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 $\epsilon$-JDP algorithm with a regret of $O(\sqrt{SAH^2T} + S^3A^2H^3/\epsilon)$ 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-
Regret under \(O(\sqrt{SAH^2T + S^2AH^3/\epsilon})\), 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 \(O(\sqrt{SAH^2T})\) [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 \(O(S^2AH^3/\epsilon)\), 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 \(\epsilon\)-JDP, DP-UCBVI achieves regret of \(O(\sqrt{SAH^2T + S^2AH^3/\epsilon})\), 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 \(\epsilon\)-LDP constraint, DP-UCBVI achieves regret of \(O(\sqrt{SAH^2T + S^2AH^3/\epsilon})\) 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] and [Garcelon et al., 2021] are given in Table 1 while we leave more discussions about results on

Table 1: Comparison of our results (in blue) to existing work regarding regret under \(\epsilon\)-joint differential privacy, regret under \(\epsilon\)-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.

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 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-OB1 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, 2022] 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\(\star\). 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

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.
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 accumulate 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]), \( b_h^k(s, a) = \tilde{O}(H \cdot \sqrt{1/N_h^k(s, a)}) \), the construction of private bonus is straightforward. We can simply replace original counts \( N_h^k(s, a) \) with private counts \( \tilde{N}_h^k(s, a) \) and add an additional term to account for the difference between these two bonuses. Next, combining the two private bonuses with the uniform upper bound of \( |\tilde{N}_h^k(s, a) - N_h^k(s, a)| \), 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 \( \text{Var}_P(b_h^{\pi_h}(s, a)|V_h^k(\cdot)) \), the first technical challenge is to replace the empirical transition kernel \( \tilde{P}_h^k(s, a) \) 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 \( |\tilde{N}_h^k(s, a) - N_h^k(s, a)| \) is the same scale compared to previous approaches. With the private transition kernel estimates \( \tilde{P}_h^k \), we can replace \( \text{Var}_{\tilde{P}_h^k(s, a)} V_h^k(\cdot) \) with \( \text{Var}_{\tilde{P}_h^k(s, a)} \tilde{V}_h^k(\cdot) \) where \( \tilde{V}_h^k(\cdot) \) 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 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 \( N \in \mathbb{Z}_+ \), \( [N] = \{1, 2, \ldots, N\} \). For any set \( W \), \( \Delta(W) \) denotes the set of all probability distributions over \( W \). Besides, we use standard notations such as \( \mathcal{O} \) and \( \Theta \) to suppress constants while \( \mathcal{O} \) and \( \Theta \) 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 \( \mathcal{M} = (S, A, H, \{P_h^H\}_{h=1}^H, \{r_h^H\}_{h=1}^H, d_1) \) [Sutton and Barto, 1998], where \( S \) is state space with \( |S| = S \), \( A \) is action space with \( |A| = A \) and \( H \) is the horizon. The non-stationary transition kernel has the form \( P_h : S \times A \times S \to [0,1] \) 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, \( r_h(s, a, h) \in \Delta([0,1]) \) 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 \( \pi_h \) maps each state \( s \in S \) to a probability distribution over actions, i.e. \( \pi_h : S \to \Delta(A) \), \( \forall h \in [H] \). 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 \( \pi \) and any \( h \in [H] \), the value function \( V_h^\pi(\cdot) \) is defined as:

\[
V_h^\pi(s) = \mathbb{E}_{\pi} (\sum_{t=h}^{H-1} r_t | s_0 = s, \pi, \alpha), \quad V_h^\pi(s) = \mathbb{E}_{\pi} (\sum_{t=h}^{H-1} r_t | s_0 = s, \alpha), \quad \forall s, \alpha \in S \times A.
\]

The optimal policy \( \pi^* \) maximizes \( V_h^\pi(s) \) for all \( s, h \in S \times [H] \) simultaneously and we denote the value function and Q-value function with respect to \( \pi^* \) by \( V_h^* \) and \( Q_h^* \). Then Bellman (optimality) equation follows \( \forall h \in [H]: \)

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

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

\[
\text{Regret}(K) := \sum_{k=1}^K [V_1^*(s_{k}^1) - V_1^{\pi_k}(s_{k}^1)],
\]

where \( s_1^k \) is the initial state and \( \pi_k \) 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 $s^k_h$ to agent $M$. $M$ sends back an action $a^k_h$, and finally $u_k$ sends her reward $r^k_h$ to $M$. Formally, we denote a sequence of $K$ users who participate in the above RL protocol by $U = (u_1, \ldots, u_K)$. 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 for any possible sequences of actions from the agent. We let $\mathcal{M}(U) = (a^1_1, \ldots, a^K_H)$ denote the whole sequence of actions chosen by agent $M$. An ideal privacy preserving agent would guarantee that $\mathcal{M}(U)$ 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 $\mathcal{M} : U \rightarrow \mathbb{A}^H$ is $(\epsilon, \delta)$-differentially private if for any possible user sequences $U$ and $U'$ differing on a single user and any subset $E \subseteq \mathbb{A}^H$,  
\[ \Pr[\mathcal{M}(U) \in E] \leq e^\epsilon \Pr[\mathcal{M}(U') \in E] + \delta. \]

If $\delta = 0$, we say that $\mathcal{M}$ is $\epsilon$-differentially private ($\epsilon$-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 $\mathcal{M} : U \rightarrow \mathbb{A}^H$ is $\epsilon$-joint differentially private if for any $k \in [K]$, any user sequences $U$, $U'$ differing on the $k$-th user and any subset $E \subseteq \mathbb{A}^{(K-1)H}$,  
\[ \Pr[\mathcal{M}_{-k}(U) \in E] \leq e^\epsilon \Pr[\mathcal{M}_{-k}(U') \in E], \]

where $\mathcal{M}_{-k}(U) \in E$ 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 $M$ sends policy $\pi_k$ to user $u_k$, after deploying $\pi_k$ and getting trajectory $X_k$, user $u_k$ privatizes her trajectory to $X'_k$ and finally sends it to $M$. We denote the privacy mechanism on user’s side by $\hat{M}$ and define local differential privacy formally below.

**Definition 2.3** (Local Differential Privacy (LDP)). For any $\epsilon > 0$, a mechanism $\hat{M}$ is $\epsilon$-local differentially private if for any possible trajectories $X, X'$ and any possible set $E \subseteq \{\mathcal{M}(X) | X \text{ is any possible trajectory}\}$,
\[ \Pr[\hat{M}(X) \in E] \leq e^\epsilon \Pr[\hat{M}(X') \in E]. \]

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 [5] which naturally implies regret bounds under both Joint DP and Local DP.

We begin with the following definition of counts. Let $N^k_h(s, a) = \sum_{i=1}^{k-1} 1(s^i_h, a^i_h = s, a)$ denote the visitation number of $(s, a)$ at step $h$ before the $k$-th episode. Similarly, $N^k_h(s, a, s') = \sum_{i=1}^{k-1} 1(s^i_h, a^i_h = s, a, s')$ and $R^k_h(s, a) = \sum_{i=1}^{k-1} 1(s^i_h, a^i_h = s, a) \cdot r^i_h$ 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 $\tilde{N}^k_h(s, a), \tilde{N}^k_h(s, a, s'), \tilde{R}^k_h(s, a)$ returned by Privatizer. We make the Assumption [5] 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 $(h, s, a, s', k) \in [H] \times S \times A \times S \times [K]$:

1. $|\tilde{N}^k_h(s, a, s') - N^k_h(s, a, s')| \leq E_{\epsilon, \beta}$. $\tilde{N}^k_h(s, a) - N^k_h(s, a)$
Algorithm 1 DP-UCBVI

1: **Input:** Number of episodes $K$, privacy budget $\epsilon$, failure probability $\beta$ and a Privatizer (can be either Central or Local).
2: **Initialize:** Private counts $\hat{R}_h^k(s, a) = \hat{N}_h^k(s, a) = \hat{N}_h^k(s, a, s') = 0$ for all $(h, s, a, s') \in [H] \times S \times A \times S$. Set up the confidence bound $E_{\epsilon, \beta}$ w.r.t the Privatizer. $t = \log(30HSA\epsilon/\beta)$.
3: for $k = 1, 2, \ldots, K$ do
4: \hspace{1cm} $V_{H+1}(s) = 0$.
5: for $h = H, H - 1, \ldots, 1$ do
6: \hspace{2cm} Compute $\bar{P}_h^k(s'|s, a)$ and $\bar{r}_h^k(s, a)$ as in (1).
7: \hspace{2cm} Calculate private bonus $b_h^k(s, a) = 2\sqrt{\frac{\text{Var}_{e \sim P_h^k(s'|s, a)} V_{h+1}(s')}{'N_h^k(s, a)}} + \sqrt{\frac{2}{N_h^k(s, a)}} + \frac{20HSA\epsilon}{N_h^k(s, a)} + 4\sqrt{t}$.
8: \hspace{2cm} for $(s, a) \in S \times A$ do
9: \hspace{3cm} $\bar{Q}_h^k(s, a) = \min \{Q_{h-1}^k(s, a), H, \bar{r}_h^k(s, a) + \sum_{s'} \bar{P}_h^k(s'|s, a) \cdot V_{h+1}(s') + b_h^k(s, a)\}$.
10: \hspace{2cm} end for
11: \hspace{2cm} $V_h^k(s) = \max_{a \in A} \bar{Q}_h^k(s, a)$.
12: \hspace{2cm} $\pi_h^k(s) = \arg \max_{a \in A} \bar{Q}_h^k(s, a)$ with ties broken arbitrarily.
13: end for
14: end for
15: Deploy policy $\pi_k = (\pi_1^k, \ldots, \pi_H^k)$ and get trajectory $(s_1^k, a_1^k, r_1^k, \ldots, s_{H+1}^k)$.
16: Update the private counts to $\hat{R}_h^{k+1}, \hat{N}_h^{k+1}$ via Privatizer.
17: end for

