Differentially Private Episodic Reinforcement Learning with Heavy-tailed Rewards

Yulian Wu [1 2] **Xingyu Zhou** [3] **Sayak Ray Chowdhury** [4] **Di Wang** [1 2]

Abstract

In this paper, we study the problem of (finite horizon tabular) Markov decision processes (MDPs) with heavy-tailed rewards under the constraint of differential privacy (DP). Compared with the previous studies for private reinforcement learning that typically assume rewards are sampled from some bounded or sub-Gaussian distributions to ensure DP, we consider the setting where reward distributions have only finite $(1+v)$-th moments with some $v \in (0, 1]$. By resorting to robust mean estimators for rewards, we first propose two frameworks for heavy-tailed MDPs, i.e., one is for value iteration and another is for policy optimization. Under each framework, we consider both joint differential privacy (JDP) and local differential privacy (LDP) models. Based on our frameworks, we provide regret upper bounds for both JDP and LDP cases and show that the moment of distribution and privacy budget both have significant impacts on regrets. Finally, we establish a lower bound of regret minimization for heavy-tailed MDPs in JDP model by reducing it to the instance-independent lower bound of heavy-tailed multi-armed bandits in DP model. We also show the lower bound for the problem in LDP by adopting some private minimax methods. Our results reveal that there are fundamental differences between the problem of private RL with sub-Gaussian and that with heavy-tailed rewards.

1. Introduction

As a fundamental paradigm in decision-making problems, *reinforcement learning* (RL), where an agent aims to maximize its long-term reward from interacting with an environment, has been widely applied in various fields such as finance(Xu & Wang, 2022), healthcare (Gottesman et al., 2019), transportation (Li et al., 2019) and online recommendations (Zong et al., 2016). However, in these applications, there are some privacy issues as they always involve sensitive data (Pan et al., 2019). And users are less willing to disclose information and are more concerned about how their personal data are used (Das et al., 2021). These privacy concerns in RL require us to design some reinforcement learning algorithms that achieve good performance while can also protect the training environment's privacy.

In order to protect the information of sensitive data, *differential privacy* (DP)(Dwork et al., 2006) has become a de facto technique for private data analysis. Over the past decade, differentially private reinforcement learning (DP-RL) has been extensively studied for various settings, including full information RL (Jain et al., 2012) and partial information RL, e.g., Mishra & Thakurta (2015); Chowdhury & Zhou (2022b;a); Vietri et al. (2020). For private RL problems, it is natural to first consider achieving privacy protection in the standard DP model in Dwork et al. (2006) where we treat each episode as a specific user and the agent collects the raw data of users and aims to achieve privacy protection for history trajectories. However, Shariff & Sheffet (2018) and Dubey (2021) show that the standard differential privacy provably incurs linear regret in contextual bandits (hence in RL as well) and such linear dependency is unavoidable. Therefore, for general DP-RL, it is more reasonable to consider a relaxation of DP, namely *joint differential privacy* (JDP) (Kearns et al., 2014), in which for each user i, knowledge of the other users does not reveal much about user i's data. Moreover, in some situations, users are unwilling to share their data with the agent, so it is also common to consider another DP model, *local differential privacy* (LDP), which perturbs users' data locally before sending them to the central server so that only the data owners can access the original data.

As we mentioned above, sensitive information is contained in the states and rewards of trajectories for RL. Several methods of preserving the privacy of rewards have been proposed in the past few years. However, these methods

[1] Provable Responsible AI and Data Analytics Lab [2]King Abdullah University of Science and Technology, Saudi Arabia [3]Wayne State University, USA [4]Microsoft Research, India. Correspondence to: Di Wang <di.wang@kaust.edu.sa>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

always need to assume the rewards are sampled from light-tailed distributions, such as sub-Gaussian distributions, to ensure DP. However, in a wide variety of real-world systems such as economics (Ibragimov et al., 2015), medicine (Zhao et al., 2020), and market crashes (Schluter & Trede, 2008), rewards are often generated by certain heavy-tailed distributions. For example, it has been shown that RL is quite suitable to recommendation systems (Afsar et al., 2022), and more and more companies such as Google (Chen et al., 2019) are utilizing the power of RL to recommend better items to their customers. However, in such scenarios, the rewards, which correspond to users' feedback such as click, not click, or rating, always contain sensitive information and always follow heavy-tailed or even long-tailed distributions (Park & Tuzhilin, 2008; Celma & Celma, 2010). Therefore, it is necessary to design private algorithms for these RL problems with heavy-tailed rewards.

Motivated by these facts, in this paper, we focus on one fundamental model in RL under DP constraints, i.e., private (finite horizon tabular) Markov decision processes (MDPs) (Sutton & Barto, 2018), with heavy-tailed rewards in that the reward distribution of each state-action pair has only bounded $(1+v)$-th moment for some $v \in (0, 1]$. To the best of our knowledge, we are the first to consider MDPs with heavy-tailed rewards in both joint and local DP models. And our contributions can be summarized as follows.

1. We start from the most commonly used method, i.e., the value-iteration (VI) algorithm. Specifically, we present a general framework, Private-Heavy-UCBVI, for designing private optimistic VI algorithms based on some robust mean estimator for heavy-tailed distributions. To guarantee privacy, we use an adaptive version of the Tree-based mechanism and a new noise allocation method to achieve JDP, and we use the Laplacian mechanism for LDP. Based on this framework, we establish regret upper bounds for the problem in both JDP and LDP models to measure the performance of the framework.

2. Based on our private mechanisms, we then consider policy optimization (PO) based algorithms and propose a framework, namely Private-Heavy-UCBPO. We also analyze the regret bounds for both JDP and LDP models based on this framework by developing some new theoretical techniques as byproducts. It is notable that this is also the first PO-based algorithm for heavy-tailed MDPs, and there is even no previous PO algorithm in the non-private case.

3. Finally, we study the lower bounds of our problem in JDP and LDP. In particular, for the JDP model, two unique challenges arise when applying the standard reduction from RL to MAB for minimax lower bounds. First, it is still open for the lower bound in the MAB case as highlighted in Tao et al. (2022b). We resolve this open problem by deriving the first lower bound for heavy-tailed MAB under the central DP model. Second, additional care is required to translate the lower bound for MAB under DP to the lower bound for RL under JDP. We resolve these challenges by using the notion of JDP with the public initial state as a bridge. For the LDP case, we derive the lower bound by providing some new hard instances of MDPs and using some private minimax methods. All the instances and methods can also be used in other private RL problems.

We summarize our theoretical results in Table 1. Due to space limitations, some additional algorithms and sections, and all proofs and experiments are included in Appendix.

2. Preliminaries

2.1. MDPs with Heavy-tailed Rewards

In a finite horizon Markov decision process (MDP), an agent needs to interact with the environment to make sequential decisions. We can formalize the problem by a tuple $(\mathcal{S}, \mathcal{A}, H, (P_h)_{h=1}^H, (r_h)_{h=1}^H)$, where $\mathcal{S}$ and $\mathcal{A}$ is the state and the action space with cardinality S and A respectively, $H \in \mathbb{N}$ is the episode length, $P_h(s'|s,a)$ is the probability of transitioning to state s' from state s provided action a is taken at step h and $r_h(s,a)$ is the mean of the reward distribution at step h. The actions are chosen following some policy $\pi = (\pi_h)_{h=1}^H$, where each π_h is a mapping from the state space $\mathcal{S}$ into a probability distribution over the action space $\mathcal{A}$ i.e. $\pi_h(a|s) \geq 0$ and $\sum_{a\in\mathcal{A}} \pi_h(a|s) = 1$ for each $s \in \mathcal{S}$. Solving a reinforcement learning task means finding a policy π that maximizes the long-term expected reward starting from every state $s \in \mathcal{S}$ and every step $h \in [H]$, defined as

$$V_h^\pi(s) := \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'}\left(s_{h'}, a_{h'}\right) \mid s_h = s, \pi\right],$$

where the expectation takes over the randomness of the transition kernel $P = (P_h)_{h=1}^H$ and the policy π. We call $V_h^\pi(s)$ as the *value function* of a state s under policy π at step h. Now, defining the *Q-function* of taking action a in state s under policy π at step h as

$$Q_h^\pi(s,a) := \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'}\left(s_{h'}, a_{h'}\right) \mid s_h = s, a_h = a, \pi\right],$$

we obtain $Q_h^\pi(s,a) = r_h(s,a) + \sum_{s'\in\mathcal{S}} V_{h+1}^\pi(s') P_h(s'|s,a)$ and $V_h^\pi(s) = \sum_{a\in\mathcal{A}} Q_h^\pi(s,a)\pi_h(a|s)$.

We call a policy π^* optimal if it maximizes the value function of all states s and steps h simultaneously, and the corresponding optimal value function is denoted by $V_h^*(s) = \max_{\pi\in\Pi} V_h^\pi(s)$ for all $h \in [H]$, where Π is the set of all non-stationary policies. The agent interacts with

Problem	Reward/ Cost	DP	Algorithm	Upper bound	Lower bound
MAB	Heavy-tailed	ϵ-DP	Robust-SE	$\widetilde{O}\left(\left(\frac{A}{\epsilon}\right)^{\frac{v}{1+v}} K^{\frac{1}{1+v}}\right)$ (Tao et al., 2022b)	$\Omega\left(\left(\frac{A}{\epsilon}\right)^{\frac{v}{1+v}} K^{\frac{1}{1+v}}\right)$
		ϵ-LDP	Robust-SE	$\widetilde{O}\left(\left(\frac{A}{\epsilon^2}\right)^{\frac{v}{1+v}} K^{\frac{1}{1+v}}\right)$ (Tao et al., 2022b)	$\Omega\left(\left(\frac{A}{\epsilon^2}\right)^{\frac{v}{1+v}} K^{\frac{1}{1+v}}\right)$ (Tao et al., 2022b)
MDPs	Bounded	ϵ-JDP	UCB-VI	$\widetilde{O}\left(\sqrt{SAH^3T} + S^2AH^3/\epsilon\right)$ (Chowdhury & Zhou, 2021)	$\Omega\left(\sqrt{HSAT} + \frac{SAH\log T}{\epsilon}\right)$ (Vietri et al., 2020)
			UCB-PO	$\widetilde{O}\left(\sqrt{S^2AH^3T} + S^2AH^3/\epsilon\right)$ (Chowdhury & Zhou, 2021)	
		ϵ-LDP	UCB-VI	$\widetilde{O}\left(\sqrt{SAH^3T} + S^2A\sqrt{H^5T}/\epsilon\right)$ (Chowdhury & Zhou, 2021)	$\Omega\left(\frac{H\sqrt{SAK}}{\min\{e^\epsilon-1,1\}}\right)$ (Garcelon et al., 2021)
			UCB-PO	$\widetilde{O}\left(\sqrt{S^2AH^3T} + S^2A\sqrt{H^5T}/\epsilon\right)$ (Chowdhury & Zhou, 2021)	
MDPs	Heavy-tailed	ϵ-JDP	UCB-VI	$\widetilde{O}\left(\sqrt{SAH^3T} + \frac{S^2AH^3}{\epsilon} + \left(\frac{SAH^2}{\epsilon}\right)^{\frac{v}{1+v}} T^{\frac{1}{1+v}}\right)$	$\Omega\left(\left(\frac{SA}{\epsilon}\right)^{\frac{v}{1+v}} K^{\frac{1}{1+v}}\right)$
			UCB-PO	$\widetilde{O}\left(\sqrt{S^2AH^3T} + \frac{S^2AH^3}{\epsilon} + \left(\frac{SAH^2}{\epsilon}\right)^{\frac{v}{1+v}} T^{\frac{1}{1+v}}\right)$	
		ϵ-LDP	UCB-VI	$\widetilde{O}\left(\sqrt{SAH^3T} + \frac{S^2A\sqrt{H^5T}}{\epsilon} + \left(\frac{H^3SA}{\epsilon^2}\right)^{\frac{v}{2(1+v)}} T^{\frac{2+v}{2(1+v)}}\right)$	$\Omega\left(\left(\frac{SA}{\epsilon^2}\right)^{\frac{v}{1+v}} K^{\frac{1}{1+v}}\right)$
			UCB-PO	$\widetilde{O}\left(\sqrt{S^2AH^3T} + \frac{S^2A\sqrt{H^5T}}{\epsilon} + \left(\frac{H^3SA}{\epsilon^2}\right)^{\frac{v}{2(1+v)}} T^{\frac{2+v}{2(1+v)}}\right)$	

Table 1: *Summary of our results and regret comparisons for private RL. All results are in the expected regret form. For the heavy-tailed reward distribution case, we assume the $(1+v)$-th raw moment of each reward distribution is bounded by 1 for some known $v \in (0,1]$. For the bounded reward case, we assume the rewards are in $[0,1]$. Here $T = KH$ is the total number of steps, where K is the total number of episodes and H is the number of steps per episode. S is the number of states and A is the number of actions. $\epsilon \in (0,1]$ is the privacy budget. $\widetilde{O}(\cdot)$ hides poly $\log(S, A, T, 1/\delta)$ factors, where $\delta \in (0,1]$ is the desired confidence level. In MAB problem, $S = 1, H = 1$. We highlight our results in blue color.*

the environment for K episodes to learn the unknown transition probabilities $P_h(s'|s,a)$ and mean rewards $r_h(s,a)$, and thus, in turn, the optimal policy π^*. At each episode k, the agent chooses a policy $\pi^k = (\pi_h^k)_{h=1}^H$ and samples a trajectory $\left\{s_1^k, a_1^k, r_1^k, \ldots, s_H^k, a_H^k, r_H^k, s_{H+1}^k\right\}$ by interacting with the MDP using this policy. Here, at a given step h, s_h^k denotes the state of the MDP, $a_h^k \sim \pi_h^k(\cdot|s_h^k)$ denotes the action taken by the agent, r_h^k denotes the random reward obtained by the agent with the mean value $r_h(s_h^k, a_h^k)$ and $s_{h+1}^k \sim P_h(\cdot|s_h^k, a_h^k)$ denotes the next state.

We consider a heavy-tailed setting in this paper where the reward distribution of each state-action pair (s,a) at step h only has the finite raw moment of order $1+v$ for some $v \in (0,1]$. Concretely, we assume that there is a constant $u > 0$ such that at step h, for each state-action pair (s,a) reward distribution $\mathcal{R}_h(s,a)$, $\mathbb{E}_{X\sim\mathcal{R}_h(s,a)}[|X|^{1+v}] \le u$. In this paper, we assume both v and u are known constants. Since this raw moment of $(1+v)$ for reward is finite, the expectation of reward random variable is also finite, and we denote $|r_h(s,a)| = |\mathbb{E}_{X\sim\mathcal{R}_h(s,a)}[X]| \le \tau$. where τ is a known constant.

We measure the agent's performance by using the cumulative regret accumulated over K episodes, which is defined as

$$Reg(T) := \sum_{k=1}^{K} \left[V_1^*\left(s_1^k\right) - V_1^{\pi^k}\left(s_1^k\right)\right],$$

where $T = KH$ denotes the total number of steps and s_1^k is the initial state.

2.2. DP in Episodic Reinforcement Learning

In the episodic RL setting described above, it is natural to view each episode $k \in [K]$ as a trajectory associated with a specific user. To this end, we let $U_k = (u_1, \ldots, u_K) \in \mathcal{U}^K$ to denote a sequence of K unique users participating in the private RL protocol with an RL agent $\mathcal{M}$, where $\mathcal{U}$ is the set of all users. Each user u_k is identified by the reward and state responses $(r_h^k, s_{h+1}^k)_{h=1}^H$ she/he gives to the action $(a_h^k)_{h=1}^H$ chosen by the agent. We let $\mathcal{M}(U_K) = (a_1^1, \ldots, a_H^K) \in \mathcal{A}^{KH}$ to denote the set of all actions chosen by the agent $\mathcal{M}$ when interacting with the user sequence U_K and let $\mathcal{M}_{-k}(U_K) := \mathcal{M}(U_K) \setminus \left(a_h^k\right)_{h=1}^H$ to denote all the actions chosen by the agent $\mathcal{M}$ excluding those recommended to u_k.

Definition 1 (Joint Differential Privacy (Kearns et al., 2014)). *For any $\epsilon \ge 0$, a mechanism $\mathcal{M} : \mathcal{U}^K \to \mathcal{A}^{KH}$ is ϵ-joint differential privacy (JDP) if for all $k \in [K]$, for all user sequences $U_K, U_K' \in \mathcal{U}^K$ differing only on the k-th user and for all set of actions $\mathcal{A}_{-k} \subset \mathcal{A}^{(K-1)H}$ given to all but the k-th user $\mathbb{P}\left[\mathcal{M}_{-k}(U_K) \in \mathcal{A}_{-k}\right] \le \exp(\varepsilon)\mathbb{P}\left[\mathcal{M}_{-k}(U_K') \in \mathcal{A}_{-k}\right]$.*

Local differential privacy is a more user-friendly model since it requires protecting each user's data $X = (s_h^k, a_h^k, r_h^k, s_{h+1}^k)_{h=1}^H$ before collection. We define local differential privacy for finite-horizon RL as follows:

Definition 2 (Local Differential Privacy (Duchi et al., 2013)). *For any $\epsilon \ge 0$, a mechanism $\mathcal{M}'$ is ϵ-local differentially private (LDP) if for all trajectories $X, X' \in \mathcal{X}$ and for all possible subsets $\mathcal{E}_0 \subset \{\mathcal{M}'(X) \mid X \in \mathcal{X}\}$ we have $\mathbb{P}\left[\mathcal{M}'(X) \in \mathcal{E}_0\right] \le \exp(\varepsilon)\mathbb{P}\left[\mathcal{M}'(X') \in \mathcal{E}_0\right]$.*

We introduce some notations for later analysis. We denote the number of times that the agent has visited the state-action pair (s,a) at step h before episode k as $N_h^k(s,a) := \sum_{k'=1}^{k-1} \mathbb{I}\{s_h^{k'} = s, a_h^{k'} = a\}$. Similarly, $N_h^k(s,a,s') := \sum_{k'=1}^{k-1} \mathbb{I}\{s_h^{k'} = s, a_h^{k'} = a, s_{h+1}^{k'} = s'\}$ denotes the count of going to state s' from s upon playing action a at step h before episode k. Finally, we denotes the total **truncated** rewards obtained by taking action a on state s and at step h before episode k as

$$R_h^k(s,a) = \sum_{k'=1}^{k-1} \mathbb{I}\{s_h^{k'} = s, a_h^{k'} = a, |r_h^{k'}| \le B_{N_h^{k'}(s,a)}\} r_h^{k'}, \tag{1}$$

where $B_{N_h^{k'}(s,a)}$ is truncation threshold which is a non-decreasing function of $N_h^{k'}(s,a)$ and to be set later. In the non-private case, these counters are sufficient to find estimates of the transition kernels $(P_h)_h$ and mean reward functions $(r_h)_h$ to design policy $(\pi_h^k)_h$ for episode k for model-based MDP by using the table lookup model (Agarwal et al., 2019). However, in the private case, the challenge is that these counters depend on users' states and reward responses to suggest further actions, which are considered as sensitive information. Therefore, we must release these counts through some privacy-preserving mechanism namely **PRIVATIZER** on which the learning agent would rely. To this end, we let $\widetilde{N}_h^k(s,a)$, $\widetilde{R}_h^k(s,a)$, and $\widetilde{N}_h^k(s,a,s')$ to denote the **privatized** version of $N_h^k(s,a)$, $R_h^k(s,a)$, and $N_h^k(s,a,s')$, respectively.

Now we make a general assumption on the counts released by the PRIVATIZER, which roughly means that the errors of private counts w.r.t the actual ones are bounded by some terms with high probability. Later on, we will show the specific PRIVATIZERs we use will automatically satisfy such an assumption.

Assumption 1 (Properties of private counts)**.** *For any $\epsilon > 0$ and $\delta \in (0,1]$, there exist functions $E_{\epsilon,\delta,1}, E_{\epsilon,\delta,k,2}, E_{\epsilon,\delta,3} > 0$ such that with probability at least $1-\delta$, uniformly over all (s,a,h,k), the private counts returned by the PRIVATIZER (both LOCAL and CENTRAL) satisfy: (i)$|\widetilde{N}_h^k(s,a) - N_h^k(s,a)| \le E_{\epsilon,\delta,1}$, (ii) $|\widetilde{R}_h^k(s,a) - R_h^k(s,a)| \le E_{\epsilon,\delta,k,2}$, and (iii)$|\widetilde{N}_h^k(s,a,s') - N_h^k(s,a,s')| \le E_{\epsilon,\delta,3}$.*

Based on the above, then we define the private mean empirical rewards and private empirical transition probabilities for all (s,a,h,k) as

$$\begin{aligned} \widetilde{r}_h^k(s,a) &= \frac{\widetilde{R}_h^k(s,a)}{1 \vee (\widetilde{N}_h^k(s,a) + E_{\varepsilon,\delta,1})}, \\ \widetilde{P}_h^k(s'|s,a) &= \frac{\widetilde{N}_h^k(s,a,s')}{1 \vee (\widetilde{N}_h^k(s,a) + E_{\varepsilon,\delta,1})}. \end{aligned} \tag{2}$$

We refer the readers to Table 2 in Appendix for all the above notations for convenience.

3. Private Value-iteration for Heavy-tailed Rewards

In the standard non-private RL setting, a straightforward way to get the optimal policy is to find the optimal value function, which can be determined by a simple iterative algorithm called *value iteration* (VI) that has been shown to converge to the correct V^* (Bellman, 1957; Beutler, 1989). Based on *Upper Confidence Bound* (UCB) philosophy, the non-private UCB-VI method is proposed by Azar et al. (2017), which takes some value-aware concentration results as the exploration bonus. Due to its simplicity, there are several private valued-based RL algorithms for private RL with bounded rewards (Chowdhury & Zhou, 2021; Vietri et al., 2020; Garcelon et al., 2021). However, there is no UCB-VI algorithm for private RL with heavy-tailed rewards. In this section, we will start by designing value iteration-based algorithms for our problem in both JDP and LDP models.

Our general framework, Private-Heavy-UCBVI algorithm, is presented in Algorithm 1. The key idea of our algorithm is that we first establish SAH parallel private counters for each tuple (s,a,h) and S^2AH parallel private counters for each tuple (s,a,h,s'), by using the (adaptive) Tree-based mechanism in JDP or the Laplacian mechanism in LDP to guarantee privacy. Based on these private counts, we design our algorithm by using a private and robust version of UCB at steps $7-9$ where the UCB bonus term is from the concentration results for our private estimators in Lemma 15 or Lemma 17. At steps $9-10$, we compute private versions of the Q-function and the value function by using the optimistic Bellman recursion. Then a greedy policy π^k is obtained by maximizing the private estimated Q-function at step 12. After rolling out the trajectory by acting the policy π^k, we truncate the rewards by an adaptive and non-decreasing truncation threshold $B_{N_h^k(s,a)}$ and translate all non-private statistics into private ones.

3.1. Heavy-tailed Value-iteration in JDP

As we mentioned earlier, different instances of PRIVATIZER correspond to different privacy models. For JDP, we will use CENTRAL-PRIVATIZER, which runs an adaptive version of the binary-tree mechanism for each count $N_h^k(s,a), R_h^k(s,a), N_h^k(s,a,s')$, i.e., it uses $2SAH + S^2AH$ counters in total. In Algorithm 2 of Appendix we provide the details of the mechanism. In total, based on the composition theorem of DP we have the following result.

