Near-Optimal Differentially Private Reinforcement Learning

Dan Qiao
Department of Computer Science
UC Santa Barbara
Santa Barbara, CA 93106
danqiao@ucsb.edu

Yu-Xiang Wang
Department of Computer Science
UC Santa Barbara
Santa Barbara, CA 93106
yuxiangw@cs.ucsb.edu

Abstract

Motivated by personalized healthcare and other applications involving sensitive data, we study online exploration in reinforcement learning with differential privacy (DP) constraints. Existing work on this problem established that no-regret learning is possible under joint differential privacy (JDP) and local differential privacy (LDP) but did not provide an algorithm with optimal regret. We close this gap for the JDP case by designing an ϵ-JDP algorithm with a regret of $\widetilde{O}(\sqrt{SAH^2T} + S^2AH^3/\epsilon)$ which matches the information-theoretic lower bound of non-private learning for all choices of $\epsilon > S^{1.5}A^{0.5}H^2/\sqrt{T}$. In the above, S, A denote the number of states and actions, H denotes the planning horizon, and T is the number of steps. To the best of our knowledge, this is the first private RL algorithm that achieves *privacy for free* asymptotically as $T \to \infty$. Our techniques — which could be of independent interest — include privately releasing Bernstein-type exploration bonuses and an improved method for releasing visitation statistics. The same techniques also imply a slightly improved regret bound for the LDP case.

1 Introduction

The wide range application of Reinforcement Learning (RL) based algorithms is becoming paramount in many personalized services, including medical care [Raghu et al., 2017], autonomous driving [Sallab et al., 2017] and recommendation systems [Afsar et al., 2021]. In these applications, the learning agent continuously improves its performance by learning from users' private feedback and data. The private data from users, however, usually contain sensitive information. Take recommendation system as an instance, the agent makes recommendation (corresponding to the action in a MDP) according to users' location, age, gender, etc. (corresponding to the state in a MDP), and improves its performance based on users' feedback (corresponding to the reward in a MDP). Unfortunately, it is shown that unless privacy protections are launched, learning agents will implicitly memorize information of individual training data points [Carlini et al., 2019], even if they are irrelevant for learning [Brown et al., 2021], which makes RL agents vulnerable to various privacy attacks.

Differential privacy (DP) [Dwork et al., 2006] has become the standard notion of privacy. The output of a differentially private RL algorithm is indistinguishable from its output returned under an alternative universe where any individual user is replaced, thereby preventing the aforementioned privacy risks. However, recent works [Shariff and Sheffet, 2018] show that standard DP is incompatible with sublinear regret bound for contextual bandits. Therefore, a relaxed variant of DP: *Joint Differential Privacy* (JDP) [Kearns et al., 2014] is considered. JDP ensures that the output of all other users will not leak much information about any specific user and such notion has been studied extensively in bandits problems[Shariff and Sheffet, 2018, Garcelon et al., 2022]. In addition, another variant of DP: *Local Differential Privacy* (LDP) [Duchi et al., 2013] has drawn more and more attention due to its stronger privacy protection. LDP requires that each user's raw data is privatized before being sent to the agent and LDP has been well studied under bandits [Basu et al., 2019, Zheng et al., 2020].

Compared to the large body of work on private bandits, existing work that studies private RL is sparser. Under the tabular MDP model, Vietri et al. [2020] first defined JDP and proposed PUCB with regret bound and JDP guarantee. Garcelon et al. [2021] introduced LDP under tabular MDP and designed LDP-OBI with regret bound and LDP guarantee. Recently, Chowdhury and Zhou [2021] provided a general framework for this problem and de-

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

Algorithms	Regret under ϵ-JDP	Regret under ϵ-LDP	Type of bonus
PUCB [Vietri et al., 2020]	$\widetilde{O}(\sqrt{S^2AH^3T} + S^2AH^3/\epsilon)^\star$	NA	Hoeffding
LDP-OBI [Garcelon et al., 2021]	NA	$\widetilde{O}(\sqrt{S^2AH^3T} + S^2A\sqrt{H^5T}/\epsilon)^\dagger$	Hoeffding
Private-UCB-PO [Chowdhury and Zhou, 2021]	$\widetilde{O}(\sqrt{S^2AH^3T} + S^2AH^3/\epsilon)$	$\widetilde{O}(\sqrt{S^2AH^3T} + S^2A\sqrt{H^5T}/\epsilon)$	Hoeffding
Private-UCB-VI [Chowdhury and Zhou, 2021]	$\widetilde{O}(\sqrt{SAH^3T} + S^2AH^3/\epsilon)$	$\widetilde{O}(\sqrt{SAH^3T} + S^2A\sqrt{H^5T}/\epsilon)$	Hoeffding
DP-UCBVI (Our Algorithm 1)	$\widetilde{O}(\sqrt{SAH^2T} + S^2AH^3/\epsilon)$	$\widetilde{O}(\sqrt{SAH^2T} + S^2A\sqrt{H^5T}/\epsilon)$	Bernstein
Lower bound without DP [Jin et al., 2018]	$\Omega(\sqrt{SAH^2T})$	$\Omega(\sqrt{SAH^2T})$	NA

Table 1: Comparison of our results (in blue) to existing work regarding regret under ϵ-joint differential privacy, regret under ϵ-local differential privacy and type of bonus. Here $T = KH$ is the number of steps, S, A, H refer to number of states, number of actions and the planning horizon. Bernstein-type bonus uses the knowledge of estimated variance while Hoeffding-type bonus directly bounds the variance by its uniform upper bound. $\star$: For more discussions about this bound, please refer to Chowdhury and Zhou [2021]. $\dagger$: The original regret bound in Garcelon et al. [2021] is achieved under stationary MDP, and can be translated to the bound stated here by adding $\sqrt{H}$ to the first term.

rived the best-known regret bounds under both JDP and LDP. However, the best known regret bound under ϵ-JDP $\widetilde{O}(\sqrt{SAH^3T} + S^2AH^3/\epsilon)$, although with the additional regret due to JDP being a lower order term, is still sub-optimal by $\sqrt{H}$ compared to the minimax optimal regret $\widetilde{O}(\sqrt{SAH^2T})$[1] [Azar et al., 2017] without constraints on DP. Therefore, if we run Algorithm 2 of Chowdhury and Zhou [2021], we not only pay for a constant additional regret $\widetilde{O}(S^2AH^3/\epsilon)$, but also suffer from a multiplicative factor of $\sqrt{H}$. Motivated by this, we want to find out whether it is possible to design an algorithm that has optimal regret bound up to lower order terms while satisfying Joint DP.

Our contributions. In this paper, we answer the above question affirmatively by constructing a general algorithm for DP RL: Algorithm 1. Our contributions are threefold.

- A new upper confidence bound (UCB) based algorithm (DP-UCBVI, Algorithm 1) that can be combined with any Privatizer (for JDP or LDP). Under the constraint of ϵ-JDP, DP-UCBVI achieves regret of $\widetilde{O}(\sqrt{SAH^2T} + S^2AH^3/\epsilon)$, which matches the minimax lower bound up to lower order terms.
- We propose a novel privatization of visitation numbers that satisfies several nice properties (see Assumption 3.1 for details). More importantly, our approach is the first to privatize Bernstein-type bonus, which helps tighten our regret bounds through law of total variance.
- Under the ϵ-LDP constraint, DP-UCBVI achieves regret of $\widetilde{O}(\sqrt{SAH^2T} + S^2A\sqrt{H^5T}/\epsilon)$ and improves the best known result [Chowdhury and Zhou, 2021].

1.1 Related work

Detailed comparisons with existing work on differentially private RL under tabular MDP [Vietri et al., 2020, Garcelon et al., 2021, Chowdhury and Zhou, 2021] are given in Table 1, while we leave more discussions about results on regret minimization to Appendix A. Notably, all existing algorithms privatize Hoeffding-type bonus and suffer from sub-optimal regret bound. In comparison, we privatize Bernstein-type bonus and the non-private part of our regret[2] matches the minimax lower bound in Jin et al. [2018].

Generally speaking, to achieve DP guarantee under RL, a common approach is to add appropriate noise to existing non-private algorithms, and derive tight regret bounds. We discuss about private algorithms under tabular MDP below and leave more discussions about algorithms under other settings to Appendix A. Under the constraint of JDP, Vietri et al. [2020] designed PUCB by privatizing UBEV [Dann et al., 2017]. Private-UCB-VI [Chowdhury and Zhou, 2021] resulted from UCBVI (with bonus 1) [Azar et al., 2017]. Under the constraint of LDP, Garcelon et al. [2021] designed LDP-OBI based on UCRL2 [Jaksch et al., 2010]. However, all these works privatized Hoeffding-type bonus, which is easier to handle, but will lead to sub-optimal regret bound. In contrast, we directly build upon the non-private algorithm with minimax optimal regret bound: UCBVI with bonus 2 [Azar et al., 2017], where the privatization of Bernstein-type bonus requires more advanced techniques.

A concurrent work [Qiao and Wang, 2022b] focused on the offline RL setting and derived a private version of APVI [Yin and Wang, 2021]. Their algorithm achieved tight sub-optimality bound of the output policy through privatization of Bernstein-type pessimism[3]. However, their analysis relied on the assumption that the visitation numbers of all (state,action) pairs are larger than some threshold. We overcome the requirement of such assumption via an improved privatization of visitation numbers. More importantly, offline RL can be viewed as one step of online RL, therefore

[1] Under the non-stationary MDP as in this paper, the result in Azar et al. [2017] will have additional $\sqrt{H}$ dependence.

[2] As shown in Table 1, the regret bounds of all DP-RL algorithms contain two parts: one results from running the non-private RL algorithms, while the other is the additional cost due to DP guarantees. Throughout the paper, we use "non-private part" to denote the regret from running the non-private RL algorithms.

[3] Pessimism is the counterpart of bonus under offline RL, which aims to discourage the choice of (s, a) pairs with large uncertainty.

privatization of Bernstein type bonus is more technically demanding. Finally, our approach actually realizes the future direction stated in the conclusion of Qiao and Wang [2022b].

1.2 A remark on technical novelty.

The general idea behind the previous differentially private algorithms under tabular MDP [Vietri et al., 2020, Garcelon et al., 2021, Chowdhury and Zhou, 2021] is to add noise to accumulative visitation numbers, and construct a private bonus based on privatized visitation numbers. Since Hoeffding-type bonus $b_h^k(s,a)$ only uses the information of visitation numbers (*e.g.*, in Azar et al. [2017], $b_h^k(s,a) = \widetilde{O}(H \cdot \sqrt{1/N_h^k(s,a)})$), the construction of private bonus is straightforward. We can simply replace original counts $N_h^k(s,a)$ with private counts $\widetilde{N}_h^k(s,a)$ and add an additional term to account for the difference between these two bonuses. Next, combining the construction of private bonuses with the uniform upper bound of $|\widetilde{N}_h^k(s,a) - N_h^k(s,a)|$, we can upper bound the private bonus by its non-private counterpart plus some additional lower order term. Therefore the proof schedule of the original non-private algorithms also applies to their private counterparts.

Unfortunately, although the idea to privatize UCBVI with bonus 2 (Bernstein-type) [Azar et al., 2017] is straightforward, the generalization of the previous approaches is technically non-trivial. Since the bonus 2 in Azar et al. [2017] includes the term $\mathrm{Var}_{\widehat{P}_h^k(\cdot|s,a)} V_{h+1}^k(\cdot)$, the first technical challenge is to replace the empirical transition kernel $\widehat{P}_h^k(s,a)$ with a private estimate. However, the private transition kernel estimates constructed in previous works are not valid probability distributions. In this paper, for both JDP and LDP, we propose a novel privatization of visitation numbers such that the private transition kernel estimates are valid probability distributions and meanwhile, the upper bound on $|\widetilde{N}_h^k(s,a) - N_h^k(s,a)|$ is the same scale compared to previous approaches. With the private transition kernel estimates $\widetilde{P}_h^k$, we can replace $\mathrm{Var}_{\widehat{P}_h^k(\cdot|s,a)} V_{h+1}^k(\cdot)$ with $\mathrm{Var}_{\widetilde{P}_h^k(\cdot|s,a)} \widetilde{V}_{h+1}^k(\cdot)$ where $\widetilde{V}_h^k(\cdot)$ is the value function calculated from value iteration with private estimates. Then the second challenge is to bound the difference between these two variances and retain the optimism. We overcome the second challenge via concentration inequalities. Briefly speaking, we add an additional term (using private statistics) to compensate for the difference of these two bonuses and recovered the proof of optimism. With all these techniques, we derive our regret bound using techniques like error decomposition and error propagation originated from Azar et al. [2017].

2 Notations and Problem Setup

Throughout the paper, for $N \in \mathbb{Z}^+$, $[N] = \{1, 2, \cdots, N\}$. For any set W, $\Delta(W)$ denotes the set of all probability distributions over W. Besides, we use standard notations such as O and Ω to suppress constants while $\widetilde{O}$ and $\widetilde{\Omega}$ absorb logarithmic factors.

Below we present the definition of episodic Markov Decision Processes and introduce differential privacy in reinforcement learning.

2.1 Markov decision processes and regret

We consider finite-horizon episodic *Markov Decision Processes* (MDP) with non-stationary transitions, denoted by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h=1}^H, \{r_h\}_{h=1}^H, d_1)$ [Sutton and Barto, 1998], where $\mathcal{S}$ is state space with $|\mathcal{S}| = S$, $\mathcal{A}$ is action space with $|\mathcal{A}| = A$ and H is the horizon. The non-stationary transition kernel has the form $P_h : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0,1]$ with $P_h(s'|s,a)$ representing the probability of transition from state s, action a to next state s' at time step h. In addition, $r_h(s,a) \in \Delta([0,1])$ denotes the corresponding distribution of reward, we overload the notation so that r also denotes the expected (immediate) reward function. Besides, d_1 is the initial state distribution. A policy can be seen as a series of mapping $\pi = (\pi_1, \cdots, \pi_H)$, where each π_h maps each state $s \in \mathcal{S}$ to a probability distribution over actions, *i.e.* $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$, $\forall\, h \in [H]$. A random trajectory $(s_1, a_1, r_1, \cdots, s_H, a_H, r_H, s_{H+1})$ is generated by the following rule: $s_1 \sim d_1$, $a_h \sim \pi_h(\cdot|s_h)$, $r_h \sim r_h(s_h, a_h)$, $s_{h+1} \sim P_h(\cdot|s_h, a_h)$, $\forall\, h \in [H]$.

Given a policy π and any $h \in [H]$, the value function $V_h^\pi(\cdot)$ and Q-value function $Q_h^\pi(\cdot,\cdot)$ are defined as: $V_h^\pi(s) = \mathbb{E}_\pi[\sum_{t=h}^H r_t | s_h = s]$, $Q_h^\pi(s,a) = \mathbb{E}_\pi[\sum_{t=h}^H r_t | s_h, a_h = s, a]$, $\forall\, s, a \in \mathcal{S} \times \mathcal{A}$. The optimal policy $\pi^\star$ maximizes $V_h^\pi(s)$ for all $s, h \in \mathcal{S} \times [H]$ simultaneously and we denote the value function and Q-value function with respect to $\pi^\star$ by $V_h^\star(\cdot)$ and $Q_h^\star(\cdot,\cdot)$. Then Bellman (optimality) equation follows $\forall\, h \in [H]$:

$$Q_h^\pi(s,a) = r_h(s,a) + P_h(\cdot|s,a)V_{h+1}^\pi, \;\; V_h^\pi = \mathbb{E}_{a\sim\pi_h}[Q_h^\pi],$$
$$Q_h^\star(s,a) = r_h(s,a) + P_h(\cdot|s,a)V_{h+1}^\star, \; V_h^\star = \max_a Q_h^\star(\cdot,a).$$

We measure the performance of online reinforcement learning algorithms by the regret. The regret of an algorithm is defined as

$$\mathrm{Regret}(K) := \sum_{k=1}^K [V_1^\star(s_1^k) - V_1^{\pi_k}(s_1^k)],$$

where s_1^k is the initial state and π_k is the policy deployed at episode k. Let K be the number of episodes that the agent plan to play and total number of steps is $T := KH$.

2.2 Differential privacy under episodic RL

Under the episodic RL setting, each trajectory represents one specific user. We first consider the following RL protocol: during the h-th step of the k-th episode, user u_k sends her state s_h^k to agent $\mathcal{M}$, $\mathcal{M}$ sends back an action a_h^k, and finally u_k sends her reward r_h^k to $\mathcal{M}$. Formally, we denote a sequence of K users who participate in the above RL protocol by $\mathcal{U} = (u_1, \cdots, u_K)$. Following the definition in Vietri et al. [2020], each user can be seen as a tree of depth H encoding the state and reward responses they would reply to all A^H possible sequences of actions from the agent. We let $\mathcal{M}(\mathcal{U}) = (a_1^1, \cdots, a_H^K)$ denote the whole sequence of actions chosen by agent $\mathcal{M}$. An ideal privacy preserving agent would guarantee that $\mathcal{M}(\mathcal{U})$ and all users but u_k together will not reveal much information about user u_k. We formalize such privacy preservation through adaptation of differential privacy [Dwork et al., 2006].