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 $\sqrt{T}$ in regret bound. Intuitively, the first term of $b_h^k$ aims to approximate the variance w.r.t $V_h^k$, 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 4 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

$$
\text{Regret}(K) \leq \tilde{O}(\sqrt{SAHT^2} + S^2A^2H^2E_{\epsilon, \beta}),
$$

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

Under Assumption 3.1, the best known regret bound is $O(\sqrt{SAH^2T} + S^2A^2H^2E_{\epsilon, \beta})$ (Theorem 4.2 of Chowdhury and Zhou (2021)). As a comparison, in our regret bound,
the term parameterized by privacy loss $\epsilon$ remains the same while the leading term is improved by a factor of $\sqrt{H}$ into $O(\sqrt{SAHT^2})$. 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 $(h, s, a)$, $\{N_{h}^{k}(s, a) = \sum_{i=1}^{k} \mathbb{1}(s_{h}^{i}, a_{h}^{i} = s, a)\}_{k \in \mathbb{K}}$ is the partial sums of data stream $\{\mathbb{1}(s_{h}^{i}, a_{h}^{i} = s, a)\}_{i \in \mathbb{K}}$. 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

(1) For all $(h, s, a, s')$, we privatize $\{N_{h}^{k}(s, a)\}_{k \in \mathbb{K}}$ and $\{N_{h}^{k}(s, a, s')\}_{k \in \mathbb{K}}$ (which is summation of bounded streams) by applying Binary Mechanism (Algorithm 2 in Chan et al. 2011) with $\epsilon' = \frac{\epsilon}{3T \log K}$. We denote the output of Binary Mechanism by $N_{h}^{k}$. 

(2) The private counts $\tilde{N}_{h}^{k}$ are solved through the procedure in Section 5.1.1 with $E_{\epsilon, \beta} = \tilde{O}(4)$. 

(3) For the counters of accumulated reward, for all $(h, s, a)$, we apply the same Binary Mechanism with $\epsilon' = \frac{\epsilon}{3T \log K}$ to privatize $R_{h}^{k}(s, a)$ and get $\hat{R}_{h}^{k}(s, a)$.

We sum up the properties of Central Privatizer below.

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

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 7) with Central Privatizer as input, with probability $1 - \beta$, it holds that:

\[
\text{Regret}(K) \leq \tilde{O}(\sqrt{\frac{SAHT^2}{\beta}} + S^2 A H^3 / \epsilon).
\]  

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 $\epsilon$ 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 $O(\sqrt{SAHT^2})$ (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 $\tilde{N}_{h}^{k}(s, a)$ and $\tilde{N}_{h}^{k}(s, a, s')$ (for all $(h, s, a, s') \in |H| \times S \times A \times S$), we construct the following private counts that satisfy Assumption 3.1. The choice of $N_{h}^{k}$ follows: for all $(h, s, a)$

\[
\{\tilde{N}_{h}^{k}(s, a, s')\}_{s' \in S} = \arg \min_{x_{s'}} \max_{s' \in S} \left| x_{s'} - \tilde{N}_{h}^{k}(s, a, s') \right|
\]

such that $\sum_{s' \in S} x_{s'} - \tilde{N}_{h}^{k}(s, a) \leq \frac{E_{\epsilon, \beta}}{4}$ and $x_{s'} \geq 0, \forall s'$.

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

\[
\tilde{N}_{h}^{k}(s, a) = \tilde{N}_{h}^{k}(s, a) + \frac{E_{\epsilon, \beta}}{2S}.
\]

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

\[
\min t, \text{ s.t. } |x_{s'} - \tilde{N}_{h}^{k}(s, a, s')| \leq t, x_{s'} \geq 0, \forall s' \in S, |\sum_{s' \in S} x_{s'} - \tilde{N}_{h}^{k}(s, a)| \leq \frac{E_{\epsilon, \beta}}{4}.
\]

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 1953] or other provably efficient algorithms [Nemhauser and Wolsey 1988]. Therefore, since during the whole process, we only solve HSAT such Linear Programming problems, our Algorithm is computationally efficient.

The properties of private counts $\tilde{N}_{h}^{k}$ is summarized below.

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

\[
|\tilde{N}_{h}^{k}(s, a, s') - N_{h}^{k}(s, a, s')| \leq \frac{E_{\epsilon, \beta}}{4},
\]

\[
|\tilde{N}_{h}^{k}(s, a) - N_{h}^{k}(s, a)| \leq \frac{E_{\epsilon, \beta}}{4},
\]
then the \( \tilde{N}_h^k \) derived from (4) and (5) satisfies Assumption 3.1.

**Remark 5.5.** Compared to the concurrent work [Qiao and Wang 2022b], our private counts \( \tilde{N}_h^k(s,a) \) 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 \( \{\sigma_h^k(s,a) = 1(s_h^k, a_h^k = s, a)\}_{(s,a)} \) has \( \ell_1 \) sensitivity \( H \). Therefore, the perturbed version of the counts above \( \tilde{\sigma}_h^k(s,a) = \sigma_h^k(s,a) + \text{Lap}(\frac{3H}{\epsilon}) \) satisfies \( \frac{1}{2} \)-LDP. In addition, similar perturbations to \( \{\text{Lap}(s_h^k, a_h^k, s_{h+1}^k = s, a, s')\}_{(s,a,s')} \) and \( \{\text{Lap}(s_h^k, a_h^k = s, a) + \text{Lap}(\frac{3H}{\epsilon}) \} \) 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 \( \tilde{\sigma}_h^k(s,a) = \text{Lap}(\frac{3H}{\epsilon}) \)
2. The noisy counts are calculated by

\[
\tilde{N}_h^k(s,a) = \sum_{i=1}^{k-1} \tilde{\sigma}_h^k(s,a),
\]

Then the private counts \( \tilde{N}_h^k(s,a) \) 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: \( \tilde{r}_h^k(s,a) = 1(s_h^k, a_h^k = s, a) \cdot r_h^k + \text{Lap}(\frac{3H}{\epsilon}) \). The accumulative statistic is calculated by \( \tilde{R}_h^k(s,a) = \sum_{i=1}^{k-1} \tilde{r}_h^k(s,a) \).

Properties of our Local Privatizer is summarized below.

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

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

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

\[
\text{Regret}(K) \leq \tilde{O}(\sqrt{\text{SAT}^2T} + S^2A\sqrt{HST}/\epsilon).
\]
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 $(i, h, s, a) \in [k] \times [H] \times \mathcal{S} \times \mathcal{A}$, $Q_h^k(s, a) \geq Q_h^i(s, a)$, we prove $(T_k = kH)$

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

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

$$\bar{V}_h^k(s) - V_h^*(s) \leq \bar{O} \left( \frac{SA^3}{N_k^k(s)} + \frac{S^2 A^2 H \epsilon, \beta}{N_k^k(s)} \right).$$

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

$$\bar{V}_h^k(s) - V_h^*(s) \leq \bar{O} \left( \frac{SA^3}{N_k^k(s)} + \frac{S^2 A^2 H \epsilon, \beta}{N_k^k(s)} \right),$$

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

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

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

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

**Proof of regret bound: (1).** We only need to prove the upper bound of $\text{Regret}(k)$, as the upper bound of $V_h^k(s) - V_h^*(s)$ 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 $\sum_{i=1}^k \sum_{h=1}^H \frac{20HSE_L}{N_h^i(s^i_h, a_h^i)}$, 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

$$\sum_{i=1}^k \sum_{h=1}^H \frac{20HSE_L}{N_h^i(s^i_h, a_h^i)} \leq \sum_{i=1}^k \sum_{h=1}^H \frac{20HSE_L}{N_h^i(s^i_h, a_h^i)} \leq \bar{O} \left( \frac{S^2 A^2 H \epsilon, \beta}{N_k^k(s)} \right). \quad (10)$$

Now we analyze the first term (which is also the main term):

$$\sum_{i=1}^k \sum_{h=1}^H \frac{\text{Var} \cdot \bar{P}_h^i(\cdot | s^i_h, a_h^i) \bar{V}_h^{i+1}(\cdot)}{N_h^i(s^i_h, a_h^i)}.$$

It holds that

$$\sum_{i=1}^k \sum_{h=1}^H \frac{1}{N_h^i(s^i_h, a_h^i)} \leq \bar{O} \left( \frac{SA^3}{N_k^k(s)} + \frac{S^2 A^2 H \epsilon, \beta}{N_k^k(s)} \right). \quad (11)$$

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

$$\sum_{i=1}^k \sum_{h=1}^H \frac{\text{Var} \cdot \bar{P}_h^i(\cdot | s^i_h, a_h^i) \bar{V}_h^{i+1}(\cdot)}{N_h^i(s^i_h, a_h^i)} \leq \bar{O} \left( \frac{SA^3}{N_k^k(s)} + \frac{S^2 A^2 H \epsilon, \beta}{N_k^k(s)} \right). \quad (12)$$

Therefore, the main term in the regret bound scales as $\bar{O} \left( \sqrt{H^2 SAT_k + H^2 S^2 A \epsilon, \beta} \right)$. 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 $b_h^k(s, a) \geq |r_h^k(s, a) - r_h(s, a)| + (\bar{P}_h^k - P_h) \cdot V_h^{i+1}(s, a)$.

