Differentially Private Regret Minimization in Episodic Markov Decision Processes

Sayak Ray Chowdhury,[1*] **Xingyu Zhou** [2*]

[1] Indian Institute of Science, Bangalore, India
[2] ECE Department, Wayne State University, Detroit, USA
sayak@iisc.ac.in, xingyu.zhou@wayne.edu

Abstract

We study regret minimization in finite horizon tabular Markov decision processes (MDPs) under the constraints of differential privacy (DP). This is motivated by the widespread applications of reinforcement learning (RL) in real-world sequential decision making problems, where protecting users' sensitive and private information is becoming paramount. We consider two variants of DP – joint DP (JDP), where a centralized agent is responsible for protecting users' sensitive data and local DP (LDP), where information needs to be protected directly on the user side. We first propose two general frameworks – one for policy optimization and another for value iteration – for designing private, optimistic RL algorithms. We then instantiate these frameworks with suitable privacy mechanisms to satisfy JDP and LDP requirements, and simultaneously obtain sublinear regret guarantees. The regret bounds show that under JDP, the cost of privacy is only a lower order additive term, while for a stronger privacy protection under LDP, the cost suffered is multiplicative. Finally, the regret bounds are obtained by a unified analysis, which, we believe, can be extended beyond tabular MDPs.

1 Introduction

Reinforcement learning (RL) is a fundamental sequential decision making problem, where an agent learns to maximize its reward in an unknown environment through trial and error. Recently, it is ubiquitous in various personalized services, including healthcare (Gottesman et al. 2019), virtual assistants (Li et al. 2016), social robots (Gordon et al. 2016) and online recommendations (Li et al. 2010). In these applications, the learning agent continuously improves its decision by learning from users' personal data and feedback. However, nowadays people are becoming increasingly concerned about potential privacy leakage in these interactions. For example, in personalized healthcare, the private data of a patient can be sensitive informations such as her age, gender, height, weight, medical history, state of the treatment, etc. Therefore, developing RL algorithms which can protect users' private data are of paramount importance in these applications.

Differential privacy (DP) (Dwork 2008) has become a standard in designing private sequential decision-making algorithms both in the full information (Jain, Kothari, and Thakurta 2012) and partial or bandit information (Mishra and Thakurta 2015; Tossou and Dimitrakakis 2016) settings. Under DP, the learning agent collects users' raw data to train its algorithm while ensuring that its output will not reveal users' sensitive information. This notion of privacy protection is suitable for situations, where a user is willing to share her own information to the agent in order to obtain a service specially tailored to her needs, but meanwhile she does not like to allow any third party to infer her private information seeing the output of the learning algorithm (e.g., Google GBoard). However, a recent body of work (Shariff and Sheffet 2018; Dubey 2021) show that the standard DP guarantee is irreconcilable with sublinear regret in contextual bandits, and thus, a variant of DP, called *joint differential privacy* (JDP) (Kearns et al. 2014) is considered. Another variant of DP, called *local differential privacy* (LDP) (Duchi, Jordan, and Wainwright 2013) has recently gained increasing popularity in personalized services due to its stronger privacy protection. It has been studied in various bandit settings recently (Ren et al. 2020; Zheng et al. 2020; Zhou and Tan 2020). Under LDP, each user's raw data is directly protected before being sent to the learning agent. Thus, the learning agent only has access to privatized data to train its algorithm, which often leads to a worse regret guarantee compared to DP or JDP.

In contrast to the vast amount of work in private bandit algorithms, much less attention are given to address privacy in RL problems. To the best of our knowledge, Vietri et al. (2020) propose the first RL algorithm – PUCB – for regret minimization with JDP guarantee in tabular finite state, finite action MDPs. On the other hand, Garcelon et al. (2020) design the first private RL algorithm – LDP-OBI – with regret and LDP guarantees. Recently, Chowdhury, Zhou, and Shroff (2021) study linear quadratic regulators under the JDP constraint. It is worth noting that all these prior work consider only value-based RL algorithms, and a study on policy-based private RL algorithms remains elusive. Recently, policy optimization (PO) has seen great success in many real-world applications, especially when coupled with deep neural networks (Silver et al. 2017; Duan et al. 2016; Wang, Li, and He 2018), and a variety of PO based algorithms have been proposed (Williams 1992; Kakade 2001; Schulman et al. 2015, 2017; Konda and Tsitsiklis 2000). The