Definition 2.1 (Differential Privacy (DP))**.** *For any $\epsilon > 0$ and $\delta \in [0, 1]$, a mechanism $\mathcal{M} : \mathcal{U} \to \mathcal{A}^{KH}$ is (ϵ, δ)-differentially private if for any possible user sequences $\mathcal{U}$ and $\mathcal{U}'$ differing on a single user and any subset E of $\mathcal{A}^{KH}$,*

$$\mathbb{P}[\mathcal{M}(\mathcal{U}) \in E] \leq e^{\epsilon}\mathbb{P}[\mathcal{M}(\mathcal{U}') \in E] + \delta.$$

If $\delta = 0$, we say that $\mathcal{M}$ is ϵ-differentially private (ϵ-DP).

However, although recommendation to other users will not affect the privacy of user u_k significantly, it is impractical to privately recommend actions to user u_k while protecting the information of her state and reward. Therefore, the notion of DP is relaxed to *Joint Differential Privacy* (JDP) [Kearns et al., 2014], which requires that for all user u_k, the recommendation to all users but u_k will not reveal much information about u_k. JDP is weaker than DP, while JDP can still provide strong privacy protection since it protects a specific user from any possible collusion of all other users against her. Formally, the definition of JDP is shown below.

Definition 2.2 (Joint Differential Privacy (JDP))**.** *For any $\epsilon > 0$, a mechanism $\mathcal{M} : \mathcal{U} \to \mathcal{A}^{KH}$ is ϵ-joint differentially private if for any $k \in [K]$, any user sequences $\mathcal{U}, \mathcal{U}'$ differing on the k-th user and any subset E of $\mathcal{A}^{(K-1)H}$,*

$$\mathbb{P}[\mathcal{M}_{-k}(\mathcal{U}) \in E] \leq e^{\epsilon}\mathbb{P}[\mathcal{M}_{-k}(\mathcal{U}') \in E],$$

where $\mathcal{M}_{-k}(\mathcal{U}) \in E$ means the sequence of actions recommended to all users but u_k belongs to set E.

JDP ensures that even if an adversary can observe the recommended actions to all users but u_k, it is impossible to identify the trajectory from u_k accurately. JDP is first defined and analyzed under RL by Vietri et al. [2020].

Although JDP provides strong privacy protection, the agent can still observe the raw trajectories from users. Under some circumstances, however, the users are not even willing to share their original data with the agent. This motivates a stronger notion of privacy which is called *Local Differential Privacy* (LDP) [Duchi et al., 2013]. Since under LDP, the agent is not allowed to directly observe the state of users, we consider the following RL protocol for LDP: during the k-th episode, the agent $\mathcal{M}$ sends policy π_k to user u_k, after deploying π_k and getting trajectory X_k, user u_k privatizes her trajectory to X_k' and finally sends it to $\mathcal{M}$. We denote the privacy mechanism on user's side by $\widetilde{\mathcal{M}}$ and define local differential privacy formally below.

Definition 2.3 (Local Differential Privacy (LDP))**.** *For any $\epsilon > 0$, a mechanism $\widetilde{\mathcal{M}}$ is ϵ-local differentially private if for any possible trajectories X, X' and any possible set $E \subseteq \{\widetilde{\mathcal{M}}(X) | X$ is any possible trajectory$\}$,*

$$\mathbb{P}[\widetilde{\mathcal{M}}(X) \in E] \leq e^{\epsilon}\mathbb{P}[\widetilde{\mathcal{M}}(X') \in E].$$

Local DP ensures that even if an adversary observes the whole reply from user u_k, it is still statistically hard to identify her trajectory. LDP is first defined and analyzed under RL by Garcelon et al. [2021].

3 Algorithm

In this section, we propose DP-UCBVI (Algorithm 1) that takes Privatizer as input, where the Privatizer can be either Central (for JDP) or Local (for LDP). We provide regret analysis for all privatizers satisfying the following Assumption 3.1, which naturally implies regret bounds under both Joint DP and Local DP.

We begin with the following definition of counts. Let $N_h^k(s, a) = \sum_{i=1}^{k-1} \mathbb{1}(s_h^i, a_h^i = s, a)$ denote the visitation number of (s, a) at step h before the k-th episode. Similarly, $N_h^k(s, a, s') = \sum_{i=1}^{k-1} \mathbb{1}(s_h^i, a_h^i, s_{h+1}^i = s, a, s')$ and $R_h^k(s, a) = \sum_{i=1}^{k-1} \mathbb{1}(s_h^i, a_h^i = s, a) \cdot r_h^i$ denote the visitation number of (h, s, a, s') and accumulative reward at (h, s, a) before the k-th episode. In non-private RL, such counts are sufficient for estimating transition kernel P_h, reward function r_h and deciding the exploration policy, as in Azar et al. [2017]. However, these counts are derived from the raw trajectories of the users, which could contain sensitive information. Therefore, under the constraint of privacy, we can only use these counts in a privacy-preserving way, *i.e.* we use the private counts $\widetilde{N}_h^k(s, a), \widetilde{N}_h^k(s, a, s'), \widetilde{R}_h^k(s, a)$ returned by Privatizer. We make the Assumption 3.1 below, which says that with high probability, the private counts are close to real ones, such assumption will be justified by our Privatizers in Section 5.

Assumption 3.1 (Private counts)**.** *We assume that for any privacy budget $\epsilon > 0$ and failure probability $\beta \in [0, 1]$, the private counts returned by Privatizer satisfies that for some $E_{\epsilon,\beta} > 0$, with probability at least $1 - \beta/3$, uniformly over all $(h, s, a, s', k) \in [H] \times \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [K]$:*
(1) $|\widetilde{N}_h^k(s, a, s') - N_h^k(s, a, s')| \leq E_{\epsilon,\beta}$, $|\widetilde{N}_h^k(s, a) -$

Algorithm 1 DP-UCBVI

Input: Number of episodes K, privacy budget ϵ, failure probability β and a Privatizer (can be either Central or Local).
Initialize: Private counts $\widetilde{R}_h^1(s,a) = \widetilde{N}_h^1(s,a) = \widetilde{N}_h^1(s,a,s') = 0$ for all $(h,s,a,s') \in [H] \times \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. Set up the confidence bound $E_{\epsilon,\beta}$ w.r.t the Privatizer. $\iota = \log(30HSAT/\beta)$.
for $k = 1, 2, \cdots, K$ **do**
  $\widetilde{V}_{H+1}^k(\cdot) = 0$.
  for $h = H, H-1, \cdots, 1$ **do**
    Compute $\widetilde{P}_h^k(s'|s,a)$ and $\widetilde{r}_h^k(s,a)$ as in (1).
    Calculate private bonus $b_h^k(s,a) = 2\sqrt{\frac{\mathrm{Var}_{s'\sim\widetilde{P}_h^k(\cdot|s,a)}\widetilde{V}_{h+1}^k(\cdot)\cdot\iota}{\widetilde{N}_h^k(s,a)}} + \sqrt{\frac{2\iota}{\widetilde{N}_h^k(s,a)}} + \frac{20HSE_{\epsilon,\beta}\cdot\iota}{\widetilde{N}_h^k(s,a)} + 4\sqrt{\iota}\cdot \sqrt{\frac{\sum_{s'}\widetilde{P}_h^k(s'|s,a)\min\left\{\frac{1000^2H^3SA\iota^2}{\widetilde{N}_{h+1}^k(s')} + \frac{1000^2H^4S^4A^2E_{\epsilon,\beta}^2\iota^4}{\widetilde{N}_{h+1}^k(s')^2} + \frac{1000^2H^6S^4A^2\iota^4}{\widetilde{N}_{h+1}^k(s')^2}, H^2\right\}}{\widetilde{N}_h^k(s,a)}}$.
    for $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
      $\widetilde{Q}_h^k(s,a) = \min\{\widetilde{Q}_h^{k-1}(s,a), H, \widetilde{r}_h^k(s,a) + \sum_{s'}\widetilde{P}_h^k(s'|s,a)\cdot\widetilde{V}_{h+1}^k(s') + b_h^k(s,a)\}$.
    end for
    for $s \in \mathcal{S}$ **do**
      $\widetilde{V}_h^k(s) = \max_{a\in\mathcal{A}}\widetilde{Q}_h^k(s,a)$.
      $\pi_h^k(s) = \arg\max_{a\in\mathcal{A}}\widetilde{Q}_h^k(s,a)$ with ties broken arbitrarily.
    end for
  end for
  Deploy policy $\pi_k = (\pi_1^k, \cdots, \pi_H^k)$ and get trajectory $(s_1^k, a_1^k, r_1^k, \cdots, s_{H+1}^k)$.
  Update the private counts to $\widetilde{R}^{k+1}, \widetilde{N}^{k+1}$ via Privatizer.
end for

$N_h^k(s,a)| \le E_{\epsilon,\beta}$ *and* $|\widetilde{R}_h^k(s,a) - R_h^k(s,a)| \le E_{\epsilon,\beta}$.
(2) $\widetilde{N}_h^k(s,a) = \sum_{s'\in\mathcal{S}}\widetilde{N}_h^k(s,a,s') \ge N_h^k(s,a)$. $\widetilde{N}_h^k(s,a,s') > 0$. *Also, we let* $\widetilde{N}_h^k(s) = \sum_{a\in\mathcal{A}}\widetilde{N}_h^k(s,a)$.

Under Assumption 3.1, for all (h,s,a,s',k), we define the private estimations of transition kernel and reward function.

$$\widetilde{P}_h^k(s'|s,a) = \frac{\widetilde{N}_h^k(s,a,s')}{\widetilde{N}_h^k(s,a)}, \qquad \widetilde{r}_h^k(s,a) = \left(\frac{\widetilde{R}_h^k(s,a)}{\widetilde{N}_h^k(s,a)}\right)_{[0,1]}. \tag{1}$$

Remark 3.2. *Different from the private empirical transition kernels in Vietri et al. [2020], Garcelon et al. [2021], Chowdhury and Zhou [2021], Assumption 3.1 implies that our estimated transition kernel* $\widetilde{P}_h^k(\cdot|s,a)$ *is a valid probability distribution, this property results from our construction of Privatizer. We truncate the empirical reward function so that it stays in* $[0,1]$ *while still preserving privacy.*

Algorithmic design. Similar to non-private algorithms [Azar et al., 2017], DP-UCBVI (Algorithm 1) follows the procedure of optimistic value iteration. More specifically, in episode k, we do value iteration based on private estimations $\widetilde{P}_h^k, \widetilde{r}_h^k$ and private bonus term b_h^k to derive private Q-value functions $\widetilde{Q}_h^k$. Next, the greedy policy π_k w.r.t $\widetilde{Q}_h^k$ is chosen and we collect one trajectory by running π_k. Finally, the Privatizer translates the non-private counts to private ones for the next episode. We highlight that, different from all previous works regarding private RL, our bonus is variance-dependent. According to Law of total variance, variance-dependent bonus can effectively save a factor of $\sqrt{H}$ in regret bound. Intuitively, the first term of b_h^k aims to approximate the variance w.r.t to $V_h^\star$, the last term accounts for the difference between these two variances and the third term is the additional bonus due to differential privacy.

4 Main results

In this section, we present our main results that formalize the algorithmic ideas discussed in previous sections. We first state a general result based on Assumption 3.1, which can be combined with any Privatizers. The proof of Theorem 4.1 is sketched in Section 6 with details in Appendix C.

Theorem 4.1. *For any privacy budget* $\epsilon > 0$*, failure probability* $0 < \beta < 1$ *and any Privatizer that satisfies Assumption 3.1, with probability at least* $1-\beta$*, the regret of DP-UCBVI (Algorithm 1) is*

$$\textit{Regret}(K) \le \widetilde{O}(\sqrt{SAH^2T} + S^2AH^2E_{\epsilon,\beta}), \tag{2}$$

where K *is the number of episodes and* $T = HK$.

Under Assumption 3.1, the best known regret bound is $\widetilde{O}(\sqrt{SAH^3T} + S^2AH^2E_{\epsilon,\beta})$ (Theorem 4.2 of Chowdhury and Zhou [2021]). As a comparison, in our regret bound,

the term parameterized by privacy loss ϵ remains the same while the leading term is improved by a factor of $\sqrt{H}$ into $\widetilde{O}(\sqrt{SAH^2T})$. More importantly, when T is sufficiently large, our result nearly matches the lower bound in Jin et al. [2018], hence is information-theoretically optimal up to a logarithmic factor.

5 Choice of Privatizers

In this section, we design Privatizers that satisfy Assumption 3.1 and different DP constraints (JDP or LDP). All the proofs in this section are deferred to Appendix D.

5.1 Central Privatizer for Joint DP

The Central Privatizer protects the information of all single users by privatizing all the counter streams $N_h^k(s,a)$, $N_h^k(s,a,s')$ and $R_h^k(s,a)$ using the Binary Mechanism [Chan et al., 2011], which focused on privately releasing data stream [Zhao et al., 2022]. More specifically, for each (h,s,a), $\{N_h^k(s,a)=\sum_{i=1}^{k-1}\mathbb{1}(s_h^i,a_h^i=s,a)\}_{k\in[K]}$ is the partial sums of data stream $\{\mathbb{1}(s_h^i,a_h^i=s,a)\}_{i\in[K]}$. Binary Mechanism works as below: for each episode k, after observing $\mathbb{1}(s_h^{k-1},a_h^{k-1}=s,a)$, the mechanism outputs private version of $\sum_{i=1}^{k-1}\mathbb{1}(s_h^i,a_h^i=s,a)$ while ensuring Differential Privacy.[4] Given privacy budget $\epsilon>0$, we construct the Central Privatizer as below:

(1) For all (h,s,a,s'), we privatize $\{N_h^k(s,a)\}_{k\in[K]}$ and $\{N_h^k(s,a,s')\}_{k\in[K]}$ (which is summation of bounded streams) by applying Binary Mechanism (Algorithm 2 in Chan et al. [2011]) with $\epsilon'=\frac{\epsilon}{3H\log K}$. We denote the output of Binary Mechanism by $\widehat{N}_h^k$.

(2) The private counts $\widetilde{N}_h^k$ are solved through the procedure in Section 5.1.1 with $E_{\epsilon,\beta}=O(\frac{H}{\epsilon}\log(HSAT/\beta)^2)$.

(3) For the counters of accumulative reward, for all (h,s,a), we apply the same Binary Mechanism with $\epsilon'=\frac{\epsilon}{3H\log K}$ to privatize $R_h^k(s,a)$ and get $\widetilde{R}_h^k(s,a)$.

We sum up the properties of Central Privatizer below.

Lemma 5.1. *For any $\epsilon>0$ and $0<\beta<1$, the Central Privatizer satisfies ϵ-JDP and Assumption 3.1 with $E_{\epsilon,\beta}=\widetilde{O}(\frac{H}{\epsilon})$.*

Therefore, combining Lemma 5.1 and Theorem 4.1, the following regret bound holds.

Theorem 5.2 (Regret under JDP)**.** *For any $\epsilon>0$ and $0<\beta<1$, running DP-UCBVI (Algorithm 1) with Central Privatizer as input, with probability $1-\beta$, it holds that:*

$$\textit{Regret}(K)\le\widetilde{O}(\sqrt{SAH^2T}+S^2AH^3/\epsilon). \tag{3}$$

[4]For more details about Binary Mechanism, please refer to Chan et al. [2011] or Kairouz et al. [2021].

Under the most prevalent regime where the privacy budget ϵ is a constant, the additional regret bound due to JDP is a lower order term. The main term of Theorem 5.2 improves the best known result $\widetilde{O}(\sqrt{SAH^3T})$ (Corollary 5.2 of Chowdhury and Zhou [2021]) by $\sqrt{H}$ and matches the minimax lower bound without constrains on DP [Jin et al., 2018].

5.1.1 A post-processing step

During the k-th episode, given the noisy counts $\widehat{N}_h^k(s,a)$ and $\widehat{N}_h^k(s,a,s')$ (for all $(h,s,a,s')\in[H]\times\mathcal{S}\times\mathcal{A}\times\mathcal{S}$), we construct the following private counts that satisfy Assumption 3.1. The choice of $\widetilde{N}_h^k$ follows: for all (h,s,a)

$$\begin{aligned}
&\{\widetilde{N}_h^k(s,a,s')\}_{s'\in\mathcal{S}}=\operatorname*{argmin}_{\{x_{s'}\}_{s'\in\mathcal{S}}}\left(\max_{s'\in\mathcal{S}}\left|x_{s'}-\widehat{N}_h^k(s,a,s')\right|\right)\\
&\text{such that }\left|\sum_{s'\in\mathcal{S}}x_{s'}-\widehat{N}_h^k(s,a)\right|\le\frac{E_{\epsilon,\beta}}{4}\text{ and }x_{s'}\ge0,\forall\,s'.\\
&\widetilde{N}_h^k(s,a)=\sum_{s'\in\mathcal{S}}\widetilde{N}_h^k(s,a,s').
\end{aligned} \tag{4}$$

Finally, for all (h,s,a), we add the following terms such that with high probability, $\widetilde{N}_h^k(s,a)$ will never underestimate.

$$\begin{aligned}
\widetilde{N}_h^k(s,a,s')&=\widetilde{N}_h^k(s,a,s')+\frac{E_{\epsilon,\beta}}{2S},\\
\widetilde{N}_h^k(s,a)&=\widetilde{N}_h^k(s,a)+\frac{E_{\epsilon,\beta}}{2}.
\end{aligned} \tag{5}$$

Remark 5.3. *The optimization problem* (4) *can be reformulated as:*

$$\begin{aligned}
&\min\ t,\ \textit{s.t.}\,|x_{s'}-\widehat{N}_h^k(s,a,s')|\le t,\ x_{s'}\ge0,\ \forall\,s'\in\mathcal{S},\\
&\left|\sum_{s'\in\mathcal{S}}x_{s'}-\widehat{N}_h^k(s,a)\right|\le\frac{E_{\epsilon,\beta}}{4}.
\end{aligned} \tag{6}$$

Note that (6) *is a* Linear Programming *problem with $O(S)$ variables and $O(S)$ linear constraints. This can be solved efficiently by the simplex method [Ficken, 2015] or other provably efficient algorithms [Nemhauser and Wolsey, 1988]. Therefore, since during the whole process, we only solve $HSAK$ such Linear Programming problems, our Algorithm 1 is computationally efficient.*

The properties of private counts $\widetilde{N}_h^k$ is summarized below.