Lemma 1 (Privacy and Utility Guarantees of CENTRAL-PRIVATIZER)**.** *For any $\epsilon > 0$, the CENTRAL-PRIVATIZER we mentioned above with Laplace noise Lap $\left(\frac{6B_k H \log K}{\epsilon}\right)$ for $\widetilde{R}_h^k(s,a)$ and Lap $\left(\frac{3H \log K}{\epsilon}\right)$ for $\widetilde{N}_h^k(s,a)$ and $\widetilde{N}_h^k(s,a,s')$ is ϵ-DP. Further-*

Algorithm 1 Private-Heavy-UCBVI

Require: Number of episodes K, time horizon H, privacy level $\epsilon > 0$, reward mean bound τ, a PRIVATIZER (Local or Central) and confidence level $\delta \in (0, 1]$

Initialize private counts $\widetilde{R}_h^1(s,a) = 0, \widetilde{N}_h^1(s,a) = 0, \widetilde{N}_h^1(s,a,s') = 0$ for all (s,a,s',h)
Set precision levels $E_{\epsilon,\delta,1}, E_{\epsilon,\delta,k,2}, E_{\epsilon,\delta,3}$ of the PRIVATIZER
for $k = 1, \ldots, K$ **do**
  Initialize private value estimates: $\widetilde{V}_{H+1}^k(s) = 0$
  for $h = H, H-1, \ldots, 1$ **do**
    Compute $\widetilde{r}_h^k(s,a)$ and $\widetilde{P}_h^k(s'|s,a)$ for $\forall(s,a,s')$ as in (2) using the private counts
    Set exploration bonus using Lemma 15: $\beta_h^k(s,a) = \beta_h^{k,r}(s,a) + \beta_h^{k,pv}(s,a) \forall (s,a)$
    Compute: $\forall(s,a)$, $\widetilde{Q}_h^k(s,a) = \max\{-(H-h+1)\tau, \min\{(H-h+1)\tau, \widetilde{r}_h^k(s,a) + \sum_{s'\in\mathcal{S}} \widetilde{V}_{h+1}^k(s')\widetilde{P}_h^k(s'|s,a) + \beta_h^k(s,a)\}\}$
    Compute private value function: $\forall s, \widetilde{V}_h^k(s) = \max_{a\in\mathcal{A}} \widetilde{Q}_h^k(s,a)$
  end for
  Compute policy: $\forall(s,h), \pi_h^k(s) = \arg\max_{a\in\mathcal{A}} \widetilde{Q}_h^k(s,a)$
  Roll out a trajectory $(s_1^k, a_1^k, r_1^k, \ldots, s_{H+1}^k)$ by acting the policy $\pi^k = (\pi_h^k)_{h=1}^H$
  Receive private counts $\widetilde{N}_h^{k+1}(s,a), \widetilde{N}_h^{k+1}(s,a,s')$ and truncated private rewards summation $\widetilde{R}_h^{k+1}(s,a)$
end for

more, for any $\delta \in (0,1]$, it satisfies Assumption 1 with $E_{\epsilon,\delta,1} = \frac{3H\log^{1.5} K \ln\frac{3SAT}{\delta}}{\epsilon}$, $E_{\epsilon,\delta,k,2} = \frac{6B_k H\log^{1.5} K\ln\frac{3SAT}{\delta}}{\epsilon}$, $E_{\epsilon,\delta,3} = \frac{3H\log^{1.5} K\ln\frac{3S^2AT}{\delta}}{\epsilon}$.

Failure of using total ℓ_1 distance of all streams to get the counter error of rewards. It is notable that our PRIVATIZER is quite different from the previous methods in Vietri et al. (2020); Chowdhury & Zhou (2021). If we adopt their methods then we will get larger errors. In detail, if we use their methods, then we need to allocate or add the same noise for each episode. For each counter, it will take the data stream $\sigma_h(s,a) \in [-B_K, B_K]^K$ as input since the B_k is non-decreasing on k, where the j-th entry

$$\sigma_h^j(s,a) := \mathbb{I}\{s_h^j = s, a_h^j = a, |r_h^j| \le B_{N_h^j(s,a)}\} r_h^j \quad (3)$$

denotes whether the pair (s,a) is encountered or not at step h of episode j and if the pair is encountered, we will take the truncated reward. Consider its one adjacent data stream $\sigma_h'(s,a) \in [-B_K, B_K]^K$ which differs from $\sigma_h(s,a)$ only in one entry, then we will have $\|\sigma_h(s,a) - \sigma_h'(s,a)\|_1 \le 2B_K$. Furthermore, since at every episode at most H state-action pairs are encountered, we obtain $\sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \|\sigma_h(s,a) - \sigma_h'(s,a)\|_1 \le 2HB_K$. Then we will get in each episode $k \in [K]$, we need to add noise Lap $\left(\frac{6B_K H\log K}{\epsilon}\right)$ which will make us get a loose error bound for reward count as in Lemma 1 we just need Lap$(\frac{6B_k H\log K}{\epsilon})$ with $k \le K$.

Based on the Billboard lemma in Hsu et al. (2016), the composition of all K episodes satisfies ϵ-JDP if the policy π^k is computed with an ϵ-DP mechanism for all $k \in [K]$.

Theorem 1. *For any $\epsilon > 0$, Algorithm 1 is ϵ-JDP if we use the CENTRAL-PRIVATIZER in Lemma 1.*

In the following, we will show the regret bound of our algorithm in the JDP model.

Theorem 2 (Regret Bound for Private-Heavy-UCBVI in JDP)**.** *For any $\epsilon \in (0,1]$ and $\delta \in (0,1]$ and take $B_n = \left(\frac{\epsilon u n}{H\log^{1.5} K\log(3SAT/\delta)}\right)^{\frac{1}{1+v}}$ in (1). Then if we use the CENTRAL-PRIVATIZER in Lemma 1, with probability $1-\delta$ the regret of Private-Heavy-UCBVI is upper bounded by*

$$\widetilde{O}\left(\sqrt{SAH^3T} + \frac{S^2AH^3}{\epsilon} + u^{\frac{1}{1+v}}\left(\frac{SAH^2}{\epsilon}\right)^{\frac{v}{1+v}} T^{\frac{1}{1+v}}\right).$$

Remark 1. *In the above bound, there are three terms. The first one corresponds to regret due to the uncertainty in transition probabilities. The second term comes from the estimation error of private counts and the third term is caused by the heavy-tailed nature of rewards. Tao et al. (2022b) gives a regret rate of $O\left(\left(\frac{A\log T}{\epsilon}\right)^{\frac{v}{1+v}} T^{\frac{1}{1+v}}\right)$ for private heavy-tailed MAB in the DP model where there are no transition probabilities and counts and $S = H = 1$. Thus, our regret in JDP matches their bound in DP.*

In the non-private case of our problem, Zhuang & Sui (2021) establishes a regret bound of $\widetilde{O}(\sqrt{H^3SAT} + H^2(SA)^{\frac{v}{1+v}}T^{\frac{1}{1+v}} + \sqrt{H^9S^3A^3} + \sqrt{H^{\frac{1+4v}{v}}S^2A^2} + H^{\frac{1+3v}{v}}\sqrt{S^3A^3})$ by proposing the Heavy-Q-Learning with UCB-Bernstein algorithm. Compared with it, we can see the price of privacy is an additional factor of $\frac{1}{\epsilon}$ for private counts and $\left(\frac{1}{\epsilon}\right)^{\frac{v}{1+v}}$ for private estimation of heavy-tailed reward distributions. Besides, for the terms related to heavy-tailed distributions estimation , the dependency on H is better in our bound by a factor of $H^{\frac{2}{1+v}}$ ($H^{\frac{2v}{1+v}}$ v.s. H^2).

Compared with the result of $\widetilde{O}\left(\sqrt{SAH^3T} + \frac{S^2AH^3}{\epsilon}\right)$ given by Chowdhury & Zhou (2021) for private MDPs with bounded rewards in JDP, we can see there is an additional term due to the assumption of the heavy-tailed reward in our problem. This is because Chowdhury & Zhou (2021) assumes the rewards are in $[0,1]$ so that these rewards have the same sensitivity as counts. Hence, the regret corresponding to estimating the mean of heavy-tailed rewards is also

bounded by S^2AH^3/ϵ. Compared with the regret bound of $\widetilde{O}(\sqrt{SAH^2T} + S^2AH^3/\epsilon)$ given by Qiao & Wang (2023) under JDP by adopting DP-UCBVI for bounded rewards, the difference of the first term comes from the type of bonus since Qiao & Wang (2023) adopts Bernstein type but we use Hoeffding type. And the additional third term in our bound is due to the heavy-tailed nature of the rewards.

3.2. Heavy-tailed Value-iteration in LDP

Next, we consider the LDP case. We first introduce its corresponding PRIVATIZER, namely LOCAL-PRIVATIZER. For each episode k, LOCAL-PRIVATIZER releases the private counts by injecting Laplace noise into each data stream. At each episode j, given privacy parameter $\epsilon' > 0$, LOCAL-PRIVATIZER perturbs each entry $\sigma_h^j(s,a)$ of the data stream $\sigma_h(s,a)$ with an independent Laplace noise $\text{Lap}(\frac{1}{\epsilon'})$, i.e., $\widetilde{\sigma}_h^j(s,a) = \sigma_h^j(s,a) + \text{Lap}(\frac{1}{\epsilon'})$, where $\sigma_h^j(s,a)$ is in (3). The private counts for the k-th episode are computed as $\widetilde{N}_h^k(s,a) = \sum_{j=1}^{k-1} \widetilde{\sigma}_h^j(s,a)$. The private counts corresponding to empirical rewards $R_h^k(s,a)$ and state transitions $N_h^k(s,a,s')$ are computed similarly. On the one hand, for $N_h^k(s,a)$ with a fixed episode $k \in [K]$, we run SAH parallel private counters, one for each tuple (s,a,h). Thus, from the DP parallel Lemma (Lemma 3), if we want to guarantee the privacy mechanism satisfying $\epsilon/3$-LDP for all $\widetilde{N}_h^k(s,a)$, we just need to make every $\widetilde{N}_h^k(s,a)$ be $\epsilon/3$-LDP. On the other hand, at each episode at most H state-action pairs are encountered, so according to the composition theorem (Lemma 4), we need to guarantee $\widetilde{\sigma}_h^j(s,a)$ is $\frac{\epsilon}{3H}$-LDP at each step h. Then from the Laplacian mechanism, we use independent noise $\text{Lap}(\frac{3H}{\epsilon})$ in $N_h^k(s,a)$ and $N_h^k(s,a,s')$. Similarly we set independent noise as $\text{Lap}(\frac{6HB_j}{\epsilon})$ to protect the privacy of $R_h^k(s,a)$. Based on the concentration property of Laplacian distributions (Lemma 10), we can get the following error bounds for counts under LOCAL-PRIVATIZER.

Lemma 2 (Privacy and utility guarantees of LOCAL-PRIVATIZER)**.** *For any $\epsilon \in (0,1]$, the LOCAL-PRIVATIZER above is ϵ-LDP. Furthermore, for any $\delta \in (0,1]$, it satisfies Assumption 1 with $E_{\epsilon,\delta,1} = \frac{6H}{\epsilon}\sqrt{K \log \frac{6SAT}{\delta}}$, $E_{\epsilon,\delta,k,2} = \frac{12HB_k}{\epsilon}\sqrt{k \log \frac{6SAT}{\delta}}$, $E_{\epsilon,\delta,3} = \frac{6H}{\epsilon}\sqrt{K \log \frac{6S^2AT}{\delta}}$*

Remark 2. *Compared with the errors of private counts under JDP which depend on $\log K$ in Lemma 1, the above errors under LDP depend on polynomial of K. The reason is that in the LDP model, we add noise to the data of each user, so we add more noise in total than it in the JDP model to guarantee stronger privacy.*

Theorem 3 (Regret Bound for Private-Heavy-UCBVI in LDP)**.** *For any $\epsilon \in (0,1]$ and $\delta \in (0,1]$ and take $B_n = \left(\frac{u\epsilon\sqrt{n}}{H\log(6SAT/\delta)}\right)^{\frac{1}{1+v}}$ in equation (1). Then if we use the LOCAL-PRIVATIZER in Lemma 2, with probability $1-\delta$ the regret $Reg(T)$ of Private-Heavy-UCBVI satisfies*

$$\widetilde{O}\left(\sqrt{SAH^3T} + \frac{S^2A\sqrt{H^5T}}{\epsilon} + u^{\frac{1}{(1+v)}}\left(\frac{H^3SA}{\epsilon^2}\right)^{\frac{v}{2(1+v)}} T^{\frac{2+v}{2(1+v)}}\right).$$

Remark 3. *Compared with the regret bound for JDP, the differences of the second term and the third term in the above bound for LDP come from the fact that the noise magnitude of private count is $\log k$ for JDP while it will be $\sqrt{k}$ for LDP. Moreover, compared with LDP heavy-tailed MAB, our bound cannot recover the optimal rate of $O\left(\left(A\log T/\epsilon^2\right)^{\frac{v}{1+v}} T^{\frac{1}{1+v}}\right)$ in Tao et al. (2022b) with $S = H = 1$. This is because to achieve the optimal rate for MAB, Tao et al. (2022b) proposes a locally private and robust version of the successive elimination algorithm while we use UCB-based method for our LDP case. Since the key ideas of these algorithms are quite different, we cannot adopt successive elimination methods for RL problems directly because of the existence of transition probabilities and states. We may get some regret bounds that match the optimal rate in LDP heavy-tailed MAB by using some variants of the UCB method, and we leave it as an open problem. Compared with the bound of $\widetilde{O}\left(\sqrt{SAH^2T} + S^2A\sqrt{H^5T}/\epsilon\right)$ provided by Qiao & Wang (2023) under LDP based on value iteration method for bounded rewards case, except for the difference in the first term as in JDP, the third term is different due to heavy-tailed rewards and the noise magnitude of private count is $\sqrt{k}$ for LDP.*

4. Private Policy Optimization for Heavy-tailed Rewards

In the previous section, we proposed a value-iteration-based framework for private heavy-tailed MDPs. However, in our algorithm, the value iteration function runs through all possible actions to find the maximum action value, i.e., Algorithm 1 is computationally heavy. In the non-private MDPs with bounded/sub-Gaussian rewards case, it is well known that policy optimization (PO) based algorithms proposed by Pashenkova et al. (1996) are more efficient than the value iteration based ones from the computational perspective. And due to this, in practice, researchers are more willing to use PO-based algorithms. Thus, a natural question is whether we can design some private PO-based algorithm for MDPs with heavy-tailed rewards. In this section, we give an affirmative answer to the question by proposing a policy-optimization-based framework, namely Private-Heavy-UCBPO, for private MDPs with heavy-tailed rewards. See Algorithm 3 for details.

The key idea of Private-Heavy-UCBPO is that we start by choosing a policy π^1 from the uniform distribution and then we iteratively evaluate and improve the policy. In the policy

evaluation stage, we use the UCB framework to compute a Q-function estimation where we use the estimation errors of private empirical reward means and private empirical transition probabilities as the exploration bonus, which is similar to Algorithm 1. Then based on Bellman expectation equation, we compute the corresponding value function at step 11. Next, we roll out a new trajectory by acting the policy and receive the private sum of truncated rewards and private counts from the same PRIVATIZER as in Algorithm 1. Finally, in the policy improvement stage, we update the policy by leveraging a standard mirror-descent step.

Similar to the previous section, we will show the regrets for Algorithm 3 in JDP and LDP models, respectively.

Theorem 4. *Given $\epsilon > 0$, by using the same CENTRAL-PRIVATIZER (LOCAL-PRIVATIZER) as in Lemma 1 (Lemma 2), Algorithm 3 is ϵ-JDP (ϵ-LDP).*

Theorem 5 (Regret bound of Private-Heavy-UCBPO in JDP)**.** *Fix any $\epsilon \in (0,1]$ and $\delta \in (0,1]$ and set $\eta = \sqrt{2\log A/(\tau^2 H^2 K)}$ and take $B_n = \left(\frac{\epsilon u n}{H \log^{1.5} K \log(3SAT/\delta)}\right)^{\frac{1}{1+v}}$ in equation (1). Then, if we use the CENTRAL-PRIVATIZER in Lemma 1, with probability at least $1-\delta$, the cumulative regret of Private-Heavy-UCBPO (Algorithm 3) is upper bounded by*

$$\widetilde{O}\left(\sqrt{S^2AH^3T} + \frac{S^2AH^3}{\epsilon} + u^{\frac{1}{1+v}}\left(\frac{SAH^2}{\epsilon}\right)^{\frac{v}{1+v}} T^{\frac{1}{1+v}}\right).$$

Theorem 6 (Regret bound of Private-Heavy-UCBPO in LDP)**.** *Fix any $\epsilon \in (0,1]$, $\delta \in (0,1]$ and set $\eta = \sqrt{2\log A/(\tau^2 H^2 K)}$. Take $B_n = \left(\frac{u\epsilon\sqrt{n}}{H \log(6SAT/\delta)}\right)^{\frac{1}{1+v}}$ in (1). Then if we use the same LOCAL-PRIVATIZER as in Lemma 2, with probability at least $1-\delta$, the cumulative regret of Private-Heavy-UCBPO (Algorithm 3) is upper bounded by*

$$\widetilde{O}\left(\sqrt{S^2AH^3T} + \frac{S^2A\sqrt{H^5T}}{\epsilon} + u^{\frac{1}{1+v}}\left(\frac{H^3SA}{\epsilon^2}\right)^{\frac{v}{2(1+v)}} T^{\frac{2+v}{2(1+v)}}\right).$$

Compared with the regret bounds of Private-Heavy-UCBVI, there is an additional factor of $\sqrt{S}$ in the leading privacy-independent term in the bounds of Private-Heavy-UCBPO. This follows the same pattern as in private bounded MDPs case (Chowdhury & Zhou, 2021). Actually, there is no previous policy-optimization-based algorithm, even for non-private heavy-tailed MDPs. Therefore, our proof techniques (such as the results of Lemma 21 in Appendix) can be considered as byproducts that can be used to design policy-optimization-based algorithms for non-private heavy-tailed RL problems.

5. Lower Bounds

5.1. Regret Lower Bound under JDP

We first focus on establishing a lower bound on the regret for heavy-tailed MDPs under JDP constraint. In the literature, a common approach for proving lower bounds in RL is via a reduction to MAB. However, in our setting, two challenges will arise when we adopt the above approach. First, the regret of our problem corresponds to the instance-independent regret in MAB. Unfortunately, there is no existing private minimax (instance-independent) regret lower bound for heavy-tailed MAB in the central DP model. In fact, such a lower bound is listed as an open problem in Tao et al. (2022b). Second, in the private case, the reduction from RL to MAB needs additional care. This is due to the difference in privacy definition, i.e., JDP for RL and DP for MAB. In other words, even if one has established a regret lower bound for MAB in the central DP model, it cannot be directly used to establish a lower bound for RL in JDP.

To address the aforementioned challenges, we take the following steps in this section. First, we establish the first private minimax lower bound for heavy-tailed MAB in central DP, hence resolving the open problem in Tao et al. (2022b). The key step behind this result is a *private* version of KL-divergence for the hard instances. Then, to tackle the second challenge, we use the notion called JDP with *public* initial state as a bridge between the classic JDP for RL and DP for MAB, which allows us to build upon our lower bound for MAB under DP to establish the lower bound for RL in JDP. Here we only show our intuition for overcoming the first challenge and our hard instances for heavy-tailed MDPs. See Appendix 5 for the details of overcoming the second challenge.

We now start with the MAB lower bound. In particular, we consider the agent interacts with the environment sequentially for K rounds. The agent is faced with a set of A independent arms $\{1,\dots,A\}$. In each round $k \in K$, the agent selects an arm $a_k \in [A]$ to pull and obtains a reward that is drawn i.i.d. from a fixed but unknown heavy-tailed distribution associated with the chosen arm.

Definition 3 (Differential Privacy (DP) for MAB (Vietri et al., 2020))**.** *For any $\epsilon \geq 0$, a mechanism $\mathcal{M}: \mathcal{U}^K \to \mathcal{A}^K$ is ϵ-DP if for all user sequences $U_K, U_K' \in \mathcal{U}^K$ differing only in a single user and for all events $E \subset \mathcal{A}^K$, we have $\mathbb{P}[\mathcal{M}(U_K) \in E] \leq e^{\varepsilon}\mathbb{P}[\mathcal{M}(U_K') \in E]$.*

Theorem 7 (Instance-independent Lower Bound for DP Heavy-tailed MAB)**.** *There exists a heavy-tailed A-armed bandit instance with the $(1+v)$-th bounded moment of each reward distribution is bounded by 1. Moreover, if K is large enough, for any ϵ-DP algorithm $\mathcal{M}$ with $\epsilon \in (0,1]$, the expected regret must satisfy*

$$Reg_K \geq \Omega\left((A/\epsilon)^{\frac{v}{1+v}} K^{\frac{1}{1+v}}\right).$$

Remark 4. *Tao et al. (2022b) gives an instance-independent upper bound of* $O\left((A\log K/\epsilon)^{\frac{v}{1+v}}K^{\frac{1}{1+v}}\right)$ *for heavy-tailed MAB in central DP by proposing a private and robust version of the successive elimination algorithm. Our above lower bound almost matches the upper bound up to a factor of* $(\log K)^{\frac{v}{1+v}}$*. Thus, the above lower bound is already near optimal.*

Now, inspired by (Vietri et al., 2020), we construct hard instances for MDPs as depicted in Figure 1. Within the class of MDPs, state space is denoted as $\mathcal{S} := [n] \cup \{+, -\}$ and action space is denoted as $\mathcal{A} := [m]$. During each episode, the agent initiates from one of the initial states, randomly selected with equal probability from the set. Now, we construct hard instances for MDPs as depicted in Figure 1. Within the class of MDPs, state space is denoted as $\mathcal{S} := [n] \cup \{+, -\}$ and action space is denoted as $\mathcal{A} := [m]$. During each episode, the agent initiates from one of the initial states, randomly selected with equal probability from the set. For each initial state, the agent faces m potential actions, and transitions can only lead it to either of the two terminal states, $\{+, -\}$.

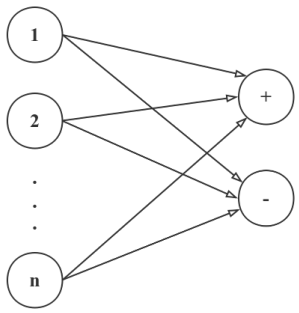

Figure 1: Hard MDP instance for JDP model

Such an instance can be regarded as the composition of n parallel MAB instances. The transition probabilities between the initial state $s \in \{1, \ldots, n\}$ and the absorbing states $\{+, -\}$ are determined by each instance of the MAB. We assign an index to each MAB instance based on its optimal arm (or action) $I_s \in \{1, \ldots, m\}$ within each initial state. Then we transfer the randomness of the Bernoulli reward distribution in the MAB to transition probabilities in the MDP. Therefore, based on the instances for proof of Theorem 7, we define the transition probabilities in Figure 1 such that in the MAB instance where the optimal arm $I_s = 1$, we have $P(+|s, 1) = \frac{5}{10}\gamma^{1+v}$, $P(-|s, 1) = 1 - P(+|s, 1)$ and $P(+|s, a) = \frac{3}{10}\gamma^{1+v}$, $P(-|s, a) = 1 - P(+|s, a)$ for $a \neq 1$. And for the MAB instance with $I_s \neq 1$, we have similar transition probabilities as above except the I_s-th arm, i.e., $P(+|s, I_s) = \frac{7}{10}\gamma^{1+v}$, $P(-|s, I_s) = 1 - P(+|s, I_s)$, where $\gamma > 0$ is a parameter to be determined later. Every action which transits to state $+$ provides reward $1/\gamma$ while actions transitioning to state $-$ provide reward 0.

Based on all the above results and instances, now we have our main theorem.

Theorem 8 (JDP Regret Lower Bound for Heavy-tailed MDPs)**.** *For any ϵ-JDP algorithm $\mathcal{M}$ there exists a heavy-tailed MDP with S states and A actions over $H(= 1)$ time steps per episode such that for any initial state $s \in \mathcal{S}$ the expected regret of $\mathcal{M}$ after K episodes satisfies*

$$Reg(T) \geq \Omega\left((SA/\epsilon)^{\frac{v}{1+v}}K^{\frac{1}{1+v}}\right).$$

Remark 5. *For episodic MDPs in JDP model with bounded rewards, the regret lower bound is* $\Omega\left(\sqrt{HSAK} + SAH\log K/\epsilon\right)$ *(Vietri et al., 2020), where the additional term of* $\Omega\left(SAH\log K/\epsilon\right)$ *is the price of privacy compared with the non-private case. However, in the case of the heavy-tailed reward, the lower bound in the above theorem shows that the price to pay for JDP is a factor of* $\left(\frac{1}{\epsilon}\right)^{\frac{v}{1+v}}$ *compared with the lower bound in the non-private case (Zhuang & Sui, 2021). Our above lower bound also matches the error due to private estimation for heavy-tailed distributions in our upper bounds (the third term of the result in Theorem 2 and the third term of the result in Theorem 5) for JDP model on parameters ϵ, S, A, K. It is also notable that here we only present the lower bound for JDP MDPs with heavy-tailed rewards under the case where there is just one step in each episode. For the multiple-step MDPs with heavy-tailed rewards case, the lower bound is more challenging. We leave the problem of tight lower bound for multiple-step MDPs with heavy-tailed rewards in JDP model as an open problem.*

5.2. Regret Lower Bound under LDP

Unlike the techniques in the JDP case which make a reduction from MDPs to MAB, here we can directly construct an instance of MDPs to get the lower bound.

Inspired by (Garcelon et al., 2021), we construct the following MDP instance for a given number of states S and actions A. Such MDP instance can be represented by a tree whose root is the initial state 0 with A actions that deterministically lead to the next state. Moreover, each node in the tree has A children and there are exactly $S - 2$ states, excluding terminal states. The leaves of the tree, denoted as $\mathcal{L} = x_1, \ldots, x_L$, represent the set of possible transitions from the intermediate states to two terminal states, labeled as $+$ and $-$. At the terminal states, the agent will receive the reward of $1/\gamma$ and 0, respectively, where $\gamma > 0$ is a parameter to be determined later. The tree without nodes $+$ and $-$ is a perfect A-ary tree. We show the instance with $S = 15$ and $A = 3$ in Figure 2 as an example.