It is clear that $|\bar{P}_h^k - P_h| \cdot V_h^{i+1}(s, a)$ 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, $(\bar{P}_h^k - P_h) \cdot V_h^{i+1}(s, a)$ can be bounded by $\bar{O} \left( \sqrt{\text{Var} \cdot \bar{P}_h^k(\cdot | s, a) V_h^{i+1}(\cdot) / N_k^k(s, a)} \right)$, which can be further bounded by $\bar{O} \left( \sqrt{\text{Var} \cdot \bar{P}_h^k(\cdot | s, a) V_h^{i+1}(\cdot) / N_k^k(s, a)} \right)$ plus a portion of the third term of $b_h^k(s, a)$. Finally, together with the upper bound of $|\bar{V}_h^k(s) - V_h^*(s)|$ (derived from condition of (2) and optimism), the last term of $b_h^k(s, a)$ compensates for the difference of $\sqrt{\text{Var} \cdot \bar{P}_h^k(\cdot | s, a) V_h^{i+1}(\cdot) / N_k^k(s, a)}$ (first term of $b_h^k(s, a)$) and $\sqrt{\text{Var} \cdot \bar{P}_h^k(\cdot | s, a) V_h^{i+1}(\cdot) / N_k^k(s, a)}$.

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.](image-url)
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 $\epsilon$) has smaller regret. However, under constraints 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 $\epsilon$-JDP, we achieved regret bound of $O(\sqrt{SAH^2T} + S^2A\sqrt{H^3T}/\epsilon)$, which matches the lower bound up to lower order terms. Meanwhile, under $\epsilon$-LDP, we derived regret upper bound of $O(\sqrt{SAH^2T} + S^2A\sqrt{H^3T}/\epsilon)$ 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], $O(\sqrt{d^3H^3T})$, 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


Near-Optimal Differentially Private Reinforcement Learning


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 $O(\sqrt{HSA\tau})$ under stationary MDP. Later, Zhang et al. [2020] achieved the optimal regret bound $O(\sqrt{H\tau^2SA})$ 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:

$$P_h^k(s'|s,a) = \frac{N_h^k(s,a,s')}{N_h^k(s,a)},$$

$$r_h^k(s,a) = \frac{R_h^k(s,a)}{N_h^k(s,a)}.$$  \hspace{1cm} (13)

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

$$\tilde{P}_h^k(s'|s,a) = \frac{\tilde{N}_h^k(s,a,s')}{\tilde{N}_h^k(s,a)},$$

$$\tilde{r}_h^k(s,a) = \left( \frac{\tilde{R}_h^k(s,a)}{\tilde{N}_h^k(s,a)} \right)_{[0,1]}.$$  \hspace{1cm} (14)

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

$$|\tilde{r}_h^k(s,a) - r_h(s,a)| \leq \sqrt{\frac{2L}{\tilde{N}_h^k(s,a)}} + \frac{2E_{r,s,\beta}}{\tilde{N}_h^k(s,a)},$$  \hspace{1cm} (15)
Proof of Lemma B.1. We have for all \( h, s, a, k \in [H] \times S \times A \times [K] \),

\[
|\tilde{r}_h^k(s, a) - r_h(s, a)| \leq \left| \frac{\bar{R}_h^k(s, a)}{N_h^k(s, a)} - r_h(s, a) \right|
\leq \frac{\bar{R}_h^k(s, a) - R_h^k(s, a)}{N_h^k(s, a)} + \frac{R_h^k(s, a) - r_h(s, a)}{N_h^k(s, a)}
\leq \frac{E_{e, \beta}}{N_h^k(s, a)} + \frac{N_h^k(s, a)}{N_h^k(s, a)} \cdot \left| \frac{R_h^k(s, a)}{N_h^k(s, a)} - r_h(s, a) \right| + \left| r_h(s, a) \left( \frac{N_h^k(s, a)}{N_h^k(s, a)} - 1 \right) \right|
\leq \frac{E_{e, \beta}}{N_h^k(s, a)} + \frac{N_h^k(s, a)}{N_h^k(s, a)} \cdot \sqrt{\frac{2\epsilon}{N_h^k(s, a)}} + \frac{E_{e, \beta}}{N_h^k(s, a)}
\]

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 \( h, s, a, k \in [H] \times S \times A \times [K] \), it holds that:

\[
\left\| \bar{P}_h^k(\cdot | s, a) - P_h(\cdot | s, a) \right\|_1 \leq 2 \sqrt{\frac{S_t}{N_h^k(s, a)}} + 2SE_{e, \beta}.
\]

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

\[
\left\| \bar{P}_h^k(\cdot | s, a) - P_h(\cdot | s, a) \right\|_1 = \sum_{s'} \left| \bar{P}_h^k(s' | s, a) - P_h(s' | s, a) \right|
\leq \sum_{s'} \left| \frac{N_h^k(s, a, s') - N_h^k(s, a, s')}{N_h^k(s, a)} \right| + \sum_{s'} \left| \frac{N_h^k(s, a, s')}{N_h^k(s, a)} - P_h(s' | s, a) \right|
\leq \frac{SE_{e, \beta}}{N_h^k(s, a)} + \sum_{s'} \left| \frac{N_h^k(s, a, s')}{N_h^k(s, a)} - P_h(s' | s, a) \right|
\leq \frac{SE_{e, \beta}}{N_h^k(s, a)} + \sum_{s'} \left| \frac{N_h^k(s, a, s') - N_h^k(s, a)}{N_h^k(s, a)} \right| + \sum_{s'} \left| P_h(s' | s, a) \left( \frac{N_h^k(s, a)}{N_h^k(s, a)} - 1 \right) \right|
\leq \frac{N_h^k(s, a)}{N_h^k(s, a)} \cdot 2 \sqrt{\frac{S_t}{N_h^k(s, a)}} + \frac{2SE_{e, \beta}}{N_h^k(s, a)}
\leq 2 \sqrt{\frac{S_t}{N_h^k(s, a)}} + \frac{2SE_{e, \beta}}{N_h^k(s, a)}
\]

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 \( h, s, a, k \in [H] \times S \times A \times [K] \),

\[
\left\| \bar{P}_h^k(\cdot | s, a) - \bar{P}_h^k(\cdot | s, a) \right\|_1 \leq 2 \sqrt{\frac{S_t}{N_h^k(s, a)}} + \frac{2SE_{e, \beta}}{N_h^k(s, a)}
\]

(19)

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.

\[\text{Dan Qiao, Yu-Xiang Wang}\]
Lemma B.4. With probability $1 - \frac{2}{15}$, for all $h, s, a, s', k \in [H] \times S \times A \times [K]$, it holds that:

$$
\left| \bar{P}_{h}^{k}(s'|s, a) - P_{h}(s'|s, a) \right| \leq \sqrt{\frac{2P_{h}(s'|s, a) }{N_{h}^{k}(s, a)}} + \frac{2E_{\epsilon,\beta}}{N_{h}^{k}(s, a)}.
$$

(20)

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

$$
\left| \bar{P}_{h}^{k}(s'|s, a) - P_{h}(s'|s, a) \right| \leq \left| \frac{N_{h}^{k}(s, a, s') - N_{h}^{k}(s, a, s')}{N_{h}^{k}(s, a)} \right| + \left| \frac{N_{h}^{k}(s, a, s') - P_{h}(s'|s, a)}{N_{h}^{k}(s, a)} \right| \\
\leq \frac{E_{\epsilon,\beta}}{N_{h}^{k}(s, a)} + \frac{N_{h}^{k}(s, a, s') - P_{h}(s'|s, a)}{N_{h}^{k}(s, a)} \\
\leq \frac{2E_{\epsilon,\beta}}{N_{h}^{k}(s, a)} + \frac{N_{h}^{k}(s, a, s') - P_{h}(s'|s, a)}{N_{h}^{k}(s, a)} + \left| \frac{N_{h}^{k}(s, a, s') - P_{h}(s'|s, a)}{N_{h}^{k}(s, a)} \right| \\
\leq \frac{2E_{\epsilon,\beta}}{N_{h}^{k}(s, a)} + \frac{N_{h}^{k}(s, a, s') - P_{h}(s'|s, a)}{N_{h}^{k}(s, a)} + \left| \frac{N_{h}^{k}(s, a, s') - P_{h}(s'|s, a)}{N_{h}^{k}(s, a)} \right|
$$

(21)

\[ \leq \frac{2P_{h}(s'|s, a) }{N_{h}^{k}(s, a)} + \frac{2E_{\epsilon,\beta}}{N_{h}^{k}(s, a)} \]

where the second, forth and last inequalities result from Assumption 3.1. The fifth inequality holds with probability $1 - \frac{2}{15}$ due to Bernstein’s inequality and union bound.

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

$$
\left| \bar{P}_{h}^{k}(s'|s, a) - \bar{P}_{h}^{k}(s'|s, a) \right| \leq \left| \frac{N_{h}^{k}(s, a, s') - N_{h}^{k}(s, a, s')}{N_{h}^{k}(s, a)} \right| + \left| \frac{N_{h}^{k}(s, a, s') - N_{h}^{k}(s, a, s')}{N_{h}^{k}(s, a)} \right| \\
\leq \frac{E_{\epsilon,\beta}}{N_{h}^{k}(s, a)} + \frac{N_{h}^{k}(s, a, s') - N_{h}^{k}(s, a, s')}{N_{h}^{k}(s, a)} \\
\leq \frac{2E_{\epsilon,\beta}}{N_{h}^{k}(s, a)}.
$$

(22)

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

$$
\left| \left( \bar{P}_{h}^{k} - P_{h} \right) \cdot V_{h+1}^{s}(s, a) \right| \leq \min \left\{ \sqrt{\frac{2\text{Var} \bar{P}_{h}^{k}(s|s, a) \cdot V_{h+1}^{s}(\cdot) }{N_{h}^{k}(s, a)}}, \sqrt{\frac{2\text{Var} \bar{P}_{h}^{k}(s|s, a) \cdot V_{h+1}^{s}(\cdot) }{N_{h}^{k}(s, a)}} \right\} + \frac{2HSE_{\epsilon,\beta}}{N_{h}^{k}(s, a)}.
$$