Lemma 5.4. *Suppose $\widehat{N}_h^k$ satisfies that with probability $1-\frac{\beta}{3}$, uniformly over all (h,s,a,s',k), it holds that*

$$|\widehat{N}_h^k(s,a,s')-N_h^k(s,a,s')|\le\frac{E_{\epsilon,\beta}}{4},$$

$$|\widehat{N}_h^k(s,a)-N_h^k(s,a)|\le\frac{E_{\epsilon,\beta}}{4},$$

then the $\widetilde{N}_h^k$ derived from (4) *and* (5) *satisfies Assumption 3.1.*

Remark 5.5. *Compared to the concurrent work [Qiao and Wang, 2022b], our private counts $\widetilde{N}_h^k(s,a)$ have additional guarantee of never underestimating the true values, which is a desirable property for analysis in Appendix B. In comparison, the analysis in Qiao and Wang [2022b] heavily relies on the assumption that the visitation number is larger than some threshold such that the scale of noise is ignorable.*

5.2 Local Privatizer for Local DP

For each episode k, the Local Privatizer privatizes each single trajectory by perturbing the statistics calculated from that trajectory. For visitation of (state,action) pairs, the original visitation number $\{\sigma_h^k(s,a) = \mathbb{1}(s_h^k, a_h^k = s,a)\}_{(h,s,a)}$ has ℓ_1 sensitivity H. Therefore, the perturbed version of the counts above $\widetilde{\sigma}_h^k(s,a) = \sigma_h^k(s,a) + \text{Lap}(\frac{3H}{\epsilon})$ satisfies $\frac{\epsilon}{3}$-LDP. In addition, similar perturbations to $\{\mathbb{1}(s_h^k, a_h^k, s_{h+1}^k = s,a,s')\}_{(h,s,a,s')}$ and $\{\mathbb{1}(s_h^k, a_h^k = s,a)\cdot r_h^k\}_{(h,s,a)}$ will lead to the same result. As a result, we construct Local Privatizer as below:

(1) For all (k,h,s,a,s'), we perturb $\sigma_h^k(s,a) = \mathbb{1}(s_h^k, a_h^k = s,a)$ and $\sigma_h^k(s,a,s') = \mathbb{1}(s_h^k, a_h^k, s_{h+1}^k = s,a,s')$ by adding independent Laplace noises:

$$\widetilde{\sigma}_h^k(s,a) = \sigma_h^k(s,a) + \text{Lap}(\frac{3H}{\epsilon}),$$

$$\widetilde{\sigma}_h^k(s,a,s') = \sigma_h^k(s,a,s') + \text{Lap}(\frac{3H}{\epsilon}).$$

(2) The noisy counts are calculated by

$$\widehat{N}_h^k(s,a) = \sum_{i=1}^{k-1} \widetilde{\sigma}_h^i(s,a),$$

$$\widehat{N}_h^k(s,a,s') = \sum_{i=1}^{k-1} \widetilde{\sigma}_h^i(s,a,s').$$

Then the private counts $\widetilde{N}_h^k$ are solved through the procedure in Section 5.1.1 with $E_{\epsilon,\beta} = O(\frac{H}{\epsilon}\sqrt{K\log(HSAT/\beta)})$.

(3) We perturb the trajectory-wise reward by adding independent Laplace noise: $\widetilde{r}_h^k(s,a) = \mathbb{1}(s_h^k, a_h^k = s,a)\cdot r_h^k + \text{Lap}(\frac{3H}{\epsilon})$. The accumulative statistic is calculated by $\widetilde{R}_h^k(s,a) = \sum_{i=1}^{k-1} \widetilde{r}_h^i(s,a)$.

Properties of our Local Privatizer is summarized below.

Lemma 5.6. *For any $\epsilon > 0$ and $0 < \beta < 1$, the Local Privatizer satisfies ϵ-LDP and Assumption 3.1 with $E_{\epsilon,\beta} = \widetilde{O}(\frac{H}{\epsilon}\sqrt{K})$.*

Therefore, combining Lemma 5.6 and Theorem 4.1, the following regret bound holds.

Theorem 5.7 (Regret under LDP)**.** *For any $\epsilon > 0$ and $0 < \beta < 1$, running DP-UCBVI (Algorithm 1) with Local Privatizer as input, with probability $1-\beta$, it holds that:*

$$\textit{Regret}(K) \le \widetilde{O}(\sqrt{SAH^2T} + S^2A\sqrt{H^5T}/\epsilon). \quad (7)$$

Theorem 5.7 improves the non-private part of regret bound in the best known result (Corollary 5.5 of Chowdhury and Zhou [2021]).

5.3 More discussions

The step (1) of our Privatizers is similar to previous works [Vietri et al., 2020, Garcelon et al., 2021, Chowdhury and Zhou, 2021]. However, different from their approaches (directly use $\widehat{N}_h^k$ as private counts), we apply the post-processing step in Section 5.1.1, which ensures that $\widetilde{P}_h^k$ is valid probability distribution while $E_{\epsilon,\beta}$ is only worse by a constant factor. Therefore, we can apply Bernstein type bonus to achieve the optimal non-private part in our regret bound.

We remark that the Laplace Mechanism can be replaced with other mechanisms, like Gaussian Mechanism [Dwork et al., 2014] for approximate DP (or zCDP). According to Theorem 4.1, the regret bounds can be easily derived by plugging in the corresponding $E_{\epsilon,\beta}$.

6 Proof Sketch

In this section, we provide a proof overview for Theorem 4.1, which can imply the results under JDP (Theorem 5.2) and LDP (Theorem 5.7). Recall that $N_h^k(s,a)$ and $N_h^k(s,a,s')$ are real visitation numbers while $\widetilde{N}_h^k$'s are private ones satisfying Assumption 3.1. Other notations like $\widetilde{P}_h^k$, $\widetilde{r}_h^k$, $\widetilde{Q}_h^k$, $\widetilde{V}_h^k$ and ι are defined in Algorithm 1. The statement "with high probability" means that the summation of all failure probabilities is bounded by β. We begin with some properties of private statistics below.

Properties of $\widetilde{P}$ and $\widetilde{r}$. Due to concentration inequalities and Assumption 3.1, we provide high probability bounds for $\left|\widetilde{r}_h^k(s,a) - r_h(s,a)\right|$, $\left\|\widetilde{P}_h^k(\cdot|s,a) - P_h(\cdot|s,a)\right\|_1$ and $\left|\widetilde{P}_h^k(s'|s,a) - P_h(s'|s,a)\right|$ in Appendix B. In addition, we bound the key term $\left|\left(\widetilde{P}_h^k - P_h\right)\cdot V_{h+1}^\star(s,a)\right|$ below.

Lemma 6.1 (Informal version of Lemma B.6)**.** *With high probability, for all (h,s,a,k), it holds that:*

$$\left|\left(\widetilde{P}_h^k - P_h\right)\cdot V_{h+1}^\star(s,a)\right| \le \widetilde{O}\left(\sqrt{\frac{\text{Var}_{\widehat{P}_h^k(\cdot|s,a)}V_{h+1}^\star(\cdot)}{\widetilde{N}_h^k(s,a)}}\right) + \widetilde{O}\left(\frac{HSE_{\epsilon,\beta}}{\widetilde{N}_h^k(s,a)}\right). \quad (8)$$

With these concentrations, we are ready to present our proof sketch. Since we apply Bernstein-type bonus, the proof of optimism is not straightforward. We prove our regret upper bound through induction, which is shown below.

Induction over episodes. Our induction is for all $k \in [K]$,

(1) Given that for all $(i,h,s,a) \in [k] \times [H] \times \mathcal{S} \times \mathcal{A}$, $\widetilde{Q}_h^i(s,a) \geq Q_h^\star(s,a)$, we prove ($T_k = kH$)

$$\text{Regret}(k) \leq \widetilde{O}\left(\sqrt{H^2SAT_k} + H^2S^2AE_{\epsilon,\beta}\right)$$

and for all $(h,s) \in [H] \times \mathcal{S}$,

$$\widetilde{V}_h^k(s) - V_h^\star(s) \leq \widetilde{O}\left(\sqrt{\frac{SAH^3}{N_h^k(s)}} + \frac{S^2AH^2E_{\epsilon,\beta}}{N_h^k(s)}\right).$$

(2) Given that for all $(h,s) \in [H] \times \mathcal{S}$,

$$\widetilde{V}_h^k(s) - V_h^\star(s) \leq \widetilde{O}\left(\sqrt{\frac{SAH^3}{N_h^k(s)}} + \frac{S^2AH^2E_{\epsilon,\beta}}{N_h^k(s)}\right),$$

we prove that for all (h,s,a), $\widetilde{Q}_h^{k+1}(s,a) \geq Q_h^\star(s,a)$.

Suppose the above induction holds, we have point (1) holds for all $k \in [K]$ and therefore,

$$\text{Regret}(K) \leq \widetilde{O}\left(\sqrt{H^2SAT} + H^2S^2AE_{\epsilon,\beta}\right). \quad (9)$$

Below we discuss about the proof of (1) and (2) separately.

Proof of regret bound: (1). We only need to prove the upper bound of Regret(k), as the upper bound of $\widetilde{V}_h^k(s) - V_h^\star(s)$ follows similarly. Using the standard technique of layer-wise error decomposition (details in Appendix C.3) and ignoring lower order terms: summation of martingale differences, we only need to bound $\sum_{i=1}^k \sum_{h=1}^H b_h^i(s_h^i, a_h^i)$ which consists of four terms according to the definition of b_h^k. First of all, the second and forth terms are dominated by the first and third terms. Next, for the third term, we have

$$\begin{aligned}\sum_{i=1}^k \sum_{h=1}^H \frac{20HSE_{\epsilon,\beta}\iota}{\widetilde{N}_h^i(s_h^i,a_h^i)} \leq & \sum_{i=1}^k \sum_{h=1}^H \frac{20HSE_{\epsilon,\beta}\iota}{N_h^i(s_h^i,a_h^i)} \\ \leq & \widetilde{O}(S^2AH^2E_{\epsilon,\beta}).\end{aligned} \quad (10)$$

Now we analyze the first term (which is also the main term): $\sum_{i=1}^k \sum_{h=1}^H \sqrt{\frac{\text{Var}_{s' \sim \widetilde{P}_h^i(\cdot|s_h^i,a_h^i)} \widetilde{V}_{h+1}^i(\cdot)}{\widetilde{N}_h^i(s_h^i,a_h^i)}}$. It holds that

$$\begin{aligned}&\sum_{i=1}^k \sum_{h=1}^H \sqrt{\frac{\text{Var}_{s' \sim \widetilde{P}_h^i(\cdot|s_h^i,a_h^i)} \widetilde{V}_{h+1}^i(\cdot)}{\widetilde{N}_h^i(s_h^i,a_h^i)}} \\ \leq & \underbrace{\sqrt{\sum_{i=1}^k \sum_{h=1}^H \frac{1}{N_h^i(s_h^i,a_h^i)}}}_{\leq \widetilde{O}(HSA)} \cdot \underbrace{\sqrt{\sum_{i=1}^k \sum_{h=1}^H \text{Var}_{\widetilde{P}_h^i(\cdot|s_h^i,a_h^i)} \widetilde{V}_{h+1}^i(\cdot)}}_{(a)}.\end{aligned} \quad (11)$$

We bound (a) below (details are deferred to Appendix C.4).

$$(a) \leq \underbrace{\sum_{i=1}^k \sum_{h=1}^H \text{Var}_{P_h(\cdot|s_h^i,a_h^i)} V_{h+1}^{\pi_i}(\cdot)}_{\leq \widetilde{O}(H^2k) \text{ w.h.p due to LTV}} + \text{lower order terms}. \quad (12)$$

Therefore, the main term in the regret bound scales as $\widetilde{O}(\sqrt{H^2SAT_k} + H^2S^2AE_{\epsilon,\beta})$. The details about lower order terms are deferred to Appendix C.3, C.4 and C.5.

Proof of optimism: (2). To prove optimism, we only need

$$b_h^k(s,a) \geq |\widetilde{r}_h^k(s,a) - r_h(s,a)| + |(\widetilde{P}_h^k - P_h) \cdot V_{h+1}^\star(s,a)|.$$

It is clear that $|\widetilde{r}_h^k(s,a) - r_h(s,a)|$ can be bounded by the second term and a portion of the third term of $b_h^k(s,a)$. Due to Lemma 6.1, $|(\widetilde{P}_h^k - P_h) \cdot V_{h+1}^\star(s,a)|$ can be bounded by $\widetilde{O}\left(\sqrt{\text{Var}_{\widehat{P}_h^k(\cdot|s,a)} V_{h+1}^\star(\cdot)/\widetilde{N}_h^k(s,a)}\right)$, which can be further bounded by $\widetilde{O}\left(\sqrt{\text{Var}_{\widetilde{P}_h^k(\cdot|s,a)} V_{h+1}^\star(\cdot)/\widetilde{N}_h^k(s,a)}\right)$ plus a portion of the third term of $b_h^k(s,a)$. Finally, together with the upper bound of $|\widetilde{V}_h^k(s) - V_h^\star(s)|$ (derived from condition of (2) and optimism), the last term of $b_h^k(s,a)$ compensates for the difference of $\sqrt{\text{Var}_{\widetilde{P}_h^k(\cdot|s,a)} \widetilde{V}_{h+1}^k(\cdot)/\widetilde{N}_h^k(s,a)}$ (first term of $b_h^k(s,a)$) and $\sqrt{\text{Var}_{\widetilde{P}_h^k(\cdot|s,a)} V_{h+1}^\star(\cdot)/\widetilde{N}_h^k(s,a)}$. More details about optimism are deferred to Appendix C.6.

7 Simulations

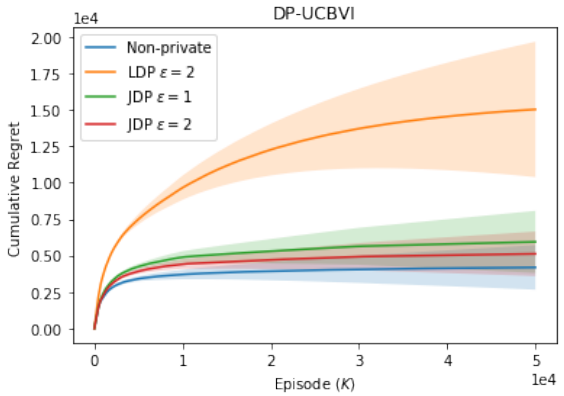

Figure 1: Comparison of cumulative regret for UCBVI and DP-UCBVI with different DP guarantees.

In this section, we run simulations to show the performance of DP-UCBVI (Algorithm 1). We run simulation on a standard benchmark for tabular MDP: Riverswim [Strehl and Littman, 2008], and Chowdhury and Zhou [2021] run simulations on the same environment. Briefly speaking, the

environment consists of six consecutive states and two actions "left" and "right". Choosing "left", the agent will tend to move towards the left side, and vice versa. The agent starts from the left side and tries to reach the right side, where she can get higher reward. For more details and illustration about this setting, please refer to Chowdhury and Zhou [2021].

Similar to Chowdhury and Zhou [2021], we set the planning horizon to be $H = 20$ and run $K = 50000$ episodes. For each algorithm, we run 5 times and derive the average performance and confidence region. We compare the performance of DP-UCBVI under constraints of JDP and LDP, and the original UCBVI. The cumulative regret for each algorithm is shown in Figure 1. Comparing the regret, it is shown that the non-private UCBVI has the best performance, while the cost of privacy under constraints of JDP is a small constant, and thus becomes negligible as the number of episodes increases. In addition, the DP-UCBVI with weaker privacy protection (i.e., larger ϵ) has smaller regret. However, under constrains of LDP, the cost of privacy remains high and it takes a much longer period for the algorithm to converge to near-optimal policies. Our simulation results are consistent with our theories which state that the cost of JDP is a constant term while the cost of LDP is multiplicative.

8 Conclusion

In this paper, we studied the well-motivated problem of differentially private reinforcement learning. Under the tabular MDP setting, we propose a general framework: DP-UCBVI (Algorithm 1) that can be combined with any Privatizers for different variants of DP. Under ϵ-JDP, we achieved regret bound of $\widetilde{O}(\sqrt{SAH^2T} + S^2AH^3/\epsilon)$, which matches the lower bound up to lower order terms. Meanwhile, under ϵ-LDP, we derived regret upper bound of $\widetilde{O}(\sqrt{SAH^2T} + S^2A\sqrt{H^5T}/\epsilon)$ and improves the best known result.

We believe our framework can be further generalized to more general settings, like the linear MDP setting. The best known result under linear MDP [Ngo et al., 2022] built upon LSVI-UCB [Jin et al., 2020], which is arguably a Hoeffding-type algorithm. The main term of regret bound in Ngo et al. [2022], $\widetilde{O}(\sqrt{d^3H^3T})$, is known to be suboptimal due to the recent work [Hu et al., 2022], which incorporates Bernstein-type self-normalized concentration. An interesting future direction is to privatize LSVI-UCB$^+$ (Algorithm 1 in Hu et al. [2022]) and derive tighter regret bounds under linear MDP and constraints of JDP. We believe the techniques in this paper (privatization of Bernstein-type bonus under tabular MDP) could serve as basic building blocks.