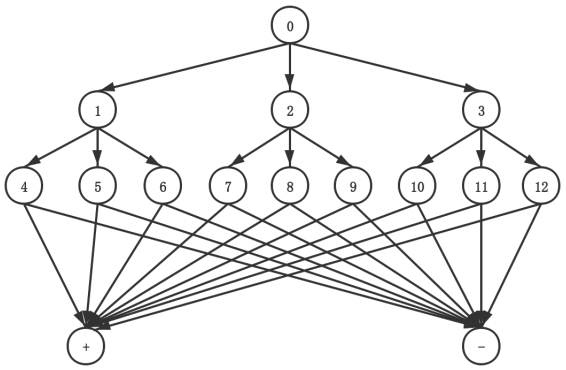

Figure 2: Hard MDP instance for LDP model

Assume $d > 0$ represents the depth of the tree where the depth of the tree with $S - 2$ nodes is $d - 1$ and nodes $+, -$ are located at depth d. Without loss of generality, we assume that all leaves, $x_1, \ldots, x_L$, are positioned at depth $d - 1$, implying that the number of leaves satisfies $L = A^{d-1} \geq (S - 2)/2$. Furthermore, we make the assumption that $H = d + 1$. That means once the agent arrives at $+$ or $-$, it arrives at the end of the episode. Next, we provide the distributions for these leaves.

We assume there exists a unique action a^* and leaf x_{i^*} such that: $P(+|x_{i^*}, a^*) = \gamma^{1+v}$ and $P(-|x_{i^*}, a^*) = 1 - \gamma^{1+v}$. where $\gamma^{1+v} \in (0, \frac{3}{4}]$. Each of the other leaves has a transition probability $P(+|x_i, a) = \frac{1}{2}\gamma^{1+v}$ and $P(-|x_i, a) = 1 - \frac{1}{2}\gamma^{1+v}$. We denote above instance by $\mathbb{P}_{(x_{i^*}, a^*)}$. In order to get the regret lower bound, we also consider another instance $\mathbb{P}_0$ where for all leaf states and any action, the transition probabilities are $P(+|x_i, a) = \frac{1}{2}\gamma^{1+v}$ and $P(-|x_i, a) = 1 - \frac{1}{2}\gamma^{1+v}$. Based on the above instances, we provide a lower bound on the regret for our problem.

Theorem 9 (LDP Regret Lower Bound)**.** *For any ϵ-LDP algorithm $\mathcal{M}$ where $\epsilon \in (0, 1]$, there exists a heavy-tailed MDP instance with $S(\geq 3)$ states, $A(\geq 2)$ actions and one-step heavy-tailed reward per episode such that the expected regret of $\mathcal{M}$ after K episodes is*

$$Reg(T) \geq \Omega\left(\left(SA/\epsilon^2\right)^{\frac{v}{1+v}} K^{\frac{1}{1+v}}\right).$$

Remark 6. *Compared with the lower bound of $\Omega\left(\frac{H\sqrt{SAK}}{\min\{e^\epsilon - 1, 1\}}\right)$ for the bounded case in Garcelon et al. (2021), we can see the price of privacy in the heavy-tailed case is a factor of $\left(\frac{1}{\epsilon^2}\right)^{\frac{v}{1+v}}$. When $v = 1$ and $H = 1$, our lower bound can recover the lower bound of the bounded case. Compared with the optimal instance-independent lower bound of $\Omega\left(\left(A/\epsilon^2\right)^{\frac{v}{1+v}} K^{\frac{1}{1+v}}\right)$ for heavy-tailed MAB in LDP model given by Tao et al. (2022b), our above lower bound with $S = 1$ can also recover the result.*

6. Conclusion and Future Work

In this work, we provided the first study on finite horizon Markov decision processes (MDPs) with heavy-tailed rewards in both joint (JDP) and local differential privacy (LDP) models. We mainly focused on the case where the reward distributions have only finite $(1 + v)$-th raw moment for $v \in (0, 1]$. We first proposed a private and robust version of both UCB-based value-iteration and policy-optimization algorithms. To guarantee privacy, we adopted the adaptive Tree-based mechanism for JDP and the Laplacian mechanism for LDP. Based on the algorithm, we established regret bounds for both JDP and LDP cases. Finally, we established lower bounds for finite-horizon MDPs with heavy-tailed rewards in both JDP and LDP models. Through these results, we also found some differences between our problem and the problem of private MDPs with bounded rewards and some differences between our problem and the problem in the non-private case. All of our ideas, techniques, and frameworks can also potentially be applied to other related private reinforcement learning problems.

There are still some open problems left. First of all, in the whole paper, we assume the parameters u and v are known in advance, but how to deal with the case where u and v are unknown which is a more practical situation in the real world. Secondly, the order of T is $\frac{2+v}{2(1+v)}$ in our regret upper bound under the LDP model, which is larger than the order of $\frac{1}{1+v}$ in our lower bound for the problem under LDP. Thus, can we further close the gap between these two bounds by designing other algorithms? Finally, in this paper, while we mainly focus on the finite-horizon problem, we proposed some private mean estimators and some new hard instances to prove the lower bounds. Can we extend these techniques and ideas to other related problems such as private MDPs with finite diameters with heavy-tailed rewards or some model-free reinforcement learning problems?

Acknowledgments

Yulian Wu and Di Wang are supported in part by BAS/1/1689-01-01, URF/1/4663-01-01, FCC/1/1976-49-01 of King Abdullah University of Science and Technology (KAUST). Di Wang is also supported by the funding of the SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI). Xingyu Zhou is supported in part by NSF CNS-2153220.

References

Afsar, M. M., Crump, T., and Far, B. Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys*, 55(7):1–38, 2022.

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. Rein-

forcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, pp. 10–4, 2019.

Agrawal, S., Juneja, S., and Glynn, P. Optimal δ-correct best-arm selection for heavy-tailed distributions. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory (ALT)*, pp. 61–110, 2020.

Agrawal, S., Juneja, S., and Koolen, W. M. Regret minimization in heavy-tailed bandits. In *Proceedings of 34th Conference on Learning Theory (COLT)*, pp. 26–62, 2021.

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

Azize, A. and Basu, D. When privacy meets partial information: A refined analysis of differentially private bandits, 2022. URL https://arxiv.org/abs/2209.02570.

Barber, R. F. and Duchi, J. C. Privacy and statistical risk: Formalisms and minimax bounds. *arXiv preprint arXiv:1412.4451*, 2014.

Bellman, R. Dynamic programming. *Princeton, USA: Princeton University Press*, 1(2):3, 1957.

Beutler, F. J. Dynamic programming: Deterministic and stochastic models (dimitri p. bertsekas). *SIAM Review*, 31(1):132, 1989.

Brunel, V.-E. and Avella-Medina, M. Propose, test, release: Differentially private estimation with high probability. *arXiv preprint arXiv:2002.08774*, 2020.

Bubeck, S., Cesa-Bianchi, N., and Lugosi, G. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59 (11):7711–7717, 2013.

Celma, Ò. and Celma, Ò. The long tail in recommender systems. *Music Recommendation and Discovery: The Long Tail, Long Fail, and Long Play in the Digital Music Space*, pp. 87–107, 2010.

Chan, T.-H. H., Shi, E., and Song, D. Private and continual release of statistics. *ACM Transactions on Information and System Security (TISSEC)*, 14(3):1–24, 2011.

Chen, M., Beutel, A., Covington, P., Jain, S., Belletti, F., and Chi, E. H. Top-k off-policy correction for a reinforce recommender system. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pp. 456–464, 2019.

Chen, X., Miao, S., and Wang, Y. Differential privacy in personalized pricing with nonparametric demand models. *arXiv preprint arXiv:2109.04615*, 2021.

Chowdhury, S. R. and Zhou, X. Differentially private regret minimization in episodic markov decision processes. *arXiv preprint arXiv:2112.10599*, 2021.

Chowdhury, S. R. and Zhou, X. Distributed differential privacy in multi-armed bandits. *arXiv preprint arXiv:2206.05772*, 2022a.

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

Das, S., Gutzwiller, R. S., Roscoe, R. D., Rajivan, P., Wang, Y., Camp, L. J., and Hoyle, R. Panel: Humans and technology for inclusive privacy and security. *arXiv preprint arXiv:2101.07377*, 2021.

Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pp. 578–598. PMLR, 2021.

Dubey, A. No-regret algorithms for private gaussian process bandit optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2062–2070. PMLR, 2021.

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

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.

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

Garivier, A., Ménard, P., and Stoltz, G. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F., and Celi, L. A. Guidelines for reinforcement learning in healthcare. *Nature medicine*, 25(1):16–18, 2019.

Hsu, J., Huang, Z., Roth, A., Roughgarden, T., and Wu, Z. S. Private matchings and allocations. *SIAM Journal on Computing*, 45(6):1953–1984, 2016.

Hu, L., Ni, S., Xiao, H., and Wang, D. High dimensional differentially private stochastic optimization with heavy-tailed data. In *Proceedings of the 41st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pp. 227–236, 2022.

Ibragimov, M., Ibragimov, R., and Walden, J. *Heavy-tailed distributions and robustness in economics and finance*, volume 214. Springer, 2015.

Jain, P., Kothari, P., and Thakurta, A. Differentially private online learning. In *Conference on Learning Theory*, pp. 24–1. JMLR Workshop and Conference Proceedings, 2012.

Jin, C., Jin, T., Luo, H., Sra, S., and Yu, T. Learning adversarial markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, pp. 4860–4869. PMLR, 2020.

Jin, C., Zhou, K., Han, B., Cheng, J., and Yang, M.-C. Efficient private sco for heavy-tailed data via clipping. *arXiv preprint arXiv:2206.13011*, 2022.

Kamath, G., Singhal, V., and Ullman, J. Private mean estimation of heavy-tailed distributions. In *Proceedings of 33rd Conference on Learning Theory (COLT)*, pp. 2204–2235, 2020.

Kamath, G., Liu, X., and Zhang, H. Improved rates for differentially private stochastic convex optimization with heavy-tailed data. In *International Conference on Machine Learning*, pp. 10633–10660. PMLR, 2022.

Kearns, M., Pai, M., Roth, A., and Ullman, J. Mechanism design in large games: Incentives and privacy. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pp. 403–410, 2014.

Lattimore, T. A scale free algorithm for stochastic bandits with bounded kurtosis. In *Proceedings of the 31st Conference on Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1583–1592, 2017.

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

Lee, K., Yang, H., Lim, S., and Oh, S. Optimal algorithms for stochastic multi-armed bandits with heavy tailed rewards. In *Proceedings of the 34th Conference on Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Lei, Y. M., Miao, S., and Momot, R. Privacy-preserving personalized revenue management. *HEC Paris Research Paper No. MOSI-2020-1391*, 2020.

Li, F., Zhou, X., and Ji, B. Differentially private linear bandits with partial distributed feedback. In *2022 20th International Symposium on Modeling and Optimization in Mobile, Ad hoc, and Wireless Networks (WiOpt)*, pp. 41–48. IEEE, 2022.

Li, F., Zhou, X., and Ji, B. (private) kernelized bandits with distributed biased feedback. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 7(1): 1–47, 2023.

Li, M., Qin, Z., Jiao, Y., Yang, Y., Wang, J., Wang, C., Wu, G., and Ye, J. Efficient ridesharing order dispatching with mean field multi-agent reinforcement learning. In *The world wide web conference*, pp. 983–994, 2019.

Li, X. and Sun, Q. Variance-aware robust reinforcement learning with linear function approximation with heavy-tailed rewards. *arXiv preprint arXiv:2303.05606*, 2023.

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

Liu, X., Kong, W., Kakade, S., and Oh, S. Robust and differentially private mean estimation. *arXiv preprint arXiv:2102.09159*, 2021.

Mishra, N. and Thakurta, A. (nearly) optimal differentially private stochastic multi-arm bandits. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pp. 592–601, 2015.

Orabona, F. A modern introduction to online learning. *arxiv.org/abs/1912.13213*, 2023.

Osband, I., Russo, D., and Van Roy, B. (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26, 2013.

Pan, X., Wang, W., Zhang, X., Li, B., Yi, J., and Song, D. How you act tells a lot: Privacy-leaking attack on deep reinforcement learning. In *AAMAS*, pp. 368–376, 2019.

Park, Y.-J. and Tuzhilin, A. The long tail of recommender systems and how to leverage it. In *Proceedings of the 2008 ACM conference on Recommender systems*, pp. 11–18, 2008.

Pashenkova, E., Rish, I., and Dechter, R. Value iteration and policy iteration algorithms for markov decision problem. In *AAAI'96: Workshop on Structural Issues in Planning and Temporal Reasoning*. Citeseer, 1996.

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

Ren, W., Zhou, X., Liu, J., and Shroff, N. B. Multi-armed bandits with local differential privacy. *arXiv preprint arXiv:2007.03121*, 2020.

Sajed, T. and Sheffet, O. An optimal private stochastic-mab algorithm based on optimal private stopping rule. In *International Conference on Machine Learning*, pp. 5579–5588. PMLR, 2019.

Schluter, C. and Trede, M. Identifying multiple outliers in heavy-tailed distributions with an application to market crashes. *Journal of Empirical Finance*, 15(4):700–713, 2008.

Shani, L., Efroni, Y., Rosenberg, A., and Mannor, S. Optimistic policy optimization with bandit feedback. In *International Conference on Machine Learning*, pp. 8604–8613. PMLR, 2020.

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

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Tao, Y., Wu, Y., Cheng, X., and Wang, D. Private stochastic convex optimization and sparse learning with heavy-tailed data revisited. International Joint Conferences on Artificial Intelligence Organization, 2022a.

Tao, Y., Wu, Y., Zhao, P., and Wang, D. Optimal rates of (locally) differentially private heavy-tailed multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 1546–1574. PMLR, 2022b.

Vakili, S., Liu, K., and Zhao, Q. Deterministic sequencing of exploration and exploitation for multi-armed bandit problems. *IEEE Journal of Selected Topics in Signal Processing*, 7(5):759–767, 2013.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

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

Wang, D. and Xu, J. Differentially private ℓ_1-norm linear regression with heavy-tailed data. *arXiv preprint arXiv:2201.03204*, 2022.

Wang, D., Gaboardi, M., and Xu, J. Empirical risk minimization in non-interactive local differential privacy revisited. In *Proceedings of the 32nd Conference on Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Wang, D., Xiao, H., Devadas, S., and Xu, J. On differentially private stochastic convex optimization with heavy-tailed data. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 10081–10091, 2020.

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

Xu, J. and Wang, Y.-X. Towards agnostic feature-based dynamic pricing: Linear policies vs linear valuation with unknown noise. In *International Conference on Artificial Intelligence and Statistics*, pp. 9643–9662. PMLR, 2022.

Yu, X., Shao, H., Lyu, M. R., and King, I. Pure exploration of multi-armed bandits with heavy-tailed payoffs. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 937–946, 2018.

Zhao, W., Khosa, S. K., Ahmad, Z., Aslam, M., and Afify, A. Z. Type-i heavy tailed family with applications in medicine, engineering and insurance. *PloS one*, 15(8): e0237462, 2020.

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

Zhou, X. Differentially private reinforcement learning with linear function approximation. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 6 (1):1–27, 2022.

Zhou, X. and Tan, J. Local differential privacy for bayesian optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11152–11159, 2021.

Zhuang, V. and Sui, Y. No-regret reinforcement learning with heavy-tailed rewards. In *International Conference on Artificial Intelligence and Statistics*, pp. 3385–3393. PMLR, 2021.

Zong, S., Ni, H., Sung, K., Ke, N. R., Wen, Z., and Kveton, B. Cascading bandits for large-scale recommendation problems. *arXiv preprint arXiv:1603.05359*, 2016.

Appendix

A. Related work and challenges

Besides the work we mentioned above, there are other numerous previous studies on either private RL/bandits with bounded/sub-Gaussian rewards (Sajed & Sheffet, 2019; Liao et al., 2021; Chen et al., 2021; Lei et al., 2020; Zheng et al., 2020; Zhou, 2022; Ren et al., 2020; Zhou & Tan, 2021; Li et al., 2022; 2023) or non-private RL/bandits with heavy-tailed rewards (Zhuang & Sui, 2021; Bubeck et al., 2013; Lee et al., 2020; Yu et al., 2018; Lattimore, 2017; Agrawal et al., 2021; Vakili et al., 2013; Agrawal et al., 2020; Li & Sun, 2023). In the following, we only discuss the work that is most related to ours.

For the studies of non-private RL with heavy-tailed rewards, Bubeck et al. (2013) first considers the finite-armed bandit problem in which the reward distributions have only finite $(1+v)$-th moments for some $v \in (0, 1]$. It develops a robust UCB algorithm by leveraging several mean estimators for heavy-tailed distributions, such as the truncated mean estimators and the median of mean estimator. Leveraging techniques from these robust mean estimators, Zhuang & Sui (2021) considers the heavy-tailed rewards in the problem of undiscounted reinforcement learning and proposes the method of Heavy-UCRL2 and Heavy-Q-learning for model-based and model-free settings respectively. It also generalizes the algorithms to deep reinforcement learning and presents Heavy-DQN as an example. Motivated by these, we use the idea of a truncated mean estimator as the backbone in our frameworks to deal with heavy-tailed rewards in our problem. However, as we mentioned earlier there are several additional challenges by injecting additional noises. Moreover, in this paper, we also propose policy-optimization-based algorithms which have not been studied before for heavy-tailed rewards even in the non-private case.

For private RL/bandits with heavy-tailed rewards, to the best of our knowledge, Tao et al. (2022b) is the only one which investigates MAB with heavy-tailed rewards in both central and local DP models. It proposes robust versions of successive elimination (SE) algorithms for the problem in central DP and local DP models and establishes (near) optimal rates. However, for the problem we studied, it is unsuitable for the central DP model since several recent works Shariff & Sheffet (2018); Dubey (2021) show that the standard DP model is irreconcilable with sub-linear regret for contextual bandits. Thus, we consider a relaxation of central DP, i.e., the joint differential privacy. Besides, we cannot use an arm elimination algorithm for our problem because of the existence of states. Since our private heavy-tailed MDPs with $H = 1$ are just the private heavy-tailed contextual bandit problem and that with $H = 1, S = 1$ is a private heavy-tailed MAB problem, our problem could be considered as a more general case. There are also some studies on private MDPs recently. However, all of them only consider the bounded reward case and cannot be extended to the heavy-tailed one. For example, Chowdhury & Zhou (2021) studies DP episodic MDPs with bounded rewards and proposes policy optimization and value iteration frameworks, and presents the regret upper bounds for these frameworks. Motivated by these frameworks, we propose Private-Heavy-UCBVI and Private-Heavy-UCBPO for our problem for both JDP and LDP. However, here we cannot directly use their Tree-based mechanism for JDP and the allocation methods for the privacy budget since the heavy-tailed rewards are now unbounded. Besides, Garcelon et al. (2021) establishes the lower bound of regret minimization for MDPs with bounded rewards in LDP model by constructing some MDP examples satisfying bounded rewards. However, we cannot directly use the MDP instance to get the regret lower bound since our rewards are heavy-tailed with bounded $(1+v)$-th moment. See Table 1 for a detailed comparison.

Robust and differentially private estimation has drawn much attention in recent years. Barber & Duchi (2014) provided the first study on private mean estimation for distributions with the bounded moment, which is extended by Kamath et al. (2020); Liu et al. (2021); Brunel & Avella-Medina (2020) recently. However, all of them need to assume the underlying distribution has the second-order moment, while in this paper we only need to assume the reward distributions have the $(1+v)$-th moment for some $v \in (0, 1]$. Moreover, all of these works only focus on the central DP model and offline setting and it is unknown whether they could be extended to the stream setting. Thus, our problem is more general. Besides the mean estimation problem, recently Wang et al. (2020); Kamath et al. (2022); Jin et al. (2022); Hu et al. (2022) study differentially private stochastic convex optimization with heavy-tailed data. However, all of them need to assume the distribution of the gradient has the second-order moment and cannot be used in the stream setting. Wang & Xu (2022) studies private ℓ_1-regression where the covariate x has bounded $(1+v)$-th moment. However, its method cannot be generalized to other problems. Tao et al. (2022a) considers the differentially private stochastic convex optimization with heavy-tailed data where the distribution of each coordinate of the gradient has bounded $(1+v)$-th moment. However, its method cannot be used in the online stream setting.

Challenges. Compared with the previous related work, there are additional challenges for private MDPs with heavy-tailed rewards, which are mainly from the following three aspects. **First**, in the non-private case, Zhuang & Sui (2021) proposes methods that truncate each reward via a certain threshold. However, due to the additional privacy constraint in our problem, we need to be more careful in choosing the threshold, which is related to three kinds of error: an error from truncating rewards, an error due to using a finite number of truncated rewards to estimate heavy-tailed reward distributions and an error due to the noise for ensuring privacy. In detail, we need to first get a general upper bound and then based on the trade-off among these three terms of errors we can determine the optimal threshold. However, getting such a trade-off in general is still more difficult than in the non-private case, as here we will have more biased estimators due to the noise we added. For example, we need to add noise to the count of each state-action pair and this will make empirical transition probabilities be biased. **Secondly**, in private MDPs with bounded rewards case, Chowdhury & Zhou (2021) assumes that all rewards are bounded by 1 and it determines and allocates the noise added for each step in an episode by using the total ℓ_1 distance of all data streams. However, we cannot adopt the same strategy as the heavy-tailed rewards now are unbounded. The problem is still challenging even if we just adopt similar reward-truncating based methods in some previous related work (Zhuang & Sui, 2021; Tao et al., 2022b), as the thresholds of truncated rewards are increasing, which makes these truncated rewards not uniformly bounded. **Thirdly**, in the stateless case, i.e., private heavy-tailed MAB, Tao et al. (2022b) achieves the optimal regret rates for both central DP and local DP models via successive elimination algorithms. However, we cannot use similar methods in the RL setting since the states and transition probabilities can also affect policy.

