the first state. The the event $\bar{E}$ is the union of all events $E_{\bar{a}}$, defined as

$$\bar{E} = \cup_{\bar{a}\in E} E_{\bar{a}} \subseteq \mathcal{A}^{H\times T} \times \Pi$$

Let $\bar{E}_{-t} \subseteq \mathcal{A}^{H\times(T-1)} \times \Pi$ be the collection of outputs from $\bar{E}$ truncated to length $T-1$ and including the output policy. Furthermore, let $\bar{E}_{<t} \subseteq \mathcal{A}^{H\times[t-1]}$ be the collection of outputs from $\bar{E}$ truncated to length $t-1$ and similarly let $\bar{E}_{>t} \subseteq \mathcal{A}^{H\times(T-t)}$ be the sequences truncated to length $T-t$. For any $\vec{a} \in \mathcal{A}^H$ we define the following notation

$$\begin{aligned}
\bar{E}_{-t}^{|\vec{a}} &= \left\{e \in \bar{E}_{-t} : (\vec{a}, e) \in \bar{E}\right\} \\
\bar{E}_{>t}^{|\vec{a}} &= \left\{e \in \bar{E}_{>t} : \exists_{b^{(1)},\ldots,b^{(t-1)}\in\mathcal{A}^H} \left(b^{(1)}, \ldots, b^{(t-1)}, \vec{a}, e\right) \in \bar{E}\right\} \\
\bar{E}_{<t}^{|\vec{a}} &= \left\{e \in \bar{E}_{<t} : \exists_{b^{(1)},\ldots,b^{(T-t)}\in\mathcal{A}^H} \left(b^{(1)}, \ldots, b^{(T-t)}, \vec{a}, e\right) \in \bar{E}\right\} \ .
\end{aligned}$$

For the remaining of the proof, denote by $\mathcal{M}_t(U, S_1)$ the output during episode t, $\mathcal{M}_{<t}(U, S_1)$ all the outputs before episode t, $\mathcal{M}_{>t}(U, S_1)$ all the outputs after episode t, and $\mathcal{M}_{-t}(U, S_1)$ are all the outputs except for the output during episode t and it includes the final output policy. For any $\vec{a} \in \mathcal{A}^H$ It is easy to show that

$$\mathcal{M}_{-t}(U, S_1) \in \bar{E}_{-t}^{|\vec{a}} \text{ if and only if } \mathcal{M}_{>t}(U, S_1) \in \bar{E}_{>t}^{|\vec{a}} \text{ and } \mathcal{M}_{<t}(U, S_1) \in \bar{E}_{<t}^{|\vec{a}} \tag{41}$$

Observe that since $\mathcal{M}$ processes its inputs incrementally we have that

$$\Pr\left[\mathcal{M}_t(U, S_1) = \vec{a} \wedge \mathcal{M}_{<t}(U, S_1) \in \bar{E}_{<t}^{|\vec{a}} \,\middle|\, \mathcal{M}_{>t}(U, S_1) \in \bar{E}_{>t}^{|\vec{a}}\right] = \Pr\left[\mathcal{M}_t(U, S_1) = \vec{a} \wedge \mathcal{M}_{<t}(U, S_1) \in \bar{E}_{<t}^{|\vec{a}}\right] \tag{42}$$

The equation (42) says that conditioning on the output of future events does not affect the probability of the present event.

Now take (U', S_1) to be a t-neighboring user-state sequence and note U'_s is a neighboring sequence of U_s in the sense used in the definition of DP for MAB mechanisms. The next equation says that the output of $\mathcal{M}_t$ on episode t, is not distinguishable on the user-state sequences (U, S_1) and (U', S_1). This is because U and U' match on all episodes before episode t and they share the same initial state on every episode. We have that

$$\Pr[\mathcal{M}_t(U, S_1)] = \Pr[\mathcal{M}_t(U', S_1)] \tag{43}$$

We will use the following simple application of Baye's Rule:

$$\Pr[A|B, C] = \frac{\Pr[A \wedge B|C]}{\Pr[B|C]} \tag{44}$$

Next We want to show that

$$\Pr\left[\mathcal{M}_t(U, S_1) = \vec{a} \,\middle|\, \mathcal{M}_{-t}(U, S_1) \in \bar{E}_{-t}^{|\vec{a}}\right] = \Pr\left[\mathcal{M}_t(U', S_1) = \vec{a} \,\middle|\, \mathcal{M}_{-t}(U', S_1) \in \bar{E}_{-t}^{|\vec{a}}\right] \tag{45}$$