(23)
Proof of Lemma B.6. We have for all $h, s, a, k \in [H] \times S \times A \times [K]$,
\[
\left| \left( \tilde{P}^k_h - P_h \right) \cdot V^*_{h+1}(s, a) \right| \leq \sum_{s'} \left( \tilde{P}^k_h(s', s, a) - P_h(s', s, a) \right) V^*_{h+1}(s')
\leq \sum_{s'} \frac{N^k_h(s, a, s') - N^k_h(s, a, s')}{N^k_h(s, a)} \cdot V^*_{h+1}(s') \leq \frac{HSE_{e, \beta}}{N^k_h(s, a)} + \sum_{s'} \frac{N^k_h(s, a, s')}{N^k_h(s, a)} - P_h(s', s, a) - 1 \right| V^-_{h+1}(s') + \sum_{s'} P_h(s', s, a) V^-_{h+1}(s') \left( \frac{N^k_h(s, a)}{N^k_h(s, a)} - 1 \right)
\leq \frac{2HSE_{e, \beta}}{N^k_h(s, a)} + \sum_{s'} \frac{N^k_h(s, a, s')}{N^k_h(s, a)} - P_h(s', s, a) - 1 \right| V^-_{h+1}(s') \leq \frac{2HSE_{e, \beta}}{N^k_h(s, a)} + \sum_{s'} \frac{2\text{Var}_{P_h(s, a)} V^-_{h+1}(s') \cdot \ell}{N^k_h(s, a)} + \frac{2HSE_{e, \beta}}{N^k_h(s, a)} + \frac{2HSE_{e, \beta}}{N^k_h(s, a)} + \frac{7Ht}{3N^k_h(s, a)}
\leq \min \left\{ \frac{2\text{Var}_{P_h(s, a)} V^-_{h+1}(s') \cdot \ell}{N^k_h(s, a)}, \frac{2\text{Var}_{P_h(s, a)} V^-_{h+1}(s') \cdot \ell}{N^k_h(s, a)} + \frac{2HSE_{e, \beta}}{N^k_h(s, a)} \right\}
\leq \frac{2HSE_{e, \beta}}{N^k_h(s, a)},
\]  
(24)

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

Remark B.7. Similarly, we have for all $h, s, a, k \in [H] \times S \times A \times [K]$,
\[
\left| \left( \tilde{P}^k_h - P_h \right) \cdot V^-_{h+1}(s, a) \right| \leq H \left\| \tilde{P}^k_h(\cdot | s, a) - P^k_h(\cdot | s, a) \right\|_1 \leq \frac{2HSE_{e, \beta}}{N^k_h(s, a)},
\]  
(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 $1 - \frac{\beta}{H}$, 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
\[
\left\| \sqrt{\text{Var}_{P^k_h(\cdot | s, a)}(V)} - \sqrt{\text{Var}_{\tilde{P}^k_h(\cdot | s, a)}(V)} \right\| \leq \sqrt{3} \left\| P^k_h(\cdot | s, a) - \tilde{P}^k_h(\cdot | s, a) \right\|, \]
(26)

In addition, according to Remark B.3 the left hand side can be further bounded by
\[
\left\| \sqrt{\text{Var}_{P^k_h(\cdot | s, a)}(V)} - \sqrt{\text{Var}_{\tilde{P}^k_h(\cdot | s, a)}(V)} \right\| \leq 3H \sqrt{\frac{\text{SE}_{e, \beta}}{N^k_h(s, a)}},
\]  
(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.
\[
V^-_{i,j} = \text{Var}_{P^j_{\pi_i}(\cdot | s, a)} V^-_{j+1}(\cdot),
\]  
(28)
According to Definition C.1 above, it is clear that

\[ V^*_{i,j} = \text{Var}_{P_j(\cdot|s'_j, a'_j)} V^*_{j+1}(\cdot). \]  

(29)

Next, recall the definition of our private bonus \( b^k_h \):

\[ b^k_h(s, a) = 2 \sqrt{ \frac{\text{Var}_{s' \sim P^k_h(\cdot|s, a)} \tilde{V}^k_{h+1}(\cdot)}{N^k_h(s, a)}} + \sqrt{\frac{2t}{N^k_h(s, a)}} + \frac{20 H S E_{\epsilon, \beta} \cdot t}{N^k_h(s, a)} \]

\[ + 4 \sqrt{t} \cdot \sqrt{ \sum_{s'} \tilde{P}^k_h(s'|s, a) \min \left\{ \frac{1000^2 H^3 S A_2^2}{N^k_{h+1}(s')}, \frac{1000^2 H^4 S^4 A^2 E_2^4}{N^k_{h+1}(s')^2}, \frac{1000^2 H^8 S^4 A^2, H^2}{N^k_{h+1}(s')^4}, \frac{H^2}{N^k_h(s, a)} \right\}}. \]

(31)

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

\[ b^k_h(s, a) \leq 2 \sqrt{ \frac{\text{Var}_{s' \sim P^k_h(\cdot|s, a)} \tilde{V}^k_{h+1}(\cdot)}{N^k_h(s, a)}} + \sqrt{\frac{2t}{N^k_h(s, a)}} + \frac{20 H S E_{\epsilon, \beta} \cdot t}{N^k_h(s, a)} \]

\[ + 4 \sqrt{t} \cdot \sqrt{ \sum_{s'} \tilde{P}^k_h(s'|s, a) \min \left\{ \frac{1000^2 H^3 S A_2^2}{N^k_{h+1}(s')}, \frac{1000^2 H^4 S^4 A^2 E_2^4}{N^k_{h+1}(s')^2}, \frac{1000^2 H^8 S^4 A^2, H^2}{N^k_{h+1}(s')^4}, \frac{H^2}{N^k_h(s, a)} \right\}}. \]

(32)

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

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

\[ c_{4, i, j} = \frac{H^2 S t}{N^k_j(s'_j, a'_j)}, \quad c_{1, i, j} = \sqrt{\frac{2V^*_i}{N^k_j(s'_j, a'_j)}}, \quad \tilde{b}_{i,j} = \tilde{b}^k_j(s'_{i,j}, a'_{i,j}). \]

(33)

C.1.2 Typical episodes

Now we define the typical episodes and the typical episodes with respect to \((h, s) \in [H] \times S \). 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 \([k]_{\text{typ}} = \{ i : i \in [k], i \geq 250 H^2 S^2 A_2 \} \). Also, we define typical episodes with respect to \((h, s) \in [H] \times S \) as:

\[ [k]_{\text{typ}, h, s} = \{ i \in [k] : N^k_i(s) \geq 250 H^2 S^2 A_2 \}, \]

where \( N^k_i(s) \) is the real visitation number of \((h, s) \) before episode \( i \).

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

\[ H \cdot [k]/[k]_{\text{typ}} \leq 250 H^3 S^2 A_2. \]

(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 \( k, h, s \in [K] \times [H] \times S \):

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

(35)
We define $\delta$ based on optimism, for all $(i, h) \in [k] \times [H]$. \(\delta_{i,h} \leq \delta_{i,h} = \bar{V}^i_h(s_h^i) - V^*_h(s_h^i) = \bar{Q}^i_h(s_h^i, a_h^i) - Q^*_h(s_h^i, a_h^i)\)

\begin{align*}
\delta_{i,h} & \leq \bar{\delta}_{i,h} = \bar{V}^i_h(s_h^i) - V^*_h(s_h^i) = \bar{Q}^i_h(s_h^i, a_h^i) - Q^*_h(s_h^i, a_h^i) \\
& \leq b_h^i(s_h^i, a_h^i) + \sqrt{\frac{2\epsilon}{N_h^i(s_h^i, a_h^i)}} + (\bar{P}_h - P_h) \cdot V^*_h(s_h^i, a_h^i) + P_h \cdot (\bar{V}^i_{h+1} - V^*_h(s_h^i, a_h^i)) \\
& \leq b_h^i(s_h^i, a_h^i) + c_{1,i,h} + (\bar{P}_h - P_h) \cdot (\bar{V}^i_{h+1} - V^*_h(s_h^i, a_h^i)) + P_h \cdot (\bar{V}^i_{h+1} - V^*_h(s_h^i, a_h^i)) - \frac{2HSE_{\epsilon,\beta}}{N_h^i(s_h^i, a_h^i)}.
\end{align*}

The second inequality is because of the definition of $\bar{Q}^i_h$. 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 $b_{i,h}$.

### 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 $\bar{V}^i_h$ is upper bound of $V^*_h$ and $\bar{P}_h$ is close to $P_h$, $\text{Var}(\bar{V}^i_h)$ is not necessarily an upper bound of $\text{Var}(P_h)$. However, we can prove by induction that $\bar{V}^i_h$ is close enough to $V^*_h$, and therefore the last term of $b_h^i$ will be sufficiently large to make $\bar{V}^i_h$ a valid upper bound of $V^*_h$. More precisely, our induction is as below:

1. Assume for all $(i, h, s, a) \in [k] \times [H] \times S \times A, \bar{Q}^i_h(s, a) \geq Q^*_h(s, a)$, we prove for all $(h, s) \in [H] \times S$, $\bar{V}^i_h(s) - V^*_h(s) \leq \bar{O}
\left(\sqrt{S/AH^3/N_h^i(s)} + S^2AH^2E_{\epsilon,\beta}/N_h^i(s) + S^2AH^3/N_h^i(s)\right)$.

2. We deduce that $b_h^i$ compensates for the possible variance difference and for all $(h, s, a) \in [H] \times S \times A, \bar{Q}^{k+1}_h(s, a) \geq Q^*_h(s, a)$.