B. Notations and Technical Lemmas

Table 2: List of Notations

Notations	Descriptions
$\mathcal{S}$	state space with cardinality S
$\mathcal{A}$	action space with cardinality A
Π	set of all non-stationary policies
K	number of episodes
H	episode length
P_h	transition kernel at step h
r_h	mean of the reward distribution at step h
π_h	policy at step h
s_h^k	state at episode k and step h
a_h^k	the action taken by agent at episode k and step h
r_h^k	random reward at episode k and step h
$V_h^\pi(s)$	value function of a state s under policy π at step h
$V_h^*(s)$	optimal value function
$Q_h^\pi(s,a)$	Q-function of taking action a in state s under policy π at step h
π^*	optimal policy
$\mathcal{R}_h(s,a)$	reward distribution for state-action pair (s,a) at step h
T	$T = KH$ is the total number of steps
$N_h^k(s,a)$	count of visiting state-action pair (s,a) at step h before episode k
$N_h^k(s,a,s')$	count of going to state s' from s upon playing action a at step h before episode k
$R_h^k(s,a)$	sum of truncated rewards obtained by taking action a on state s at step h before episode k
ϵ	privacy budget
$\widetilde{N}_h^k(s,a)$	the privatized version of $N_h^k(s,a)$
$\widetilde{N}_h^k(s,a,s')$	the privatized version of $N_h^k(s,a,s')$
$\widetilde{R}_h^k(s,a)$	the privatized version of $R_h^k(s,a)$
$\widetilde{r}_h^k(s,a)$	the private empirical mean estimation of truncated rewards for state-action pair (s,a) at step h before episode k
$\widetilde{P}_h^k(s'\|s,a)$	the private empirical transition probability
$[n]$	the set of $\{1,2,\ldots,n\}$
$a \vee b$	the maximal value between a and b

Lemma 3 (Parallel Composition). *Suppose there are k number of ϵ-differentially private mechanisms $\{\mathcal{M}_i\}_{i=1}^k$ and k disjoint datasets denoted by $\{D_i\}_{i=1}^k$. Then the algorithm, which applies each $\mathcal{M}_i$ on the corresponding D_i, preserves ϵ-DP in total.*

Lemma 4 (Composition Theorem (Dwork et al., 2014)). *Let $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_h$ be a sequence of randomized algorithms, where $\mathcal{M}_1 : \mathcal{X}^n \to \mathcal{Y}_1$, $\mathcal{M}_2 : \mathcal{Y}_1 \times \mathcal{X}^n \to \mathcal{Y}_2, \ldots,$ $\mathcal{M}_h : \mathcal{Y}_1 \times \mathcal{Y}_2 \times \mathcal{Y}_{h-1} \times \mathcal{X}^n \to \mathcal{Y}_h$. Suppose for every $i \in [h]$ and $y_1 \in \mathcal{Y}_1, y_2 \in \mathcal{Y}_2, \ldots, y_h \in \mathcal{Y}_h$, we have $\mathcal{M}_i(y_1, \ldots, y_{i-1}, \cdot) : \mathcal{X}^n \to \mathcal{Y}_i$ is ϵ_i-DP. Then the algorithm $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}_1 \times \mathcal{Y}_2 \times \ldots \mathcal{Y}_h$ that runs the algorithm $\mathcal{M}_i$ sequentially is ϵ-DP for $\epsilon = \sum_{i=1}^h \epsilon_i$.*

Lemma 5 (Laplace Mechanism). *Given a dataset $D \in \mathcal{X}^n$ and a function $q : \mathcal{X}^n \to \mathbb{R}^d$, the Laplace Mechanism is defined as $q(D) + (Y_1, Y_2, \cdots, Y_d)$, where each Y_i is i.i.d. sampled from the Laplace Distribution $\mathrm{Lap}(\frac{\Delta_1(q)}{\epsilon})$, where $\Delta_1(q)$ is the ℓ_1-sensitivity of the function q,* i.e., *$\Delta_1(q) = \sup_{D \sim D'} ||q(D) - q(D')||_1$. The density of the Laplace distribution with parameter λ is $\mathrm{Lap}(\lambda)(x) = \frac{1}{2\lambda} \exp(-\frac{x}{\lambda})$. Laplace mechanism preserves ϵ-DP.*

Lemma 6 (Hoeffding's inequality). *Let $X_1, \ldots, X_n$ be independent random variables such that $a_i \le X_i \le b_i$ almost surely. Consider the sum of these random variables, $S_n = X_1 + \cdots + X_n$, then for all $t > 0$, we have*

$$\mathbb{P}\left(|S_n - \mathbb{E}\left[S_n\right]| \ge t\right) \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Lemma 7 (Bernstein's Inequality (Vershynin, 2018)). *Let $X_1, \cdots X_n$ be n independent zero-mean random variables. Suppose $|X_i| \le M$ and $\mathbb{E}[X_i^2] \le s$ for all $i \in [n]$. Then for any $t > 0$, we have*

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n X_i \ge t\right\} \le \exp\left(-\frac{\frac{1}{2}t^2 n}{s + \frac{1}{3}Mt}\right)$$

Lemma 8 (Lemma 7 in (Tao et al., 2022b)). *Given a random variable X with $\mathbb{E}[|X|^{1+v}] \le u$ for some $v \in (0, 1]$, for any $B > 0$ we have*

$$\mathbb{E}\left[X \cdot \mathbb{I}_{|X|>B}\right] \le \frac{u}{B^v}.$$

Lemma 9 (Hölder's inequality). *For $p, q \in (1, \infty)$ with $1/p + 1/q = 1$,*

$$\sum_{k=1}^n |x_k y_k| \le \left(\sum_{k=1}^n |x_k|^p\right)^{\frac{1}{p}} \left(\sum_{k=1}^n |y_k|^q\right)^{\frac{1}{q}}.$$

Lemma 10 (Concentration of Laplace Variables (Wang et al., 2018)). *If $X_1, \cdots X_n \sim \mathrm{Lap}(s/\epsilon)$, then with probability at least $1 - \beta$, we have*

$$\left|\frac{1}{n}\sum_{i=1}^n X_i\right| \le \frac{2s}{\epsilon\sqrt{n}}\sqrt{\log\frac{2}{\beta}}.$$

Lemma 11 (Markov Inequality). *If X is a non-negative random variable and $a > 0$, then the probability that X is at least a is at most the expectation of X divided by a:*

$$P(X \ge a) \le \frac{\mathbb{E}(X)}{a}$$

Lemma 12 (Lemma 1 in (Garivier et al., 2019)). *Consider a measurable space $(\Omega, \mathcal{F})$ equipped with two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$. For any $\mathcal{F}$-measurable function $Z : \Omega \to [0, 1]$, we have*

$$\mathrm{KL}\left(\mathbb{P}_1, \mathbb{P}_2\right) \ge \mathrm{kl}\left(\mathbb{E}_1[Z], \mathbb{E}_2[Z]\right)$$

where $\mathbb{E}_1$ and $\mathbb{E}_2$ are the expectations under $\mathbb{P}_1$ and $\mathbb{P}_2$ respectively.

C. Algorithms and Proofs of Section 3

C.1. Details of the Tree-based Mechanism

Note that our CENTRAL-PRIVATIZER is established by an adaptive version of the Tree-based mechanism, we first give details of this mechanism.

Algorithm 2 (Adaptive) Tree-based Mechanism

Require: time horizon K, privacy budget ϵ, a stream σ.
Ensure: A private version $\widehat{S}(k)$ for $S(k) = \sum_{i=1}^{k} \sigma(i)$ at each $k \in [K]$
Initialize each p-sum α_i and noisy p-sum $\widehat{\alpha}_i$ to 0.
$\epsilon' \leftarrow \epsilon / \log K$.
for $k = 1, \dots, K$ **do**
  Express k in binary form: $k = \sum_j \mathrm{Bin}_j(k) \cdot 2^j$.
  $i \leftarrow \min\{j : \mathrm{Bin}_j(k) \neq 0\}$.
  $\alpha_i \leftarrow \sum_{j<i} \alpha_j + \sigma(k)$.
  for $j = 0, \dots, i-1$ **do**
    $\alpha_j \leftarrow 0$, $\widehat{\alpha}_j \leftarrow 0$.
  end for
  $\widehat{\alpha}_i \leftarrow \alpha_i + \mathrm{Lap}(2B_k/\epsilon')$.
  Return $\hat{S}(k) \leftarrow \sum_{j:\mathrm{Bin}_j(k)=1} \widehat{\alpha}_j$.
end for

Definition 4 (p-sum). *A p-sum is a partial sum of consecutive data items. Let $1 \le i \le j$. For a data stream σ of length K, we use $\sigma(k)$ to denote the data item at time $k \in [K]$ and $\sum[i, j] \triangleq \sum_{k=i}^{j} \sigma(k)$ to denote a partial sum involving data items i through j. We use the notation α_i^k to denote the p-sum $\sum[k - 2^i + 1, k]$.*

Lemma 13 ((Adaptive) Tree-based Mechanism (Tao et al., 2022b)). *Given a stream σ such that $\sigma(k) \in [-B_k, B_k]$ for $\forall k \in [K]$, where B_k is non-decreasing with k, we want to privately and continually release the sum of the stream $S(k) \triangleq \sum_{i=1}^{k} \sigma(i)$ for each $k \in [K]$. The (adaptive) tree-based Mechanism (Algorithm 2) outputs an estimation $\widehat{S}(k)$ for $S(k)$ at each $k \in [K]$ such that $\widehat{S}(k)$ preserves ϵ-differential privacy and guarantees the following noise bound with probability at least $1 - \delta$ for any $\delta > 0$,*

$$\left|\widehat{S}(k) - S(k)\right| \le \frac{2B_k}{\epsilon} \log^{1.5} K \cdot \ln \frac{1}{\delta}. \tag{4}$$

It is notable that when each $B_k = m$ for some m then Algorithm 2 is just the standard m-bounded tree-based mechanism in (Chan et al., 2011). Thus, Algorithm 2 is a generalization compared with the standard one. To ensure ϵ'-DP for some given ϵ', for counter $R_h^k(s, a)$, we use adaptive tree-based mechanism and add $\mathrm{Lap}(\frac{2B_k}{\epsilon'})$ for some non-decreasing B_k to every p-sum before releasing them. Then we get private count $\widetilde{R}_h^k(s, a)$. For $N_h^k(s, a)$ and $N_h^k(s, a, s')$, we use 1-bounded binary-tree mechanisms with Laplace noise $\mathrm{Lap}(\frac{1}{\epsilon'})$ to release the respective private counts $\widetilde{N}_h^k(s, a)$ and $\widetilde{N}_h^k(s, a, s')$.

C.2. Proof of Lemma 1

Proof of Lemma 1. Therefore, we first focus on $E_{\epsilon,\delta,2}$ which is the error bound between the private count for the sum of rewards $\widetilde{R}_h^k(s, a)$ and the non-private count $R_h^k(s, a)$.

We start with the privacy guarantee of the CENTRAL-PRIVATIZER. First, note that there are SAH many counters for the sum of rewards $R_h^k(s, a)$, and each counter is a K-bounded adaptive tree-based mechanism. For a fixed tuple $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, the private count $\widetilde{R}_h^k(s, a)$ is the sum of at most $\log K$ noisy p-sums, where each p-sum is corrupted by an independent Laplace noise $\mathrm{Lap}\left(\frac{6B_k H \log K}{\epsilon}\right)$. By Lemma 13, the private counts $\{\widetilde{R}_h^k(s, a)\}_{k\in[K]}$ satisfy $\frac{\epsilon}{3H}$-DP. We leverage the fact that the total change across all counters a user can have scales with the length of the episode H to get the composition of the SAH $\frac{\epsilon}{3H}$-DP counters for $\widetilde{R}(\cdot)$ satisfy $\frac{\epsilon}{3}$-DP.

Using similar arguments of Lemma 5.1 in (Chowdhury & Zhou, 2021), one can show that composition of the counters for $\widetilde{N}_h^k(s, a)$ and $\widetilde{N}_h^k(s, a, s')$ satisfy $\frac{\epsilon}{3}$-DP respectively. Finally, employing the composition property of DP in (Dwork et al., 2014), we obtain that the CENTRAL-PRIVATIZER is ϵ-DP.

Now we focus on the utility of CENTRAL-PRIVATIZER. First, for a fixed tuple $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, we consider the private counts $\widetilde{R}_h^k(s, a)$ corresponding to the sum of rewards $R_h^k(s, a)$. Note that, at each episode $k \in [K]$, the noise bound $|\widetilde{R}_h^k(s, a) - R_h^k(s, a)|$ is the sum of at most $\log K$ random variables drawn from the Laplace distribution. From Lemma 13,

we have

$$\mathbb{P}\left[\left|\widetilde{R}_h^k(s,a) - R_h^k(s,a)\right| \leq \frac{6B_k H}{\epsilon}\log^{1.5} K\ln(3SAT/\delta)\right] \geq 1-\delta/(3SAT)$$

By a union bound argument, we can obtain

$$\mathbb{P}\left[\forall(s,a,k,h), \left|\widetilde{R}_h^k(s,a) - R_h^k(s,a)\right| \leq \frac{6B_k H}{\epsilon}\log^{1.5} K\ln(3SAT/\delta)\right] \geq 1-\delta/3$$

By using the same arguments, we can get the values of error bounds $E_{\epsilon,\delta,1}$ and $E_{\epsilon,\delta,3}$.

$$\mathbb{P}\left[\forall(s,a,k,h), \left|\widetilde{N}_h^k(s,a) - N_h^k(s,a)\right| \leq \frac{3H}{\epsilon}\log^{1.5} K\ln(3SAT/\delta)\right] \geq 1-\delta/3$$

$$\mathbb{P}\left[\forall(s,a,k,h), \left|\widetilde{N}_h^k(s,a,s') - N_h^k(s,a,s')\right| \leq \frac{3H}{\epsilon}\log^{1.5} K\ln(3S^2AT/\delta)\right] \geq 1-\delta/3$$

Combining all three guarantees together using a union bound, we obtain that CENTRAL-PRIVATIZER satisfies Assumption 1. □

C.3. Proof of Theorem 1

Proof of Theorem 1. To prove the JDP guarantee, we use the billboard lemma in (Hsu et al., 2016, Lemma 9) which states that an algorithm is JDP if the output sent to each user is a function of the user's private data and a common quantified computed with standard differential privacy. The formal lemma is stated as follows:

Lemma 14 (Billboard lemma (Hsu et al., 2016))**.** *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. then 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.*

Note that by the privacy guarantee in Lemma 1 and the post-processing property of DP in (Dwork et al., 2014), the policies $(\pi^k)_k$ are computed with a ϵ-DP. Therefore, by the above billboard lemma, the composition of the output of all K episodes satisfies ϵ-JDP.

□

C.4. Proof of Theorem 2

Before showing the proof of Theorem 2 we first consider the following two lemmas.

The following lemma shows the estimation errors of our private mean empirical rewards and private empirical transition probabilities, which are in step 7 of the algorithm.

Lemma 15 (Concentration bounds of private estimators)**.** *Fix any $\epsilon \in (0,1]$ and $\delta \in (0,1)$ and take $B_n = \left(\frac{\epsilon un}{H\log^{1.5} K\log(3SAT/\delta)}\right)^{\frac{1}{1+v}}$ in equation (1). Then, under Assumption 1, with probability at least $1-3\delta$, uniformly over all (s,a,h,k) we have*

$$|\widetilde{r}_h^k(s,a) - r_h(s,a)| \leq \beta_h^{k,r}(s,a), \left|\left(\widetilde{P}_h^k - P_h\right)V_{h+1}^*(s,a)\right| \leq \beta_h^{k,pv}(s,a),$$

$\left|P_h(s'|s,a) - \widetilde{P}_h^k(s'|s,a)\right| \leq C\sqrt{\frac{P_h(s'|s,a)\ln\frac{2SAT}{\delta}}{(\widetilde{N}_h^k(s,a)+E_{\epsilon,\delta,1})\vee 1}} + \frac{C\ln\frac{2SAT}{\delta}+2E_{\epsilon,\delta,1}+E_{\epsilon,\delta,3}}{(\widetilde{N}_h^k(s,a)+E_{\epsilon,\delta,1})\vee 1}$, *where C is a constant, $a\vee b = \max\{a,b\}$,* $(PV_{h+1})(s,a) = \sum_{s'} P(s'|s,a)V_{h+1}(s')$, *and* $\beta_h^{k,r}(s,a) = \frac{2\tau E_{\varepsilon,\delta,1}}{\left(\widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1}\right)\vee 1} + 10u^{\frac{1}{1+v}}\left(\frac{H\log^{1.5} K\log(3SAT/\delta)}{\epsilon\left(\left(\widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1}\right)\vee 1\right)}\right)^{\frac{v}{1+v}}$,

$\beta_h^{k,pv}(s,a) = \tau H\sqrt{\frac{2\ln(4SAT/\delta)}{(\widetilde{N}_h^k(s,a)+E_{\epsilon,\delta,1})\vee 1}} + \frac{\tau H(2E_{\epsilon,\delta,1}+SE_{\epsilon,\delta,3})}{(\widetilde{N}_h^k(s,a)+E_{\epsilon,\delta,1})\vee 1}$.

Remark 7. *Actually, all the above errors are determined by three terms: an error from truncating rewards, an error due to using the finite number of truncated rewards to estimate heavy-tailed reward distributions and an error due to the noise for ensuring privacy. In order to balance these three terms, we need to take the truncation threshold*

$B_n = \left(\frac{\epsilon un}{H \log^{1.5} K \log(3SAT/\delta)}\right)^{\frac{1}{1+v}}$. *Compared with the truncation value* $B_n = \left(\frac{\epsilon un}{\log^{1.5} T}\right)^{\frac{1}{1+v}}$ *in (Tao et al., 2022b) where the authors focus on DP-MAB with heavy-tailed rewards, the difference comes from that the privacy budget in our problem is* $\frac{\epsilon}{H}$ *for each step* h *based on Composition Theorem in Lemma 4, while* $H = 1$ *in MAB case. Compared with the truncation threshold* $B_n = \left(\frac{un}{\log(2SAT/\delta)}\right)^{\frac{1}{1+v}}$ *in non-private RL with heavy-tailed rewards (Zhuang & Sui, 2021), we can see the efficient number of samples in our problem becomes to* $n\epsilon$ *due to privacy. From our truncation threshold value* $B_n = \left(\frac{\epsilon un}{H \log^{1.5} K \log(3SAT/\delta)}\right)^{\frac{1}{1+v}}$*, we also can check the trade-off between utility and privacy: larger* ϵ *value provides weaker privacy guarantee but we will truncate the data with a larger range, so more data information will be used, then the bias will be smaller so that we get better utility .*

Proof of Lemma 15. We first define the non-private mean empirical rewards: $\bar{r}_h^k(s,a) := \frac{R_h^k(s,a)}{N_h^k(s,a)\vee 1}$. We then get the non-private estimation error. For a fixed tuple $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, at every episode $k \in [K]$, from Bernstein's inequality in Lemma 7 for bounded random variables, with probability at least $1 - \frac{\delta}{2SAT}$, noting that

$$\mathbb{E}(X^2 \cdot \mathbb{I}_{|X|\le B}) = \mathbb{E}(X^{1+v}X^{1-v}\cdot \mathbb{I}_{|X|\le B}) \le B^{1-v}\mathbb{E}(X^{1+v}\cdot \mathbb{I}_{|X|\le B}) \le uB^{1-v} \tag{5}$$

and

$$\mathbb{E}[X\mathbb{I}_{|X|>B}] \le \mathbb{E}[|X|^{1+v}|X|^{-v}\mathbb{I}_{|X|>B}] \le \frac{u}{B^v} \tag{6}$$

if $\mathbb{E}|X|^{1+v} \le u$,

$$\begin{aligned}
&\left|\bar{r}_h^k(s,a) - r_h(s,a)\right| \\
=&\left|\frac{R_h^k(s,a)}{N_h^k(s,a)\vee 1} - r_h(s,a)\right| \\
=&|\frac{1}{N_h^k(s,a)\vee 1}\sum_{i=1}^{N_h^k(s,a)}[r_{h,i}(s,a)\mathbb{I}\{|r_{h,i}(s,a)|\le B_i\} - \mathbb{E}(X_h(s,a)\mathbb{I}\{|X_h(s,a)|\le B_i\}) \\
&+\mathbb{E}(X_h(s,a)\mathbb{I}\{|X_h(s,a)|\le B_i\}) - \mathbb{E}(X_h(s,a))]| \\
\le&\left|\frac{1}{N_h^k(s,a)\vee 1}\sum_{i=1}^{N_h^k(s,a)}[r_{h,i}(s,a)\mathbb{I}\{|r_{h,i}(s,a)|\le B_i\} - \mathbb{E}(X_h(s,a)\mathbb{I}\{|X_h(s,a)|\le B_i\})\right| \\
&+\left|\frac{1}{N_h^k(s,a)\vee 1}\sum_{i=1}^{N_h^k(s,a)}\mathbb{E}(X_h(s,a)\mathbb{I}\{|X_h(s,a)|\ge B_i\})\right| \\
\le&\sqrt{\frac{2uB^{1-v}_{N_h^k(s,a)}\log(\frac{2SAT}{\delta})}{N_h^k(s,a)\vee 1}} + \frac{B_{N_h^k(s,a)}\log\frac{2SAT}{\delta}}{3(N_h^k(s,a)\vee 1)} + \frac{1}{N_h^k(s,a)\vee 1}\sum_{i=1}^{N_h^k(s,a)}\frac{u}{B_i^v}
\end{aligned} \tag{7}$$

Let $L_{r,k} = \sqrt{\frac{2uB^{1-v}_{N_h^k(s,a)}\log(\frac{2SAT}{\delta})}{N_h^k(s,a)\vee 1}} + \frac{B_{N_h^k(s,a)}\log\frac{2SAT}{\delta}}{3(N_h^k(s,a)\vee 1)} + \frac{1}{N_h^k(s,a)\vee 1}\sum_{i=1}^{N_h^k(s,a)}\frac{u}{B_i^v}$, then we denote

$$F^c = \left\{\exists s,a,h,k : \left|\bar{r}_h^k(s,a) - r_h(s,a)\right| \ge L_{r,k}\right\} \tag{8}$$

Then, by using union bound over all s,a,h,k, we have $\mathbb{P}(F^c) \le \frac{\delta}{2}$. So,

$$\begin{aligned}
\mathbb{P}(F) &= \mathbb{P}\left(\left\{\forall s,a,h,k : \left|\bar{r}_h^k(s,a) - r_h(s,a)\right| \le L_{r,k}\right\}\right) \\
&= 1 - \mathbb{P}(F^c) \\
&= 1 - \frac{\delta}{2}.
\end{aligned} \tag{9}$$

We first study the concentration of the private reward estimate. Note that under the event in Assumption 1,

$$\left|\frac{\widetilde{R}_h^k(s,a)}{\left(\widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1}\right)\vee 1}-\frac{R_h^k(s,a)}{\left(\widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1}\right)\vee 1}\right|\leq\frac{E_{\varepsilon,\delta,k,2}}{\left(\widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1}\right)\vee 1}, \tag{10}$$

since $\widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1}\geq N_h^k(s,a)\geq 0$ and $|\widetilde{R}_h^k(s,a)-R_h^k(s,a)|\leq E_{\varepsilon,\delta,k,2}$.

$$\begin{aligned}
&\left|\frac{R_h^k(s,a)}{\left(\widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1}\right)\vee 1}-r_h(s,a)\right|\\
\leq&\left|r_h(s,a)\left(\frac{N_h^k(s,a)\vee 1}{\left(\widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1}\right)\vee 1}-1\right)\right|+\left|\frac{N_h^k(s,a)\vee 1}{\left(\widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1}\right)\vee 1}\left(\frac{R_h^k(s,a)}{N_h^k(s,a)\vee 1}-r_h(s,a)\right)\right|\\
\leq&|r_h(s,a)|\left|1-\frac{N_h^k(s,a)\vee 1}{\left(\widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1}\right)\vee 1}\right|+\frac{N_h^k(s,a)\vee 1}{\left(\widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1}\right)\vee 1}L_{r,k}\\
\leq&\frac{2\tau E_{\varepsilon,\delta,1}}{\left(\widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1}\right)\vee 1}+\frac{L_{r,k}(N_h^k(s,a)\vee 1)}{\left(\widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1}\right)\vee 1}
\end{aligned} \tag{11}$$

We put two pieces together, then get

$$|\widetilde{r}_h^k(s,a)-r_h(s,a)|\leq\frac{2\tau E_{\varepsilon,\delta,1}}{\left(\widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1}\right)\vee 1}+\frac{L_{r,k}(N_h^k(s,a)\vee 1)}{\left(\widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1}\right)\vee 1}+\frac{E_{\varepsilon,\delta,k,2}}{\left(\widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1}\right)\vee 1} \tag{12}$$

where $E_{\epsilon,\delta,k,2}=\frac{6HB_{N_h^k(s,a)}}{\epsilon}\log^{1.5}K\ln(3SAT/\delta)$. In order to get the trade-off between mean estimation error and private count error, we set the truncation threshold $B_n=\left(\frac{\epsilon un}{H\log^{1.5}K\log(3SAT/\delta)}\right)^{\frac{1}{1+v}}$. Then we can bound each term in $L_{r,k}$ for $\epsilon\in(0,1]$ and $K\geq 2$:

$$\begin{aligned}
\sqrt{\frac{2uB_{N_h^k(s,a)}^{1-v}\log(\frac{2SAT}{\delta})}{N_h^k(s,a)\vee 1}}&\leq\frac{\sqrt{2}u^{\frac{1}{1+v}}\epsilon^{\frac{1-v}{2(1+v)}}(\log(3SAT/\delta))^{\frac{v}{1+v}}}{(N_h^k(s,a)\vee 1)^{\frac{v}{1+v}}(H\log^{1.5}K)^{\frac{1-v}{2(1+v)}}}\\
&\leq\sqrt{2}u^{\frac{1}{1+v}}\left(\frac{H\log^{1.5}K\log(3SAT/\delta)}{\epsilon(N_h^k(s,a)\vee 1)}\right)^{\frac{v}{1+v}}
\end{aligned},$$

$$\frac{B_{N_h^k(s,a)}\log\frac{2SAT}{\delta}}{3(N_h^k(s,a)\vee 1)}\leq\frac{u^{\frac{1}{1+v}}(\log(3SAT/\delta))^{\frac{v}{1+v}}}{3(N_h^k(s,a)\vee 1)^{\frac{v}{1+v}}}\left(\frac{\epsilon}{H\log^{1.5}K}\right)^{\frac{1}{1+v}}\leq\frac{u^{\frac{1}{1+v}}}{3}\left(\frac{H\log^{1.5}K\log(3SAT/\delta)}{\epsilon(N_h^k(s,a)\vee 1)}\right)^{\frac{v}{1+v}},$$

where the last inequality is based on the fact that $x^{\frac{1}{1+v}}\leq x^{-\frac{v}{1+v}}$ for $x\in(0,1]$,

$$\begin{aligned}
\frac{1}{N_h^k(s,a)\vee 1}\sum_{i=1}^{N_h^k(s,a)}\frac{u}{B_i^v}&\leq\frac{u}{N_h^k(s,a)\vee 1}\sum_{i=1}^{N_h^k(s,a)}\left(\frac{H\log^{1.5}K\log(3SAT/\delta)}{\epsilon ui}\right)^{\frac{v}{1+v}}\\
&\leq\frac{u^{\frac{1}{1+v}}}{N_h^k(s,a)\vee 1}\left(\frac{H\log^{1.5}K\log(3SAT/\delta)}{\epsilon}\right)^{\frac{v}{1+v}}\sum_{i=1}^{N_h^k(s,a)}i^{-\frac{v}{1+v}}\\
&\leq\frac{u^{\frac{1}{1+v}}}{N_h^k(s,a)\vee 1}\left(\frac{H\log^{1.5}K\log(3SAT/\delta)}{\epsilon}\right)^{\frac{v}{1+v}}\cdot(1+v)\cdot(N_h^k(s,a))^{\frac{1}{1+v}}\\
&\leq 2u^{\frac{1}{1+v}}\left(\frac{H\log^{1.5}K\log(3SAT/\delta)}{\epsilon(N_h^k(s,a)\vee 1)}\right)^{\frac{v}{1+v}}
\end{aligned}.$$

Thus, $L_{r,k} \leq 4u^{\frac{1}{1+v}} \left(\frac{H \log^{1.5} K \log(3SAT/\delta)}{\epsilon(N_h^k(s,a)\vee 1)}\right)^{\frac{v}{1+v}}$.