Let fix one $\vec{a}$ for now, then from (41), (42), (43), and (44) we have

$$\Pr\left[\mathcal{M}_t(U, S_1) = \vec{a} \,\middle|\, \mathcal{M}_{-t}(U, S_1) \in \bar{E}_{-t}^{|\vec{a}}\right]$$

$$
\begin{aligned}
&= \Pr\left[\mathcal{M}_t(U,S_1) = \vec{a} \,\middle|\, \mathcal{M}_{>t}(U,S_1) \in \bar{E}_{>t}^{|\vec{a}} \wedge \mathcal{M}_{<t}(U,S_1) \in \bar{E}_{<t}^{|\vec{a}}\right] \quad \text{equation (41)}\\
&= \frac{\Pr\left[\mathcal{M}_t(U,S_1) = \vec{a} \wedge \mathcal{M}_{<t}(U,S_1) \in \bar{E}_{<t}^{|\vec{a}} \,\middle|\, \mathcal{M}_{>t}(U,S_1) \in \bar{E}_{>t}^{|\vec{a}}\right]}{\Pr\left[\mathcal{M}_{<t}(U,S_1) \in \bar{E}_{<t}^{|\vec{a}} \,\middle|\, \mathcal{M}_{>t}(U,S_1) \in \bar{E}_{>t}^{|\vec{a}}\right]} \quad \text{equation (44)}\\
&= \frac{\Pr\left[\mathcal{M}_t(U,S_1) = \vec{a} \wedge \mathcal{M}_{<t}(U,S_1) \in \bar{E}_{<t}^{|\vec{a}}\right]}{\Pr\left[\mathcal{M}_{<t}(U,S_1) \in \bar{E}_{<t}^{|\vec{a}}\right]} \quad \text{equation (42)}\\
&= \frac{\Pr\left[\mathcal{M}_t(U',S_1) = \vec{a} \wedge \mathcal{M}_{<t}(U',S_1) \in \bar{E}_{<t}^{|\vec{a}}\right]}{\Pr\left[\mathcal{M}_{<t}(U',S_1) \in \bar{E}_{<t}^{|\vec{a}}\right]} \quad \text{equation (43)}\\
&= \frac{\Pr\left[\mathcal{M}_t(U',S_1) = \vec{a} \wedge \mathcal{M}_{<t}(U',S_1) \in \bar{E}_{<t}^{|\vec{a}} \,\middle|\, \mathcal{M}_{>t}(U',S_1) \in \bar{E}_{>t}^{|\vec{a}}\right]}{\Pr\left[\mathcal{M}_{<t}(U',S_1) \in \bar{E}_{<t}^{|\vec{a}} \,\middle|\, \mathcal{M}_{>t}(U',S_1) \in \bar{E}_{>t}^{|\vec{a}}\right]} \quad \text{equation (42)}\\
&= \Pr\left[\mathcal{M}_t(U',S_1) = \vec{a} \,\middle|\, \mathcal{M}_{>t}(U',S_1) \in \bar{E}_{>t}^{|\vec{a}} \wedge \mathcal{M}_{<t}(U',S_1) \in \bar{E}_{<t}^{|\vec{a}}\right] \quad \text{equation (44)}\\
&= \Pr\left[\mathcal{M}_t(U',S_1) = \vec{a} \,\middle|\, \mathcal{M}_{-t}(U',S_1) \in \bar{E}_{-t}^{|\vec{a}}\right]
\end{aligned}
$$

Combined with the ε-JDP assumption on $\mathcal{M}$ this implies that

$$
\begin{aligned}
&\Pr\left[\mathcal{M}(U,S_1) \in \bar{E}\right]\\
&= \sum_{\vec{a}\in\mathcal{A}^H} \Pr\left[\mathcal{M}_t(U,S_1) = \vec{a} \wedge \mathcal{M}_{-t}(U,S_1) \in \bar{E}_{-t}^{|\vec{a}}\right]\\
&= \sum_{\vec{a}\in\mathcal{A}^H} \Pr\left[\mathcal{M}_t(U,S_1) = \vec{a} \,\middle|\, \mathcal{M}_{-t}(U,S_1) \in \bar{E}_{-t}^{|\vec{a}}\right] \Pr\left[\mathcal{M}_{-t}(U,S_1) \in \bar{E}_{-t}^{|\vec{a}}\right]\\
&= \sum_{\vec{a}\in\mathcal{A}^H} \Pr\left[\mathcal{M}_t(U',S_1) = \vec{a} \,\middle|\, \mathcal{M}_{-t}(U',S_1) \in \bar{E}_{-t}^{|\vec{a}}\right] \Pr\left[\mathcal{M}_{-t}(U,S_1) \in \bar{E}_{-t}^{|\vec{a}}\right] \qquad \text{Equation (45)}\\
&\le \sum_{\vec{a}\in\mathcal{A}^H} \Pr\left[\mathcal{M}_t(U',S_1) = \vec{a} \,\middle|\, \mathcal{M}_{-t}(U',S_1) \in \bar{E}_{-t}^{|\vec{a}}\right] e^{\varepsilon} \Pr\left[\mathcal{M}_{-t}(U',S_1) \in \bar{E}_{-t}^{|\vec{a}}\right] \qquad \mathcal{M} \text{ satisfies } \varepsilon\text{-JDP}\\
&= e^{\varepsilon} \Pr\left[\mathcal{M}(U',S_1) \in \bar{E}\right] \ .
\end{aligned}
$$