Acknowledgements

The research is partially supported by NSF Awards #2007117 and #2048091.

References

M Mehdi Afsar, Trafford Crump, and Behrouz Far. Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys (CSUR)*, 2021.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR, 2020.

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

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

Gavin Brown, Mark Bun, Vitaly Feldman, Adam Smith, and Kunal Talwar. When is memorization of irrelevant training data necessary for high-accuracy learning? In *ACM SIGACT Symposium on Theory of Computing*, pages 123–132, 2021.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security Symposium (USENIX Security 19)*, pages 267–284, 2019.

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

Sayak Ray Chowdhury and Xingyu Zhou. Differentially private regret minimization in episodic markov decision processes. *arXiv preprint arXiv:2112.10599*, 2021.

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

Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pages 1507–1516. PMLR, 2019.

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

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

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

Frederick Arthur Ficken. *The simplex method of linear programming*. Courier Dover Publications, 2015.

Evrard Garcelon, Vianney Perchet, Ciara Pike-Burke, and Matteo Pirotta. Local differential privacy for regret minimization in reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Evrard Garcelon, Kamalika Chaudhuri, Vianney Perchet, and Matteo Pirotta. Privacy amplification via shuffling for linear contextual bandits. In *International Conference on Algorithmic Learning Theory*, pages 381–407. PMLR, 2022.

Justin Hsu, Zhiyi Huang, Aaron Roth, Tim Roughgarden, and Zhiwei Steven Wu. Private matchings and allocations. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 21–30, 2014.

Pihe Hu, Yu Chen, and Longbo Huang. Nearly minimax optimal reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 8971–9019. PMLR, 2022.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

Peter Kairouz, Brendan McMahan, Shuang Song, Om Thakkar, Abhradeep Thakurta, and Zheng Xu. Practical and private (deep) learning without sampling or shuffling. In *International Conference on Machine Learning*, pages 5213–5225. PMLR, 2021.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49 (2-3):209–232, 2002.

Michael Kearns, Mallesh Pai, Aaron Roth, and Jonathan Ullman. Mechanism design in large games: Incentives and privacy. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pages 403–410, 2014.

Chonghua Liao, Jiafan He, and Quanquan Gu. Locally differentially private reinforcement learning for linear mixture markov decision processes. *arXiv preprint arXiv:2110.10133*, 2021.

George Nemhauser and Laurence Wolsey. Polynomial-time algorithms for linear programming. *Integer and Combinatorial Optimization*, pages 146–181, 1988.

Dung Daniel T Ngo, Giuseppe Vietri, and Steven Wu. Improved regret for differentially private exploration in linear mdp. In *International Conference on Machine Learning*, pages 16529–16552. PMLR, 2022.

Dan Qiao and Yu-Xiang Wang. Near-optimal deployment efficiency in reward-free reinforcement learning with linear function approximation. *arXiv preprint arXiv:2210.00701*, 2022a.

Dan Qiao and Yu-Xiang Wang. Offline reinforcement learning with differential privacy. *arXiv preprint arXiv:2206.00810*, 2022b.

Dan Qiao, Ming Yin, Ming Min, and Yu-Xiang Wang. Sample-efficient reinforcement learning with loglog(T) switching cost. In *International Conference on Machine Learning*, pages 18031–18061. PMLR, 2022.

Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. In *Machine Learning for Healthcare Conference*, pages 147–163, 2017.

Ahmad EL Sallab, Mohammed Abdou, Etienne Perot, and Senthil Yogamani. Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, 2017 (19):70–76, 2017.

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

Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74 (8):1309–1331, 2008.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Giuseppe Vietri, Borja Balle, Akshay Krishnamurthy, and Steven Wu. Private reinforcement learning with pac and regret guarantees. In *International Conference on Machine Learning*, pages 9754–9764. PMLR, 2020.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

Jianyu Xu and Yu-Xiang Wang. Logarithmic regret in feature-based dynamic pricing. *Advances in Neural Information Processing Systems*, 34:13898–13910, 2021.

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

Jianyu Xu, Dan Qiao, and Yu-Xiang Wang. Doubly fair dynamic pricing. *arXiv preprint arXiv:2209.11837*, 2022.

Ming Yin and Yu-Xiang Wang. Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems*, 34, 2021.

Ming Yin, Yaqi Duan, Mengdi Wang, and Yu-Xiang Wang. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. *arXiv preprint arXiv:2203.05804*, 2022.

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

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learning via reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33:15198–15207, 2020.

Fuheng Zhao, Dan Qiao, Rachel Redberg, Divyakant Agrawal, Amr El Abbadi, and Yu-Xiang Wang. Differentially private linear sketches: Efficient implementations and applications. *arXiv preprint arXiv:2205.09873*, 2022.

Kai Zheng, Tianle Cai, Weiran Huang, Zhenguo Li, and Liwei Wang. Locally differentially private (contextual) bandits learning. *Advances in Neural Information Processing Systems*, 33:12300–12310, 2020.

Xingyu Zhou. Differentially private reinforcement learning with linear function approximation. *arXiv preprint arXiv:2201.07052*, 2022.

A Extended related works

Regret minimization under tabular MDP. Under the most fundamental setting of tabular MDP, regret minimization has been widely studied by a long stream of works [Kearns and Singh, 2002, Jaksch et al., 2010, Jin et al., 2018, Xu and Wang, 2021, Qiao et al., 2022, Xu et al., 2022, Qiao and Wang, 2022a, Xu and Wang, 2022]. Among the optimal results, Azar et al. [2017] designed an UCB-based algorithm: UCBVI and derived the minimax optimal regret bound $\widetilde{O}(\sqrt{HSAT})$ under stationary MDP. Later, Zhang et al. [2020] achieved the optimal regret bound $\widetilde{O}(\sqrt{H^2SAT})$ under non-stationary MDP through Q-learning type algorithm: UCB-ADVANTAGE. Meanwhile, in addition to stating optimal regret bound, Dann et al. [2019] also provided policy certificates via their algorithm: ORLC. Different from the minimax optimal algorithms above, Zanette and Brunskill [2019] designed an algorithm: EULER and derived the first problem-dependent regret bound, which can imply the minimax optimal regret.

Other differentially private reinforcement learning algorithms. In this paragraph, we discuss about algorithms under linear MDP or linear mixture MDP. Under linear MDP, the only algorithm with JDP guarantee: Private LSVI-UCB [Ngo et al., 2022] is private version of LSVI-UCB [Jin et al., 2020], while LDP under linear MDP still remains open. Under linear mixture MDP, LinOpt-VI-Reg [Zhou, 2022] generalized UCRL-VTR [Ayoub et al., 2020] to guarantee JDP. In addition, Liao et al. [2021] also privatized UCRL-VTR for LDP guarantee. On the offline side, Qiao and Wang [2022b] provided the first result under linear MDP based on VAPVI [Yin et al., 2022].

B Properties of private estimations

In this section, we present some useful concentrations about our private estimations that hold with high probability. Throughout the proof, we denote the non-private estimations by:

$$\begin{aligned} \widehat{P}_h^k(s'|s,a) &= \frac{N_h^k(s,a,s')}{N_h^k(s,a)}, \\ \widehat{r}_h^k(s,a) &= \frac{R_h^k(s,a)}{N_h^k(s,a)}. \end{aligned} \tag{13}$$

In addition, recall that our private estimations are defined as:

$$\begin{aligned} \widetilde{P}_h^k(s'|s,a) &= \frac{\widetilde{N}_h^k(s,a,s')}{\widetilde{N}_h^k(s,a)}, \\ \widetilde{r}_h^k(s,a) &= \left(\frac{\widetilde{R}_h^k(s,a)}{\widetilde{N}_h^k(s,a)}\right)_{[0,1]}. \end{aligned} \tag{14}$$

Lemma B.1. *With probability* $1-\frac{\beta}{15}$*, for all* $h,s,a,k \in [H]\times\mathcal{S}\times\mathcal{A}\times[K]$*, it holds that:*

$$\left|\widetilde{r}_h^k(s,a) - r_h(s,a)\right| \le \sqrt{\frac{2\iota}{\widetilde{N}_h^k(s,a)}} + \frac{2E_{\epsilon,\beta}}{\widetilde{N}_h^k(s,a)}. \tag{15}$$

Proof of Lemma B.1. We have for all $h, s, a, k \in [H] \times \mathcal{S} \times \mathcal{A} \times [K]$,

$$\begin{aligned}
&\left|\widetilde{r}_h^k(s,a) - r_h(s,a)\right| \le \left|\frac{\widetilde{R}_h^k(s,a)}{\widetilde{N}_h^k(s,a)} - r_h(s,a)\right| \\
\le& \left|\frac{\widetilde{R}_h^k(s,a)}{\widetilde{N}_h^k(s,a)} - \frac{R_h^k(s,a)}{\widetilde{N}_h^k(s,a)}\right| + \left|\frac{R_h^k(s,a)}{\widetilde{N}_h^k(s,a)} - r_h(s,a)\right| \\
\le& \frac{E_{\epsilon,\beta}}{\widetilde{N}_h^k(s,a)} + \left|\frac{N_h^k(s,a)}{\widetilde{N}_h^k(s,a)}\left(\frac{R_h^k(s,a)}{N_h^k(s,a)} - r_h(s,a)\right)\right| + \left|r_h(s,a)\left(\frac{N_h^k(s,a)}{\widetilde{N}_h^k(s,a)} - 1\right)\right| \\
\le& \frac{E_{\epsilon,\beta}}{\widetilde{N}_h^k(s,a)} + \frac{N_h^k(s,a)}{\widetilde{N}_h^k(s,a)} \cdot \sqrt{\frac{2\iota}{N_h^k(s,a)}} + \frac{E_{\epsilon,\beta}}{\widetilde{N}_h^k(s,a)} \\
\le& \sqrt{\frac{2\iota}{\widetilde{N}_h^k(s,a)}} + \frac{2E_{\epsilon,\beta}}{\widetilde{N}_h^k(s,a)},
\end{aligned} \tag{16}$$

where the third and last inequalities are because of Assumption 3.1. The forth inequality holds with probability $1 - \frac{\beta}{15}$ due to Hoeffding's inequality and union bound over h, s, a, k. □

Lemma B.2. *With probability* $1 - \frac{\beta}{15}$*, for all* $h, s, a, k \in [H] \times \mathcal{S} \times \mathcal{A} \times [K]$*, it holds that:*

$$\left\|\widetilde{P}_h^k(\cdot|s,a) - P_h(\cdot|s,a)\right\|_1 \le 2\sqrt{\frac{S\iota}{\widetilde{N}_h^k(s,a)}} + \frac{2SE_{\epsilon,\beta}}{\widetilde{N}_h^k(s,a)}. \tag{17}$$

Proof of Lemma B.2. We have for all $h, s, a, k \in [H] \times \mathcal{S} \times \mathcal{A} \times [K]$,

$$\begin{aligned}
&\left\|\widetilde{P}_h^k(\cdot|s,a) - P_h(\cdot|s,a)\right\|_1 = \sum_{s'}\left|\widetilde{P}_h^k(s'|s,a) - P_h(s'|s,a)\right| \\
\le& \sum_{s'}\left|\frac{\widetilde{N}_h^k(s,a,s') - N_h^k(s,a,s')}{\widetilde{N}_h^k(s,a)}\right| + \sum_{s'}\left|\frac{N_h^k(s,a,s')}{\widetilde{N}_h^k(s,a)} - P_h(s'|s,a)\right| \\
\le& \frac{SE_{\epsilon,\beta}}{\widetilde{N}_h^k(s,a)} + \sum_{s'}\left|\frac{N_h^k(s,a,s')}{N_h^k(s,a)} \cdot \frac{N_h^k(s,a)}{\widetilde{N}_h^k(s,a)} - P_h(s'|s,a)\right| \\
\le& \frac{SE_{\epsilon,\beta}}{\widetilde{N}_h^k(s,a)} + \sum_{s'}\left|\left(\frac{N_h^k(s,a,s')}{N_h^k(s,a)} - P_h(s'|s,a)\right) \cdot \frac{N_h^k(s,a)}{\widetilde{N}_h^k(s,a)}\right| + \sum_{s'}\left|P_h(s'|s,a)\left(\frac{N_h^k(s,a)}{\widetilde{N}_h^k(s,a)} - 1\right)\right| \\
\le& \frac{SE_{\epsilon,\beta}}{\widetilde{N}_h^k(s,a)} + \frac{N_h^k(s,a)}{\widetilde{N}_h^k(s,a)}\left\|\widehat{P}_h^k(\cdot|s,a) - P_h(\cdot|s,a)\right\|_1 + \sum_{s'}\left|P_h(s'|s,a)\frac{E_{\epsilon,\beta}}{\widetilde{N}_h^k(s,a)}\right| \\
\le& \frac{N_h^k(s,a)}{\widetilde{N}_h^k(s,a)} \cdot 2\sqrt{\frac{S\iota}{N_h^k(s,a)}} + \frac{2SE_{\epsilon,\beta}}{\widetilde{N}_h^k(s,a)} \\
\le& 2\sqrt{\frac{S\iota}{\widetilde{N}_h^k(s,a)}} + \frac{2SE_{\epsilon,\beta}}{\widetilde{N}_h^k(s,a)},
\end{aligned} \tag{18}$$

where the second, forth and last inequalities hold since Assumption 3.1. The fifth inequality holds with probability $1 - \frac{\beta}{15}$ according to Theorem 2.1 of Weissman et al. [2003] and union bound. □

Remark B.3. *Similarly, we have for all* $h, s, a, k \in [H] \times \mathcal{S} \times \mathcal{A} \times [K]$*,*

$$\begin{aligned}
\left\|\widetilde{P}_h^k(\cdot|s,a) - \widehat{P}_h^k(\cdot|s,a)\right\|_1 \le& \sum_{s'}\left|\frac{\widetilde{N}_h^k(s,a,s')}{\widetilde{N}_h^k(s,a)} - \frac{N_h^k(s,a,s')}{\widetilde{N}_h^k(s,a)}\right| + \sum_{s'}\left|\frac{N_h^k(s,a,s')}{\widetilde{N}_h^k(s,a)} - \frac{N_h^k(s,a,s')}{N_h^k(s,a)}\right| \\
\le& \frac{2SE_{\epsilon,\beta}}{\widetilde{N}_h^k(s,a)}.
\end{aligned} \tag{19}$$

Lemma B.4. *With probability* $1-\frac{\beta}{15}$*, for all* $h,s,a,s',k\in[H]\times\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[K]$*, it holds that:*

$$\left|\widetilde{P}_h^k(s'|s,a)-P_h(s'|s,a)\right|\le\sqrt{\frac{2P_h(s'|s,a)\iota}{\widetilde{N}_h^k(s,a)}}+\frac{2E_{\epsilon,\beta}\iota}{\widetilde{N}_h^k(s,a)}. \tag{20}$$

Proof of Lemma B.4. We have for all $h,s,a,s',k\in[H]\times\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[K]$,

$$\begin{aligned}
&\left|\widetilde{P}_h^k(s'|s,a)-P_h(s'|s,a)\right|\le\left|\frac{\widetilde{N}_h^k(s,a,s')-N_h^k(s,a,s')}{\widetilde{N}_h^k(s,a)}\right|+\left|\frac{N_h^k(s,a,s')}{\widetilde{N}_h^k(s,a)}-P_h(s'|s,a)\right|\\
\le&\frac{E_{\epsilon,\beta}}{\widetilde{N}_h^k(s,a)}+\left|\frac{N_h^k(s,a,s')}{N_h^k(s,a)}\cdot\frac{N_h^k(s,a)}{\widetilde{N}_h^k(s,a)}-P_h(s'|s,a)\right|\\
\le&\frac{E_{\epsilon,\beta}}{\widetilde{N}_h^k(s,a)}+\left|\left(\frac{N_h^k(s,a,s')}{N_h^k(s,a)}-P_h(s'|s,a)\right)\cdot\frac{N_h^k(s,a)}{\widetilde{N}_h^k(s,a)}\right|+\left|P_h(s'|s,a)\left(\frac{N_h^k(s,a)}{\widetilde{N}_h^k(s,a)}-1\right)\right|\\
\le&\frac{2E_{\epsilon,\beta}}{\widetilde{N}_h^k(s,a)}+\frac{N_h^k(s,a)}{\widetilde{N}_h^k(s,a)}\cdot\left|\widehat{P}_h^k(s'|s,a)-P_h(s'|s,a)\right|\\
\le&\frac{2E_{\epsilon,\beta}}{\widetilde{N}_h^k(s,a)}+\frac{N_h^k(s,a)}{\widetilde{N}_h^k(s,a)}\cdot\left(\sqrt{\frac{2P_h(s'|s,a)\iota}{N_h^k(s,a)}}+\frac{2\iota}{3N_h^k(s,a)}\right)\\
\le&\sqrt{\frac{2P_h(s'|s,a)\iota}{\widetilde{N}_h^k(s,a)}}+\frac{2E_{\epsilon,\beta}\iota}{\widetilde{N}_h^k(s,a)},
\end{aligned} \tag{21}$$

where the second, forth and last inequalities result from Assumption 3.1. The fifth inequality holds with probability $1-\frac{\beta}{15}$ due to Bernstein's inequality and union bound. □