And

$$E_{\epsilon,\delta,k,2} = \frac{6HB_{N_h^k(s,a)}}{\epsilon} \log^{1.5} K \ln(3SAT/\delta) = 6(uN_h^k(s,a))^{\frac{1}{1+v}} \left(\frac{H \log^{1.5} K \log(3SAT/\delta)}{\epsilon}\right)^{\frac{v}{1+v}}.$$

Since $\widetilde{N}_h^k(s,a) + E_{\varepsilon,\delta,1} \geq N_h^k(s,a) \geq 0$, we obtain

$$|\widetilde{r}_h^k(s,a) - r_h(s,a)| \leq \frac{2\tau E_{\varepsilon,\delta,1}}{\left(\widetilde{N}_h^k(s,a) + E_{\varepsilon,\delta,1}\right) \vee 1} + 10u^{\frac{1}{1+v}} \left(\frac{H \log^{1.5} K \log(3SAT/\delta)}{\epsilon\left((\widetilde{N}_h^k(s,a) + E_{\varepsilon,\delta,1}) \vee 1\right)}\right)^{\frac{v}{1+v}}. \tag{13}$$

To show the second result, we first note that V^* is fixed and $V_h^*(s) \leq \tau H$ for all h and s. Based on Hoeffding's inequality in Lemma 6, we can use similar proof as Lemma 4.1 in (Chowdhury & Zhou, 2021) to get with probability at least $1 - \delta/2$

$$\left|\left(\widetilde{P}_h^k - P_h\right) V_{h+1}^*(s,a)\right| \leq \tau H \sqrt{\frac{2\ln(4SAT/\delta)}{(\widetilde{N}_h^k(s,a) + E_{\epsilon,\delta,1}) \vee 1}} + \frac{\tau H(2E_{\epsilon,\delta,1} + SE_{\epsilon,\delta,3})}{(\widetilde{N}_h^k(s,a) + E_{\epsilon,\delta,1}) \vee 1}$$

and with probability at least $1 - \delta$,

$$\left|P_h\left(s' \mid s,a\right) - \widetilde{P}_h^k\left(s' \mid s,a\right)\right| \leq C\sqrt{\frac{P_h(s'|s,a)\ln(2SAT/\delta)}{(\widetilde{N}_h^k(s,a) + E_{\epsilon,\delta,1}) \vee 1}} + \frac{C\ln(2SAT/\delta) + 2E_{\epsilon,\delta,1} + E_{\epsilon,\delta,3}}{(\widetilde{N}_h^k(s,a) + E_{\epsilon,\delta,1}) \vee 1},$$

which is based on the application of empirical Bernstein inequality. □

The next lemma claims that the value function maintained in Algorithm 1 is optimistic.

Lemma 16. *Fix some $\delta \in (0,1]$, with probability at least $1 - 3\delta$, $\widetilde{V}_h^k(s) \geq V_h^*(s)$ for all (k,h,s).*

Proof of Lemma 16. For a fixed k, consider $h = H+1, H, \ldots, 1$. In the base case $h = H+1$, it trivially holds since $\widetilde{V}_{H+1}^k(s) = 0 = V_{H+1}^*(s)$. Assume that $\widetilde{V}_{h+1}^k(s) \geq V_{h+1}^*(s)$ for all s. Then, by the update rule, we have

$$\widetilde{Q}_h^k(s,a) = \max\{-(H-h+1)\tau, \min\{(H-h+1)\tau, \widetilde{r}_h^k(s,a) + (\widetilde{P}_h^k\widetilde{V}_{h+1}^k)(s,a) + \beta_h^k(s,a)\}\}$$

First, we would like to show that the truncation at $-(H-h+1)\tau$ does not affect the analysis. To see this, first, observe that under Lemma 15

$$\begin{aligned}
\widetilde{r}_h^k(s,a) + (\widetilde{P}_h^k\widetilde{V}_{h+1}^k)(s,a) + \beta_h^k(s,a) &\overset{(a)}{\geq} r_h(s,a) + (\widetilde{P}_h^k\widetilde{V}_{h+1}^k)(s,a) + \beta_h^{k,pv}(s,a) \\
&\overset{(b)}{\geq} r_h(s,a) + (\widetilde{P}_h^k V_{h+1}^*)(s,a) + \beta_h^{k,pv}(s,a) \\
&\overset{(c)}{\geq} r_h(s,a) + (P_h V_{h+1}^*)(s,a) = Q_h^*(s,a) \\
&\geq -(H-h+1)\tau
\end{aligned} \tag{14}$$

where (a) holds by the first result in Lemma 15; (b) holds by induction; (c) holds by the second result in Lemma 15. This directly implies that

$$\widetilde{Q}_h^k(s,a) = \min\{(H-h+1)\tau, \widetilde{r}_h^k(s,a) + (\widetilde{P}_h^k\widetilde{V}_{h+1}^k)(s,a) + \beta_h^k(s,a)\}$$

Hence, if the maximum is attained at $(H-h+1)\tau$, then $\widetilde{Q}_h^k(s,a) \geq Q_h^*(s,a)$ trivially holds since $Q_h^*(s,a) \in [-(H-h+1)\tau, (H-h+1)\tau]$. Otherwise, by Eq. (14), we also have $\widetilde{Q}_h^k(s,a) \geq Q_h^*(s,a)$. Therefore, we have $\widetilde{Q}_h^k(s,a) \geq Q_h^*(s,a)$, and hence $\widetilde{V}_h^k(s) \geq V_h^*(s)$. □

Proof of Theorem 2. By the optimistic result in Lemma 16, we have

$$Reg(T) = \sum_{k=1}^{K}(V_1^*(s_1) - V_1^{\pi_k}(s_1)) \leq \sum_{k=1}^{K}(\widetilde{V}_1^k(s_1) - V_1^{\pi_k}(s_1)) \tag{15}$$

Now, we turn to upper bound $\widetilde{V}_h^k(s_h^k) - V_h^{\pi_k}(s_h^k)$ by a recursive form.

First, observe that

$$(\widetilde{V}_h^k - V_h^{\pi_k})(s_h^k) = (\widetilde{Q}_h^k - Q_h^{\pi_k})(s_h^k, a_h^k),$$

which holds since the action executed by π_k at step h, and the action used to update $\widetilde{V}_h^k$ is the same. Now, to bound the Q-value difference, we have

$$\begin{aligned}
&(\widetilde{Q}_h^k - Q_h^{\pi_k})(s_h^k, a_h^k) \\
&\overset{(a)}{\leq} 2\beta_h^{k,r}(s_h^k, a_h^k) + (\widetilde{P}_h^k \widetilde{V}_{h+1}^k - P_h V_{h+1}^{\pi_k})(s_h^k, a_h^k) + \beta_h^{k,pv}(s,a) \\
&= \left[(\widetilde{P}_h^k - P_h)\widetilde{V}_{h+1}^k\right](s_h^k, a_h^k) + \left[P_h(\widetilde{V}_{h+1}^k - V_{h+1}^{\pi_k})\right](s_h^k, a_h^k) + 2\beta_h^{k,r}(s_h^k, a_h^k) + \beta_h^{k,pv}(s,a) \\
&= \left[(\widetilde{P}_h^k - P_h)V_{h+1}^*\right](s_h^k, a_h^k) + \left[(P_h - \widetilde{P}_h^k)(V_{h+1}^* - \widetilde{V}_{h+1}^k)\right](s_h^k, a_h^k) + \left[P_h(\widetilde{V}_{h+1}^k - V_{h+1}^{\pi_k})\right](s_h^k, a_h^k) \\
&\quad + 2\beta_h^{k,r}(s_h^k, a_h^k) + \beta_h^{k,pv}(s,a) \\
&\overset{(b)}{\leq} \left[(\widetilde{P}_h^k - P_h)(\widetilde{V}_{h+1}^k - V_{h+1}^*)\right](s_h^k, a_h^k) + \left[P_h(\widetilde{V}_{h+1}^k - V_{h+1}^{\pi_k})\right](s_h^k, a_h^k) + 2\beta_h^{k,r}(s_h^k, a_h^k) + 2\beta_h^{k,pv}(s,a)
\end{aligned} \tag{16}$$

where (a) we have used the reward concentration result in Lemma 15; (b) holds by the transition concentration result in Lemma 15. Thus, so far we have arrived at

$$\begin{aligned}
(\widetilde{V}_h^k - V_h^{\pi_k})(s_h^k) \leq &\left[(\widetilde{P}_h^k - P_h)(\widetilde{V}_{h+1}^k - V_{h+1}^*)\right](s_h^k, a_h^k) + \left[P_h(\widetilde{V}_{h+1}^k - V_{h+1}^{\pi_k})\right](s_h^k, a_h^k) \\
&+ 2\beta_h^{k,r}(s_h^k, a_h^k) + 2\beta_h^{k,pv}(s_h^k, a_h^k).
\end{aligned} \tag{17}$$

We will first carefully analyze the first term. In particular, let $G := (\widetilde{V}_{h+1}^k - V_{h+1}^*)$, we have

$$\begin{aligned}
&\left[(\widetilde{P}_h^k - P_h)(\widetilde{V}_{h+1}^k - V_{h+1}^*)\right](s_h^k, a_h^k) \\
&= \sum_{s'}\left(\widetilde{P}_h^k(s'|s_h^k, a_h^k) - P_h(s'|s_h^k, a_h^k)\right)G(s') \\
&\overset{(a)}{\leq} c\sum_{s'}\left(\sqrt{\frac{\ln(2SAT/\delta)P_h(s'|s_h^k, a_h^k)}{(\widetilde{N}_h^k(s,a) + E_{\epsilon,\delta,1}) \vee 1}} + \frac{\ln(2SAT/\delta)}{(\widetilde{N}_h^k(s,a) + E_{\epsilon,\delta,1}) \vee 1} + \frac{2E_{\epsilon,\delta,1} + E_{\epsilon,\delta,3}}{(\widetilde{N}_h^k(s,a) + E_{\epsilon,\delta,1}) \vee 1}\right)G(s') \\
&\overset{(b)}{\leq} \sum_{s'}\left(\frac{P_h(s'|s_h^k, a_h^k)}{H}G(s')\right) + c\sum_{s'}\left(\frac{H\ln(2SAT/\delta)}{(\widetilde{N}_h^k(s,a) + E_{\epsilon,\delta,1}) \vee 1}\right)G(s') \\
&\quad + c\sum_{s'}\left(\frac{\ln(2SAT/\delta)}{(\widetilde{N}_h^k(s,a) + E_{\epsilon,\delta,1}) \vee 1}\right)G(s') + c\sum_{s'}\left(\frac{2E_{\epsilon,\delta,1} + E_{\epsilon,\delta,3}}{(\widetilde{N}_h^k(s,a) + E_{\epsilon,\delta,1}) \vee 1}\right)G(s') \\
&\overset{(c)}{\leq} \sum_{s'}\left(\frac{P_h(s'|s_h^k, a_h^k)}{H}G(s')\right) + \sum_{s'}\left(\frac{c'H^2\tau\ln(2SAT/\delta)}{(\widetilde{N}_h^k(s,a) + E_{\epsilon,\delta,1}) \vee 1}\right) + c\sum_{s'}\left(\frac{4\tau HE_{\epsilon,\delta,1} + 2\tau HE_{\epsilon,\delta,3}}{(\widetilde{N}_h^k(s,a) + E_{\epsilon,\delta,1}) \vee 1}\right),
\end{aligned} \tag{18}$$

where (a) holds by the third result in Lemma 15 and c is some absolute constant; (b) holds by $\sqrt{xy} \leq x + y$ for positive numbers x, y; (c) holds since $G(s') \leq 2H\tau$ by the boundedness of V-value. Now, plugging the definition for $G(s')$ into (18),

yields

$$\begin{aligned}
&\left[(\widetilde{P}_h^k - P_h)(\widetilde{V}_{h+1}^k - V_{h+1}^*)\right](s_h^k, a_h^k) \\
\leq &\frac{1}{H}\left[P_h(\widetilde{V}_{h+1}^k - V_{h+1}^*)\right](s_h^k, a_h^k) + \frac{c'\tau SH^2\ln(2SAT/\delta)}{(\widetilde{N}_h^k(s,a) + E_{\epsilon,\delta,1}) \vee 1} + \frac{4c\tau SHE_{\epsilon,\delta,1} + 2c\tau SHE_{\epsilon,\delta,3}}{(\widetilde{N}_h^k(s,a) + E_{\epsilon,\delta,1}) \vee 1} \\
\overset{(a)}{\leq} &\frac{1}{H}\left[P_h(\widetilde{V}_{h+1}^k - V_{h+1}^*)\right](s_h^k, a_h^k) + \xi_h^k + \zeta_h^k \\
\overset{(b)}{\leq} &\frac{1}{H}\left[P_h(\widetilde{V}_{h+1}^k - V_{h+1}^{\pi_k})\right](s_h^k, a_h^k) + \xi_h^k + \zeta_h^k,
\end{aligned} \tag{19}$$

where (a) holds by definitions $\xi_h^k =: \frac{c'\tau SH^2\ln(2SAT/\delta)}{(\widetilde{N}_h^k(s,a)+E_{\epsilon,\delta,1})\vee 1}$ and $\zeta_h^k := \frac{4c\tau SHE_{\epsilon,\delta,1}+2c\tau SHE_{\epsilon,\delta,3}}{(\widetilde{N}_h^k(s,a)+E_{\epsilon,\delta,1})\vee 1}$; (b) holds since $V_{h+1}^{\pi_k} \leq V_{h+1}^*$. Plugging (19) into (17), yields the following recursive formula.

$$\begin{aligned}
(\widetilde{V}_h^k - V_h^{\pi_k})(s_h^k) &\overset{(a)}{\leq} \left(1 + \frac{1}{H}\right)\left[P_h(\widetilde{V}_{h+1}^k - V_{h+1}^{\pi_k})\right](s_h^k, a_h^k) + \xi_h^k + \zeta_h^k + 2\beta_h^k \\
&\overset{(b)}{=} \left(1 + \frac{1}{H}\right)\left[(\widetilde{V}_{h+1}^k - V_{h+1}^{\pi_k})(s_{h+1}^k) + \chi_h^k\right] + \xi_h^k + \zeta_h^k + 2\beta_h^k
\end{aligned}$$

where in (a), we let $\beta_h^k := \beta_h^{k,r}(s_h^k, a_h^k) + \beta_h^{k,pv}(s_h^k, a_h^k)$ for notation simplicity; (b) holds by definition $\chi_h^k := \left[P_h(\widetilde{V}_{h+1}^k - V_{h+1}^{\pi_k})\right](s_h^k, a_h^k) - (\widetilde{V}_{h+1}^k - V_{h+1}^{\pi_k})(s_{h+1}^k)$. Based on this, we have the following bound on $(\widetilde{V}_1^k - V_1^{\pi_k})(s_1^k)$,

$$\begin{aligned}
&(\widetilde{V}_1^k - V_1^{\pi_k})(s_1^k) \\
\leq &\left(1 + \frac{1}{H}\right)(\chi_1^k + \xi_1^k + \zeta_1^k + 2\beta_1^k) + \left(1 + \frac{1}{H}\right)^2(\chi_2^k + \xi_2^k + \zeta_2^k + 2\beta_2^k) + \dots \\
&+ \left(1 + \frac{1}{H}\right)^H(\chi_H^k + \xi_H^k + \zeta_H^k + 2\beta_H^k) \\
\leq &3\sum_{h=1}^H(\chi_h^k + \xi_h^k + \zeta_h^k + 2\beta_h^k).
\end{aligned} \tag{20}$$

Therefore, plugging (20) into (15), we have the regret decomposition as follows.

$$Reg(T) \leq 3\sum_{k=1}^K\sum_{h=1}^H(\chi_h^k + \xi_h^k + \zeta_h^k + 2\beta_h^k)$$

We are only left to bound each of them. To start with, we focus on the bonus term. We first focus on $\beta_h^{k,r}(s,a)$ as shown in Lemma 15. By definition, we have

$$\sum_{k=1}^K\sum_{h=1}^H\beta_h^{k,r}(s,a) = \underbrace{\sum_{k=1}^K\sum_{h=1}^H\frac{2\tau E_{\varepsilon,\delta,1}}{\left(\widetilde{N}_h^k(s_h^k,a_h^k) + E_{\varepsilon,\delta,1}\right)\vee 1}}_{\mathcal{O}_1} + \underbrace{10u^{\frac{1}{1+v}}\sum_{k=1}^K\sum_{h=1}^H\left(\frac{H\log^{1.5}K\log(3SAT/\delta)}{\epsilon\left((\widetilde{N}_h^k(s_h^k,a_h^k) + E_{\varepsilon,\delta,1})\vee 1\right)}\right)^{\frac{v}{1+v}}}_{\mathcal{O}_2},$$

The first term can be upper bounded as follows ($T = KH$) under assumption 1

$$\begin{aligned}
\mathcal{O}_1 &\leq 2\tau E_{\epsilon,\delta,1} \sum_{k=1}^{K}\sum_{h=1}^{H} \frac{1}{N_h^k(s_h^k,a_h^k)\vee 1} \\
&= 2\tau E_{\epsilon,\delta,1} \sum_{h,s,a} \sum_{i=1}^{N_h^K(s,a)} \frac{1}{i} \\
&\leq c' E_{\epsilon,\delta,1} \sum_{h,s,a} \ln(N_h^K(s,a)) \\
&= \widetilde{O}\left(HSAE_{\epsilon,\delta,1}\right).
\end{aligned}$$

where $\widetilde{O}(\cdot)$ hides $polylog(S, A, T, 1/\delta)$ factors.

The second term can be upper bounded as follows under Assumption 15.

$$\begin{aligned}
\mathcal{O}_2 &\leq cu^{\frac{1}{1+v}} \left(\frac{H\log^{1.5} K\log(3SAT/\delta)}{\epsilon}\right)^{\frac{v}{1+v}} \sum_{k=1}^{K}\sum_{h=1}^{H} \left(\frac{1}{N_h^k(s_h^k,a_h^k)\vee 1}\right)^{\frac{v}{1+v}} \\
&= cu^{\frac{1}{1+v}} \left(\frac{H\log^{1.5} K\log(3SAT/\delta)}{\epsilon}\right)^{\frac{v}{1+v}} \sum_{h,s,a} \sum_{i=1}^{N_h^K(s,a)} \frac{1}{i^{\frac{v}{1+v}}} \\
&\leq c' u^{\frac{1}{1+v}} \left(\frac{H\log^{1.5} K\log(3SAT/\delta)}{\epsilon}\right)^{\frac{v}{1+v}} \sum_{h,s,a} (N_h^K(s,a))^{\frac{1}{1+v}} \\
&\overset{(a)}{\leq} c' u^{\frac{1}{1+v}} \left(\frac{H\log^{1.5} K\log(3SAT/\delta)}{\epsilon}\right)^{\frac{v}{1+v}} \left(\sum_{h,s,a} 1\right)^{\frac{v}{1+v}} \left(\sum_{h,s,a} N_h^K(s,a)\right)^{\frac{1}{1+v}} \\
&\leq \widetilde{O}\left(u^{\frac{1}{1+v}} \left(\frac{H^2SA}{\epsilon}\right)^{\frac{v}{1+v}} T^{\frac{1}{1+v}}\right)
\end{aligned}$$

where (a) is based on Hölder's inequality with $x_k = 1, y_k = (N_h^K(s,a))^{\frac{1}{1+v}}, q = 1+v$ in Lemma 9.

Putting them together, we have the upper bound for the summation over $\beta_h^{k,r}(s,a)$,

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \beta_h^{k,r}(s,a) = \widetilde{O}\left(HSAE_{\epsilon,\delta,1} + u^{\frac{1}{1+v}} \left(\frac{H^2SA}{\epsilon}\right)^{\frac{v}{1+v}} T^{\frac{1}{1+v}}\right)$$

Now, we study the upper bound of $\beta_h^{k,pv}(s_h^k, a_h^k)$. By definition, we have

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H} \beta_h^{k,pv}(s_h^k,a_h^k) \\
=&\underbrace{\tau H \sum_{k=1}^{K}\sum_{h=1}^{H} \sqrt{\frac{2\ln(4SAT/\delta)}{(\widetilde{N}_h^k(s_h^k,a_h^k) + E_{\epsilon,\delta,1})\vee 1}}}_{\mathcal{T}_1} + \underbrace{\tau H \sum_{k=1}^{K}\sum_{h=1}^{H} \frac{(SE_{\epsilon,\delta,3} + 2E_{\epsilon,\delta,1})}{(\widetilde{N}_h^k(s_h^k,a_h^k) + E_{\epsilon,\delta,1})\vee 1}}_{\mathcal{T}_2}.
\end{aligned}$$

The first term can be upper bounded as follows ($T := KH$) under Assumption 1.

$$\begin{aligned}
\mathcal{T}_1 &\leq \tau H\sqrt{2\ln(4SAT/\delta)}\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{1}{N_h^k(s,a)\vee 1}} \\
&= \tau\sqrt{2\ln(4SAT/\delta)}\sum_{h,s,a}\sum_{i=1}^{N_h^K(s,a)}\frac{1}{\sqrt{i}} \\
&\leq c'H\sqrt{2\ln(4SAT/\delta)}\sum_{h,s,a}\sqrt{N_h^K(s,a)} \\
&\leq c'H\sqrt{2\ln(4SAT/\delta)}\sqrt{\left(\sum_{h,s,a}1\right)\left(\sum_{h,s,a}N_h^K(s,a)\right)} \\
&= \widetilde{O}\left(\sqrt{H^3SAT}\right).
\end{aligned}$$

The second term can be upper bounded as follows under Assumption 1.

$$\begin{aligned}
\mathcal{T}_2 &\leq cH(SE_{\epsilon,\delta,3}+E_{\epsilon,\delta,1})\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{N_h^k(s_h^k,a_h^k)\vee 1} \\
&= cH(SE_{\epsilon,\delta,3}+E_{\epsilon,\delta,1})\sum_{h,s,a}\sum_{i=1}^{N_h^K(s,a)}\frac{1}{i} \\
&\leq c'H(SE_{\epsilon,\delta,3}+E_{\epsilon,\delta,1})\sum_{h,s,a}\ln(N_h^K(s,a)) \\
&= \widetilde{O}\left(H^2S^2AE_{\epsilon,\delta,3}+H^2SAE_{\epsilon,\delta,1}\right).
\end{aligned}$$

Putting them together, we have the following bound on the summation over β_h^k.

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\beta_h^k = \widetilde{O}\left(\sqrt{H^3SAT}+H^2S^2AE_{\epsilon,\delta,3}+H^2SAE_{\epsilon,\delta,1}+u^{\frac{1}{1+v}}\left(\frac{H^2SA}{\epsilon}\right)^{\frac{v}{1+v}}T^{\frac{1}{1+v}}\right).$$

By following the same analysis as in $\mathcal{T}_2$, we can bound the summation over $\xi_h^k =: \frac{c'\tau SH^2\ln(2SAT/\delta)}{(\widetilde{N}_h^k(s,a)+E_{\epsilon,\delta,1})\vee 1}$ and $\zeta_h^k := \frac{4c\tau SHE_{\epsilon,\delta,1}+2c\tau SHE_{\epsilon,\delta,3}}{(\widetilde{N}_h^k(s,a)+E_{\epsilon,\delta,1})\vee 1}$ as follows.