Finally, using the inequality above, and by the construction of $\mathcal{M}_{1,s}$ and $\bar{E}$ we have

$$
\begin{aligned}
\Pr\left[\mathcal{M}_{1,s}(U,S_1) \in E\right] &= \Pr\left[\mathcal{M}(U,S_1) \in \bar{E}\right]\\
&\le e^{\varepsilon} \Pr\left[\mathcal{M}(U',S_1) \in \bar{E}\right]\\
&= e^{\varepsilon} \Pr\left[\mathcal{M}_{1,s}(U',S_1) \in E\right] \ .
\end{aligned}
$$

□

To prove the lower bound we consider the class of MDPs shown in Figure 2. An MDP in this class has state space $\mathcal{S} := [n] \cup \{+,-\}$ and action space $\mathcal{A} := \{0,\dots,m\}$. On each episode, the agent starts on one of the initial states $\{1,\dots,n\}$ chosen uniformly at random. The state labelled 0 is a dummy state which represents the initial transition to any state $s \in \{1,\dots,n\}$ with uniform probability. On each of the initial states the agent has $m+1$ possible actions and transitions can only take it to one of two possible absorbing states $\{+,-\}$. Lastly, if the current state is either one of $\{+,-\}$ then the only possible transition is a self loop, hence the agent is stays in that state until the end of the episode. We assume in these absorbing states the agent can only take a fixed action. Every action which transitions to state + provides reward 1 while actions transitioning to state − provide reward 0. In particular, in each episode the agent either receives reward H or 0.

Such an MDP can be seen as consisting of n parallel MAB problems. Each MAB problem determines the transition probabilities between the initial state $s \in \{1,\dots,n\}$ and the absorbing states $\{+,-\}$. We index the possible MAB problems in each initial state by their optimal arm, which is always one of $\{0,\dots,m\}$. We write $I_s \in \{0,\dots,m\}$ to denote the MAB instance in initial state s, and define the transition probabilities such that $\Pr[+|s,0] = 1/2+\alpha'/2$ and $\Pr[+|s,a'] = 1/2$ for $a' \neq I_s$ for all I_s, and for $I_s \neq 0$ we also have $\Pr[+|s,I_s] = 1/2+\alpha'$. Here α' is a free parameter to be determined later. We succinctly represent an MDP in the class by identifying the optimal action (i.e. arm) in each initial state: $I := (I_1,\dots,I_n)$.

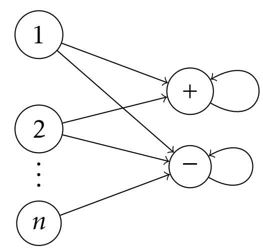

Figure 2: Hard MDP

Proof of Lemma 10. We start by noting that the first term in the lower bound comes from the corresponding lower bound for the non-private episodic RL setting [DB15, Theorem 2], which also holds for our case.

Now let $I = (I_1,\dots,I_n)$ encode an MDP from the class above with $n+2$ states and $m+1$ actions. The optimal policy on this MDP is given by $\pi^*(s) = I_s$ for $s \in [n]$, and we write ρ_I^* to denote the total expected reward of the optimal policy on a single episode.