Remark B.5. *Similarly, we have for all* $h,s,a,s',k\in[H]\times\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[K]$*,*

$$\begin{aligned}
\left|\widetilde{P}_h^k(s'|s,a)-\widehat{P}_h^k(s'|s,a)\right|\le&\left|\frac{\widetilde{N}_h^k(s,a,s')}{\widetilde{N}_h^k(s,a)}-\frac{N_h^k(s,a,s')}{\widetilde{N}_h^k(s,a)}\right|+\left|\frac{N_h^k(s,a,s')}{\widetilde{N}_h^k(s,a)}-\frac{N_h^k(s,a,s')}{N_h^k(s,a)}\right|\\
\le&\frac{E_{\epsilon,\beta}}{\widetilde{N}_h^k(s,a)}+\frac{N_h^k(s,a,s')E_{\epsilon,\beta}}{\widetilde{N}_h^k(s,a)\cdot N_h^k(s,a)}\\
\le&\frac{2E_{\epsilon,\beta}}{\widetilde{N}_h^k(s,a)}.
\end{aligned} \tag{22}$$

Lemma B.6. *With probability* $1-\frac{2\beta}{15}$*, for all* $h,s,a,k\in[H]\times\mathcal{S}\times\mathcal{A}\times[K]$*, it holds that:*

$$\left|\left(\widetilde{P}_h^k-P_h\right)\cdot V_{h+1}^\star(s,a)\right|\le\min\left\{\sqrt{\frac{2\mathrm{Var}_{P_h(\cdot|s,a)}V_{h+1}^\star(\cdot)\cdot\iota}{\widetilde{N}_h^k(s,a)}},\sqrt{\frac{2\mathrm{Var}_{\widehat{P}_h^k(\cdot|s,a)}V_{h+1}^\star(\cdot)\cdot\iota}{\widetilde{N}_h^k(s,a)}}\right\}+\frac{2HSE_{\epsilon,\beta}\iota}{\widetilde{N}_h^k(s,a)}. \tag{23}$$

Proof of Lemma B.6. We have for all $h, s, a, k \in [H] \times \mathcal{S} \times \mathcal{A} \times [K]$,

$$\begin{aligned}
&\left|\left(\widetilde{P}_h^k - P_h\right) \cdot V_{h+1}^\star(s,a)\right| \leq \left|\sum_{s'}\left(\widetilde{P}_h^k(s'|s,a) - P_h(s'|s,a)\right) V_{h+1}^\star(s')\right| \\
\leq &\left|\sum_{s'} \frac{\widetilde{N}_h^k(s,a,s') - N_h^k(s,a,s')}{\widetilde{N}_h^k(s,a)} \cdot V_{h+1}^\star(s')\right| + \left|\sum_{s'}\left(\frac{N_h^k(s,a,s')}{\widetilde{N}_h^k(s,a)} - P_h(s'|s,a)\right) V_{h+1}^\star(s')\right| \\
\leq &\frac{HSE_{\epsilon,\beta}}{\widetilde{N}_h^k(s,a)} + \left|\frac{N_h^k(s,a)}{\widetilde{N}_h^k(s,a)} \cdot \sum_{s'}\left(\frac{N_h^k(s,a,s')}{N_h^k(s,a)} - P_h(s'|s,a)\right) V_{h+1}^\star(s')\right| + \left|\sum_{s'} P_h(s'|s,a) V_{h+1}^\star(s')\left(\frac{N_h^k(s,a)}{\widetilde{N}_h^k(s,a)} - 1\right)\right| \\
\leq &\frac{2HSE_{\epsilon,\beta}}{\widetilde{N}_h^k(s,a)} + \left|\frac{N_h^k(s,a)}{\widetilde{N}_h^k(s,a)} \cdot \sum_{s'}\left(\frac{N_h^k(s,a,s')}{N_h^k(s,a)} - P_h(s'|s,a)\right) V_{h+1}^\star(s')\right| \\
\leq &\frac{2HSE_{\epsilon,\beta}}{\widetilde{N}_h^k(s,a)} + \frac{N_h^k(s,a)}{\widetilde{N}_h^k(s,a)} \cdot \min\left\{\sqrt{\frac{2\mathrm{Var}_{P_h(\cdot|s,a)} V_{h+1}^\star(\cdot)\cdot\iota}{N_h^k(s,a)}} + \frac{2H\iota}{3N_h^k(s,a)}, \sqrt{\frac{2\mathrm{Var}_{\widehat{P}_h^k(\cdot|s,a)} V_{h+1}^\star(\cdot)\cdot\iota}{N_h^k(s,a)}} + \frac{7H\iota}{3N_h^k(s,a)}\right\} \\
\leq &\min\left\{\sqrt{\frac{2\mathrm{Var}_{P_h(\cdot|s,a)} V_{h+1}^\star(\cdot)\cdot\iota}{\widetilde{N}_h^k(s,a)}}, \sqrt{\frac{2\mathrm{Var}_{\widehat{P}_h^k(\cdot|s,a)} V_{h+1}^\star(\cdot)\cdot\iota}{\widetilde{N}_h^k(s,a)}}\right\} + \frac{2HSE_{\epsilon,\beta}\iota}{\widetilde{N}_h^k(s,a)},
\end{aligned} \tag{24}$$

where the third, forth and last inequalities come from Assumption 3.1. The fifth inequality holds with probability $1 - \frac{2\beta}{15}$ because of Bernstein's inequality, Empirical Bernstein's inequality and union bound. □

Remark B.7. *Similarly, we have for all $h, s, a, k \in [H] \times \mathcal{S} \times \mathcal{A} \times [K]$,*

$$\left|\left(\widetilde{P}_h^k - \widehat{P}_h^k\right) \cdot V_{h+1}^\star(s,a)\right| \leq H \cdot \left\|\widetilde{P}_h^k(\cdot|s,a) - \widehat{P}_h^k(\cdot|s,a)\right\|_1 \leq \frac{2HSE_{\epsilon,\beta}}{\widetilde{N}_h^k(s,a)}, \tag{25}$$

where the last inequality results from Remark B.3.

Combining all the concentrations, we have the following lemma.

Lemma B.8. *Under the high probability event that Assumption 3.1 holds, with probability at least $1 - \frac{\beta}{3}$, the conclusions in Lemma B.1, Lemma B.2, Lemma B.4, Lemma B.6, Remark B.3, Remark B.5 and Remark B.7 hold simultaneously.*

In the following proof, we will prove under the high probability event where Assumption 3.1 and Lemma B.8 hold. Lastly, we state the following lemma regarding difference of variance.

Lemma B.9 (Lemma C.5 of Qiao and Wang [2022b])**.** *For any function $V \in \mathbb{R}^S$ such that $\|V\|_\infty \leq H$, it holds that*

$$\left|\sqrt{\mathrm{Var}_{\widetilde{P}_h^k(\cdot|s,a)}(V)} - \sqrt{\mathrm{Var}_{\widehat{P}_h^k(\cdot|s,a)}(V)}\right| \leq \sqrt{3}H \cdot \sqrt{\left\|\widetilde{P}_h^k(\cdot|s,a) - \widehat{P}_h^k(\cdot|s,a)\right\|_1}. \tag{26}$$

In addition, according to Remark B.3, the left hand side can be further bounded by

$$\left|\sqrt{\mathrm{Var}_{\widetilde{P}_h^k(\cdot|s,a)}(V)} - \sqrt{\mathrm{Var}_{\widehat{P}_h^k(\cdot|s,a)}(V)}\right| \leq 3H\sqrt{\frac{SE_{\epsilon,\beta}}{\widetilde{N}_h^k(s,a)}}. \tag{27}$$

C Proof of Theorem 4.1

In this section, we assume the conclusions in Assumption 3.1 and Lemma B.8 hold and prove the regret bound.

C.1 Some preparations

C.1.1 Notations

For all $i, j \in [K] \times [H]$, we define the following variances we will use throughout the proof.

$$V_{i,j}^\pi = \mathrm{Var}_{P_j(\cdot|s_j^i, a_j^i)} V_{j+1}^{\pi_i}(\cdot). \tag{28}$$

$$V^{\star}_{i,j} = \mathrm{Var}_{P_j(\cdot|s^i_j,a^i_j)} V^{\star}_{j+1}(\cdot). \tag{29}$$

$$\widetilde{V}_{i,j} = \mathrm{Var}_{\widetilde{P}^i_j(\cdot|s^i_j,a^i_j)} \widetilde{V}^i_{j+1}(\cdot). \tag{30}$$

Next, recall the definition of our private bonus b^k_h.

$$\begin{aligned} b^k_h(s,a) =& 2\sqrt{\frac{\mathrm{Var}_{s'\sim\widetilde{P}^k_h(\cdot|s,a)}\widetilde{V}^k_{h+1}(\cdot)\cdot\iota}{\widetilde{N}^k_h(s,a)}} + \sqrt{\frac{2\iota}{\widetilde{N}^k_h(s,a)}} + \frac{20HSE_{\epsilon,\beta}\cdot\iota}{\widetilde{N}^k_h(s,a)} \\ &+4\sqrt{\iota}\cdot\sqrt{\frac{\sum_{s'}\widetilde{P}^k_h(s'|s,a)\min\left\{\frac{1000^2H^3SA\iota^2}{\widetilde{N}^k_{h+1}(s')} + \frac{1000^2H^4S^4A^2E^2_{\epsilon,\beta}\iota^4}{\widetilde{N}^k_{h+1}(s')^2} + \frac{1000^2H^6S^4A^2\iota^4}{\widetilde{N}^k_{h+1}(s')^2}, H^2\right\}}{\widetilde{N}^k_h(s,a)}}. \end{aligned} \tag{31}$$

According to Assumption 3.1, the private visitation numbers will never underestimate the real ones, therefore it holds that

$$\begin{aligned} b^k_h(s,a) \le& \underbrace{2\sqrt{\frac{\mathrm{Var}_{s'\sim\widetilde{P}^k_h(\cdot|s,a)}\widetilde{V}^k_{h+1}(\cdot)\cdot\iota}{N^k_h(s,a)}} + \sqrt{\frac{2\iota}{N^k_h(s,a)}} + \frac{20HSE_{\epsilon,\beta}\cdot\iota}{N^k_h(s,a)}}_{\mathrm{b}^{\mathrm{k}}_{\mathrm{h},1}(\mathrm{s,a})} \\ &\underbrace{+4\sqrt{\iota}\cdot\sqrt{\frac{\sum_{s'}\widetilde{P}^k_h(s'|s,a)\min\left\{\frac{1000^2H^3SA\iota^2}{N^k_{h+1}(s')} + \frac{1000^2H^4S^4A^2E^2_{\epsilon,\beta}\iota^4}{N^k_{h+1}(s')^2} + \frac{1000^2H^6S^4A^2\iota^4}{N^k_{h+1}(s')^2}, H^2\right\}}{N^k_h(s,a)}}}_{b^k_{h,2}(s,a)}. \end{aligned} \tag{32}$$

For the analysis later, we define $\widehat{b}^k_h(s,a) := 2b^k_{h,1}(s,a) + b^k_{h,2}(s,a)$.

In addition, we define the following three terms for all $(i,j) \in [K]\times[H]$:

$$c_{4,i,j} = \frac{H^2S\iota}{N^i_j(s^i_j,a^i_j)},\ c_{1,i,j} = \sqrt{\frac{2V^{\star}_{i,j}\iota}{N^i_j(s^i_j,a^i_j)}},\ \widehat{b}_{i,j} = \widehat{b}^i_j(s^i_j,a^i_j). \tag{33}$$

C.1.2 Typical episodes

Now we define the typical episodes and the typical episodes with respect to $(h,s) \in [H]\times\mathcal{S}$. Briefly speaking, typical episodes ensure that the number of total episodes or visitation number to some state is large enough.

Definition C.1 (Typical episodes)**.** *We define the general typical episodes as* $[k]_{typ} = \{i : i \in [k],\ i \ge 250H^2S^2A\iota^2\}$. *Also, we define typical episodes with respect to* $(h,s) \in [H]\times\mathcal{S}$ *as:*

$$[k]_{typ,h,s} = \{i \in [k] : N^i_h(s) \ge 250H^2S^2A\iota^2\},$$

where $N^i_h(s)$ *is the real visitation number of* (h,s) *before episode* i.

According to Definition C.1 above, it is clear that

$$H\cdot|[k]/[k]_{typ}| \le 250H^3S^2A\iota^2. \tag{34}$$

In the following proof, when we consider summation over episodes, we can consider only the typical episodes since all episodes that are not typical only contribute to a constant term in final regret bound.

Finally, we define the following summations for all $k,h,s \in [K]\times[H]\times\mathcal{S}$:

$$C_k = \sum_{i=1}^{k}\mathbb{1}(i\in[k]_{\mathrm{typ}})\sum_{j=1}^{H}(c_{1,i,j}+c_{4,i,j}). \tag{35}$$

$$B_k = \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \widehat{b}_{i,j}. \tag{36}$$

$$C_{k,h,s} = \sum_{i=1}^{k} \mathbb{1}(s_h^i = s,\, i \in [k]_{\text{typ},h,s}) \sum_{j=h}^{H} (c_{1,i,j} + c_{4,i,j}). \tag{37}$$

$$B_{k,h,s} = \sum_{i=1}^{k} \mathbb{1}(s_h^i = s,\, i \in [k]_{\text{typ},h,s}) \sum_{j=h}^{H} \widehat{b}_{i,j}. \tag{38}$$

C.2 Our induction

Since we apply Bernstein-type bonus, different from Chowdhury and Zhou [2021], optimism is not very straightforward. This is because even if $\widetilde{V}_h^k$ is upper bound of $V_h^\star$ and $\widetilde{P}_h^k$ is close to $\widehat{P}_h^k$, $\mathrm{Var}_{\widetilde{P}_h^k(\cdot|s,a)} \widetilde{V}_{h+1}^k(\cdot)$ is not necessarily an upper bound of $\mathrm{Var}_{\widehat{P}_h^k(\cdot|s,a)} V_{h+1}^\star(\cdot)$. However, we can prove by induction that $\widetilde{V}_{h+1}^k$ is close enough to $V_{h+1}^\star$, and therefore the last term of b_h^k will be sufficiently large to make $\widetilde{V}_h^k$ a valid upper bound of $V_h^\star$. More precisely, our induction is as below:

1. Assume for all $(i,h,s,a) \in [k] \times [H] \times \mathcal{S} \times \mathcal{A}$, $\widetilde{Q}_h^i(s,a) \geq Q_h^\star(s,a)$, we prove for all $(h,s) \in [H] \times \mathcal{S}$,

$$\widetilde{V}_h^k(s) - V_h^\star(s) \leq \widetilde{O}\left(\sqrt{SAH^3/N_h^k(s)} + S^2AH^2E_{\epsilon,\beta}/N_h^k(s) + S^2AH^3/N_h^k(s) \right).$$

2. We deduce that the last term of b_h^k compensates for the possible variance difference and for all $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$, $\widetilde{Q}_h^{k+1}(s,a) \geq Q_h^\star(s,a)$.

Next, we will first prove the point 1 above under optimism in Section C.3, Section C.4 and Section C.5, and then prove optimism (point 2 above) based on point 1 in Section C.6.

C.3 Error decomposition

We define $\delta_{i,h} := V_h^\star(s_h^i) - V_h^{\pi_i}(s_h^i)$ and $\widetilde{\delta}_{i,h} := \widetilde{V}_h^i(s_h^i) - V_h^{\pi_i}(s_h^i)$. Now we provide the error decomposition below, based on optimism, for all $(i,h) \in [k] \times [H]$,

$$\begin{aligned}
&\delta_{i,h} \leq \widetilde{\delta}_{i,h} = \widetilde{V}_h^i(s_h^i) - V_h^{\pi_i}(s_h^i) = \widetilde{Q}_h^i(s_h^i,a_h^i) - Q_h^{\pi_i}(s_h^i,a_h^i) \\
\leq& \widetilde{r}_h^i(s_h^i,a_h^i) + \widetilde{P}_h^i \cdot \widetilde{V}_{h+1}^i(s_h^i,a_h^i) + b_h^i(s_h^i,a_h^i) - r_h(s_h^i,a_h^i) - P_h \cdot V_{h+1}^{\pi_i}(s_h^i,a_h^i) \\
\leq& b_h^i(s_h^i,a_h^i) + \sqrt{\frac{2\iota}{N_h^i(s_h^i,a_h^i)}} + \frac{2E_{\epsilon,\beta}}{N_h^i(s_h^i,a_h^i)} + (\widetilde{P}_h^i - P_h) \cdot V_{h+1}^\star(s_h^i,a_h^i) + (\widetilde{P}_h^i - P_h) \cdot (\widetilde{V}_{h+1}^i - V_{h+1}^\star)(s_h^i,a_h^i) \\
&+ P_h \cdot (\widetilde{V}_{h+1}^i - V_{h+1}^{\pi_i})(s_h^i,a_h^i) \\
\leq& b_h^i(s_h^i,a_h^i) + b_{h,1}^i(s_h^i,a_h^i) + c_{1,i,h} + (\widetilde{P}_h^i - P_h) \cdot (\widetilde{V}_{h+1}^i - V_{h+1}^\star)(s_h^i,a_h^i) + P_h \cdot (\widetilde{V}_{h+1}^i - V_{h+1}^{\pi_i})(s_h^i,a_h^i) - \frac{2HSE_{\epsilon,\beta}\iota}{N_h^i(s_h^i,a_h^i)} \\
\leq& \widehat{b}_{i,h} + c_{1,i,h} + (\widetilde{P}_h^i - P_h) \cdot (\widetilde{V}_{h+1}^i - V_{h+1}^\star)(s_h^i,a_h^i) + P_h \cdot (\widetilde{V}_{h+1}^i - V_{h+1}^{\pi_i})(s_h^i,a_h^i) - \frac{2HSE_{\epsilon,\beta}\iota}{N_h^i(s_h^i,a_h^i)}.
\end{aligned} \tag{39}$$

The second inequality is because of the definition of $\widetilde{Q}_h^i$. The third inequality results from Lemma B.1. The forth inequality holds since Lemma B.6 and the definition of $b_{h,1}^i$, $c_{1,i,h}$. The last inequality holds due to definition of $\widehat{b}_{i,h}$.