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}\xi_h^k &= \widetilde{O}\left(H^3S^2A\right) \\
\sum_{k=1}^{K}\sum_{h=1}^{H}\zeta_h^k &= \widetilde{O}\left(H^2S^2A(E_{\epsilon,\delta,3}+E_{\epsilon,\delta,1})\right).
\end{aligned}$$

Finally, we are going to bound the summation over $\chi_h^k := \left[P_h(\widetilde{V}_{h+1}^k - V_{h+1}^{\pi_k})\right](s_h^k,a_h^k) - (\widetilde{V}_{h+1}^k - V_{h+1}^{\pi_k})(s_{h+1}^k)$, which turns out to be a martingale difference sequence. In particular, we define a filtration $\mathcal{F}_h^k$ that includes all the randomness up to the k-th episode and the h-th step. Then, we have $\mathcal{F}_1^1 \subset \mathcal{F}_2^1 \ldots \subset \mathcal{F}_H^1 \subset \mathcal{F}_1^2 \subset \mathcal{F}_2^2 \ldots$. Also, we have $(\widetilde{V}_{h+1}^k - V_{h+1}^{\pi_k}) \in \mathcal{F}_1^k \subset \mathcal{F}_h^k$ since they are decided by data collected up to episode $k-1$. A bit abuse of notation, we define $Y_{h+1}^k := \chi_h^k$. Then, we have

$$\mathbb{E}\left[Y_{h+1}^k | \mathcal{F}_h^k\right] = 0. \tag{21}$$

This holds since the expectation only captures randomness over s_{h+1}^k. Thus, Y_{h+1}^k is a martingale difference sequence. Moreover, we have $|Y_{h+1}^k| \leq 4H\tau$ a.s. By Azuma-Hoeffding inequality, we have with probability at least $1-\delta$

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\chi_h^k = \sum_{k=1}^{K}\sum_{h=1}^{H}Y_{h+1}^k = c'\sqrt{H^2T\ln(2/\delta)} = \widetilde{O}\left(\sqrt{H^2T}\right)$$

Putting everything together, and applying union bound on all high-probability events, we have shown that with probability at least $1-\delta$,

$$Reg(T) = \widetilde{O}\left(\sqrt{SAH^3T} + S^2AH^3 + S^2AH^2E_{\epsilon,\delta,1} + S^2AH^2E_{\epsilon,\delta,3} + u^{\frac{1}{1+v}}\left(\frac{H^2SA}{\epsilon}\right)^{\frac{v}{1+v}}T^{\frac{1}{1+v}}\right).$$

Based on the value of $E_{\epsilon,\delta,1}, E_{\epsilon,\delta,3}$ in Lemma 1, we obtain

$$Reg(T) = \widetilde{O}\left(\sqrt{SAH^3T} + S^2AH^3/\epsilon + u^{\frac{1}{1+v}}\left(\frac{SAH^2}{\epsilon}\right)^{\frac{v}{1+v}}T^{\frac{1}{1+v}}\right).$$

□

C.5. Proof of Theorem 3

Similar to the JDP case we first provide some concentration bounds.

Lemma 17 (Concentration bounds of locally private estimators)**.** *Fix any $\epsilon \in (0,1]$ and $\delta \in (0,1)$ and take $B_n = \left(\frac{u\epsilon\sqrt{n}}{H\log(6SAT/\delta)}\right)^{\frac{1}{1+v}}$ in equation (1). Then, under Assumption 1, with probability at least $1-3\delta$, uniformly over all (s,a,h,k), $|\widetilde{r}_h^k(s,a) - r_h(s,a)| \leq \beta_h^{k,r}(s,a)$, $\left|\left(\widetilde{P}_h^k - P_h\right)V_{h+1}^*(s,a)\right| \leq \beta_h^{k,pv}(s,a)$, $\left|P_h\left(s' \mid s,a\right) - \widetilde{P}_h^k\left(s' \mid s,a\right)\right| \leq C\sqrt{\frac{P_h(s'|s,a)\ln(2SAT/\delta)}{(\widetilde{N}_h^k(s,a)+E_{\epsilon,\delta,1})\vee 1}} + \frac{C\ln(2SAT/\delta)+2E_{\epsilon,\delta,1}+E_{\epsilon,\delta,3}}{(\widetilde{N}_h^k(s,a)+E_{\epsilon,\delta,1})\vee 1}$ where C is a positive constant,$(PV_{h+1})(s,a) = \sum_{s'}P(s'|s,a)V_{h+1}(s')$, $\beta_h^{k,r}(s,a) = \frac{2\tau E_{\varepsilon,\delta,1}}{\left(\widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1}\right)\vee 1} + 16u^{\frac{1}{1+v}}\left(\frac{H\log(6SAT/\delta)}{\epsilon\sqrt{(\widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1})\vee 1}}\right)^{\frac{v}{1+v}}$, and $\beta_h^{k,pv}(s,a)$ is defined in Lemma 15.*

Remark 8. *Compared with the JDP case, we can see there are several differences. Firstly, due to the error caused by the noise we added to guarantee privacy becomes larger, in the LDP case we need to make B_n smaller than it in the JDP case and finally, we can get the optimal one $B_n = \left(\frac{u\epsilon\sqrt{n}}{H\log(6SAT/\delta)}\right)^{\frac{1}{1+v}}$. It also indicates that the bound for LDP will be less than that for JDP as we leverage less data information. Secondly, due to the stronger privacy guarantee in the local model, we can see the second term of in $\beta_h^{k,r}(s,a)$ in above Lemma is worse than it in Lemma 15 by a factor of $\left(\sqrt{(\widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1})\vee 1}\right)^{-\frac{v}{1+v}}$.*

Proof of Lemma 17. Following the similar argue as proof of Lemma 15, we can get

$$|\widetilde{r}_h^k(s,a) - r_h(s,a)| \leq \frac{2\tau E_{\varepsilon,\delta,1}}{\left(\widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1}\right)\vee 1} + \frac{L_{r,k}(N_h^k(s,a)\vee 1)}{\left(\widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1}\right)\vee 1} + \frac{E_{\varepsilon,\delta,k,2}}{\left(\widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1}\right)\vee 1} \tag{22}$$

where $E_{\epsilon,\delta,k,2} = \frac{12HB_{N_h^k(s,a)}}{\epsilon}\sqrt{N_h^k(s,a)\log(6SAT/\delta)}$ and $L_{r,k} = \sqrt{\frac{2uB_{N_h^k(s,a)}^{1-v}\log(\frac{2SAT}{\delta})}{N_h^k(s,a)\vee 1}} + \frac{B_{N_h^k(s,a)}\log\frac{2SAT}{\delta}}{3(N_h^k(s,a)\vee 1)} + \frac{1}{N_h^k(s,a)\vee 1}\sum_{i=1}^{N_h^k(s,a)}\frac{u}{B_i^v}$.

Let $B_n = \left(\frac{u\epsilon\sqrt{n}}{H\log(6SAT/\delta)}\right)^{\frac{1}{1+v}}$, then

$$L_{r,k} = \sqrt{\frac{2uB_{N_h^k(s,a)}^{1-v}\log(\frac{2SAT}{\delta})}{N_h^k(s,a)\vee 1}} + 16u^{\frac{1}{1+v}}\left(\frac{H\log(6SAT/\delta)}{\epsilon\sqrt{(\widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1})\vee 1}}\right)^{\frac{v}{1+v}}.$$

□

Proof of Theorem 3. Following the same idea in the proof of Theorem 2, we need to calculate $\beta_h^{k,r}(s,a)$ as shown in Lemma 17. By definition, we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\beta_h^{k,r}(s,a) = \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{2\tau E_{\varepsilon,\delta,1}}{\left(\widetilde{N}_h^k(s_h^k,a_h^k)+E_{\varepsilon,\delta,1}\right)\vee 1}}_{\mathcal{O}_1} + \underbrace{16u^{\frac{1}{1+v}}\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\frac{H\log(6SAT/\delta)}{\epsilon\sqrt{(\widetilde{N}_h^k(s_h^k,a_h^k)+E_{\varepsilon,\delta,1})\vee 1}}\right)^{\frac{v}{1+v}}}_{\mathcal{O}_2},$$

The first term can be upper bounded as follows ($T=KH$) under assumption 1

$$\mathcal{O}_1 = \widetilde{O}\left(HSAE_{\epsilon,\delta,1}\right).$$

where $\widetilde{O}(\cdot)$ hides $polylog(S,A,T,1/\delta)$ factors.

The second term can be upper bounded as follows under Assumption 15.

$$\begin{aligned}
\mathcal{O}_2 &\le cu^{\frac{1}{1+v}}\left(\frac{H\log(6SAT/\delta)}{\epsilon}\right)^{\frac{v}{1+v}}\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\frac{1}{\sqrt{N_h^k(s_h^k,a_h^k)\vee 1}}\right)^{\frac{v}{1+v}}\\
&= cu^{\frac{1}{1+v}}\left(\frac{H\log(6SAT/\delta)}{\epsilon}\right)^{\frac{v}{1+v}}\sum_{h,s,a}\sum_{i=1}^{N_h^K(s,a)}\frac{1}{i^{\frac{v}{2(1+v)}}}\\
&\le c'u^{\frac{1}{1+v}}\left(\frac{H\log(6SAT/\delta)}{\epsilon}\right)^{\frac{v}{1+v}}\sum_{h,s,a}(N_h^K(s,a))^{\frac{2+v}{2(1+v)}}\\
&\overset{(a)}{\le} c'u^{\frac{1}{1+v}}\left(\frac{H\log(3SAT/\delta)}{\epsilon}\right)^{\frac{v}{1+v}}\left(\sum_{h,s,a}1\right)^{\frac{v}{2(1+v)}}\left(\sum_{h,s,a}N_h^K(s,a)\right)^{\frac{2+v}{2(1+v)}}\\
&\le \widetilde{O}\left(u^{\frac{1}{(1+v)}}\left(\frac{H^3SA}{\epsilon^2}\right)^{\frac{v}{2(1+v)}}T^{\frac{2+v}{2(1+v)}}\right)
\end{aligned}$$

where (a) is based on Hölder's inequality with $x_k=1, y_k=(N_h^K(s,a))^{\frac{2+v}{2(1+v)}}, q=\frac{2(1+v)}{2+v}$ in Lemma 9.

Hence,

$$Reg(T) = \widetilde{O}\left(\sqrt{SAH^3T}+S^2AH^3+S^2AH^2E_{\epsilon,\delta,1}+S^2AH^2E_{\epsilon,\delta,3}+u^{\frac{1}{(1+v)}}\left(\frac{H^3SA}{\epsilon^2}\right)^{\frac{v}{2(1+v)}}T^{\frac{2+v}{2(1+v)}}\right).$$

Plugging the value of error bound in Lemma 2 to the regret bound, we obtain

$$Reg(T) = \widetilde{O}\left(\sqrt{SAH^3T}+\frac{S^2A\sqrt{H^5T}}{\epsilon}+u^{\frac{1}{(1+v)}}\left(\frac{H^3SA}{\epsilon^2}\right)^{\frac{v}{2(1+v)}}T^{\frac{2+v}{2(1+v)}}\right).$$

□

D. Algorithm and Proofs of Section 4

D.1. Proof of Theorem 5

Before showing the proof of Theorem 5, we first prove the following lemma.

Algorithm 3 Private-Heavy-UCBPO

Require: Number of episodes K, time horizon H, privacy level $\epsilon > 0$, a PRIVATIZER (LOCAL or CENTRAL), confidence level $\delta \in (0, 1]$ and parameter $\eta > 0$

Initialize policy $\pi_h^1(a|s) = 1/A$ for all (s, a, h)
Initialize private counts $\widetilde{R}_h^1(s, a) = 0$, $\widetilde{N}_h^1(s, a) = 0$ and $\widetilde{N}_h^1(s, a, s') = 0$ for all (s, a, s', h)
Set precision levels $E_{\epsilon,\delta,1}, E_{\epsilon,\delta,2}, E_{\epsilon,\delta,3}$ of the PRIVATIZER
for $k = 1, 2, 3, \ldots, K$ **do**
  Initialize private value estimates: $\widetilde{V}_{H+1}^k(s) = 0$
  for $h = H, H-1, \ldots, 1$ **do**
    Compute $\widetilde{r}_h^k(s, a)$ and $\widetilde{P}_h^k(s, a)$ $\forall(s, a)$ as in (2) using the private counts
    Set exploration bonus using Lemma 18: $\beta_h^k(s, a) = \beta_h^{k,r}(s, a) + \tau H \beta_h^{k,p}(s, a)$ $\forall(s, a)$
    Compute: $\forall(s, a)$,

$$\widetilde{Q}_h^k(s, a) = \max\{-(H - h + 1)\tau, \min\{(H - h + 1)\tau, \widetilde{r}_h^k(s, a) + \sum_{s' \in \mathcal{S}} \widetilde{V}_{h+1}^k(s')\widetilde{P}_h^k(s'|s, a) + \beta_h^k(s, a)\}\}$$

    Compute private value estimates: $\forall s$, $\widetilde{V}_h^k(s) = \sum_{a \in \mathcal{A}} \widetilde{Q}_h^k(s, a)\pi_h^k(a|s)$
  end for
  Roll out a trajectory $(s_1^k, a_1^k, r_1^k, \ldots, s_{H+1}^k)$ by acting the policy $\pi^k = (\pi_h^k)_{h=1}^H$
  Receive private counts $\widetilde{R}_h^{k+1}(s, a)$, $\widetilde{N}_h^{k+1}(s, a)$, $\widetilde{N}_h^{k+1}(s, a, s')$ from the PRIVATIZER
  Update policy: $\forall(s, a, h)$, $\pi_h^{k+1}(a|s) = \frac{\pi_h^k(a|s)\exp(-\eta \widetilde{Q}_h^k(s,a))}{\sum_{a \in \mathcal{A}} \pi_h^k(a|s)\exp(-\eta \widetilde{Q}_h^k(s,a))}$
end for

Lemma 18 (Concentration bounds of private estimators)**.** *Fix any $\epsilon \in (0, 1]$ and $\delta \in (0, 1)$ and take $B_n = \left(\frac{\epsilon u n}{H \log^{1.5} K \log(3SAT/\delta)}\right)^{\frac{1}{1+v}}$ in equation (1). Then, under Assumption 1, with probability at least $1 - 2\delta$, uniformly over all (s, a, h, k),*

$$|\widetilde{r}_h^k(s, a) - r_h(s, a)| \le \beta_h^{k,r}(s, a),\ \|P_h(\cdot|s, a) - \widetilde{P}_h^k(\cdot|s, a)\|_1 \le \beta_h^{k,p}(s, a),$$

where

$$\beta_h^{k,r}(s, a) = \frac{2\tau E_{\varepsilon,\delta,1}}{\left(\widetilde{N}_h^k(s, a) + E_{\varepsilon,\delta,1}\right) \vee 1} + 10u^{\frac{1}{1+v}} \left(\frac{H \log^{1.5} K \log(3SAT/\delta)}{\epsilon\left((\widetilde{N}_h^k(s, a) + E_{\varepsilon,\delta,1}) \vee 1\right)}\right)^{\frac{v}{1+v}},$$

$$\beta_h^{k,p}(s, a) := \frac{\sqrt{4S \ln(6AT/\delta)}}{\sqrt{(\widetilde{N}_h^k(s, a) + E_{\varepsilon,\delta,1}) \vee 1}} + \frac{SE_{\epsilon,\delta,3} + 2E_{\epsilon,\delta,1}}{(\widetilde{N}_h^k(s, a) + E_{\varepsilon,\delta,1}) \vee 1}.$$

Proof of Lemma 18. $\beta_h^{k,r}(s, a)$ is defined in Lemma 15. Now we prove the concentrated upper bound for transition probability. From Theorem 2.1 in (Weissman et al., 2003) and union bound over all s, a, h, k, we obtain with probability at least $1 - \delta/2$

$$\|P_h(\cdot|s, a) - \bar{P}_h^k(\cdot|s, a)\|_1 \le \sqrt{\frac{4S \ln(6AT/\delta)}{N_h^k(s, a) \vee 1}} \tag{23}$$

where $\bar{P}_h^k(s'|s, a) := \frac{N_h^k(s,a,s')}{N_h^k(s,a) \vee 1}$ is non-private empirical transition probability.

Now, we turn to bound the transition dynamics. The error between the true transition probability and the private estimate

can be decomposed as

$$\begin{aligned}
&\sum_{s'} |P_h(s'|,s,a) - \widetilde{P}_h^k(s'|s,a)| \\
=&\sum_{s'} \left| \frac{\widetilde{N}_h^k(s,a,s')}{(\widetilde{N}_h^k(s,a)+E_{\epsilon,\delta,1})\vee 1} - P_h(s'|s,a) \right| \\
\leq& \underbrace{\sum_{s'} \left| \frac{N_h^k(s,a,s')}{(\widetilde{N}_h^k(s,a)+E_{\epsilon,\delta,1})\vee 1} - P_h(s'|s,a) \right|}_{\mathcal{P}_1} + \underbrace{\sum_{s'} \left| \frac{\widetilde{N}_h^k(s,a,s') - N_h^k(s,a,s')}{(\widetilde{N}_h^k(s,a)+E_{\epsilon,\delta,1})\vee 1} \right|}_{\mathcal{P}_2}.
\end{aligned}$$

For $\mathcal{P}_1$, we have

$$\begin{aligned}
\mathcal{P}_1 =& \sum_{s'} \left| \frac{N_h^k(s,a,s')}{N_h^k(s,a)\vee 1} \frac{N_h^k(s,a)\vee 1}{(\widetilde{N}_h^k(s,a)+E_{\epsilon,\delta,1})\vee 1} - P_h(s'|s,a) \right| \\
=& \sum_{s'} \left| \left(\frac{N_h^k(s,a,s')}{N_h^k(s,a)\vee 1} - P_h(s'|s,a) \right) \frac{N_h^k(s,a)\vee 1}{(\widetilde{N}_h^k(s,a)+E_{\epsilon,\delta,1})\vee 1} + P_h(s'|s,a) \left(\frac{N_h^k(s,a)\vee 1}{(\widetilde{N}_h^k(s,a)+E_{\epsilon,\delta,1})\vee 1} - 1 \right) \right| \\
\leq& \frac{N_h^k(s,a)\vee 1}{(\widetilde{N}_h^k(s,a)+E_{\epsilon,\delta,1})\vee 1} \|\bar{P}_h^k(\cdot|s,a) - P_h(\cdot|s,a)\|_1 + \sum_{s'} \left(P_h(s'|s,a) \frac{2E_{\epsilon,\delta,1}}{(\widetilde{N}_h^k(s,a)+E_{\epsilon,\delta,1})\vee 1} \right) \\
\overset{(a)}{\leq}& \frac{N_h^k(s,a)\vee 1}{(\widetilde{N}_h^k(s,a)+E_{\epsilon,\delta,1})\vee 1} \frac{\sqrt{4S\ln(6AT/\delta)}}{\sqrt{N_h^k(s,a)\vee 1}} + \frac{2E_{\epsilon,\delta,1}}{(\widetilde{N}_h^k(s,a)+E_{\epsilon,\delta,1})\vee 1} \\
\leq& \frac{\sqrt{4S\ln(6AT/\delta)}}{\sqrt{(\widetilde{N}_h^k(s,a)+E_{\epsilon,\delta,1})\vee 1}} + \frac{2E_{\epsilon,\delta,1}}{(\widetilde{N}_h^k(s,a)+E_{\epsilon,\delta,1})\vee 1},
\end{aligned}$$

where (a) holds by the concentration of transition probability in inequality (23). For $\mathcal{P}_2$, we have

$$\mathcal{P}_2 \leq \sum_{s'} \frac{|E_{\epsilon,\delta,3}|}{(\widetilde{N}_h^k(s,a)+E_{\epsilon,\delta,1})\vee 1} = \frac{SE_{\epsilon,\delta,3}}{(\widetilde{N}_h^k(s,a)+E_{\epsilon,\delta,1})\vee 1}.$$

Putting together $\mathcal{P}_1$ and $\mathcal{P}_2$, yields

$$\|P_h(\cdot|s,a) - \widetilde{P}_h^k(\cdot|s,a)\|_1 \leq \frac{\sqrt{4S\ln(6AT/\delta)}}{\sqrt{(\widetilde{N}_h^k(s,a)+E_{\epsilon,1})\vee 1}} + \frac{SE_{\epsilon,\delta,3}+2E_{\epsilon,\delta,1}}{(\widetilde{N}_h^k(s,a)+E_{\epsilon,\delta,1})\vee 1}.$$

□

Proof of Theorem 5. we first decompose the regret by using the extended value difference lemma (Shani et al., 2020, Lemma 1).

$$\begin{aligned}
Reg(T) =& \sum_{k=1}^K \left(V_1^{\pi^*}(s_1^k) - V_1^{\pi^k}(s_1^k) \right) = \sum_{k=1}^K \left(V_1^{\pi^*}(s_1^k) - \widetilde{V}_1^k(s_1^k) + \widetilde{V}_1^k(s_1^k) - V_1^{\pi^k}(s_1^k) \right) \\
=& \underbrace{\sum_{k=1}^K \sum_{h=1}^H \mathbb{E}\left[\langle \widetilde{Q}_h^k(s_h,\cdot), \pi_h^*(\cdot|s_h) - \pi_h^k(\cdot|s_h) \rangle | s_1^k, \pi^* \right]}_{\mathcal{T}_1} \\
&+ \underbrace{\sum_{k=1}^K \sum_{h=1}^H \mathbb{E}\left[r_h(s_h,a_h) + P_h(\cdot|s_h,a_h)\widetilde{V}_{h+1}^k - \widetilde{Q}_h^k(s_h,a_h) | s_1^k, \pi^* \right]}_{\mathcal{T}_2} \\
&+ \underbrace{\sum_{k=1}^K \left(\widetilde{V}_1^k(s_1^k) - V_1^{\pi^k}(s_1^k) \right)}_{\mathcal{T}_3}.
\end{aligned}$$

We then need to bound each of the three terms.