Define G_s to be the event that policy π produced by algorithm $\mathcal{M}$ finds the optimal arm in state s, that is $\pi(s) = I_s$. We denote by ρ_I^π the total expected reward per episode of this policy. Then, for any episode, the difference $\rho_I^* - \rho_I^\pi$ between total rewards is at least

$$\rho_I^* - \rho_I^\pi \geq H\left(1 - \frac{1}{n}\sum_{s=1}^{n} \mathbb{1}\{G_s\}\right)\alpha'/2 \ .$$

Thus, π cannot by α-optimal unless we have:

$$\begin{aligned} \alpha &\geq H\left(1 - \frac{1}{n}\sum_{s=1}^{n} \mathbb{1}\{G_s\}\right)\alpha'/2 \\ \iff \frac{2\alpha}{H\alpha'} &\geq \left(1 - \frac{1}{n}\sum_{s=1}^{n} \mathbb{1}\{G_s\}\right) \\ \iff \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{G_s\} &\geq \left(1 - \frac{2\alpha}{H\alpha'}\right) := \phi \ . \end{aligned} \tag{46}$$

Here choose $\phi = 6/7$ and set $\alpha' = \frac{14\alpha}{H}$. Equation (46) says that in order to make π an α-optimal policy we must solve at least a ϕ fraction of the MAB instances.

Hence, to get an α-optimal with probability at least $1-\beta$ we require

$$1 - \beta \leq \Pr_I\left[\rho_I^* - \rho_I^\pi \leq \alpha\right] \leq \Pr_I\left[\frac{1}{n}\sum_{s=1}^{n} \mathbb{1}\{G_s\} \geq \phi\right] \ ,$$

and by Markov's inequality we have

$$\Pr_I\left[\frac{1}{n}\sum_{s=1}^{n}\mathbb{1}\{G_s\}\geq\phi\right]\leq\frac{1}{n\phi}\sum_{s=1}^{n}\Pr_I\left[G_s\right]\ .$$

Each $\mathbb{1}\{G_s\}$ is independent from each other be construction of the MDP. Now letting β_s be an upper bound for the fail probability of each $\{G_s\}$, the derivation above implies that $1-\beta\leq\frac{1}{n\phi}\sum_{s=1}^{n}(1-\beta_s)$, or, equivalently, that $\sum_s\beta_s\leq n(1-\phi(1-\beta))$.

Now note that Lemma 9 implies that all interactions between $\mathcal{M}$ and I that start on state s constitute the execution of an (ε,δ)-DP algorithm on the MAB instance at state s. Hence, by Lemma 8 we can only have $\Pr_I[G_s]\geq 1-\beta_s$ for some $\beta_s<1/4$ if the number of episodes starting at s where $\mathcal{M}$ chooses an α'-suboptimal arm satisfies

$$\begin{aligned}\mathbb{E}[n_s]&>\frac{(A-1)}{24\varepsilon\alpha'}\ln\left(\frac{1}{4\beta_s}\right)\mathbb{1}[\beta_s<1/4]\\&=\frac{H(A-1)}{336\varepsilon\alpha}\ln\left(\frac{1}{4\beta_s}\right)\mathbb{1}[\beta_s<1/4]\\&\geq\frac{H(A-1)}{336\varepsilon\alpha}\ln\left(\frac{1}{4\beta_s}\right)\mathbb{1}[\beta_s\leq 1-\phi(1-\beta)]\ ,\end{aligned}$$

where we used that $\phi=6/7$ and $\beta<1/8$ imply $1-\phi(1-\beta)<1/4$, and that each MAB instance has $A-1$ arms which are α'-suboptimal.

Thus, we can find a lower bound $\mathbb{E}[n_{\mathcal{M}}]\geq\sum_s\mathbb{E}[n_s]$ by minimizing the sum of the lower bound on $\mathbb{E}[n_s]$ under the constraint that $\sum_s\beta_s\leq n(1-\phi(1-\beta))$. Here we can apply the argument from [DB15, Lemma D.1] to see that the optimal choice of probabilities is given by $\beta_s=1-\phi(1-\beta)$ for all s. Plugging this choice in the lower bound leads to

$$\mathbb{E}[n_{\mathcal{M}}]\geq\frac{HS(A-1)}{336\varepsilon\alpha}\ln\left(\frac{7}{4+24\beta}\right)\ .$$

□

Lemma (Lemma 11)**.** *Any RL agent $\mathcal{M}$ satisfying ε-JDP also satisfies ε-JDP in the public state setting.*

Proof. Suppose that algorithm $\mathcal{M}$ satisfies ε-JDP. Let (U,S_1) and (U',S_1') be two t-neighboring user-state sequences such that $S_1=S_1'$. Then for all events $E\subseteq\mathcal{A}^{H\times[T-1]}\times\Pi$ we have

$$\Pr[\mathcal{M}_{-t}(U,S_1)\in E]\leq e^{\varepsilon}\Pr[\mathcal{M}_{-t}(U',S_1')\in E]$$

Therefore $\mathcal{M}$ satisfies the condition for ε-JDP in the public state setting as in definition 5. □

C.4 Regret Lower Bound. Proof of theorem 6

In this section we provide the complete lower bound regret analysis of algorithm PUCB from theorem 6. We restate the argument here:

Theorem (Private Regret Lower Bound. Theorem 6)**.** *For any ε JDP-algorithm $\mathcal{M}$ there exist an MDP M with S states A actions over H time steps per episode such that the expected regret after T steps is*

$$\mathbb{E}[\mathrm{Regret}(T)]=\Omega\left(\sqrt{HSAT}+\frac{SAH\log(T)}{\varepsilon}\right)$$

for any $T\geq S^{1.1}$.