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

### C.3 Error decomposition

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

\begin{align*}
\delta_{i,h} & \leq \bar{\delta}_{i,h} = \bar{V}^i_h(s_h^i) - V^*_h(s_h^i) = \bar{Q}^i_h(s_h^i, a_h^i) - Q^*_h(s_h^i, a_h^i) \\
& \leq b_h^i(s_h^i, a_h^i) + \bar{\theta}_h(s_h^i, a_h^i) + 2E_{\epsilon,\beta}/N_h^i(s_h^i, a_h^i) + (\bar{P}_h - P_h) \cdot V^*_h(s_h^i, a_h^i) + P_h \cdot (\bar{V}^i_{h+1} - V^*_h(s_h^i, a_h^i)) \\
& \leq b_h^i(s_h^i, a_h^i) + c_{1,i,h} + (\bar{P}_h - P_h) \cdot (\bar{V}^i_{h+1} - V^*_h(s_h^i, a_h^i)) + P_h \cdot (\bar{V}^i_{h+1} - V^*_h(s_h^i, a_h^i)) - \frac{2HSE_{\epsilon,\beta}}{N_h^i(s_h^i, a_h^i)}.
\end{align*}

The second inequality is because of the definition of $\bar{Q}^i_h$. 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 $b_{i,h}$. 

\[ B_k = \sum_{i=1}^{k} \mathbb{I}(i \in [k] \backslash \{0\}) \sum_{j=1}^{H} \hat{b}_{i,j}. \]  

\[ C_{k,h,s} = \sum_{i=1}^{k} \mathbb{I}(i \in [k] \backslash \{0\}) \cdot \sum_{j=1}^{H} (c_{1,i,j} + c_{4,i,j}). \]  

\[ B_{k,h,s} = \sum_{i=1}^{k} \mathbb{I}(i \in [k] \backslash \{0\}) \cdot \sum_{j=1}^{H} \hat{b}_{i,j}. \]
In addition, we have:

\[
(P_h - P_h) \cdot (\tilde{V}_{h+1}^i - V_{h+1}^*) (s_h^i, a_h^i) = \sum_{s'} (P_h(s'|s_h^i, a_h^i) - P_h(s'|s_h^i, a_h^i)) \cdot (\tilde{V}_{h+1}^i(s') - V_{h+1}^*(s'))
\]

\[
\leq \sum_{s'} \left( \frac{2P_h(s'|s_h^i, a_h^i)}{N_h(s_h^i, a_h^i)} + \frac{2E_{e,\beta t}}{N_h(s_h^i, a_h^i)} \right) \cdot (\tilde{V}_{h+1}^i(s') - V_{h+1}^*(s'))
\]

\[
\leq \sum_{s'} \left( \frac{P_h(s'|s_h^i, a_h^i)}{H} + \frac{H}{N_h(s_h^i, a_h^i)} + \frac{2E_{e,\beta t}}{N_h(s_h^i, a_h^i)} \right) \cdot (\tilde{V}_{h+1}^i(s') - V_{h+1}^*(s'))
\]

\[
\leq \frac{1}{H} P_h \cdot (\tilde{V}_{h+1}^i - V_{h+1}^*) (s_h^i, a_h^i) + \frac{H^2 S_t}{N_h(s_h^i, a_h^i)} + \frac{2HSE_{e,\beta t}}{N_h(s_h^i, a_h^i)}
\]  

(40)

The first inequality is because of Lemma B.3. The second inequality holds since AM-GM inequality. The last inequality results from the fact that $V_{h+1}^* \geq \tilde{V}_{h+1}^i$.

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

\[
\delta_{i,h} \leq \delta_{i,h} \leq \tilde{\delta}_{i,h} \leq \tilde{\delta}_{i,h} \leq \tilde{\delta}_{i,h} = (1 + \frac{1}{H}) \tilde{\delta}_{i,h} + c_{1,i,h} + c_{4,i,h} + (1 + \frac{1}{H}) \epsilon_{i,h},
\]  

(41)

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

Recursively applying (41), we have:

\[
\tilde{\delta}_{i,h} \leq 3 \sum_{j=h}^{H} [\tilde{\delta}_{i,j} + c_{1,i,j} + c_{4,i,j} + \epsilon_{i,j}].
\]  

(42)

Summing over episodes, we have

\[
\sum_{i=1}^{k} \delta_{i,h} \leq \sum_{i=1}^{k} \tilde{\delta}_{i,h} \leq 3 \sum_{i=1}^{k} \sum_{j=h}^{H} [\tilde{\delta}_{i,j} + c_{1,i,j} + c_{4,i,j} + \epsilon_{i,j}].
\]  

(43)

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

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

\[
\sum_{i=1}^{k} \mathbb{I}(i \in [k]) \mathbb{P}(\sum_{j=h}^{H} \epsilon_{i,j} \leq H \sqrt{T_k} t).
\]  

(44)

\[
\sum_{i=1}^{k} \mathbb{I}(s_h^i = s, i \in [k], h) \sum_{j=h}^{H} \epsilon_{i,j} \leq H \sqrt{H N_k^H(s)_t}.
\]

We define $U_{k,h} = 3 \sum_{i=1}^{k} \mathbb{I}(i \in [k]) \sum_{j=h}^{H} [\tilde{\delta}_{i,j} + c_{1,i,j} + c_{4,i,j}] + 3H \sqrt{T_k} t$ and

\[
U_{k,h} = 3 \sum_{i=1}^{k} \mathbb{I}(s_h^i = s, i \in [k], h) \sum_{j=h}^{H} [\tilde{\delta}_{i,j} + c_{1,i,j} + c_{4,i,j}] + 3H \sqrt{H N_k^H(s)_t}.
\]

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 $h, s \in \mathcal{H} \times \mathcal{S}$,

\[
\sum_{i=1}^{k} \mathbb{I}(i \in [k]) \delta_{i,h} \leq \sum_{i=1}^{k} \mathbb{I}(i \in [k]) \tilde{\delta}_{i,h} \leq U_{k,h} \leq U_{k,1},
\]  

\[
\sum_{i=1}^{k} \mathbb{I}(i \in [k]) \sum_{j=h}^{H} \delta_{i,j} \leq H U_{k,1}.
\]  

(45)
At the same time, for all \( h, s \in [H] \times \mathcal{S} \) and \( j \geq h \),
\[
\sum_{i=1}^{k} \mathbb{I}(s^i_h = s, i \in [k]_{\text{typ}, h, s}) \delta_{i,j} \leq \sum_{i=1}^{k} \mathbb{I}(s^i_h = s, i \in [k]_{\text{typ}, h, s}) \tilde{\delta}_{i,j} \leq U_{k,h,s},
\]

\[
\sum_{i=1}^{k} \mathbb{I}(s^i_h = s, i \in [k]_{\text{typ}, h, s}) \sum_{j=h}^{H} \delta_{i,j} \leq HU_{k,h,s}.
\]

### 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 \( V_{i,j}^{\pi} = \text{Var}_{p_j(\cdot | s^i_j, a_j)} V_{j+1}^{\pi_i}(\cdot) \) and \( T_k = Hk \).

**Lemma C.4** (Lemma 8 of [Azar et al. 2017]). With probability \( 1 - \frac{\alpha}{17} \), for all \( k, h, s \in [K] \times [H] \times \mathcal{S} \), it holds that
\[
\sum_{i=1}^{k} \mathbb{I}(i \in [k]_{\text{typ}}) \sum_{j=h}^{H} V_{i,j}^{\pi} \leq H T_k + 2 \sqrt{H^3 T_k} + H_3 \leq 2HT_k,
\]
\[
\sum_{i=1}^{k} \mathbb{I}(s^i_h = s) \sum_{j=h}^{H} V_{i,j}^{\pi} \leq H^2 N_h^k(s) + 2 \sqrt{H^5 N_h^k(s)} + H_3 \leq 2H^2 N_h^k(s).
\]

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 \( V_{i,j}^{\pi} = \text{Var}_{p_j(\cdot | s^i_j, a_j)} V_{j+1}^{\pi_i}(\cdot) \).

**Lemma C.5.** Under the high probability event in Lemma C.2, it holds that
\[
\sum_{i=1}^{k} \mathbb{I}(i \in [k]_{\text{typ}}) \sum_{j=h}^{H} (V_{i,j}^{\pi} - V_{i,j}^{\pi_i}) \leq 2H^2 U_{k,1} + 2H^2 \sqrt{T_k}.
\]

Similarly, under the same high probability event, for all \( \mathbb{H} = [H] \times \mathcal{S} \),
\[
\sum_{i=1}^{k} \mathbb{I}(i \in [k]_{\text{typ}}) \sum_{j=h}^{H} (V_{i,j}^{\pi} - V_{i,j}^{\pi_i}) \leq 2H^2 U_{k,h,s} + 2H^2 \sqrt{H N_h^k(s)}.
\]

**Proof of Lemma C.5** We only prove the first conclusion, the second one can be proven in identical way.
\[
\sum_{i=1}^{k} \mathbb{I}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} (V_{i,j}^{\pi} - V_{i,j}^{\pi_i}) \sum_{j=1}^{H} \mathbb{E}_{s' \sim p_j(\cdot | s^i_j, a_j)} [V_{j+1}^{\pi_i}(s')^2 - V_{j+1}^{\pi_i}(s')^2] \\
\leq 2H \sum_{i=1}^{k} \mathbb{I}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \mathbb{E}_{s' \sim p_j(\cdot | s^i_j, a_j)} [V_{j+1}^{\pi_i}(s') - V_{j+1}^{\pi_i}(s')^2] \\
\leq 2H \left( \sum_{i=1}^{k} \mathbb{I}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \tilde{\delta}_{i,j+1} + H \sqrt{T_k} \right) \\
\leq 2H^2 U_{k,1} + 2H^2 \sqrt{T_k}.
\]