Analysis of $\mathcal{T}_1$. To start with, we can bound $\mathcal{T}_1$ by following standard mirror descent analysis under KL divergence. Specifically, by (Orabona, 2023, Lemma 6.7), we have for any $h \in [H]$, $s \in \mathcal{S}$ and any policy π

$$\begin{aligned}\sum_{k=1}^{K}\langle \widetilde{Q}_h^k(s,\cdot), \pi_h^*(\cdot|s) - \pi_h^k(\cdot|s)\rangle &\le \frac{\log A}{\eta} + \frac{\eta}{2}\sum_{k=1}^{K}\|\widetilde{Q}_h^k(s,a))\|_\infty^2 \\ &\overset{(a)}{\le} \frac{\log A}{\eta} + \frac{\eta\tau^2 H^2 K}{2},\end{aligned}$$

where (a) holds by $\widetilde{Q}_h^k(s,a) \in [-\tau H, \tau H]$ for any $a \in \mathcal{A}$, which follows from the truncated update of Q-value in Algorithm 3 (line 10). Thus, we can bound $\mathcal{T}_1$ as follows.

$$\mathcal{T}_1 = \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[\langle \widetilde{Q}_h^k(s_h^k,\cdot), \pi_h^*(\cdot|s_h^k) - \pi_h^k(\cdot|s_h^k)\rangle | s_1^k, \pi^*\right] \le \frac{H\log A}{\eta} + \frac{\eta\tau^2 H^3 K}{2}.$$

Choosing $\eta = \sqrt{2\log A/(\tau^2 H^2 K)}$, yields

$$\mathcal{T}_1 \le \sqrt{2\tau^2 H^4 K \log A}. \tag{24}$$

Analysis of $\mathcal{T}_2$. First, by the update rule of Q-value in Algorithm 3 and $P_h(\cdot|s,a)V_{h+1} := \sum_{s'} P_h(s'|s,a)V_{h+1}(s')$, we have

$$\begin{aligned}\widetilde{Q}_h^k(s,a) &= \max\{-(H-h+1)\tau, \min\{(H-h+1)\tau, \widetilde{r}_h^k(s,a) + \sum_{s'\in\mathcal{S}}\widetilde{V}_{h+1}^k(s')\widetilde{P}_h^k(s'|s,a) + \beta_h^k(s,a)\}\} \\ &= \max\{-(H-h+1)\tau, \min\{(H-h+1)\tau, \widetilde{r}_h^k(s,a) + \widetilde{P}_h(\cdot|s_h^k,a_h^k)\widetilde{V}_{h+1}^k + \beta_h^k(s,a)\}\} \\ &\le \max\left\{-(H-h+1)\tau, \widetilde{r}_h^k(s,a) + \beta_h^{k,r}(s,a) + \widetilde{P}_h(\cdot|s_h^k,a_h^k)\widetilde{V}_{h+1}^k + \tau H\beta_h^{k,p}(s,a)\right\} \\ &\le \max\left\{0, \widetilde{r}_h^k(s,a) + \beta_h^{k,r}(s,a) + \widetilde{P}_h(\cdot|s_h^k,a_h^k)\widetilde{V}_{h+1}^k + \tau H\beta_h^{k,p}(s,a)\right\} \\ &\overset{(a)}{\le} \max\left\{0, \widetilde{r}_h^k(s,a) + \beta_h^{k,r}(s,a)\right\} + \max\left\{0, \widetilde{P}_h(\cdot|s_h^k,a_h^k)\widetilde{V}_{h+1}^k + \tau H\beta_h^{k,p}(s,a)\right\}\end{aligned}$$

where (a) holds since for any a, b, $\max\{a+b, 0\} \le \max\{a,0\} + \max\{b,0\}$. Thus, for any (k,h,s,a), we have

$$\begin{aligned}&r_h(s,a) + P_h(\cdot|s,a)\widetilde{V}_{h+1}^k - \widetilde{Q}_h^k(s,a) \\ \le& r_h(s,a) + P_h(\cdot|s,a)\widetilde{V}_{h+1}^k - \max\left\{0, \widetilde{r}_h^k(s,a) + \beta_h^{k,r}(s,a)\right\} - \max\left\{0, \widetilde{P}_h^k(\cdot|s,a)\widetilde{V}_h^k + \tau H\beta_h^{k,p}(s,a)\right\} \\ =& r_h(s,a) + P_h(\cdot|s,a)\widetilde{V}_{h+1}^k + \min\left\{0, -\widetilde{r}_h^k(s,a) - \beta_h^{k,r}(s,a)\right\} + \min\left\{0, -\widetilde{P}_h^k(\cdot|s,a)\widetilde{V}_h^k - \tau H\beta_h^{k,p}(s,a)\right\} \\ =& \min\left\{r_h(s,a), r_h(s,a) - \widetilde{r}_h^k(s,a) - \beta_h^{k,r}(s,a)\right\} \qquad (25) \\ &+ \min\left\{P_h(\cdot|s,a)\widetilde{V}_{h+1}^k, P_h(\cdot|s,a)\widetilde{V}_{h+1}^k - \widetilde{P}_h^k(\cdot|s,a)\widetilde{V}_h^k - \tau H\beta_h^{k,p}(s,a)\right\}. \qquad (26)\end{aligned}$$

We are going to show that both (25) and (26) are less than zero for all (k,h,s,a) with high probability by Lemma 18. First, conditioned on the first result in Lemma 18, we have

$$r_h(s,a) - \widetilde{r}_h^k(s,a) - \beta_h^{k,r}(s,a) \le 0,$$

so (25) is less than zero. Further, we have conditioned on the second result in Lemma 18

$$\begin{aligned}&P_h(\cdot|s,a)\widetilde{V}_{h+1}^k - \widetilde{P}_h^k(\cdot|s,a)\widetilde{V}_h^k - \tau H\beta_h^{k,p}(s,a) \\ \overset{(a)}{\le}& \|\widetilde{P}_h^k(\cdot|s,a) - P_h(\cdot|s,a)\|_1\|\widetilde{V}_{h+1}^k\|_\infty - \tau H\beta_h^{k,p}(s,a) \\ \overset{(b)}{\le}& \tau H\|\widetilde{P}_h^k(\cdot|s,a) - P_h(\cdot|s,a)\|_1 - \tau H\beta_h^{k,p}(s,a) \\ \overset{(c)}{\le}& 0\end{aligned} \tag{27}$$

where (a) holds by Holder's inequality; (b) holds since $-\tau H \leq \widetilde{V}_{h+1}^k \leq \tau H$ based on our update rule; (c) holds by Lemma 18, so (26) is less than 0. Thus, we have shown that

$$\mathcal{T}_2 \leq 0. \tag{28}$$

Analysis of $\mathcal{T}_3$. Assume the event in Assumption 1 hold (which implies the concentration results in Lemma 18). We have

$$\begin{aligned}
\mathcal{T}_3 =& \sum_{k=1}^{K}\left(\widetilde{V}_1^k(s_1) - V_1^{\pi_k}(s_1)\right) \\
\overset{(a)}{=}& \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[\widetilde{Q}_{h+1}^k(s_h,a_h) - r_h(s_h,a_h) - P_h(\cdot|s_h,a_h)\widetilde{V}_{h+1}^k | s_1^k, \pi_k\right] \\
\overset{(b)}{=}& \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[\min\left\{\widetilde{r}_h^k(s_h,a_h) + \beta_h^{k,r}(s_h,a_h) + \widetilde{P}_h^k(\cdot|s_h,a_h)\widetilde{V}_{h+1}^k + \tau H\beta_h^{k,p}(s_h,a_h), (H-h+1)\tau\right\} | s_1^k, \pi_k\right] \\
&- \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[r_h(s_h,a_h) + P_h(\cdot|s_h,a_h)\widetilde{V}_{h+1}^k | s_1^k, \pi_k\right]
\end{aligned} \tag{29}$$

where (a) holds by the extended value difference lemma (Shani et al., 2020, Lemma 1); (b) holds by the update rue of Q-value in Algorithm 3. Note that here we can directly remove the truncation at $-(H-h+1)\tau$ since by Lemma 18, $\widetilde{r}_h^k(s_h,a_h) + \beta_h^{k,r}(s_h,a_h) + \widetilde{p}(\cdot|s_h,a_h)\widetilde{V}_{h+1}^k + \tau H\beta_h^{k,p}(s_h,a_h) \geq r_h(s,a) + P_h(\cdot|s,a)\widetilde{V}_{h+1}^k \geq -(1+H-h)\tau$.

Now, observe that for any (k,h,s,a), we have

$$\begin{aligned}
&\min\left\{\widetilde{r}_h^k(s_h,a_h) + \beta_h^{k,r}(s,a) + \widetilde{P}_h^k(\cdot|s,a)\widetilde{V}_{h+1}^k + \tau H\beta_h^{k,p}(s,a), (H-h+1)\tau\right\} - r_h(s,a) - P_h(\cdot|s,a)\widetilde{V}_{h+1}^k \\
&\leq \widetilde{r}_h^k(s,a) - r_h(s,a) + \beta_h^{k,r}(s,a) + \widetilde{P}_h^k(\cdot|s,a)\widetilde{V}_{h+1}^k - P_h(\cdot|s,a)\widetilde{V}_{h+1}^k + \tau H\beta_h^{k,p}(s,a) \\
&\overset{(a)}{\leq} 2\beta_h^{k,r}(s,a) + 2\tau H\beta_h^{k,p}(s,a),
\end{aligned} \tag{30}$$

where (a) holds by Lemma 18 and a similar analysis as in (27). Plugging (30) into (29), yields

$$\mathcal{T}_3 \leq \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[2\beta_h^{k,r}(s_h,a_h)|s_1^k,\pi^k\right]}_{\text{Term(i)}} + \underbrace{\tau H\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[2\beta_h^{k,p}(s_h,a_h)|s_1^k,\pi^k\right]}_{\text{Term(ii)}} \tag{31}$$

By the definition of $\beta_h^{k,r}$ and $\beta_h^{k,p}$ in Lemma 18 and Assumption 1, we have with probability $1-2\delta$,

$$\text{Term(i)} \leq 2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[\frac{2\tau E_{\varepsilon,\delta,1}}{N_h^k(s,a)\vee 1} + 10u^{\frac{1}{1+v}}\left(\frac{H\log^{1.5}K\log(3SAT/\delta)}{\epsilon\left(N_h^k(s,a)\vee 1\right)}\right)^{\frac{v}{1+v}}\right], \tag{32}$$

$$\text{Term(ii)} \leq 2\tau H\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[\frac{\sqrt{4S\ln(6AT/\delta)}}{\sqrt{N_h^k(s,a)\vee 1}} + \frac{SE_{\epsilon,\delta,3} + 2E_{\epsilon,\delta,1}}{N_h^k(s,a)\vee 1}\right]. \tag{33}$$

In order to bound the two terms above, we use the following lemmas.

Lemma 19. *With probability $1-2\delta$, we have*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[\frac{1}{N_h^k(s_h,a_h)\vee 1}|\mathcal{F}_{k-1}\right] = O\left(SAH\ln(KH) + H\ln(H/\delta)\right),$$

and

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[\frac{1}{\sqrt{N_h^k(s_h,a_h)\vee 1}}|\mathcal{F}_{k-1}\right]=O\left(\sqrt{SAH^2K}+SAH\ln KH+H\ln(H/\delta)\right),$$

where the filtration $\mathcal{F}_k$ includes all the events until the end of episode k.

The results of Lemma 19 have been proved in Lemma A.2 of (Chowdhury & Zhou, 2021).

In order to bound (32), we use the following standard Bernstein-type concentration inequality for martingale from Lemma 9 in (Jin et al., 2020).

Lemma 20. *Let $Y_1,\dots,Y_K$ be a martingale difference sequence with respect to a filtration $\mathcal{F}_0,\mathcal{F}_1,\dots,\mathcal{F}_K$. Assume $Y_k\le R$ a.s. for all i. Then, for any $\delta\in(0,1)$ and $\lambda\in[0,1/R]$, with probability $1-\delta$, we have*

$$\sum_{k=1}^{K}Y_k\le\lambda\sum_{k=1}^{K}\mathbb{E}\left[Y_k^2|\mathcal{F}_{k-1}\right]+\frac{\ln(1/\delta)}{\lambda}.$$

Now we can use the above lemma to prove the following lemma which is the key point to bound (32).

Lemma 21. *With probability $1-\delta$, we have*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[\frac{1}{(N_h^k(s_h,a_h)\vee 1)^{\frac{v}{1+v}}}|\mathcal{F}_{k-1}\right]=O\left((SAH)^{\frac{v}{1+v}}T^{\frac{1}{1+v}}+H\ln(H/\delta)\right),$$

where the filtration $\mathcal{F}_k$ includes all the events until the end of episode k.

Proof of Lemma 21. Let $\mathcal{I}_h^k(s,a)$ be the indicator of whether the pair (s,a) at step h and episode k so that $\mathbb{E}\left[\mathcal{I}_h^k(s,a)|\mathcal{F}_{k-1}\right]=w_h^k(s,a)$, which is the probability of visiting state-action pair (s,a) at step h and episode k. First note that

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[\frac{1}{(N_h^k(s_h,a_h)\vee 1)^{\frac{v}{1+v}}}|\mathcal{F}_{k-1}\right]\\
=&\sum_{k=1}^{K}\sum_{h,s,a}w_h^k(s,a)\frac{1}{(N_h^k(s,a)\vee 1)^{\frac{v}{1+v}}}\\
=&\sum_{k=1}^{K}\sum_{h,s,a}\frac{\mathcal{I}_h^k(s,a)}{(N_h^k(s,a)\vee 1)^{\frac{v}{1+v}}}+\sum_{k=1}^{K}\sum_{h,s,a}\frac{w_h^k(s,a)-\mathcal{I}_h^k(s,a)}{(N_h^k(s,a)\vee 1)^{\frac{v}{1+v}}}.
\end{aligned}$$

The first term can be bounded as follows.

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h,s,a}\frac{\mathcal{I}_h^k(s,a)}{(N_h^k(s,a)\vee 1)^{\frac{v}{1+v}}}&\le\sum_{h,s,a}\sum_{k=1}^{K}\frac{1}{(N_h^k(s,a)\vee 1)^{\frac{v}{1+v}}}\\
&=\sum_{h,s,a}\sum_{i=1}^{N_h^K(s,a)}\frac{1}{i^{\frac{v}{1+v}}}\\
&\le c'\sum_{h,s,a}\left(N_h^K(s,a)\right)^{\frac{1}{1+v}}\\
&\overset{(a)}{\le}\left(\sum_{h,s,a}1\right)^{\frac{v}{1+v}}\left(\sum_{h,s,a}N_h^K(s,a)\right)^{\frac{1}{1+v}}\\
&=O\left((SAH)^{\frac{v}{1+v}}T^{\frac{1}{1+v}}\right).
\end{aligned}$$

To bound the second term, we will use Lemma 20. In particular, consider $Y_{k,h} := \sum_{s,a} \frac{w_h^k(s,a)-\mathcal{I}_h^k(s,a)}{(N_h^k(s,a)\vee 1)^{\frac{v}{1+v}}} \leq 1$, $\lambda = 1$, and the fact that for any fixed h,

$$\begin{aligned}
\mathbb{E}\left[Y_{k,h}^2 | \mathcal{F}_{k-1}\right] &\leq \mathbb{E}\left[\left(\sum_{s,a} \frac{\mathcal{I}_h^k(s,a)}{(N_h^k(s,a)\vee 1)^{\frac{v}{1+v}}}\right)^2 \mid \mathcal{F}_{k-1}\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\sum_{s,a} \frac{\mathcal{I}_h^k(s,a)}{(N_h^k(s,a)\vee 1)^{\frac{2v}{1+v}}} \mid \mathcal{F}_{k-1}\right] \\
&\leq \sum_{s,a} \frac{w_h^k(s,a)}{(N_h^k(s,a)\vee 1)^{\frac{v}{1+v}}}.
\end{aligned}$$

where (a) is based on $\mathcal{I}_h^k(s,a)\mathcal{I}_h^k(s',a') = 0$ for $s \neq s'$ or $a \neq a'$. Then, via Lemma 20, we have with probability at least $1-\delta$,

$$\begin{aligned}
\sum_{k=1}^K \sum_{h,s,a} \frac{w_h^k(s,a)-\mathcal{I}_h^k(s,a)}{(N_h^k(s,a)\vee 1)^{\frac{v}{1+v}}} = \sum_{h=1}^H \sum_{k=1}^K Y_{k,h} &\leq \sum_{h=1}^H \sum_{k=1}^K \sum_{s,a} \frac{w_h^k(s,a)}{(N_h^k(s,a)\vee 1)^{\frac{v}{1+v}}} + H\ln(H/\delta) \\
&= O\left((SAH)^{\frac{v}{1+v}} T^{\frac{1}{1+v}} + H\ln(H/\delta)\right),
\end{aligned}$$

Then we complete the proof of the lemma. □

From Lemma 19 and Lemma 21, we can get the upper bounds for Term(i) and Term (ii):

$$\text{Term(i)} = \widetilde{O}\left(SAHE_{\epsilon,\delta,1} + u^{\frac{1}{1+v}}\left(\frac{SAH^2}{\epsilon}\right)^{\frac{v}{1+v}} T^{\frac{1}{1+v}}\right) \tag{34}$$

$$\text{Term(ii)} = \widetilde{O}\left(\sqrt{S^2AH^4K} + \sqrt{S^3A^2H^4} + E_{\epsilon,\delta,3}S^2AH^2 + E_{\epsilon,\delta,1}SAH^2\right) \tag{35}$$

Hence, $\mathcal{T}_3 = \widetilde{O}\left(\sqrt{S^2AH^3T} + \frac{S^2AH^3}{\epsilon} + u^{\frac{1}{1+v}}\left(\frac{SAH^2}{\epsilon}\right)^{\frac{v}{1+v}} T^{\frac{1}{1+v}}\right)$

Finally, we can get the upper bound of regret. □

Lemma 22 (Concentration bounds of locally private estimators)**.** *Fix any $\epsilon \in (0,1]$ and $\delta \in (0,1)$ and take $B_n = \left(\frac{u\epsilon\sqrt{n}}{H\log(6SAT/\delta)}\right)^{\frac{1}{1+v}}$ in equation (1). Then, under Assumption 1, with probability at least $1-2\delta$, uniformly over all (s,a,h,k),*

$$|\widetilde{r}_h^k(s,a) - r_h(s,a)| \leq \beta_h^{k,r}(s,a),\ \|P_h(\cdot|s,a) - \widetilde{P}_h^k(\cdot|s,a)\|_1 \leq \beta_h^{k,p}(s,a),$$

where

$$\begin{aligned}
\beta_h^{k,r}(s,a) &= \frac{2\tau E_{\varepsilon,\delta,1}}{\left(\widetilde{N}_h^k(s,a) + E_{\varepsilon,\delta,1}\right)\vee 1} + 16u^{\frac{1}{1+v}}\left(\frac{H\log(6SAT/\delta)}{\epsilon\sqrt{(\widetilde{N}_h^k(s,a) + E_{\varepsilon,\delta,1})\vee 1}}\right)^{\frac{v}{1+v}}, \\
\beta_h^{k,p}(s,a) &:= \frac{\sqrt{4S\ln(6AT/\delta)}}{\sqrt{(\widetilde{N}_h^k(s,a) + E_{\varepsilon,\delta,1})\vee 1}} + \frac{SE_{\epsilon,\delta,3} + 2E_{\epsilon,\delta,1}}{(\widetilde{N}_h^k(s,a) + E_{\varepsilon,\delta,1})\vee 1}.
\end{aligned}$$

In fact, $\beta_h^{k,r}(s,a)$ is the same as the form defined in Lemma 17 since we use the same mean estimation for truncated heavy-tailed rewards in the LDP model. Moreover, $\beta_h^{k,p}(s,a)$ is the same as the one in Lemma 18.

D.2. Proof of Theorem 6

Proof of Theorem 6. Similar to the idea of Theorem 5's proof, we first decompose the regret by using the extended value difference lemma (Shani et al., 2020, Lemma 1).

$$\begin{aligned}
Reg(T) &= \sum_{k=1}^{K}\left(V_1^{\pi^*}(s_1^k) - V_1^{\pi^k}(s_1^k)\right) = \sum_{k=1}^{K}\left(V_1^{\pi^*}(s_1^k) - \widetilde{V}_1^k(s_1^k) + \widetilde{V}_1^k(s_1^k) - V_1^{\pi^k}(s_1^k)\right) \\
&= \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[\langle \widetilde{Q}_h^k(s_h,\cdot), \pi_h^*(\cdot|s_h) - \pi_h^k(\cdot|s_h)\rangle | s_1^k, \pi^*\right]}_{\mathcal{T}_1} \\
&\quad + \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[r_h(s_h,a_h) + P_h(\cdot|s_h,a_h)\widetilde{V}_{h+1}^k - \widetilde{Q}_h^k(s_h,a_h)|s_1^k,\pi^*\right]}_{\mathcal{T}_2} \\
&\quad + \underbrace{\sum_{k=1}^{K}\left(\widetilde{V}_1^k(s_1^k) - V_1^{\pi^k}(s_1^k)\right)}_{\mathcal{T}_3}.
\end{aligned}$$

We then need to bound each of the three terms.

By (Shani et al., 2020, Lemma 17) and choosing $\eta = \sqrt{2\log A/(\tau^2 H^2 K)}$, we obtain $\mathcal{T}_1 \le \sqrt{2\tau^2 H^4 K \log A}$.. Furthermore, due to update rule of Q-function and Lemma 18, we have $\mathcal{T}_2 \le 0$.

Now we focus on bounding $\mathcal{T}_3$. By the extended value difference lemma (Shani et al., 2020, Lemma 1), the update rue of Q-value in Algorithm 3 and the bonus term according to Lemma 18 , we can decompose the term into two parts:

$$\mathcal{T}_3 \le \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[2\beta_h^{k,r}(s_h,a_h)|s_1^k,\pi^k\right]}_{\text{Term(i)}} + \underbrace{\tau H\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[2\beta_h^{k,p}(s_h,a_h)|s_1^k,\pi^k\right]}_{\text{Term(ii)}} \tag{36}$$

where

$$\text{Term(i)} \le 2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[\frac{2\tau E_{\varepsilon,\delta,1}}{N_h^k(s,a)\vee 1} + 16u^{\frac{1}{1+v}}\left(\frac{H\log(6SAT/\delta)}{\epsilon\sqrt{N_h^k(s,a)\vee 1}}\right)^{\frac{v}{1+v}}\right], \tag{37}$$

$$\text{Term(ii)} \le 2\tau H\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[\frac{\sqrt{4S\ln(6AT/\delta)}}{\sqrt{N_h^k(s,a)\vee 1}} + \frac{SE_{\epsilon,\delta,3} + 2E_{\epsilon,\delta,1}}{N_h^k(s,a)\vee 1}\right]. \tag{38}$$

From the same argument in the proof of Theorem 5, we can obtain

$$\text{Term(ii)} = \widetilde{O}\left(\sqrt{S^2AH^4K} + \sqrt{S^3A^2H^4} + E_{\epsilon,\delta,3}S^2AH^2 + E_{\epsilon,\delta,1}SAH^2\right).$$

Then the only thing left is to bound the Term(i). Using a similar idea in the proof of Lemma 21 we can get

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[\frac{1}{(N_h^k(s_h,a_h)\vee 1)^{\frac{v}{2(1+v)}}}|\mathcal{F}_{k-1}\right] = O\left((SAH)^{\frac{v}{2(1+v)}}T^{\frac{2+v}{2(1+v)}} + H\ln(H/\delta)\right),$$

Based on the first result in Lemma 19, we have

$$\text{Term(i)} = \widetilde{O}\left(SAHE_{\epsilon,\delta,1} + u^{\frac{1}{1+v}}\left(\frac{SAH^3}{\epsilon^2}\right)^{\frac{v}{2(1+v)}}T^{\frac{2+v}{2(1+v)}}\right).$$

Finally, based on the results of Lemma 2, we can derive the result of regret:

$$Reg(T) = \widetilde{O}\left(\sqrt{S^2AH^3T} + \frac{S^2A\sqrt{H^5T}}{\epsilon} + u^{\frac{1}{1+v}}\left(\frac{H^3SA}{\epsilon^2}\right)^{\frac{v}{2(1+v)}}T^{\frac{2+v}{2(1+v)}}\right).$$

□

E. Proofs of Section 5

E.1. Proof of Theorem 7

Proof of Theorem 7. Firstly, we construct the environments which are hard to distinguish. We define the instance $\bar{P}_1$ in which the optimal arm (denote by a_1) follows the reward distribution

$$\nu_1 = \left(1 - \frac{\gamma^{1+v}}{2}\right)\delta_0 + \frac{\gamma^{1+v}}{2}\delta_{1/\gamma},$$

where $\gamma = (5\Delta)^{\frac{1}{v}}$ with Δ is a constant to be specified later and $\Delta \in \left(0, \frac{1}{5}\right)$, and δ_x is the Dirac distribution on x and the distribution $p \cdot \delta_x + (1-p) \cdot \delta_y$ takes the value x with probability p and the value y with probability $1-p$. It is easy to verify that $\mathbb{E}[\nu_1] = \frac{5}{2}\Delta$, and the $(1+v)$-th raw moment of v_1 is $u(\nu_1) = \frac{1}{2} \leq 1$.

Any other sub-optimal arm $a \neq a_1$ in $\bar{P}_1$ follows the same reward distribution

$$\nu_a = \left(1 - \frac{\gamma^{1+v}}{2} + \Delta\gamma\right)\delta_0 + \left(\frac{\gamma^{1+v}}{2} - \Delta\gamma\right)\delta_{1/\gamma}.$$

Note that for all $a \neq a_1$ $\mathbb{E}[\nu_a] = \frac{3}{2}\Delta$, $u(\nu_a) = \frac{1}{2} - \frac{1}{5} = \frac{3}{10} < 1$.

For algorithm $\mathcal{M}$ and instance $\bar{P}_1$, we denote $i = \arg\min_{a \in \{2,\cdots,A\}} \mathbb{E}_{\mathcal{M}\bar{P}_1}[N_a(K)]$. where $\mathbb{E}_{\mathcal{M}P}$ is the expectation over the the probability measure $\mathbb{P}_{\mathcal{M}P}$ induced by the algorithm $\mathcal{M}$ and the instance P. Thus, $\mathbb{E}_{\mathcal{M}\bar{P}_1}[N_i(K)] \leq \frac{K}{A-1}$.

Now, consider another instance $\bar{P}_i$ where $\nu_1, \cdots, \nu_A$ are the same as those in $\bar{P}_1$ except the i-th arm such that

$$\nu_i' = \left(1 - \frac{\gamma^{1+v}}{2} - \Delta\gamma\right)\delta_0 + \left(\frac{\gamma^{1+v}}{2} + \Delta\gamma\right)\delta_{1/\gamma}.$$

Note that now $\mathbb{E}[\nu_i'] = \frac{7}{2}\Delta$, $u(\nu_i') = \frac{7}{10} < 1$. Then in $\bar{P}_i$, the arm i is optimal.