Proof. of theorem 6 The first term in the bound comes from the non-private regret due [JOA10], which states that the expected regret is lower-bounded by

$$\Omega\left(\sqrt{HSAT}\right)$$

Next, we analyze the regret lower bound due to privacy. Like section 6.2, we first consider the regret lower bound of any ε-differentially private algorithm under the public-initial-state setting. We also utilize the same construction of hard MDP instances, as depicted in figure 2.

Let $\mathcal{M}$ be an RL agent and (U, S_1) a user-state input sequence with initial state from some set S_1. Let $\mathcal{M}(U, S_1) = (\vec{a}^{(1)}, \ldots, \vec{a}^{(T)}, \pi) \in \mathcal{A}^{H\times T} \times \Pi$ be the collection of all outputs produced by the agent on inputs U and S_1. For every $s \in \mathcal{S}_1$ we write $\mathcal{M}_{1,s}(U, S_1)$ to denote the restriction of the previous trace to contain just the first action from all episodes starting with s together with the action predicted by the policy at states s:

$$\mathcal{M}_{1,s}(U, S_1) := \left(a_1^{(t_{s,1})}, \ldots, a_1^{(t_{s,T_s})}, \pi(s)\right) \ ,$$

where T_s is the number of occurrences of s in S_1 and $t_{s,1}, \ldots, t_{s,T_s}$ are the indices of these occurrences. Furthermore, given $s \in \mathcal{S}_1$ we write $U_s = (u_{t_{s,1}}, \ldots, u_{t_{s,T_s}})$ to denote the set of users whose initial state equals s. Then from lemma 9 we have that the trace $\mathcal{M}_{1,s}(u, s_1)$ is the output of a MAB algorithm satisfying ε-DP.

Thus, we have reduced the problem to learning $n = S - 2$ MAB instances satisfying ε-DP where each MAB instances is visited T_s many times, for all $s \in [n-2]$. Now we can use the result from [SS18] which states that the regret of any ε-DP algorithm for the MAB problem with A arms is lower bounded by $\Omega\left(\frac{A\log(T)}{\varepsilon}\right)$ where T is the total number of arm pulls. By our MDP construction, a state is selected uniformly at random at the beginning of the episode. Then the learner takes a single action and receives a reward in $\{0, H-1\}$, for this reason the regret of each MAB learner is scaled by H in our setting.

Hence, for each initial state $s \in \{1, \ldots, n\}$, the trace $\mathcal{M}_{1,s}(u, s_1)$ produces a sequence of actions satisfying ε-DP and with regret at least $\Omega\left(\frac{A\log(T_s)}{\varepsilon}\right)$. Combining the regret corresponding to each initial state $s \in \{1, \ldots, n\}$, the regret of the agent must be at least

$$\Omega\left(\frac{AH}{\varepsilon}\sum_{s\in\mathcal{S}}\log(T_s)\right)$$

where T_s is a random variable. Next we use the Markov inequality to lower bound the term $\sum_{s\in\mathcal{S}}\log(T_s)$ by

$$\sum_{s\in\mathcal{S}}\log(T_s) = S\mathbb{E}[\log(T_s)] \geq S\log(\tfrac{T}{S})\Pr\left[\log(T_s) \geq \log(\tfrac{T}{S})\right]$$

The event $\log(T_s) \geq \log(\frac{T}{S})$ happens only when $T_s \geq \frac{T}{S}$. Since each $s \in \{1, \ldots, n\}$ is selected with equal probability at the beggining of the episodes, in expectation the number of pulls is $\mathbb{E}[T_s] = \frac{T}{S}$. Thus, each random variable T_s follows a binomial distribution with mean $\frac{T}{S}$ therefore the probability that $T_s \geq \frac{T}{S}$ is $\frac{1}{2}$. Replacing the probability term we get that the total regret of the RL algorithm is lower bounded by:

$$\mathbb{E}[\text{Regret}(T)] = \Omega\left(\frac{AHS\log(\frac{T}{S})}{\varepsilon}\right)$$

□