The first inequality is because \( V_{j+1}^{\pi_i} \geq V_{j+1}^{\pi_i} \). The second inequality results from the fact that \( V_{j+1}^{\pi_i} + V_{j+1}^{\pi_i} \leq 2H \). 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 \( \tilde{V}_{i,j}^{\pi} = \text{Var}_{p_j(\cdot | s^i_j, a_j)} \tilde{V}_{j+1}(\cdot) \).
The second inequality holds with probability \(1 - \frac{\beta}{12K}\), it holds that
\[
\sum_{i=1}^{k} \mathbb{I}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} (\bar{V}_{i,j} - V_{i,j}^*) \leq 2H^2U_{k,1} + 8H^2S\sqrt{HAT_{k,t}} + 6H^3S^2AE_{\epsilon,\beta}t. \tag{51}
\]
Similarly, under the same high probability event, for all \((h, s) \in [H] \times S,
\[
\sum_{i=1}^{k} \mathbb{I}(s_h^f = s, i \in [k]_{\text{typ},h,s}) s_{i,j} \sum_{j=h}^{H} (\bar{V}_{i,j} - V_{i,j}^*) \leq 2H^2U_{k,h,s} + 8H^3S\sqrt{AN_h^f(s)t} + 6H^3S^2AE_{\epsilon,\beta}t. \tag{52}
\]

\textbf{Proof of Lemma C.6.} We only prove the first conclusion, the second one can be proven in identical way. We have
\[
\begin{align*}
&\sum_{i=1}^{k} \mathbb{I}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} (\bar{V}_{i,j} - V_{i,j}^*) \leq \sum_{i=1}^{k} \mathbb{I}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \left[ \mathbb{E}_{s' \sim P_j(\cdot|s_j^i, a_j^i)} \bar{V}^i_{j+1}(s')^2 - \mathbb{E}_{s' \sim P_j(\cdot|s_j^i, a_j^i)} V^i_{j+1}(s')^2 \right] \\
&\quad + \sum_{i=1}^{k} \mathbb{I}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \mathbb{E}_{s' \sim P_j(\cdot|s_j^i, a_j^i)} \left[ V^i_{j+1}(s')^2 - V^i_{j+1}(s') \right] \\
&\quad + \sum_{i=1}^{k} \mathbb{I}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \left[ \left( \mathbb{E}_{s' \sim P_j(\cdot|s_j^i, a_j^i)} V^i_{j+1}(s') \right)^2 - \left( \mathbb{E}_{s' \sim P_j(\cdot|s_j^i, a_j^i)} V^i_{j+1}(s') \right)^2 \right].
\end{align*}
\tag{53}
\]

The inequality holds because of direct calculation and the fact that \(V^i_{j+1} \leq \bar{V}^i_{j+1} \leq V^i_{j+1}\). Next we bound these terms separately. First of all,
\[
(a) \leq \sum_{i=1}^{k} \mathbb{I}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} H^2 \cdot \left\| \bar{P}^i_j(\cdot|s_j^i, a_j^i) - P_j(\cdot|s_j^i, a_j^i) \right\|_1
\leq \sum_{i=1}^{k} \mathbb{I}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} H^2 \left( 2 \sqrt{\frac{S_j}{N_j^f(s_j^i, a_j^i)}} + \frac{2SE_{\epsilon,\beta}}{N_j^f(s_j^i, a_j^i)} \right)
\leq 2H^2S\sqrt{HAT_{k,t}} + 2H^3S^2AE_{\epsilon,\beta}t. \tag{54}
\]

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
\[
(b) \leq 2H \sum_{i=1}^{k} \mathbb{I}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \delta_{i,j+1} + H \sqrt{T_{k,t}} \leq 2H^2U_{k,1} + 2H^2\sqrt{T_{k,t}}. \tag{55}
\]
Lastly, for (c), we have
\[
(c) \leq \sum_{i=1}^{k} \mathbb{I}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \left[ \left( \mathbb{E}_{s' \sim P_j(\cdot|s_j^i, a_j^i)} V^i_{j+1}(s') \right)^2 - \left( \mathbb{E}_{s' \sim P_j(\cdot|s_j^i, a_j^i)} V^i_{j+1}(s') \right)^2 + 2H^2 \cdot \left\| \left( \bar{P}^i_j - P^i_j \right)(\cdot|s_j^i, a_j^i) \right\|_1 \right]
\leq \sum_{i=1}^{k} \mathbb{I}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \left( 2H \cdot 2H \sqrt{\frac{l}{N_j^f(s_j^i, a_j^i)}} + 2H^2 \cdot \frac{2SE_{\epsilon,\beta}}{N_j^f(s_j^i, a_j^i)} \right)
\leq 4H^2\sqrt{HSA\sqrt{AT_{k,t}}} + 4H^3S^2AE_{\epsilon,\beta}t. \tag{56}
\]

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

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

$$C_k = \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} (c_{1,i,j} + c_{4,i,j}) \leq 3\sqrt{H^2 SAT_k t^2} + 2\sqrt{H^3 S AU_{k,1} t^2} + H^3 S^2 A t^2. \quad (57)$$

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

$$C_{k,h,s} \leq 3\sqrt{H^3 S A N_k(s) t^2} + 2\sqrt{H^3 S AU_{k,h,s} t^2} + H^3 S^2 A t^2. \quad (58)$$

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

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

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

$$\begin{align*}
(a) = \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} 2V_{i,j}^* & \leq \sqrt{2k} \cdot \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \sqrt{\frac{N_j (s_j, a_j) \sum_{j=1}^{H} 1 \cdot N_j (s_j, a_j)}{V_{i,j}^*}}. \\
& \leq \sqrt{2k} \cdot \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \frac{V_{i,j}^*}{N_j (s_j, a_j)}. \quad (60)
\end{align*}$$

Due to direct calculation, we have (c) $\leq H S A t$ and in addition,

$$\begin{align*}
(d) & \leq \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} V_{i,j} - \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} (V_{i,j}^* - V_{i,j}) \\
& \leq 2HT_k + 2H^2 U_{k,1} + 2H^2 \sqrt{1/t_k},
\end{align*}$$

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

$$\begin{align*}
(a) & \leq 3\sqrt{H^2 SAT_k t^2} + 2\sqrt{H^3 S AU_{k,1} t^2}. \quad (62)
\end{align*}$$

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

$$\begin{align*}
(b) & \leq \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \frac{H^2 S l \sum_{j=1}^{H} V_{i,j}^*}{N_j (s_j, a_j)} \leq H^3 S^2 A t^2. \quad (63)
\end{align*}$$

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

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

**Lemma C.8.** Under the high probability events in Lemma C.2, Lemma C.4, and Lemma C.5, with probability $1 - \frac{\beta}{10K}$,

$$B_k \leq 16\sqrt{H^2 SAT_k t^2} + 6\sqrt{H^3 S AU_{k,1} t^2} + 250H^2S^2AE_{\epsilon,\beta}t^2 + 250H^3S^2A t^2. \quad (64)$$

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

$$B_{k,h,s} \leq 16\sqrt{H^3 S AN^k(s) t^2} + 6\sqrt{H^3 S AU_{k,h,s} t^2} + 250H^2S^2AE_{\epsilon,\beta}t^2 + 250H^3S^2A t^2. \quad (65)$$
Proof of Lemma C.8. We only prove the first conclusion, the second one can be proven in identical way. We have

\begin{equation}
B_k \leq \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \sqrt{\frac{V_i,j_k}{N_j^i(s_j^i, a_j^i)}} + \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} 2 \sqrt{\frac{N_j^i(s_j^i, a_j^i)}{N_j^i(s_j^i, a_j^i)}} + \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \frac{40HSE_{\epsilon,\beta t}}{N_j^i(s_j^i, a_j^i)}
\end{equation}

\begin{equation}
+ \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \sqrt{\sum_{s'} \tilde{P}_j^i(s'|s_j^i, a_j^i)} \cdot \min \left\{ \frac{1000^2 H^3 S^2 A^2 \epsilon^2 \beta t^4}{N_j^i(s_j^i, a_j^i)} + \frac{1000^2 H^4 S^4 A^2 \epsilon^2 \beta t^4}{N_j^i(s_j^i, a_j^i)} + \frac{1000^2 H^6 S^4 A^2 \beta t^4}{N_j^i(s_j^i, a_j^i)} \right\}.
\end{equation}

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

\begin{equation}
(a) \leq 4\sqrt{\ell} \cdot \sqrt{HSA_t} \cdot \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} \bar{V}_{i,j}}
\end{equation}

\begin{equation}
\leq 4\sqrt{\ell} \cdot \sqrt{HSA_t} \cdot 2H^2 T_k + 2H^2 U_{k,1} + 8H^2 S \sqrt{HAT_{k,t}} + 6H^3 S^2 A E_{\epsilon,\beta t}
\end{equation}

\begin{equation}
\leq 8\sqrt{H^2 S A T_{k,t}} + 4\sqrt{H^2 S A T_{k,t}} \leq 2HT_k.
\end{equation}

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, $8H^2 S \sqrt{HAT_{k,t}} \leq 2HT_k$.