Now by the classic regret decomposition, we obtain

$$Reg^{\mathcal{M}}_{K,\bar{P}_1} = (K - \mathbb{E}_{\mathcal{M}\bar{P}_1}[N_1(K)])\Delta \geq \mathbb{P}^K_{\mathcal{M}\bar{P}_1}\left[N_1(K) \leq \frac{K}{2}\right]\frac{K\Delta}{2}.$$

$$Reg^{\mathcal{M}}_{K,\bar{P}_i} = \Delta\mathbb{E}_{\mathcal{M}\bar{P}_i}[N_1(K)] + \sum_{a \notin \{1,i\}} 2\Delta\mathbb{E}_{\mathcal{M}\bar{P}_i}[N_a(K)] \geq \mathbb{P}^K_{\mathcal{M}\bar{P}_i}\left[N_1(K) \geq \frac{K}{2}\right]\frac{K\Delta}{2}.$$

By applying the Bretagnolle–Huber inequality ((Lattimore & Szepesvári, 2020), Theorem 14.2), we have

$$\begin{aligned} Reg^{\mathcal{M}}_{K,\bar{P}_1} + Reg^{\mathcal{M}}_{K,\bar{P}_i} &\geq \frac{K\Delta}{2}\left(\mathbb{P}^K_{\mathcal{M}\bar{P}_1}\left[N_1(K) \leq \frac{K}{2}\right] + \mathbb{P}^K_{\mathcal{M}\bar{P}_i}\left[N_1(K) \geq \frac{K}{2}\right]\right). \\ &\geq \frac{K\Delta}{4}\exp\left(-\mathrm{KL}\left(\mathbb{P}^K_{\mathcal{M}\bar{P}_1} \| \mathbb{P}^K_{\mathcal{M}\bar{P}_i}\right)\right) \end{aligned}$$

Lemma 23 (Upper Bound on KL-divergence for Bandits with ϵ-DP (Azize & Basu, 2022))**.** *If $\mathcal{M}$ is a mechanism satisfying ϵ-DP, then for two instances $P_1 = (\nu_a : a \in [A])$ and $P_2 = (\nu_a' : a \in [A])$ we have*

$$KL\left(\mathbb{P}^K_{\mathcal{M}P_1} \| \mathbb{P}^K_{\mathcal{M}P_2}\right) \leq 6\epsilon\mathbb{E}_{\mathcal{M}P_1}\left[\sum_{t=1}^{K} TV(\nu_{a_t} \| \nu'_{a_t})\right]$$

where $TV(\nu_a \| \nu_a')$ is the total-variation distance between ν_a and ν_a'.

Based on the above lemma, we can get the upper bound of the KL-Divergence between the marginals.

$$\begin{aligned} \mathrm{KL}\left(\mathbb{P}^K_{\mathcal{M}\bar{P}_1} \| \mathbb{P}^K_{\mathcal{M}\bar{P}_i}\right) &\leq 6\epsilon\mathbb{E}_{\mathcal{M}P_1}\left[\sum_{t=1}^{K} \mathrm{TV}(\nu_{a_t} \| \nu'_{a_t})\right] \\ &\leq 6\epsilon\mathbb{E}_{\mathcal{M}P_1}[N_i(K)]\mathrm{TV}(\nu_i \| \nu_i') \end{aligned}$$

since $\bar{P}_1$ and $\bar{P}_i$ only differ in the arm i.

Thus,

$$Reg^{\mathcal{M}}_{K,\bar{P}_1} + Reg^{\mathcal{M}}_{K,\bar{P}_i} \geq \frac{K\Delta}{4} \exp\left(-6\epsilon \mathbb{E}_{\mathcal{M}P_1}[N_i(K)] \cdot 2\Delta\gamma\right)$$
$$\geq \frac{K\Delta}{4} \exp\left(-\frac{12 \cdot 5^{\frac{1}{v}} \epsilon K \Delta^{\frac{1+v}{v}}}{A-1}\right).$$

Taking $\Delta = \left(\frac{A-1}{K\epsilon}\right)^{\frac{v}{1+v}}$, we get the result

$$Reg^{\mathcal{M}}_{K,\bar{P}_1} \geq \Omega\left(\left(\frac{A}{\epsilon}\right)^{\frac{v}{1+v}} K^{\frac{1}{1+v}}\right).$$

□

E.2. Proof of Theorem 8

In order to give a lower bound of our problem in JDP, we first construct hard instances of MDPs as shown in Figure 1. Based on these instances and inspired by (Vietri et al., 2020), we provide the lower bound by leveraging the lower bound in the above Theorem 7. The key idea of the reduction from MDPs in JDP to MAB in DP is that we consider a setting where the initial state of each episode is public information. This means each user k will release her/his first state s_1^k in addition to sending it to the agent. Below we first define JDP algorithms for such a setting.

Definition 5 (ϵ-JDP for RL with public initial state (Vietri et al., 2020))**.** *We first define two sequences of inputs (U_K, S_1) and (U'_K, S'_1) for the RL agent as k-neighboring user-state sequences if $u_{k'} = u'_{k'}$ for all $k' \neq k$ and $S_1 = S'_1$ where $S_1 = (s_1^1, \ldots, s_1^K)$ is the sequence of initial states. Then a randomized RL mechanism $\mathcal{M}$ is ϵ-JDP under continual observation in the public initial state setting if for all $k \in [K]$, all k-neighboring user-state sequences $(U_K, S_1), (U'_K, S'_1)$ and all events $\mathcal{A}_{-k} \subset \mathcal{A}^{(K-1)H}$, we have $\mathbb{P}\left[\mathcal{M}_{-k}\left(U_K, S_1\right) \in \mathcal{A}_{-k}\right] \leq e^{\varepsilon}\mathbb{P}\left[\mathcal{M}_{-k}\left(U'_K, S'_1\right) \in \mathcal{A}_{-k}\right]$.*

Lemma 24 (Lemma 11 in (Vietri et al., 2020))**.** *Any RL mechanism $\mathcal{M}$ satisfying ϵ-JDP also satisfies ϵ-JDP in the public initial state setting.*

Based on the above lemma, the RL with heavy-tailed rewards under ϵ-JDP problem is converted to the problem under ϵ-JDP in the public initial state setting.

The relationship between ϵ-DP MAB mechanisms and ϵ-JDP MDP in the public initial state setting mechanisms is the following: We collect the first actions taken by the agent in all episodes k with a fixed initial state $s_1^k = s \in [n]$ from an ϵ-JDP mechanism for MDPs in the public initial state setting. And such an operation simulates the execution of an ϵ-DP MAB algorithm. Specifically, let $\mathcal{M}$ be a JDP mechanism for MDPs with a public initial state and (U, S_1) be a user-state sequence with initial states from some set S_1. Let $\mathcal{M}(U, S_1) = (\vec{a}^1, \ldots, \vec{a}^K) \in \mathcal{A}^{KH}$ be the collection of all outputs produced by the mechanism on inputs U and S_1. For every $s \in S_1$ we denote trace $\mathcal{M}_{1,s}(U, S_1)$ as the restriction of the previous $\mathcal{M}(U, S_1)$ which just contains the first actions from all episodes starting with s together with the actions predicted by the policy at states s: $\mathcal{M}_{1,s}\left(U, S_1\right) := \left(a_1^{k_{s,1}}, \ldots, a_1^{k_{s,K_s}}\right)$, where K_s is the number of occurrences of s in S_1 and $k_{s,1}, \ldots, k_{s,K_s}$ are the indices of these occurrences. Furthermore, given $s \in S_1$ we write $U_s = (u_{k_{s,1}}, \ldots, u_{k_{s,K_s}})$ to denote the set of users whose initial state equals to s. Then we have the following result.

Lemma 25 (Lemma 9 in (Vietri et al., 2020))**.** *Let (U, S_1) be a user-state input sequence with initial states from some set S_1. Suppose $\mathcal{M}$ is an RL mechanism that satisfies ϵ-JDP in the public initial state setting. Then, for any $s \in S_1$ the trace $\mathcal{M}_{1,s}\left(U, S_1\right)$ is the output of an ϵ-DP MAB mechanism on input U_s.*

Proof of Theorem 8. We utilize the construction of hard MDP instances in Figure 1. From Lemma 24 and Lemma 25, we reduce the problem to learning $n = S - 2$ MAB instances satisfying ϵ-DP where each MAB is visited K_s many times for all $s \in [S-2]$. Now we can use the result in Theorem 7 which states that for each initial state $s \in [n]$, the lower bound for the regret of any ϵ-DP algorithm for the MAB problem with A arms can be expressed as $\Omega\left(\left(\frac{A}{\epsilon}\right)^{\frac{v}{1+v}} K_s^{\frac{1}{1+v}}\right)$ where K_s is the total number of arm pulls. Considering our construction of the MDP, a state is chosen uniformly at random at the start of the

episode. By combining the regret corresponding to each initial state $s \in [n]$, the regret of the RL mechanism must be at least

$$\Omega\left(\left(\frac{A}{\epsilon}\right)^{\frac{v}{1+v}} \sum_{s\in[S-2]} K_s^{\frac{1}{1+v}}\right)$$

where K_s is a random variable. To establish a lower bound for the term $\sum_{s\in[S-2]} K_s^{\frac{1}{1+v}}$, we utilize the Markov inequality from Lemma 11, resulting in:

$$\sum_{s\in[S-2]} K_s^{\frac{1}{1+v}} = (S-2)\mathbb{E}[K_s^{\frac{1}{1+v}}] \geq (S-2)\left(\frac{K}{S-2}\right)^{\frac{1}{1+v}} P\left[K_s^{\frac{1}{1+v}} \geq \left(\frac{K}{S-2}\right)^{\frac{1}{1+v}}\right].$$

The event $K_s^{\frac{1}{1+v}} \geq \left(\frac{K}{S-2}\right)^{\frac{1}{1+v}}$ occurs only when $K_s \geq \frac{K}{S-2}$. Since each $s \in [n]$ is chosen with equal probability at the beginning of the episodes, the expected number of pulls is $\mathbb{E}[K_s] = \frac{K}{S-2}$. Thus, each random variable K_s follows a binomial distribution $Bin(K, \frac{1}{S-2})$ with mean $\frac{K}{S-2}$ therefore the probability that $K_s \geq \frac{K}{S-2}$ is $\frac{1}{2}$. By substituting this probability term, we can deduce that the total regret of the RL algorithm is lower bounded by:

$$\Omega\left(\left(\frac{SA}{\epsilon}\right)^{\frac{v}{1+v}} K^{\frac{1}{1+v}}\right).$$

□

E.3. Proof of Theorem 9

Proof of Theorem 9. As in the case of Figure 2, we have the transition probabilities for a unique action a^* and leaf x_{i^*} such that:

$$P(+|x_{i^*}, a^*) = \gamma^{1+v} \text{ and } P(-|x_{i^*}, a^*) = 1 - \gamma^{1+v}. \tag{39}$$

where $\gamma^{1+v} \in (0, \frac{3}{4}]$. Each of the other leaves has transition probability

$$P(+|x_i, a) = \frac{1}{2}\gamma^{1+v} \text{ and } P(-|x_i, a) = 1 - \frac{1}{2}\gamma^{1+v}. \tag{40}$$

We denote above instance by $\mathbb{P}_{(x_{i^*}, a^*)}$.

In order to get the regret lower bound, we also consider another instance $\mathbb{P}_0$ where for all leaf states and any action, the transition probabilities are

$$P(+|x_i, a) = \frac{1}{2}\gamma^{1+v} \text{ and } P(-|x_i, a) = 1 - \frac{1}{2}\gamma^{1+v}. \tag{41}$$

Based on the above transition probabilities, it's easy to check for each state-action pair, the $(1+v)$-th moment of reward is bounded by 1 since the agent will receive the reward of $1/\gamma$ or 0 at state $+$ or $-$ respectively. All other states have a reward of 0 and every other transition is deterministic.

Then for a policy π, the value function can be written:

$$V^\pi(0) = \frac{1}{\gamma}P(s_{d+1} = +) = \frac{1}{\gamma}\left(\frac{1}{2}\gamma^{1+v} + \frac{1}{2}\gamma^{1+v}P(s_d = x_{i^*}, a_d = a^*)\right).$$

Since (x_{i^*}, a^*) is the optimal state-action pair, the regret can be written as:

$$Reg(T) = \frac{1}{2}\gamma^v\left(K - \sum_{k=1}^{K} P(s_d^k = x_{i^*}, a_d^k = a^*)\right) = \frac{1}{2}\gamma^v K\left(1 - \frac{1}{K}\sum_{k=1}^{K} P(s_d^k = x_{i^*}, a_d^k = a^*)\right)$$

where $\sum_{k=1}^{K} P(s_d = x_{i^*}, a_d = a^*) = \mathbb{E}_{(x_{i^*}, a^*)}\left[N_d^K(x_{i^*}, a^*)\right] = \mathbb{E}_{(x_{i^*}, a^*)}\left[\sum_{k=1}^{K} \mathbb{I}(s_d^k = x_{i^*}, a_d^k = a^*)\right]$ and $\mathbb{E}_{(x_{i^*}, a^*)}$ is the expectation on the instance described in equations 39 and 40. Thus, we have

$$Reg(T) = \frac{1}{2}\gamma^v K\left(1 - \frac{1}{K}\mathbb{E}_{(x_{i^*}, a^*)}\left[N_d^K(x_{i^*}, a^*)\right]\right). \tag{42}$$

$N_d^K(x_{i^*}, a^*)$ is a function of the history observed by the algorithm. Since we consider the LDP setting, this history can be written as:

$$\mathcal{M}(\mathcal{H}_K) = \{\mathcal{M}(X_\ell) | \ell \leq K\}$$

where $X_\ell = \{(s_{\ell,h}, a_{\ell,h}, r_{\ell,h}) | h \leq H\}$ is the trajectory observed by the user for episode ℓ and $\mathcal{M}$ is a privacy mechanism which maintains ϵ-LDP. Thus, $N_d^K(x_{i^*}, a^*)$ is a function of $\mathcal{M}(\mathcal{H}_K)$.

Now we focus on getting upper bound on $\mathbb{E}_{(x_{i^*}, a^*)}\left[N_d^K(x_{i^*}, a^*)\right]$. Since $N_d^K(x_{i^*}, a^*)$ is a function of $\mathcal{M}(\mathcal{H}_K)$ and $N_d^K(x_{i^*}, a^*)/K \in [0, 1]$, Lemma 12 gives us

$$\mathrm{kl}\left(\frac{1}{K}\mathbb{E}_0\left[N_d^K(x_{i^*}, a^*)\right], \frac{1}{K}\mathbb{E}_{(x_{i^*}, a^*)}\left[N_d^K(x_{i^*}, a^*)\right]\right) \leq \mathrm{KL}\left(\mathbb{P}_0(\mathcal{M}(\mathcal{H}_K)) \| \mathbb{P}_{(x_{i^*}, a^*)}(\mathcal{M}(\mathcal{H}_K))\right)$$

where $\mathbb{E}_0$ is the expectation on the instance where for all leaf states and any action, the transition probabilities are

$$P(+|x_i, a) = \frac{1}{2}\gamma^{1+v} \text{ and } P(-|x_i, a) = 1 - \frac{1}{2}\gamma^{1+v}. \tag{43}$$

By Pinsker's inequality, $(p-q)^2 \leq \frac{1}{2}\mathrm{kl}(p, q)$, it implies

$$\frac{1}{K}\mathbb{E}_{(x_{i^*}, a^*)}\left[N_d^K(x_{i^*}, a^*)\right] \leq \frac{1}{K}\mathbb{E}_0\left[N_d^K(x_{i^*}, a^*)\right] + \sqrt{\frac{1}{2}\mathrm{KL}\left(\mathbb{P}_0(\mathcal{M}(\mathcal{H}_K)) \| \mathbb{P}_{(x_{i^*}, a^*)}(\mathcal{M}(\mathcal{H}_K))\right)}.$$

Using the chain rule we have:

$$\mathrm{KL}\left(\mathbb{P}_0\left(\mathcal{M}\left(\mathcal{H}_K\right)\right) \| \mathbb{P}_{(x_{i^*}, a^*)}\left(\mathcal{M}\left(\mathcal{H}_K\right)\right)\right) = \sum_{k=1}^{K} \mathbb{E}_{\mathcal{H}_{k-1} \sim \mathbb{P}_0}\left(\mathrm{KL}\left(\mathbb{P}_0\left(\cdot \mid \mathcal{M}\left(\mathcal{H}_{k-1}\right)\right) \| \mathbb{P}_{(x_{i^*}, a^*)}\left(\cdot \mid \mathcal{M}\left(\mathcal{H}_{k-1}\right)\right)\right)\right).$$

where $\mathcal{M}\left(\mathcal{H}_{k-1}\right)$ means the results of privacy mechanism on history $\mathcal{H}_{k-1}$.

Because $\mathcal{M}$ is an ϵ-LDP mechanism, from Theorem 1 in (Duchi et al., 2013) we have

$$\mathrm{KL}\left(\mathbb{P}_0\left(\cdot \mid \mathcal{M}\left(\mathcal{H}_{k-1}\right)\right) \| \mathbb{P}_{(x_{i^*}, a^*)}\left(\cdot \mid \mathcal{M}\left(\mathcal{H}_{k-1}\right)\right)\right) \leq 4(\exp(\varepsilon) - 1)^2 \mathrm{KL}\left(\mathbb{P}_0\left(\cdot \mid \mathcal{H}_{k-1}\right) \| \mathbb{P}_{(x_{i^*}, a^*)}\left(\cdot \mid \mathcal{H}_{k-1}\right)\right).$$

Thus

$$\mathrm{KL}\left(\mathbb{P}_0\left(\mathcal{M}\left(\mathcal{H}_K\right)\right) \| \mathbb{P}_{(x_{i^*}, a^*)}\left(\mathcal{M}\left(\mathcal{H}_K\right)\right)\right) \leq 4(\exp(\varepsilon) - 1)^2 \mathrm{KL}\left(\mathbb{P}_0\left(\mathcal{H}_K\right) \| \mathbb{P}_{(x_{i^*}, a^*)}\left(\mathcal{H}_K\right)\right)$$

Lemma 5 in (Domingues et al., 2021) ensures that:

$$\mathrm{KL}\left(\mathbb{P}_0\left(\mathcal{H}_K\right) \| \mathbb{P}_{(x_{i^*}, a^*)}\left(\mathcal{H}_K\right)\right) = \mathbb{E}_0\left[N_d^K(x_{i^*}, a^*)\right] \mathrm{KL}(P_0(\cdot|x_{i^*}, a^*) \| P_{(x_{i^*}, a^*)}(\cdot|x_{i^*}, a^*)).$$

By using $\mathrm{KL}(\mathrm{Ber}(p) \| \mathrm{Ber}(q)) \leq \frac{(p-q)^2}{q(1-q)}$, we obtain

$$\mathrm{KL}(P_0(\cdot|x_{i^*}, a^*) \| P_{(x_{i^*}, a^*)}(\cdot|x_{i^*}, a^*)) = \mathrm{KL}\left(\mathrm{Ber}\left(\frac{\gamma^{1+v}}{2}\right) \| \mathrm{Ber}\left(\gamma^{1+v}\right)\right) \leq \frac{\gamma^{1+v}}{4(1-\gamma^{1+v})} \leq \gamma^{1+v}$$

where the last inequality holds when $\gamma^{1+v} \in (0, \frac{3}{4}]$. According to the fact that $e^\epsilon - 1 \approx \epsilon$ when ϵ is small, we have

$$\frac{1}{K}\mathbb{E}_{(x_{i^*}, a^*)}\left[N_d^K(x_{i^*}, a^*)\right] \leq \frac{1}{K}\mathbb{E}_0\left[N_d^K(x_{i^*}, a^*)\right] + \sqrt{2\epsilon^2\gamma^{1+v}\mathbb{E}_0\left[N_d^K(x_{i^*}, a^*)\right]}.$$

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

$$\mathbb{E}_I\mathbb{E}_0\left[N_d^K(x_{i^*}, a^*)\right] = \mathbb{E}_0 \sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{a=1}^{A} \frac{1}{LA}\mathbb{I}\{s_d^k = x_l, a_d^k = a\} = \frac{K}{LA}.$$

Then thanks to Jensen's inequality the regret in (42) is lower bound by

$$\mathbb{E}_I[Reg(T)] \geq \frac{1}{2}\gamma^v K\left(1 - \frac{1}{LA} - \sqrt{\frac{2K\epsilon^2\gamma^{1+v}}{LA}}\right).$$

Take $\gamma = \left(\frac{LA}{32K\epsilon^2}\right)^{\frac{1}{1+v}}$, we have

$$\max_{I\in\{x_1,\ldots,x_L\}\times[A]} Reg(T) \geq \mathbb{E}_I[Reg(T)] \geq \Omega\left(\left(\frac{SA}{\epsilon^2}\right)^{\frac{v}{1+v}} K^{\frac{1}{1+v}}\right)$$

where the last inequality holds since $L \geq (S-2)/2$. □

F. Experiments

In this section, we conduct proof-of-concept numerical experiments to verify our theoretical results for both policy-based and value-based algorithms.

F.1. Setting

We consider the standard tabular MDP environment `RiverSwim` (Osband et al., 2013), illustrated in Fig. 3. It consists of six states and two actions 'left' and 'right', i.e., $S = 6$ and $A = 2$. An agent starts with the left side and tries to reach the right side. At each step, if the agent chooses action 'left', she will always succeed (the dotted arrow). Otherwise, the agent often fails (the solid arrow). The agent only receives a small reward of 0.005 if she reaches the leftmost side, but obtains a large reward of 1 once she arrives at the rightmost state. The agent gets a reward of 0 for the intermediate states. Thus, this MDP naturally requires sufficient exploration to obtain the optimal policy.

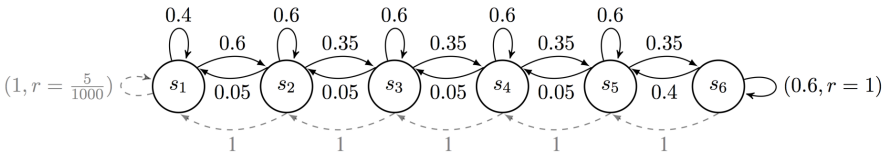

Figure 3: `RiverSwim` MDP – solid and dotted arrows denote the transitions under actions 'right' and 'left', respectively (Osband et al., 2013).

To generate heavy-tailed rewards, we use symmetric α-stable Levy distribution as in (Zhuang & Sui, 2021). The heaviness of the tail is controlled by the parameter α. In particular, α'-th moments of α-stable distributions are bounded for any $\alpha' \leq \alpha$. We denote this distribution as $\mathcal{L}(\alpha, \beta, \mu, \sigma)$, where β is the skewness parameter, μ is the mean, and σ is the shape parameter. In all experiments, we set $\alpha = 2$ (i.e., the second moment of rewards is bounded). We consider only symmetric distributions (i.e., $\beta = 0$) with unit shape (i.e., $\sigma = 1$). We generate the heavy-tailed rewards corresponding to mean values $\mu \in \{0, 1, 0.005\}$ as specified in the `RiverSwim` environment.

F.2. Results

We evaluate both Private-Heavy-UCBVI and Private-Heavy-UCBPO under different privacy budgets ϵ. As baselines, we design non-private UCBVI (Azar et al., 2017) and OPPO (Shani et al., 2020) algorithms under heavy-tailed noise following the high-level approach of (Zhuang & Sui, 2021). We set all the parameters in our proposed algorithms in the same order as the theoretical results. We tune the learning rate η and the scaling of the confidence interval to obtain the best results. We run 10 independent experiments, each consisting of $K = 2 \cdot 10^4$ episodes. Each episode is reset every $H = 20$ step. We plot the average cumulative regret along with the standard deviation for each setting, as shown in Fig. 4

As suggested by our theoretical results, in both PO and VI experiments, we see that the cost of privacy under JDP becomes negligible as the number of episodes increases (since JDP doesn't increase the order of regret). However, under the stricter

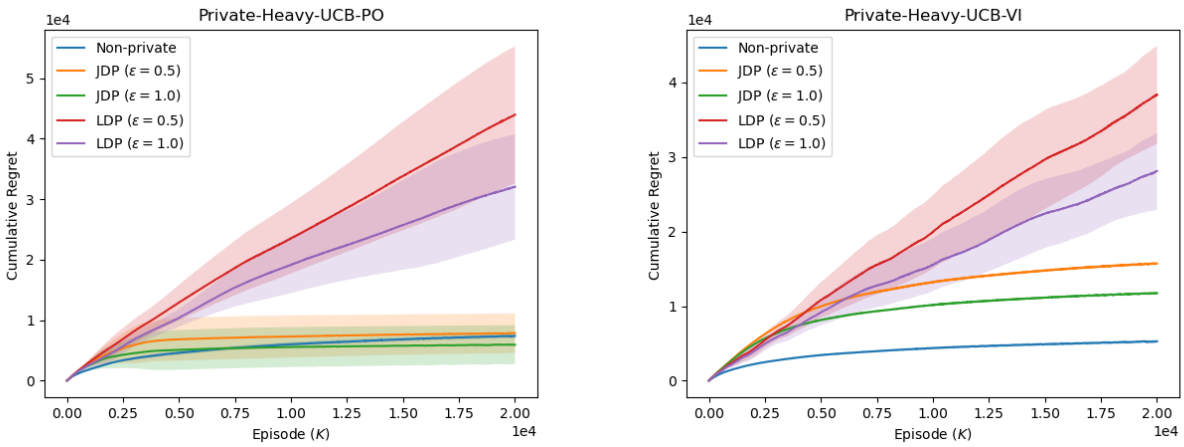

Figure 4: Cumulative regret vs. Episode for policy optimization and value iteration under heavy-tailed rewards with varying privacy levels $\epsilon \in \{0.5, 1\}$.

LDP requirement, the cost of privacy remains high (since LDP results in a higher-order term in regret). Furthermore, it is worth noting that the cost of privacy increases as the protection level increases, i.e., the value of ϵ decreases.