In addition, we have:

$$\begin{aligned}
&(\widetilde{P}_h^i - P_h)\cdot(\widetilde{V}_{h+1}^i - V_{h+1}^\star)(s_h^i,a_h^i) = \sum_{s'}\left(\widetilde{P}_h^i(s'|s_h^i,a_h^i) - P_h(s'|s_h^i,a_h^i)\right)\cdot\left(\widetilde{V}_{h+1}^i(s') - V_{h+1}^\star(s')\right)\\
\leq&\sum_{s'}\left(\sqrt{\frac{2P_h(s'|s_h^i,a_h^i)\iota}{N_h^i(s_h^i,a_h^i)}} + \frac{2E_{\epsilon,\beta}\iota}{N_h^i(s_h^i,a_h^i)}\right)\cdot\left(\widetilde{V}_{h+1}^i(s') - V_{h+1}^\star(s')\right)\\
\leq&\sum_{s'}\left(\frac{P_h(s'|s_h^i,a_h^i)}{H} + \frac{H\iota}{N_h^i(s_h^i,a_h^i)} + \frac{2E_{\epsilon,\beta}\iota}{N_h^i(s_h^i,a_h^i)}\right)\cdot\left(\widetilde{V}_{h+1}^i(s') - V_{h+1}^\star(s')\right)\\
\leq&\frac{1}{H}P_h\cdot(\widetilde{V}_{h+1}^i - V_{h+1}^{\pi_i})(s_h^i,a_h^i) + \frac{H^2S\iota}{N_h^i(s_h^i,a_h^i)} + \frac{2HSE_{\epsilon,\beta}\iota}{N_h^i(s_h^i,a_h^i)}.
\end{aligned} \tag{40}$$

The first inequality is because of Lemma B.4. The second inequality holds since AM-GM inequality. The last inequality results from the fact that $V_{h+1}^\star \geq V_{h+1}^{\pi_i}$.

Plugging (40) into (39), we have:

$$\begin{aligned}
\delta_{i,h} \leq \widetilde{\delta}_{i,h} \leq&\widehat{b}_{i,h} + c_{1,i,h} + c_{4,i,h} + (1+\frac{1}{H})P_h\cdot(\widetilde{V}_{h+1}^i - V_{h+1}^{\pi_i})(s_h^i,a_h^i)\\
=&(1+\frac{1}{H})\widetilde{\delta}_{i,h+1} + \widehat{b}_{i,h} + c_{1,i,h} + c_{4,i,h} + (1+\frac{1}{H})\epsilon_{i,h},
\end{aligned} \tag{41}$$

where $\epsilon_{i,h}$ is martingale difference that is bounded in $[-H,H]$.

Recursively applying (41), we have:

$$\widetilde{\delta}_{i,h} \leq 3\sum_{j=h}^{H}\left[\widehat{b}_{i,j} + c_{1,i,j} + c_{4,i,j} + \epsilon_{i,j}\right]. \tag{42}$$

Summing over episodes, we have

$$\sum_{i=1}^{k}\delta_{i,h} \leq \sum_{i=1}^{k}\widetilde{\delta}_{i,h} \leq 3\sum_{i=1}^{k}\sum_{j=h}^{H}\left[\widehat{b}_{i,j} + c_{1,i,j} + c_{4,i,j} + \epsilon_{i,j}\right]. \tag{43}$$

According to Azuma-Hoeffding inequality and union bound, we can bound the partial sum of martingale differences below.

Lemma C.2. *Let $T_k := Hk$ be the number of steps until episode k. Then with probability $1-\frac{\beta}{12}$, the following inequalities hold for all $k,h,s \in [K]\times[H]\times\mathcal{S}$ and $h' \geq h$:*

$$\begin{aligned}
\sum_{i=1}^{k}\mathbb{1}(i\in[k]_{typ})\sum_{j=h}^{H}\epsilon_{i,j} \leq H\sqrt{T_k\iota}.\\
\sum_{i=1}^{k}\mathbb{1}(s_h^i = s,\, i\in[k]_{typ,h,s})\sum_{j=h'}^{H}\epsilon_{i,j} \leq H\sqrt{HN_h^k(s)\iota}.
\end{aligned} \tag{44}$$

We define $U_{k,h} = 3\sum_{i=1}^{k}\mathbb{1}(i\in[k]_{\text{typ}})\sum_{j=h}^{H}\left[\widehat{b}_{i,j} + c_{1,i,j} + c_{4,i,j}\right] + 3H\sqrt{T_k\iota}$ and

$$U_{k,h,s} = 3\sum_{i=1}^{k}\mathbb{1}(s_h^i = s,\, i\in[k]_{\text{typ},h,s})\sum_{j=h}^{H}\left[\widehat{b}_{i,j} + c_{1,i,j} + c_{4,i,j}\right] + 3H\sqrt{HN_h^k(s)\iota}.$$

Therefore, combining (42) and Lemma C.2, we have the following key lemma that upper bounds summation of $\delta_{i,j}$.

Lemma C.3. *Under the high probability event in Lemma C.2, for all $h,s\in[H]\times\mathcal{S}$,*

$$\begin{aligned}
\sum_{i=1}^{k}\mathbb{1}(i\in[k]_{typ})\delta_{i,h} \leq \sum_{i=1}^{k}\mathbb{1}(i\in[k]_{typ})\widetilde{\delta}_{i,h} \leq U_{k,h} \leq U_{k,1},\\
\sum_{i=1}^{k}\mathbb{1}(i\in[k]_{typ})\sum_{j=h}^{H}\widetilde{\delta}_{i,j} \leq HU_{k,1}.
\end{aligned} \tag{45}$$

At the same time, for all $h, s \in [H] \times \mathcal{S}$ and $j \geq h$,

$$\sum_{i=1}^{k} \mathbb{1}(s_h^i = s,\, i \in [k]_{typ,h,s}) \delta_{i,j} \leq \sum_{i=1}^{k} \mathbb{1}(s_h^i = s,\, i \in [k]_{typ,h,s}) \widetilde{\delta}_{i,j} \leq U_{k,h,s},$$
$$\sum_{i=1}^{k} \mathbb{1}(s_h^i = s,\, i \in [k]_{typ,h,s}) \sum_{j=h}^{H} \widetilde{\delta}_{i,j} \leq H U_{k,h,s}. \tag{46}$$

C.4 Upper bounds of variance

From the analysis above, it suffices to derive upper bounds for C_k, B_k, $C_{k,h,s}$ and $B_{k,h,s}$. As a middle step, we discuss several upper bounds about summation of variances. We begin with the following lemma from Azar et al. [2017]. Recall that $V_{i,j}^{\pi} = \mathrm{Var}_{P_j(\cdot|s_j^i,a_j^i)} V_{j+1}^{\pi_i}(\cdot)$ and $T_k = Hk$.

Lemma C.4 (Lemma 8 of Azar et al. [2017])**.** *With probability $1 - \frac{\beta}{12}$, for all $k, h, s \in [K] \times [H] \times \mathcal{S}$, it holds that*

$$\sum_{i=1}^{k} \mathbb{1}(i \in [k]_{typ}) \sum_{j=h}^{H} V_{i,j}^{\pi} \leq H T_k + 2\sqrt{H^4 T_k \iota} + H^3 \iota \leq 2 H T_k,$$
$$\sum_{i=1}^{k} \mathbb{1}(s_h^i = s,\, i \in [k]_{typ,h,s}) \sum_{j=h}^{H} V_{i,j}^{\pi} \leq H^2 N_h^k(s) + 2\sqrt{H^5 N_h^k(s) \iota} + H^3 \iota \leq 2 H^2 N_h^k(s). \tag{47}$$

The proof results from a combination of Law of total variance, Freedman's inequality, union bound and our definition of typical episodes, more details can be found in Azar et al. [2017]. Next, we provide another upper bound for further bounding C_k (and $C_{k,h,s}$). Recall that $V_{i,j}^{\star} = \mathrm{Var}_{P_j(\cdot|s_j^i,a_j^i)} V_{j+1}^{\star}(\cdot)$.

Lemma C.5. *Under the high probability event in Lemma C.2, it holds that*

$$\sum_{i=1}^{k} \mathbb{1}(i \in [k]_{typ}) \sum_{j=1}^{H} \left(V_{i,j}^{\star} - V_{i,j}^{\pi}\right) \leq 2 H^2 U_{k,1} + 2 H^2 \sqrt{T_k \iota}. \tag{48}$$

Similarly, under the same high probability event, for all $(h, s) \in [H] \times \mathcal{S}$,

$$\sum_{i=1}^{k} \mathbb{1}(s_h^i = s,\, i \in [k]_{typ,h,s}) \sum_{j=h}^{H} \left(V_{i,j}^{\star} - V_{i,j}^{\pi}\right) \leq 2 H^2 U_{k,h,s} + 2 H^2 \sqrt{H N_h^k(s) \iota}. \tag{49}$$

Proof of Lemma C.5. We only prove the first conclusion, the second one can be proven in identical way.

$$\begin{aligned}
&\sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \left(V_{i,j}^{\star} - V_{i,j}^{\pi}\right) \leq \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \mathbb{E}_{s' \sim P_j(\cdot|s_j^i,a_j^i)} \left[V_{j+1}^{\star}(s')^2 - V_{j+1}^{\pi_i}(s')^2\right] \\
\leq & 2H \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \mathbb{E}_{s' \sim P_j(\cdot|s_j^i,a_j^i)} \left[V_{j+1}^{\star}(s') - V_{j+1}^{\pi_i}(s')\right] \\
\leq & 2H \left(\sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \widetilde{\delta}_{i,j+1} + H\sqrt{T_k \iota} \right) \\
\leq & 2H^2 U_{k,1} + 2H^2 \sqrt{T_k \iota}.
\end{aligned} \tag{50}$$

The first inequality is because $V_{j+1}^{\star} \geq V_{j+1}^{\pi_i}$. The second inequality results from the fact that $V_{j+1}^{\star} + V_{j+1}^{\pi_i} \leq 2H$. The third inequality holds since Lemma C.2. The last inequality is due to Lemma C.3. □

Lastly, we prove the following lemma for bounding B_k (and $B_{k,h,s}$). Recall that $\widetilde{V}_{i,j} = \mathrm{Var}_{\widetilde{P}_j^i(\cdot|s_j^i,a_j^i)} \widetilde{V}_{j+1}^i(\cdot)$.

Lemma C.6. *Under the high probability event in Lemma C.2, with probability* $1-\frac{\beta}{12K}$*, it holds that*

$$\sum_{i=1}^{k}\mathbb{1}(i\in[k]_{typ})\sum_{j=1}^{H}\left(\widetilde{V}_{i,j}-V_{i,j}^{\pi}\right)\le 2H^2U_{k,1}+8H^2S\sqrt{HAT_k\iota}+6H^3S^2AE_{\epsilon,\beta}\iota. \tag{51}$$

Similarly, under the same high probability event, for all $(h,s)\in[H]\times\mathcal{S}$*,*

$$\sum_{i=1}^{k}\mathbb{1}(s_h^i=s,\,i\in[k]_{typ,h,s})\sum_{j=h}^{H}\left(\widetilde{V}_{i,j}-V_{i,j}^{\pi}\right)\le 2H^2U_{k,h,s}+8H^3S\sqrt{AN_h^k(s)\iota}+6H^3S^2AE_{\epsilon,\beta}\iota. \tag{52}$$

Proof of Lemma C.6. We only prove the first conclusion, the second one can be proven in identical way. We have

$$\begin{aligned}
\sum_{i=1}^{k}\mathbb{1}(i\in[k]_{\text{typ}})\sum_{j=1}^{H}\left(\widetilde{V}_{i,j}-V_{i,j}^{\pi}\right)\le&\underbrace{\sum_{i=1}^{k}\mathbb{1}(i\in[k]_{\text{typ}})\sum_{j=1}^{H}\left[\mathbb{E}_{s'\sim\widetilde{P}_j^i(\cdot|s_j^i,a_j^i)}\widetilde{V}_{j+1}^i(s')^2-\mathbb{E}_{s'\sim P_j(\cdot|s_j^i,a_j^i)}\widetilde{V}_{j+1}^i(s')^2\right]}_{\text{(a)}}\\
&+\underbrace{\sum_{i=1}^{k}\mathbb{1}(i\in[k]_{\text{typ}})\sum_{j=1}^{H}\mathbb{E}_{s'\sim P_j(\cdot|s_j^i,a_j^i)}\left[\widetilde{V}_{j+1}^i(s')^2-V_{j+1}^{\pi_i}(s')^2\right]}_{\text{(b)}}\\
&+\underbrace{\sum_{i=1}^{k}\mathbb{1}(i\in[k]_{\text{typ}})\sum_{j=1}^{H}\left[\left(\mathbb{E}_{s'\sim P_j(\cdot|s_j^i,a_j^i)}V_{j+1}^{\star}(s')\right)^2-\left(\mathbb{E}_{s'\sim\widetilde{P}_j^i(\cdot|s_j^i,a_j^i)}V_{j+1}^{\star}(s')\right)^2\right]}_{\text{(c)}}.
\end{aligned} \tag{53}$$

The inequality holds because of direct calculation and the fact that $V_{j+1}^{\pi_i}\le V_{j+1}^{\star}\le\widetilde{V}_{j+1}^i$. Next we bound these terms separately. First of all,

$$\begin{aligned}
\text{(a)}\le&\sum_{i=1}^{k}\mathbb{1}(i\in[k]_{\text{typ}})\sum_{j=1}^{H}H^2\cdot\left\|\widetilde{P}_j^i(\cdot|s_j^i,a_j^i)-P_j(\cdot|s_j^i,a_j^i)\right\|_1\\
\le&\sum_{i=1}^{k}\mathbb{1}(i\in[k]_{\text{typ}})\sum_{j=1}^{H}H^2\left(2\sqrt{\frac{S\iota}{N_j^i(s_j^i,a_j^i)}}+\frac{2SE_{\epsilon,\beta}}{N_j^i(s_j^i,a_j^i)}\right)\\
\le&2H^2S\sqrt{HAT_k\iota}+2H^3S^2AE_{\epsilon,\beta}\iota.
\end{aligned} \tag{54}$$

The second inequality comes from Lemma B.2. The last inequality is because of direct calculation. For (b), according to Lemma C.2, similar to the proof of Lemma C.5, it holds that

$$\text{(b)}\le 2H\left(\sum_{i=1}^{k}\mathbb{1}(i\in[k]_{\text{typ}})\sum_{j=1}^{H}\widetilde{\delta}_{i,j+1}+H\sqrt{T_k\iota}\right)\le 2H^2U_{k,1}+2H^2\sqrt{T_k\iota}. \tag{55}$$

Lastly, for (c), we have

$$\begin{aligned}
\text{(c)}\le&\sum_{i=1}^{k}\mathbb{1}(i\in[k]_{\text{typ}})\sum_{j=1}^{H}\left[\left(\mathbb{E}_{s'\sim P_j(\cdot|s_j^i,a_j^i)}V_{j+1}^{\star}(s')\right)^2-\left(\mathbb{E}_{s'\sim\widehat{P}_j^i(\cdot|s_j^i,a_j^i)}V_{j+1}^{\star}(s')\right)^2+2H^2\cdot\left\|\left(\widetilde{P}_j^i-\widehat{P}_j^i\right)(\cdot|s_j^i,a_j^i)\right\|_1\right]\\
\le&\sum_{i=1}^{k}\mathbb{1}(i\in[k]_{\text{typ}})\sum_{j=1}^{H}\left(2H\cdot 2H\sqrt{\frac{\iota}{N_j^i(s_j^i,a_j^i)}}+2H^2\cdot\frac{2SE_{\epsilon,\beta}}{N_j^i(s_j^i,a_j^i)}\right)\\
\le&4H^2\sqrt{HSAT_k\iota}+4H^3S^2AE_{\epsilon,\beta}\iota.
\end{aligned} \tag{56}$$

The second inequality holds with probability $1-\frac{\beta}{12K}$ due to Hoeffding's inequality and Remark B.3. The last inequality holds because of direct calculation.

Combining the upper bounds of $\text{(a)},\text{(b)},\text{(c)}$, the proof is complete. □

C.5 Upper bound of regret

With the upper bounds in Section C.4, we are ready to bound C_k, B_k and the regret. In this section, we assume the high probability events in Lemma C.2 and Lemma C.4 hold. We begin with the upper bound of C_k and $C_{k,h,s}$. Recall that $C_k = \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} (c_{1,i,j} + c_{4,i,j})$ and $C_{k,h,s} = \sum_{i=1}^{k} \mathbb{1}(s_h^i = s,\, i \in [k]_{\text{typ},h,s}) \sum_{j=h}^{H} (c_{1,i,j} + c_{4,i,j})$.