According to direct calculation,

\begin{equation}
(b) \leq 2\sqrt{2\ell} \cdot HSA_t \leq \sqrt{H^2 S A T_{k,t}}^2
\end{equation}

For (c), it holds that

\begin{equation}
(c) \leq 40HSE_{\epsilon,\beta t} \cdot HSA_t \leq 40H^2 S^2 A E_{\epsilon,\beta t}^2.
\end{equation}

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

\begin{equation}
(d) \leq \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \sqrt{\sum_{s'} \tilde{P}_j^i(s'|s_j^i, a_j^i)} \cdot \min \left\{ \frac{1000^2 H^3 S^2 A^2 \epsilon^2 \beta t^4}{N_j^i(s_j^i, a_j^i)} + \frac{1000^2 H^4 S^4 A^2 \epsilon^2 \beta t^4}{N_j^i(s_j^i, a_j^i)} + \frac{1000^2 H^6 S^4 A^2 \beta t^4}{N_j^i(s_j^i, a_j^i)} \right\}
\end{equation}

\begin{equation}
\leq 4\sqrt{\ell} \cdot \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} \bar{V}_{i,j}}
\end{equation}

\begin{equation}
\leq 4\sqrt{\ell} \cdot \sum_{i=1}^{k} \mathbb{1}(i \in [k]_{\text{typ}}) \sum_{j=1}^{H} \sum_{s'} \tilde{P}_j^i(s'|s_j^i, a_j^i) \min \left\{ \frac{1000^2 H^3 S^2 A^2 \epsilon^2 \beta t^4}{N_j^i(s_j^i, a_j^i)} + \frac{1000^2 H^4 S^4 A^2 \epsilon^2 \beta t^4}{N_j^i(s_j^i, a_j^i)} + \frac{1000^2 H^6 S^4 A^2 \beta t^4}{N_j^i(s_j^i, a_j^i)} \right\}.
\end{equation}
Note that \((e) \leq HS\epsilon\). For \((f)\), we have with probability \(1 - \frac{\beta}{12K}\),

\[
(f) \leq \sum_{i=1}^{k} \sum_{j=1}^{H} H^2 \left\| \hat{P}_j^i (\cdot | s_j^i, a_j^i) - P_j (\cdot | s_j^i, a_j^i) \right\|_1 \left( \frac{1000^2 H^3 S A_i^2}{N_j^i (s_j^i)^2} + \frac{1000^2 H^4 S^4 A^2 E_{e,\beta} t^4}{N_j^{i+1} (s_j^{i+1})^2} + \frac{1000^2 H^6 S^4 A^2 t^4}{N_j^{i+1} (s_j^{i+1})^2} \right) + H^2 \sqrt{T_k}
\]

\[
\leq \sum_{i=1}^{k} \sum_{j=1}^{H} H^2 \left\| \hat{P}_j^i (\cdot | s_j^i, a_j^i) - P_j (\cdot | s_j^i, a_j^i) \right\|_1 + H^2 \sqrt{T_k}
\]

\[
\leq 3H^2 S \sqrt{H A T_k} + 2H^3 S^2 A E_{e,\beta} t + 1000H^4 S^3 A E_{e,\beta} t^2 + 2000H^5 S^3 A_i^2,
\]

(71)

where the first inequality holds because \(\min \{\cdot, H^2\} \leq H^2\). 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

\[
(d) \leq 7 \sqrt{H^3 S^2 A_i^2 + H A T_k} + 120 \sqrt{H^5 S^4 A^2 E_{e,\beta} t^4} + 200H^3 S^2 A_i^2.
\]

(72)

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

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

\[
\text{Regret}(k) := \sum_{i=1}^{k} \delta_{i,1}, \quad \tilde{\text{Regret}}(k) := \sum_{i=1}^{k} \tilde{\delta}_{i,1}.
\]

(73)

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

\[
\text{Regret}(k, h, s) := \sum_{i=1}^{k} \mathbb{I}(s_h^i = s) \cdot \delta_{i,h}, \quad \tilde{\text{Regret}}(k, h, s) := \sum_{i=1}^{k} \mathbb{I}(s_h^i = s) \cdot \tilde{\delta}_{i,h}.
\]

(74)

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

\[
\text{Regret}(k) \leq \tilde{\text{Regret}}(k) \leq 1000 \left( \sqrt{H^2 S A_i t^2 + H^2 S^2 A E_{e,\beta} t^2 + H^3 S^2 A_i^2} \right).
\]

(75)

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

\[
\text{Regret}(k, h, s) \leq \tilde{\text{Regret}}(k, h, s) \leq 1000 \left( \sqrt{H^2 S A_i N_h^k(s)^2 + H^2 S^2 A E_{e,\beta} t^2 + H^3 S^2 A_i^2} \right).
\]

(76)

\textbf{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 \(C.4\) (for all \(k \in [K]\)) and Lemma \(C.6\) (for all \(k \in [K]\)) hold. The failure probability is bounded by

\[
\frac{\beta}{3} + \frac{\beta}{12} + \frac{\beta}{12} + K \cdot \frac{\beta}{12K} + K \cdot \frac{\beta}{12K} \leq \beta.
\]

(77)
We only prove the first conclusion, the second one can be proven in identical way. It holds that
\[
\text{Regret}(k) \leq \sum_{i=1}^{k} \delta_{i,1} \\
\leq U_{k,1} + 250H^3S^2A_t^2 \\
\leq 3B_k + 3C_k + 3H\sqrt{T_k}\epsilon + 250H^3S^2A_t^2 \\
\leq 60\sqrt{H^2SA_kt^2} + 24\sqrt{H^3SA_kh^2} + 750H^2S^2A_{e,\beta}t^2 + 1000H^3S^2A_t^2 \\
\leq 1000 \left( \sqrt{H^2SA_kt^2} + H^2S^2A_{e,\beta}t^2 + H^3S^2A_t^2 \right),
\] (78)
where the second inequality is because Lemma \[C.3\] and the fact that \( H \cdot \left| [k]/[k]_{\text{typ}} \right| \leq 250H^3S^2A_t^2 \). 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
\[
1000 \left( \sqrt{H^3SAN^k_h(s)\epsilon^2} + H^2S^2A_{e,\beta}t^2 + H^3S^2A_t^2 \right) \geq \text{Regret}(k, h, s)
\]
\[
\geq \sum_{i=1}^{k} \mathbb{1}(s_h^i = s) \cdot \delta_{i,h} \\
\geq \sum_{i=1}^{k} \mathbb{1}(s_h^i = s) \left( \widehat{V}_h^i(s) - V_h^*\epsilon(s) \right) \\
\geq \sum_{i=1}^{k} \mathbb{1}(s_h^i = s) \left( \widehat{V}_h^i(s) - V_h^*(s) \right) \\
\geq N_h^k(s) \cdot \left( \widehat{V}_h^*(s) - V_h^*(s) \right),
\] (79)
where the first three inequalities are due to definitions of \( \text{Regret}(k, h, s) \) and \( \delta_{i,h} \). The forth inequality holds because \( V_h^* \geq V_h^{\pi^*} \). The last inequality results from our algorithmic design that \( \widehat{V}_h^i(s) \) is non-increasing (line 9 of Algorithm \[7\]).

Therefore, we have \( \widehat{V}_h^k(s) - V_h^*(s) \leq 1000 \left( \sqrt{H^3SAN^k_h(s)\epsilon^2} / N_h^k(s) + H^2S^2A_{e,\beta}t^2 / N_h^k(s) + H^3S^2A_t^2 / N_h^k(s) \right) \).

Now we have proven the first point of our induction. Together with the point 2 (which we will prove in Section \[C.9\]), 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 \( (i, h, s, a) \in [k] \times [H] \times \mathcal{S} \times \mathcal{A}, Q_h^i(s, a) \geq Q_h^*(s, a) \), we prove
\[
\text{Regret}(k) \leq \text{Regret}(k) \leq 1000 \left( \sqrt{H^2SA_kt^2} + H^2S^2A_{e,\beta}t^2 + H^3S^2A_t^2 \right)
\]
and for all \( (h, s) \in [H] \times \mathcal{S} \),
\[
\widehat{V}_h^k(s) - V_h^*(s) \leq 1000 \left( \sqrt{SAH^3t^2 / N_h^k(s)} + S^2AH^2E_{e,\beta}t^2 / N_h^k(s) + S^2AH^3t^2 / N_h^k(s) \right).
\]

2. Given that for all \( (h, s) \in [H] \times \mathcal{S} \),
\[
\widehat{V}_h^k(s) - V_h^*(s) \leq 1000 \left( \sqrt{SAH^3t^2 / N_h^k(s)} + S^2AH^2E_{e,\beta}t^2 / N_h^k(s) + S^2AH^3t^2 / N_h^k(s) \right),
\]
we prove that for all \( (h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}, Q_h^{k+1}(s, a) \geq Q_h^*(s, a) \).

Therefore, with probability \( 1 - \beta \), we have \( (T = KH) \)
\[
\text{Regret}(K) \leq \text{Regret}(K) \leq \tilde{O} \left( \sqrt{H^2SAT} + H^2S^2A_{e,\beta} + H^3S^2A \right).
\] (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 \((h, s) \in [H] \times S,\)
\[
\tilde{V}_{h+1}^k(s) - V^*_h(s) \leq 1000 \left( \sqrt{SAH^3\epsilon^2/N_{h+1}^{k+1}(s)} + S^2AH^2E_{\epsilon,\beta}^2/N_{h+1}^{k+1}(s) + S^2AH^3\epsilon^2/N_{h+1}^{k+1}(s) \right),
\]
we prove for all \((h, s, a) \in [H] \times S \times A, \tilde{Q}^k_{h+1}(s, a) \geq Q^*_h(s, a)\) 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 \((j, s, a) \in [H] \times S \times A, \tilde{Q}^k_{j}(s, a) \geq Q^*_j(s, a).\)
2. For all \(s \in S,\)
\[
0 \leq \tilde{V}_{h+1}^k(s) - V^*_h(s) \leq 1000 \left( \sqrt{SAH^3\epsilon^2/N_{h+1}^{k+1}(s)} + S^2AH^2E_{\epsilon,\beta}^2/N_{h+1}^{k+1}(s) + S^2AH^3\epsilon^2/N_{h+1}^{k+1}(s) \right),
\]

Then we have for all \((s, a) \in S \times A, \tilde{Q}_{h+1}^k(s, a) \geq Q^*_h(s, a).\)