Lemma C.7. *Under the high probability events in Lemma C.2 and Lemma C.4, we have*

$$C_k = \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{typ}) \sum_{j=1}^{H} (c_{1,i,j} + c_{4,i,j}) \le 3\sqrt{H^2SAT_k\iota^2} + 2\sqrt{H^3SAU_{k,1}\iota^2} + H^3S^2A\iota^2. \tag{57}$$

Similarly, under the same high probability event, for all $h, s \in [H] \times \mathcal{S}$,

$$C_{k,h,s} \le 3\sqrt{H^3SAN_h^k(s)\iota^2} + 2\sqrt{H^3SAU_{k,h,s}\iota^2} + H^3S^2A\iota^2. \tag{58}$$

Proof of Lemma C.7. We only prove the first conclusion, the second one can be proven in identical way. We have

$$C_k = \underbrace{\sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} c_{1,i,j}}_{\text{(a)}} + \underbrace{\sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} c_{4,i,j}}_{\text{(b)}}. \tag{59}$$

For (a), due to Cauchy-Schwarz inequality, it holds that

$$\begin{aligned} (\mathrm{a}) =& \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \sqrt{\frac{2V_{i,j}^{\star}\iota}{N_j^i(s_j^i, a_j^i)}} \\ \le& \sqrt{2\iota} \cdot \sqrt{\underbrace{\sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \frac{1}{N_j^i(s_j^i, a_j^i)}}_{\text{(c)}}} \cdot \sqrt{\underbrace{\sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} V_{i,j}^{\star}}_{\text{(d)}}}. \end{aligned} \tag{60}$$

Due to direct calculation, we have $(\mathrm{c}) \le HSA\iota$ and in addition,

$$\begin{aligned} (\mathrm{d}) \le& \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} V_{i,j}^{\pi} + \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \left(V_{i,j}^{\star} - V_{i,j}^{\pi}\right) \\ \le& 2HT_k + 2H^2U_{k,1} + 2H^2\sqrt{T_k\iota}, \end{aligned} \tag{61}$$

where the second inequality holds due to Lemma C.4 and Lemma C.5. Therefore, We have

$$(\mathrm{a}) \le 3\sqrt{H^2SAT_k\iota^2} + 2\sqrt{H^3SAU_{k,1}\iota^2}. \tag{62}$$

For (b), according to direct calculation, it holds that

$$(\mathrm{b}) \le \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \frac{H^2S\iota}{N_j^i(s_j^i, a_j^i)} \le H^3S^2A\iota^2. \tag{63}$$

Combining (62) and (63), the proof is complete. □

Next, we bound B_k and $B_{k,h,s}$. Recall that $B_k = \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \widehat{b}_{i,j}$ and $B_{k,h,s} = \sum_{i=1}^{k} \mathbb{1}(s_h^i = s,\, i \in [k]_{\text{typ},h,s}) \sum_{j=h}^{H} \widehat{b}_{i,j}$.

Lemma C.8. *Under the high probability events in Lemma C.2, Lemma C.4 and Lemma C.6, with probability $1 - \frac{\beta}{12K}$,*

$$B_k \le 16\sqrt{H^2SAT_k\iota^2} + 6\sqrt{H^3SAU_{k,1}\iota^2} + 250H^2S^2AE_{\epsilon,\beta}\iota^2 + 250H^3S^2A\iota^2. \tag{64}$$

Similarly, under the same high probability event, for all $h, s \in [H] \times \mathcal{S}$,

$$B_{k,h,s} \le 16\sqrt{H^3SAN_h^k(s)\iota^2} + 6\sqrt{H^3SAU_{k,h,s}\iota^2} + 250H^2S^2AE_{\epsilon,\beta}\iota^2 + 250H^3S^2A\iota^2. \tag{65}$$

Proof of Lemma C.8. We only prove the first conclusion, the second one can be proven in identical way. We have

$$B_k \leq \underbrace{\sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} 4\sqrt{\frac{\widetilde{V}_{i,j}\iota}{N_j^i(s_j^i, a_j^i)}}}_{\text{(a)}} + \underbrace{\sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} 2\sqrt{\frac{2\iota}{N_j^i(s_j^i, a_j^i)}}}_{\text{(b)}} + \underbrace{\sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \frac{40HSE_{\epsilon,\beta}\iota}{N_j^i(s_j^i, a_j^i)}}_{\text{(c)}}$$
$$+ \underbrace{\sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} 4\sqrt{\iota} \cdot \sqrt{\frac{\sum_{s'} \widetilde{P}_j^i(s'|s_j^i, a_j^i) \min\left\{\frac{1000^2 H^3 SA\iota^2}{N_{j+1}^i(s')} + \frac{1000^2 H^4 S^4 A^2 E_{\epsilon,\beta}^2 \iota^4}{N_{j+1}^i(s')^2} + \frac{1000^2 H^6 S^4 A^2 \iota^4}{N_{j+1}^i(s')^2}, H^2\right\}}{N_j^i(s_j^i, a_j^i)}}}_{\text{(d)}}. \tag{66}$$

We bound these terms separately, we first bound (a) below.

$$\begin{aligned}
(\mathrm{a}) \leq & 4\sqrt{\iota} \cdot \sqrt{\sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \frac{1}{N_j^i(s_j^i, a_j^i)}} \cdot \sqrt{\sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \widetilde{V}_{i,j}} \\
\leq & 4\sqrt{\iota} \cdot \sqrt{HSA\iota} \cdot \sqrt{\sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} V_{i,j}^{\pi} + \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \left(\widetilde{V}_{i,j} - V_{i,j}^{\pi}\right)} \\
\leq & 4\iota \cdot \sqrt{HSA} \sqrt{2HT_k + 2H^2 U_{k,1} + 8H^2 S\sqrt{HAT_k\iota} + 6H^3 S^2 AE_{\epsilon,\beta}\iota} \\
\leq & 8\sqrt{H^2 SAT_k\iota^2} + 4\sqrt{2H^3 SAU_{k,1}\iota^2} + 10\sqrt{H^4 S^3 A^2 E_{\epsilon,\beta}\iota^3}.
\end{aligned} \tag{67}$$

The first inequality is because of Cauchy-Schwarz inequality. The second inequality results from direct calculation. The third inequality is due to Lemma C.4 and Lemma C.6. The last inequality holds because for typical episodes, $8H^2 S\sqrt{HAT_k\iota} \leq 2HT_k$.

According to direct calculation,

$$(\mathrm{b}) \leq 2\sqrt{2\iota \cdot HSAT_k} \leq \sqrt{H^2 SAT_k\iota^2} \tag{68}$$

For (c), it holds that

$$(\mathrm{c}) \leq 40HSE_{\epsilon,\beta}\iota \cdot HSA\iota \leq 40H^2 S^2 AE_{\epsilon,\beta}\iota^2. \tag{69}$$

Finally, we bound the most complex term (d) as below. Because of Cauchy-Schwarz inequality,

$$\begin{aligned}
(\mathrm{d}) \leq & \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} 4\sqrt{\iota} \cdot \sqrt{\frac{\sum_{s'} \widetilde{P}_j^i(s'|s_j^i, a_j^i) \min\left\{\frac{1000^2 H^3 SA\iota^2}{N_{j+1}^i(s')} + \frac{1000^2 H^4 S^4 A^2 E_{\epsilon,\beta}^2 \iota^4}{N_{j+1}^i(s')^2} + \frac{1000^2 H^6 S^4 A^2 \iota^4}{N_{j+1}^i(s')^2}, H^2\right\}}{N_j^i(s_j^i, a_j^i)}} \\
\leq & 4 \sqrt{\underbrace{\iota \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \frac{1}{N_j^i(s_j^i, a_j^i)}}_{\text{(e)}}} \cdot \\
& \sqrt{\underbrace{\sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \sum_{s'} \widetilde{P}_j^i(s'|s_j^i, a_j^i) \min\left\{\frac{1000^2 H^3 SA\iota^2}{N_{j+1}^i(s')} + \frac{1000^2 H^4 S^4 A^2 E_{\epsilon,\beta}^2 \iota^4}{N_{j+1}^i(s')^2} + \frac{1000^2 H^6 S^4 A^2 \iota^4}{N_{j+1}^i(s')^2}, H^2\right\}}_{\text{(f)}}}.
\end{aligned} \tag{70}$$

Note that $(\mathrm{e}) \le HSA\iota$. For (f), we have with probability $1-\frac{\beta}{12K}$,

$$
\begin{aligned}
(\mathrm{f}) \le & \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} H^2 \left\| \widetilde{P}_j^i(\cdot|s_j^i, a_j^i) - P_j(\cdot|s_j^i, a_j^i) \right\|_1 \\
& + \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \sum_{s'} P_j(s'|s_j^i, a_j^i) \min\left\{ \frac{1000^2 H^3 SA\iota^2}{N_{j+1}^i(s')} + \frac{1000^2 H^4 S^4 A^2 E_{\epsilon,\beta}^2 \iota^4}{N_{j+1}^i(s')^2} + \frac{1000^2 H^6 S^4 A^2 \iota^4}{N_{j+1}^i(s')^2}, H^2 \right\} \\
\le & \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} H^2 \left\| \widetilde{P}_j^i(\cdot|s_j^i, a_j^i) - P_j(\cdot|s_j^i, a_j^i) \right\|_1 + H^2\sqrt{T_k\iota} \\
& + \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \min\left\{ \frac{1000^2 H^3 SA\iota^2}{N_{j+1}^i(s_{j+1}^i)} + \frac{1000^2 H^4 S^4 A^2 E_{\epsilon,\beta}^2 \iota^4}{N_{j+1}^i(s_{j+1}^i)^2} + \frac{1000^2 H^6 S^4 A^2 \iota^4}{N_{j+1}^i(s_{j+1}^i)^2}, H^2 \right\} \\
\le & \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} H^2 \left(2\sqrt{\frac{S\iota}{N_j^i(s_j^i, a_j^i)}} + \frac{2SE_{\epsilon,\beta}}{N_j^i(s_j^i, a_j^i)} \right) + \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \min\left\{ \frac{1000^2 H^3 SA\iota^2}{N_{j+1}^i(s_{j+1}^i)}, H^2 \right\} \\
& + \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \min\left\{ \frac{1000^2 H^4 S^4 A^2 E_{\epsilon,\beta}^2 \iota^4}{N_{j+1}^i(s_{j+1}^i)^2}, H^2 \right\} + \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \min\left\{ \frac{1000^2 H^6 S^4 A^2 \iota^4}{N_{j+1}^i(s_{j+1}^i)^2}, H^2 \right\} \\
& + H^2\sqrt{T_k\iota} \\
\le & 3H^2 S\sqrt{HAT_k\iota} + 2H^3 S^2 AE_{\epsilon,\beta}\iota + 1000H^4 S^3 AE_{\epsilon,\beta}\iota^2 + 2000H^5 S^3 A\iota^2,
\end{aligned} \tag{71}
$$

where the first inequality holds because $\min\{\cdot, H^2\} \le H^2$. The second inequality holds with probability $1-\frac{\beta}{12K}$ due to Azuma-Hoeffding inequality. The third inequality results from Lemma B.2. The last inequality comes from direct calculation. Therefore, we have

$$
(\mathrm{d}) \le 7\sqrt{H^3 S^2 A\iota^2 \sqrt{HAT_k\iota}} + 120\sqrt{H^5 S^4 A^2 E_{\epsilon,\beta}\iota^4} + 200H^3 S^2 A\iota^2. \tag{72}
$$

Combining (67), (68), (69) and (72), the proof is complete. □

Now we are ready to bound the regret until episode k based on optimism. We define the following regret functions.

$$
\text{Regret}(k) := \sum_{i=1}^{k} \delta_{i,1}, \ \widetilde{\text{Regret}}(k) := \sum_{i=1}^{k} \widetilde{\delta}_{i,1}. \tag{73}
$$

In addition, we define the regret with respect to $(h, s) \in [H] \times \mathcal{S}$:

$$
\text{Regret}(k, h, s) := \sum_{i=1}^{k} \mathbb{1}(s_h^i = s) \cdot \delta_{i,h}, \ \widetilde{\text{Regret}}(k, h, s) := \sum_{i=1}^{k} \mathbb{1}(s_h^i = s) \cdot \widetilde{\delta}_{i,h}. \tag{74}
$$

Lemma C.9. *With probability $1-\beta$, for all $k \in [K]$, as well as the optimism holds (*i.e. *for all $(i, h, s, a) \in [k] \times [H] \times \mathcal{S} \times \mathcal{A}$, $\widetilde{Q}_h^i(s, a) \ge Q_h^\star(s, a)$), it holds that*

$$
\textit{Regret}(k) \le \widetilde{\textit{Regret}}(k) \le 1000\left(\sqrt{H^2 SAT_k\iota^2} + H^2 S^2 AE_{\epsilon,\beta}\iota^2 + H^3 S^2 A\iota^2 \right). \tag{75}
$$

In addition, for all $(h, s) \in [H] \times \mathcal{S}$, we have

$$
\textit{Regret}(k, h, s) \le \widetilde{\textit{Regret}}(k, h, s) \le 1000\left(\sqrt{H^3 SAN_h^k(s)\iota^2} + H^2 S^2 AE_{\epsilon,\beta}\iota^2 + H^3 S^2 A\iota^2 \right). \tag{76}
$$

Proof of Lemma C.9. For the proof of this lemma, we assume the high probability events in Assumption 3.1, Lemma B.8, Lemma C.2, Lemma C.4, Lemma C.6 (for all $k \in [K]$) and Lemma C.8 (for all $k \in [K]$) hold. The failure probability is bounded by

$$
\frac{\beta}{3} + \frac{\beta}{3} + \frac{\beta}{12} + \frac{\beta}{12} + K \cdot \frac{\beta}{12K} + K \cdot \frac{\beta}{12K} \le \beta. \tag{77}
$$

We only prove the first conclusion, the second one can be proven in identical way. It holds that

$$\begin{aligned}
&\text{Regret}(k) \le \widetilde{\text{Regret}}(k) = \sum_{i=1}^{k} \widetilde{\delta}_{i,1} \\
\le& U_{k,1} + 250H^3S^2A\iota^2 \\
\le& 3B_k + 3C_k + 3H\sqrt{T_k\iota} + 250H^3S^2A\iota^2 \\
\le& 60\sqrt{H^2SAT_k\iota^2} + 24\sqrt{H^3SAU_{k,1}\iota^2} + 750H^2S^2AE_{\epsilon,\beta}\iota^2 + 1000H^3S^2A\iota^2 \\
\le& 1000\left(\sqrt{H^2SAT_k\iota^2} + H^2S^2AE_{\epsilon,\beta}\iota^2 + H^3S^2A\iota^2\right),
\end{aligned} \tag{78}$$

where the second inequality is because Lemma C.3 and the fact that $H \cdot |[k]/[k]_{typ}| \le 250H^3S^2A\iota^2$. The forth inequality is by combining Lemma C.8 and Lemma C.7. The last inequality results from solving the inequality with respect to $U_{k,1}$. □

Corollary C.10. *Under the event in Lemma C.9, we have*

$$\begin{aligned}
&1000\left(\sqrt{H^3SAN_h^k(s)\iota^2} + H^2S^2AE_{\epsilon,\beta}\iota^2 + H^3S^2A\iota^2\right) \ge \widetilde{Regret}(k,h,s) \\
\ge& \sum_{i=1}^{k} \mathbb{1}(s_h^i = s) \cdot \widetilde{\delta}_{i,h} \\
\ge& \sum_{i=1}^{k} \mathbb{1}(s_h^i = s)\left(\widetilde{V}_h^i(s) - V_h^{\pi_i}(s)\right) \\
\ge& \sum_{i=1}^{k} \mathbb{1}(s_h^i = s)\left(\widetilde{V}_h^i(s) - V_h^{\star}(s)\right) \\
\ge& N_h^k(s) \cdot \left(\widetilde{V}_h^k(s) - V_h^{\star}(s)\right),
\end{aligned} \tag{79}$$

where the first three inequalities are due to definitions of $\widetilde{Regret}(k,h,s)$ and $\widetilde{\delta}_{i,h}$. The forth inequality holds because $V_h^{\star} \ge V_h^{\pi_i}$. The last inequality results from our algorithmic design that $\widetilde{V}_h^i(s)$ is non-increasing (line 9 of Algorithm 1).

Therefore, we have $\widetilde{V}_h^k(s) - V_h^{\star}(s) \le 1000\left(\sqrt{H^3SA\iota^2/N_h^k(s)} + H^2S^2AE_{\epsilon,\beta}\iota^2/N_h^k(s) + H^3S^2A\iota^2/N_h^k(s)\right)$.

Now we have proven the first point of our induction. Together with the point 2 (which we will prove in Section C.6), we have with probability $1-\beta$, the whole induction process is valid.

For clarity, we restate the induction process under the high probability event in Lemma C.9. For all $k \in [K]$,
1. Given that for all $(i,h,s,a) \in [k] \times [H] \times \mathcal{S} \times \mathcal{A}$, $\widetilde{Q}_h^i(s,a) \ge Q_h^{\star}(s,a)$, we prove

$$\text{Regret}(k) \le \widetilde{\text{Regret}}(k) \le 1000\left(\sqrt{H^2SAT_k\iota^2} + H^2S^2AE_{\epsilon,\beta}\iota^2 + H^3S^2A\iota^2\right)$$

and for all $(h,s) \in [H] \times \mathcal{S}$,

$$\widetilde{V}_h^k(s) - V_h^{\star}(s) \le 1000\left(\sqrt{SAH^3\iota^2/N_h^k(s)} + S^2AH^2E_{\epsilon,\beta}\iota^2/N_h^k(s) + S^2AH^3\iota^2/N_h^k(s)\right).$$

2. Given that for all $(h,s) \in [H] \times \mathcal{S}$,
$\widetilde{V}_h^k(s) - V_h^{\star}(s) \le 1000\left(\sqrt{SAH^3\iota^2/N_h^k(s)} + S^2AH^2E_{\epsilon,\beta}\iota^2/N_h^k(s) + S^2AH^3\iota^2/N_h^k(s)\right)$, we prove that for all $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$, $\widetilde{Q}_h^{k+1}(s,a) \ge Q_h^{\star}(s,a)$.