**Proof of Lemma C.11.** For all \((s, a) \in S \times A, \tilde{Q}_{h+1}^k(s, a) \geq Q^*_h(s, a),\) it suffices to prove that \(\tilde{V}_{h+1}^k(s, a) + b_{h+1}^k(s, a) \geq Q^*_h(s, a).\) We have
\[
\tilde{V}_{h+1}^k(s, a) + \tilde{P}_{h+1}^k(s, a) + b_{h+1}^k(s, a) - Q^*_h(s, a) \\
\geq (\tilde{V}_{h}^k - r_h)(s, a) + (\tilde{P}_{h+1}^k - P_h) \cdot V^*_h(s, a) + b_{h+1}^k(s, a) \\
\geq (\tilde{V}_{h}^k - P_h) \cdot V^*_h(s, a) + 2\sqrt{\frac{\text{Var}_{s'} \tilde{P}_{h+1}^k(s'|s, a) \tilde{V}_{h+1}^k(\cdot)}{N_{h+1}^k(s, a)}} + \frac{19HSE_{\epsilon,\beta}}{N_{h+1}^k(s, a)} \\
+ 4\sqrt{\frac{\sum_{s'} \tilde{P}_{h+1}^k(s'|s, a) \min \left\{ \frac{1000^2H^3SA_3}{N_{h+1}^k(s')} + \frac{1000^2H^4S^4A_2E_{\epsilon,\beta}^2}{N_{h+1}^k(s')^2} + \frac{1000^2H^2S^4A_2^2}{N_{h+1}^k(s')^2} \right\}, H^2} \}}{N_{h}^k(s, a)} \\
\geq - \sqrt{2\text{Var}_{s'} \tilde{P}_{h+1}^k(s'|s, a) V^*_h(\cdot)} \cdot \epsilon \frac{\text{Var}_{s'} \tilde{P}_{h+1}^k(s'|s, a) \tilde{V}_{h+1}^k(\cdot) \cdot \epsilon}{N_{h}^k(s, a)} + 2\sqrt{\frac{\text{Var}_{s'} \tilde{P}_{h+1}^k(s'|s, a) \tilde{V}_{h+1}^k(\cdot) \cdot \epsilon}{N_{h}^k(s, a)}} + \frac{17HSE_{\epsilon,\beta}}{N_{h+1}^k(s, a)} \\
+ 4\sqrt{\frac{\sum_{s'} \tilde{P}_{h+1}^k(s'|s, a) \min \left\{ \frac{1000^2H^3SA_3}{N_{h+1}^k(s')} + \frac{1000^2H^4S^4A_2E_{\epsilon,\beta}^2}{N_{h+1}^k(s')^2} + \frac{1000^2H^2S^4A_2^2}{N_{h+1}^k(s')^2} \right\}, H^2} \}}{N_{h}^k(s, a)} \\
\geq - \sqrt{2\text{Var}_{s'} \tilde{P}_{h+1}^k(s'|s, a) V^*_h(\cdot) \cdot \epsilon} \frac{\text{Var}_{s'} \tilde{P}_{h+1}^k(s'|s, a) \tilde{V}_{h+1}^k(\cdot) \cdot \epsilon}{N_{h}^k(s, a)} + 2\sqrt{\frac{\text{Var}_{s'} \tilde{P}_{h+1}^k(s'|s, a) \tilde{V}_{h+1}^k(\cdot) \cdot \epsilon}{N_{h}^k(s, a)}} \\
+ 4\sqrt{\frac{\sum_{s'} \tilde{P}_{h+1}^k(s'|s, a) \min \left\{ \frac{1000^2H^3SA_3}{N_{h+1}^k(s')} + \frac{1000^2H^4S^4A_2E_{\epsilon,\beta}^2}{N_{h+1}^k(s')^2} + \frac{1000^2H^2S^4A_2^2}{N_{h+1}^k(s')^2} \right\}, H^2} \}}{N_{h}^k(s, a)} \\
\geq \epsilon \cdot \frac{19HSE_{\epsilon,\beta}}{N_{h}^k(s, a)} + 17HSE_{\epsilon,\beta} \cdot \epsilon + \frac{17HSE_{\epsilon,\beta}}{N_{h+1}^k(s, a)} \\
\geq 0,
\]
where the first inequality is because of Bellman equation and condition 2. The second inequality holds since the definition of \(b_{h+1}^k\) 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 \(\sqrt{\text{Var}(X)} \leq \sqrt{2\text{Var}(Y)} + \sqrt{2\text{Var}(X - Y)}\) (Lemma 2 of Azar et al. [2017]), we have
\[
\sqrt{\text{Var}_{s'} \tilde{P}_{h+1}^k(s'|s, a) V^*_h(\cdot)} \leq \sqrt{2\text{Var}_{s'} \tilde{P}_{h+1}^k(s'|s, a) \tilde{V}_{h+1}^k(\cdot)} + \sqrt{2\text{Var}_{s'} \tilde{P}_{h+1}^k(s'|s, a) \tilde{V}_{h+1}^k(\cdot) - V^*_h(\cdot)},
\]
(82)
In addition,
\[
(a) \leq \sqrt{2 \sum_{s'} \tilde{P}^{k+1}_h(s'|s, a) \left( \tilde{V}^{k+1}_{h+1}(s') - V^{k+1}_{h+1}(s') \right)^2 } \\
\leq \sqrt{6 \sum_{s'} \tilde{P}^{k+1}_h(s'|s, a) \left( \frac{1000^2 H^3 S A^2}{N^{k+1}_{h+1}(s')} + \frac{1000^2 H^4 S A^2 E_{c, a, h}^4}{N^{k+1}_{h+1}(s')^2} + \frac{1000^2 H^6 S^4 A^2 E_{c, a, h}^4}{N^{k+1}_{h+1}(s')^2} \right),}
\]
(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^2 + b^2 + 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 \(N^k_h\) is the original count, \(\tilde{N}^k_h\) is the noisy count after step (1) of both Privatizers and \(\bar{N}^k_h\) is the final private counts.

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

For utility analysis, because of Theorem 3.6 of [Chen et al. 2011], our choice \(\epsilon' = \frac{\epsilon}{3H \log R}\) in Binary Mechanism and a union bound, with probability \(1 - \frac{\beta}{3}\), for all \((h, k, s, a, s')\),
\[
|\tilde{N}^k_h(s, a, s') - N^k_h(s, a, s')| \leq O\left( \frac{H}{\epsilon} \log(HSAT/\beta) \right), \quad |\tilde{N}^k_h(s, a) - N^k_h(s, a)| \leq O\left( \frac{H}{\epsilon} \log(HSAT/\beta)^2 \right) = O\left( \frac{H}{\epsilon} \log(HSAT/\beta)^2 \right),
\]
(84)
Together with Lemma 5.4, the Central Privatizer satisfies Assumption 3.1 with \(E_{c, \beta} = O\left( \frac{H}{\epsilon} \right)\).

**Proof of Theorem 5.2** The proof directly results from plugging \(E_{c, \beta} = O\left( \frac{H}{\epsilon} \right)\) into Theorem 4.1.

**Proof of Lemma 5.4** For clarity, we denote the solution of (4) by \(\tilde{N}^k_h\) and therefore \(\tilde{N}^k_h(s, a, s') = N^k_h(s, a, s') + E_{c, \beta}\), \(\tilde{N}^k_h(s, a) = \tilde{N}^k_h(s, a) + E_{c, \beta}\).

When the condition (two inequalities) in Lemma 5.4 holds, the original counts \(\{N^k_h(s, a, s')\}_{s' \in S}\) is a feasible solution to the optimization problem, which means that
\[
\max_{s'} |\tilde{N}^k_h(s, a, s') - N^k_h(s, a, s')| \leq \max_{s'} |\tilde{N}^k_h(s, a, s') - N^k_h(s, a, s')| \leq \frac{E_{c, \beta}}{4}.
\]
Combining with the condition in Lemma 5.4 with respect to \(\tilde{N}^k_h(s, a, s')\), it holds that
\[
|\tilde{N}^k_h(s, a, s') - N^k_h(s, a, s')| \leq |\tilde{N}^k_h(s, a, s') - N^k_h(s, a, s')| + |\tilde{N}^k_h(s, a, s') - N^k_h(s, a, s')| \leq \frac{E_{c, \beta}}{2}.
\]
Since $\bar{N}_k^h(s, a, s') = \bar{N}_k^h(s, a, s') + \frac{E_{\epsilon, \beta}}{2S}$ and $\bar{N}_k^h(s, a, s') \geq 0$, we have
\[
\bar{N}_k^h(s, a, s') > 0, \quad |\bar{N}_k^h(s, a, s') - N_k^h(s, a, s')| \leq E_{\epsilon, \beta}.
\] (85)

For $\bar{N}_k^h(s, a)$, according to the constraints in the optimization problem (4), it holds that
\[
|\bar{N}_k^h(s, a) - \bar{N}_k^h(s, a)| \leq \frac{E_{\epsilon, \beta}}{4}.
\]

Combining with the condition in Lemma 5.4 with respect to $\bar{N}_k^h(s, a)$, it holds that
\[
|\bar{N}_k^h(s, a) - N_k^h(s, a)| \leq |\bar{N}_k^h(s, a) - \bar{N}_k^h(s, a)| + |\bar{N}_k^h(s, a) - N_k^h(s, a)| \leq \frac{E_{\epsilon, \beta}}{2}.
\]

Since $\bar{N}_k^h(s, a) = N_k^h(s, a) + \frac{E_{\epsilon, \beta}}{2}$, we have
\[
N_k^h(s, a) \leq \bar{N}_k^h(s, a) \leq N_k^h(s, a) + E_{\epsilon, \beta}.
\] (86)

According to the last line of the optimization problem (4), we have $\bar{N}_k^h(s, a) = \sum_{s' \in S} \bar{N}_k^h(s, a, s')$ and therefore,
\[
\bar{N}_k^h(s, a) = \sum_{s' \in S} \bar{N}_k^h(s, a, s').
\] (87)

The proof is complete by combining (85), (86) and (87). \hfill \Box

**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 $(k, h, s, a, s')$,
\[
|\bar{N}_k^h(s, a, s') - N_k^h(s, a, s')| \leq O\left(\frac{H}{\epsilon} \sqrt{K \log(HSAT/\beta)}\right), \quad |\bar{N}_k^h(s, a) - N_k^h(s, a)| \leq O\left(\frac{H}{\epsilon} \sqrt{K \log(HSAT/\beta)}\right), \quad |\bar{R}_k^h(s, a) - R_k^h(s, a)| \leq O\left(\frac{H}{\epsilon} \sqrt{K \log(HSAT/\beta)}\right).
\] (88)

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

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