Therefore, with probability $1-\beta$, we have ($T = KH$)

$$\text{Regret}(K) \le \widetilde{\text{Regret}}(K) \le \widetilde{O}\left(\sqrt{H^2SAT} + H^2S^2AE_{\epsilon,\beta} + H^3S^2A\right). \tag{80}$$

This completes the proof of Theorem 4.1.

C.6 Proof of optimism

In this part, we prove optimism. Given the condition that for all $(h,s) \in [H] \times \mathcal{S}$,

$$\widetilde{V}_h^{k+1}(s) - V_h^\star(s) \le 1000\left(\sqrt{SAH^3\iota^2/N_h^{k+1}(s)} + S^2AH^2E_{\epsilon,\beta}\iota^2/N_h^{k+1}(s) + S^2AH^3\iota^2/N_h^{k+1}(s)\right),$$

we prove for all $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$, $\widetilde{Q}_h^{k+1}(s,a) \ge Q_h^\star(s,a)$ through backward induction (induction from $H+1$ to 1). Since the conclusion holds trivially for $H+1$, it suffices to prove the following Lemma C.11.

Lemma C.11. *Under the high probability event in Assumption 3.1 and Lemma B.8, if it holds that*

1. *For all $(j,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$, $\widetilde{Q}_j^k(s,a) \ge Q_j^\star(s,a)$.*
2. *For all $s \in \mathcal{S}$,*
   $0 \le \widetilde{V}_{h+1}^{k+1}(s) - V_{h+1}^\star(s) \le 1000\left(\sqrt{SAH^3\iota^2/N_{h+1}^{k+1}(s)} + S^2AH^2E_{\epsilon,\beta}\iota^2/N_{h+1}^{k+1}(s) + S^2AH^3\iota^2/N_{h+1}^{k+1}(s)\right)$.

Then we have for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, $\widetilde{Q}_h^{k+1}(s,a) \ge Q_h^\star(s,a)$.

Proof of Lemma C.11. For all $(s,a) \in \mathcal{S} \times \mathcal{A}$, since $\widetilde{Q}_h^k(s,a), H \ge Q_h^\star(s,a)$, it suffices to prove that $\widetilde{r}_h^{k+1}(s,a) + \widetilde{P}_h^{k+1} \cdot \widetilde{V}_{h+1}^{k+1}(s,a) + b_h^{k+1}(s,a) \ge Q_h^\star(s,a)$. We have

$$\begin{aligned}
&\widetilde{r}_h^{k+1}(s,a) + \widetilde{P}_h^{k+1} \cdot \widetilde{V}_{h+1}^{k+1}(s,a) + b_h^{k+1}(s,a) - Q_h^\star(s,a) \\
\ge& \left(\widetilde{r}_h^{k+1} - r_h\right)(s,a) + \left(\widetilde{P}_h^{k+1} - P_h\right) \cdot V_{h+1}^\star(s,a) + b_h^{k+1}(s,a) \\
\ge& \left(\widetilde{P}_h^{k+1} - P_h\right) \cdot V_{h+1}^\star(s,a) + 2\sqrt{\frac{\mathrm{Var}_{s' \sim \widetilde{P}_h^{k+1}(\cdot|s,a)}\widetilde{V}_{h+1}^{k+1}(\cdot) \cdot \iota}{\widetilde{N}_h^{k+1}(s,a)}} + \frac{19HSE_{\epsilon,\beta} \cdot \iota}{\widetilde{N}_h^{k+1}(s,a)} \\
&+4\sqrt{\iota} \cdot \sqrt{\frac{\sum_{s'} \widetilde{P}_h^{k+1}(s'|s,a) \min\left\{\frac{1000^2H^3SA\iota^2}{\widetilde{N}_{h+1}^{k+1}(s')} + \frac{1000^2H^4S^4A^2E_{\epsilon,\beta}^2\iota^4}{\widetilde{N}_{h+1}^{k+1}(s')^2} + \frac{1000^2H^6S^4A^2\iota^4}{\widetilde{N}_{h+1}^{k+1}(s')^2}, H^2\right\}}{\widetilde{N}_h^k(s,a)}} \\
\ge& -\sqrt{\frac{2\mathrm{Var}_{s' \sim \widehat{P}_h^{k+1}(\cdot|s,a)}V_{h+1}^\star(\cdot) \cdot \iota}{\widetilde{N}_h^{k+1}(s,a)}} + 2\sqrt{\frac{\mathrm{Var}_{s' \sim \widetilde{P}_h^{k+1}(\cdot|s,a)}\widetilde{V}_{h+1}^{k+1}(\cdot) \cdot \iota}{\widetilde{N}_h^{k+1}(s,a)}} + \frac{17HSE_{\epsilon,\beta} \cdot \iota}{\widetilde{N}_h^{k+1}(s,a)} \\
&+4\sqrt{\iota} \cdot \sqrt{\frac{\sum_{s'} \widetilde{P}_h^{k+1}(s'|s,a) \min\left\{\frac{1000^2H^3SA\iota^2}{\widetilde{N}_{h+1}^{k+1}(s')} + \frac{1000^2H^4S^4A^2E_{\epsilon,\beta}^2\iota^4}{\widetilde{N}_{h+1}^{k+1}(s')^2} + \frac{1000^2H^6S^4A^2\iota^4}{\widetilde{N}_{h+1}^{k+1}(s')^2}, H^2\right\}}{\widetilde{N}_h^k(s,a)}} \\
\ge& -\sqrt{\frac{2\mathrm{Var}_{s' \sim \widehat{P}_h^{k+1}(\cdot|s,a)}V_{h+1}^\star(\cdot) \cdot \iota}{\widetilde{N}_h^{k+1}(s,a)}} + 2\sqrt{\frac{\mathrm{Var}_{s' \sim \widehat{P}_h^{k+1}(\cdot|s,a)}\widetilde{V}_{h+1}^{k+1}(\cdot) \cdot \iota}{\widetilde{N}_h^{k+1}(s,a)}} \\
&+4\sqrt{\iota} \cdot \sqrt{\frac{\sum_{s'} \widehat{P}_h^{k+1}(s'|s,a) \min\left\{\frac{1000^2H^3SA\iota^2}{\widetilde{N}_{h+1}^{k+1}(s')} + \frac{1000^2H^4S^4A^2E_{\epsilon,\beta}^2\iota^4}{\widetilde{N}_{h+1}^{k+1}(s')^2} + \frac{1000^2H^6S^4A^2\iota^4}{\widetilde{N}_{h+1}^{k+1}(s')^2}, H^2\right\}}{\widetilde{N}_h^k(s,a)}} \\
\ge& 0,
\end{aligned} \tag{81}$$

where the first inequality is because of Bellman equation and condition 2. The second inequality holds since the definition of b_h^{k+1} and Lemma B.1. The third inequality results from Lemma B.6. The forth inequality comes from Lemma B.9 and Remark B.3. The last inequality is due to the following analysis.

Because $\sqrt{\mathrm{Var}(X)} \le \sqrt{2\mathrm{Var}(Y)} + \sqrt{2\mathrm{Var}(X-Y)}$ (Lemma 2 of Azar et al. [2017]), we have

$$\sqrt{\mathrm{Var}_{s' \sim \widehat{P}_h^{k+1}(\cdot|s,a)}V_{h+1}^\star(\cdot)} \le \sqrt{2\mathrm{Var}_{s' \sim \widehat{P}_h^{k+1}(\cdot|s,a)}\widetilde{V}_{h+1}^{k+1}(\cdot)} + \underbrace{\sqrt{2\mathrm{Var}_{s' \sim \widehat{P}_h^{k+1}(\cdot|s,a)}\left(\widetilde{V}_{h+1}^{k+1}(\cdot) - V_{h+1}^\star(\cdot)\right)}}_{\text{(a)}}. \tag{82}$$

In addition,

$$\begin{aligned}
(\mathrm{a}) \leq & \sqrt{2\sum_{s'} \widehat{P}_h^{k+1}(s'|s,a)\left(\widetilde{V}_{h+1}^{k+1}(s') - V_{h+1}^{\star}(s')\right)^2} \\
\leq & \sqrt{6\sum_{s'} \widehat{P}_h^{k+1}(s'|s,a)\min\left\{\frac{1000^2H^3SA\iota^2}{N_{h+1}^{k+1}(s')} + \frac{1000^2H^4S^4A^2E_{\epsilon,\beta}^2\iota^4}{N_{h+1}^{k+1}(s')^2} + \frac{1000^2H^6S^4A^2\iota^4}{N_{h+1}^{k+1}(s')^2}, H^2\right\}} \\
\leq & 2\sqrt{2\sum_{s'} \widehat{P}_h^{k+1}(s'|s,a)\min\left\{\frac{1000^2H^3SA\iota^2}{\widetilde{N}_{h+1}^{k+1}(s')} + \frac{1000^2H^4S^4A^2E_{\epsilon,\beta}^2\iota^4}{\widetilde{N}_{h+1}^{k+1}(s')^2} + \frac{1000^2H^6S^4A^2\iota^4}{\widetilde{N}_{h+1}^{k+1}(s')^2}, H^2\right\}},
\end{aligned} \tag{83}$$

where the first inequality results from the definition of variance. The second inequality holds because of condition 2 and the fact that $\min\{(a+b+c)^2, H^2\} \leq 3\min\{a^2+b^2+c^2, H^2\}$. The last inequality holds according to Assumption 3.1.

Finally, plugging (82) and (83) into (81), we have the last inequality of (81) holds. Therefore, the proof of Lemma C.11 is complete. □

D Missing proofs in Section 5

In this section, we state the missing proofs in Section 5. Recall that N_h^k is the original count, $\widehat{N}_h^k$ is the noisy count after step (1) of both Privatizers and $\widetilde{N}_h^k$ is the final private counts.

Proof of Lemma 5.1. Due to Theorem 3.5 of Chan et al. [2011] and Lemma 34 of Hsu et al. [2014], the release of $\{\widehat{N}_h^k(s,a)\}_{(h,s,a,k)}$ satisfies $\frac{\epsilon}{3}$-DP. Similarly, the releases of $\{\widehat{N}_h^k(s,a,s')\}_{(k,h,s,a,s')}$ and $\{\widetilde{R}_h^k(s,a)\}_{(k,h,s,a)}$ both satisfy $\frac{\epsilon}{3}$-DP. Therefore, the release of the following private counters $\{\widehat{N}_h^k(s,a)\}_{(h,s,a,k)}$, $\{\widehat{N}_h^k(s,a,s')\}_{(k,h,s,a,s')}$ and $\{\widetilde{R}_h^k(s,a)\}_{(k,h,s,a)}$ satisfy ϵ-DP. Due to post-processing (Lemma 2.3 of Bun and Steinke [2016]), the release of all private counts $\{\widetilde{N}_h^k(s,a)\}_{(h,s,a,k)}$, $\{\widetilde{N}_h^k(s,a,s')\}_{(k,h,s,a,s')}$ and $\{\widetilde{R}_h^k(s,a)\}_{(k,h,s,a)}$ also satisfies ϵ-DP. Then it holds that the release of all π_k is ϵ-DP according to post-processing. Finally, the guarantee of ϵ-JDP results from Billboard Lemma (Lemma 9 of Hsu et al. [2014]).

For utility analysis, because of Theorem 3.6 of Chan et al. [2011], our choice $\epsilon' = \frac{\epsilon}{3H\log K}$ in Binary Mechanism and a union bound, with probability $1-\frac{\beta}{3}$, for all (k,h,s,a,s'),

$$\begin{aligned}
|\widehat{N}_h^k(s,a,s') - N_h^k(s,a,s')| \leq O(\frac{H}{\epsilon}\log(HSAT/\beta)^2), \quad |\widehat{N}_h^k(s,a) - N_h^k(s,a)| \leq O(\frac{H}{\epsilon}\log(HSAT/\beta)^2), \\
|\widetilde{R}_h^k(s,a) - R_h^k(s,a)| \leq O(\frac{H}{\epsilon}\log(HSAT/\beta)^2).
\end{aligned} \tag{84}$$

Together with Lemma 5.4, the Central Privatizer satisfies Assumption 3.1 with $E_{\epsilon,\beta} = \widetilde{O}(\frac{H}{\epsilon})$. □

Proof of Theorem 5.2. The proof directly results from plugging $E_{\epsilon,\beta} = \widetilde{O}(\frac{H}{\epsilon})$ into Theorem 4.1. □

Proof of Lemma 5.4. For clarity, we denote the solution of (4) by $\bar{N}_h^k$ and therefore $\widetilde{N}_h^k(s,a,s') = \bar{N}_h^k(s,a,s') + \frac{E_{\epsilon,\beta}}{2S}$, $\widetilde{N}_h^k(s,a) = \bar{N}_h^k(s,a) + \frac{E_{\epsilon,\beta}}{2}$.

When the condition (two inequalities) in Lemma 5.4 holds, the original counts $\{N_h^k(s,a,s')\}_{s'\in\mathcal{S}}$ is a feasible solution to the optimization problem, which means that

$$\max_{s'}|\bar{N}_h^k(s,a,s') - \widehat{N}_h^k(s,a,s')| \leq \max_{s'}|N_h^k(s,a,s') - \widehat{N}_h^k(s,a,s')| \leq \frac{E_{\epsilon,\beta}}{4}.$$

Combining with the condition in Lemma 5.4 with respect to $\widehat{N}_h^k(s,a,s')$, it holds that

$$|\bar{N}_h^k(s,a,s') - N_h^k(s,a,s')| \leq |\bar{N}_h^k(s,a,s') - \widehat{N}_h^k(s,a,s')| + |\widehat{N}_h^k(s,a,s') - N_h^k(s,a,s')| \leq \frac{E_{\epsilon,\beta}}{2}.$$

Since $\widetilde{N}_h^k(s,a,s') = \bar{N}_h^k(s,a,s') + \frac{E_{\epsilon,\beta}}{2S}$ and $\bar{N}_h^k(s,a,s') \geq 0$, we have

$$\widetilde{N}_h^k(s,a,s') > 0, \quad |\widetilde{N}_h^k(s,a,s') - N_h^k(s,a,s')| \leq E_{\epsilon,\beta}. \tag{85}$$

For $\bar{N}_h^k(s,a)$, according to the constraints in the optimization problem (4), it holds that

$$|\bar{N}_h^k(s,a) - \widehat{N}_h^k(s,a)| \leq \frac{E_{\epsilon,\beta}}{4}.$$

Combining with the condition in Lemma 5.4 with respect to $\widehat{N}_h^k(s,a)$, it holds that

$$|\bar{N}_h^k(s,a) - N_h^k(s,a)| \leq |\bar{N}_h^k(s,a) - \widehat{N}_h^k(s,a)| + |\widehat{N}_h^k(s,a) - N_h^k(s,a)| \leq \frac{E_{\epsilon,\beta}}{2}.$$

Since $\widetilde{N}_h^k(s,a) = \bar{N}_h^k(s,a) + \frac{E_{\epsilon,\beta}}{2}$, we have

$$N_h^k(s,a) \leq \widetilde{N}_h^k(s,a) \leq N_h^k(s,a) + E_{\epsilon,\beta}. \tag{86}$$

According to the last line of the optimization problem (4), we have $\bar{N}_h^k(s,a) = \sum_{s' \in \mathcal{S}} \bar{N}_h^k(s,a,s')$ and therefore,

$$\widetilde{N}_h^k(s,a) = \sum_{s' \in \mathcal{S}} \widetilde{N}_h^k(s,a,s'). \tag{87}$$

The proof is complete by combining (85), (86) and (87). □

Proof of Lemma 5.6. The privacy guarantee directly results from properties of Laplace Mechanism and composition of DP [Dwork et al., 2014].

For utility analysis, because of Corollary 12.4 of Dwork et al. [2014] and a union bound, with probability $1 - \frac{\beta}{3}$, for all (k,h,s,a,s'),

$$|\widehat{N}_h^k(s,a,s') - N_h^k(s,a,s')| \leq O(\frac{H}{\epsilon}\sqrt{K\log(HSAT/\beta)}), \quad |\widehat{N}_h^k(s,a) - N_h^k(s,a)| \leq O(\frac{H}{\epsilon}\sqrt{K\log(HSAT/\beta)}),$$
$$|\widetilde{R}_h^k(s,a) - R_h^k(s,a)| \leq O(\frac{H}{\epsilon}\sqrt{K\log(HSAT/\beta)}). \tag{88}$$

Together with Lemma 5.4, the Local Privatizer satisfies Assumption 3.1 with $E_{\epsilon,\beta} = \widetilde{O}(\frac{H}{\epsilon}\sqrt{K})$. □

Proof of Theorem 5.7. The proof directly results from plugging $E_{\epsilon,\beta} = \widetilde{O}(\frac{H}{\epsilon}\sqrt{K})$ into Theorem 4.1. □