[*] Equal contribution

	Algorithm	Regret (ε-JDP)	Regret (ε-LDP)
PO	PRIVATE-UCB-PO	$\widetilde{O}\left(\sqrt{S^2AH^3T} + S^2AH^3/\varepsilon\right)$	$\widetilde{O}\left(\sqrt{S^2AH^3T} + S^2A\sqrt{H^5T}/\varepsilon\right)$
VI	PRIVATE-UCB-VI	$\widetilde{O}\left(\sqrt{SAH^3T} + S^2AH^3/\varepsilon\right)$	$\widetilde{O}\left(\sqrt{SAH^3T} + S^2A\sqrt{H^5T}/\varepsilon\right)$
	PUCB (Vietri et al. 2020)	$\widetilde{O}\left(\sqrt{S^2AH^3T} + S^2AH^3/\varepsilon\right)$ [1]	NA
	LDP-OBI (Garcelon et al. 2020)	NA	$\widetilde{O}\left(\sqrt{S^2AH^3T} + S^2A\sqrt{H^5T}/\varepsilon\right)$ [2]

Table 1: Regret comparisons for private RL algorithms on episodic tabular MDP. $T = KH$ is total number of steps, where K is the total number of episodes and H is the number of steps per episode. S is the number of states, and A is the number of actions. $\varepsilon > 0$ is the desired privacy level. $\widetilde{O}(\cdot)$ hides $\mathrm{polylog}\,(S, A, T, 1/\delta)$ factors, where $\delta \in (0, 1]$ is the desired confidence level.

theoretical understandings of PO have also been studied in both computational (i.e., convergence) perspective (Liu et al. 2019; Wang et al. 2019a) and statistical (i.e., regret) perspective (Cai et al. 2020; Efroni et al. 2020). Thus, one fundamental question to ask is how to build on existing understandings of non-private PO algorithms to design sample-efficient policy-based RL algorithms with general privacy guarantees (e.g., JDP and LDP), which is the main motivation behind this work. Also, the existing regret bounds in both Vietri et al. (2020) and Garcelon et al. (2020) for private valued-iteration (VI) based RL algorithms are loose. Moreover, the algorithm design and regret analysis under JDP in Vietri et al. (2020) and the ones under LDP in Garcelon et al. (2020) follow different approaches (e.g., choice of exploration bonus terms and corresponding analysis). Thus, another important question to ask is whether one can obtain tighter regret bounds for VI based private RL algorithms via a unified framework under general privacy requirements.

Contributions. Motivated by the two questions above, we make the following contributions.

- We present a general framework – PRIVATE-UCB-PO – for designing private policy-based optimistic RL algorithms in tabular MDPs. This framework enables us to establish the first regret bounds for PO under both JDP and LDP requirements by instantiating it with suitable private mechanisms – the CENTRAL-PRIVATIZER and the LOCAL-PRIVATIZER – respectively.
- We revisit private optimistic value-iteration in tabular MDPs by proposing a general framework – PRIVATE-UCB-VI – for it. This framework allows us to improve upon the existing regret bounds under both JDP and LDP constraints using a unified analysis technique.
- Our regret bounds show that for both policy-based and value-based private RL algorithms, the cost of JDP guarantee is only a lower-order additive term compared to the non-private regret. In contrast, under the stringer LDP requirement, the cost suffered is multiplicative and is of the same order. Our regret bounds and their comparison to the existing ones is summarised in Table 1.

[1] Vietri et al. (2020) claim a $\widetilde{O}\left(\sqrt{SAH^3T} + S^2AH^3/\varepsilon\right)$ regret bound for PUCB. However, to the best of our understanding, we believe the current analysis has gaps (see Section 4), and the best achievable regret for PUCB should have an additional $\sqrt{S}$ factor in the first term.

[2] Garcelon et al. (2020) consider stationary transition kernels, and show a $\widetilde{O}\left(\sqrt{S^2AH^2T} + S^2A\sqrt{H^5T}/\varepsilon\right)$ regret bound for LDP-OBI. For non-stationary transitions, as considered in this work, an additional multiplicative $\sqrt{H}$ factor would appear in the first term of the bound.

Related work. Beside the papers mentioned above, there are other related work on differentially private online learning (Guha Thakurta and Smith 2013; Agarwal and Singh 2017) and multi-armed bandits (Tossou and Dimitrakakis 2017; Hu, Huang, and Mehta 2021; Sajed and Sheffet 2019; Gajane, Urvoy, and Kaufmann 2018; Chen et al. 2020). In the RL setting, in addition to Vietri et al. (2020); Garcelon et al. (2020) that focus on value-iteration based regret minimization algorithms under privacy constraints, Balle, Gomrokchi, and Precup (2016) considers private policy evaluation with linear function approximation. For MDPs with continuous state spaces, Wang and Hegde (2019) proposes a variant of Q-learning to protect the rewards information by directly injecting noise into value functions. Recently, a distributed actor-critic RL algorithm under LDP is proposed in Ono and Takahashi (2020) but without any regret guarantee. While there are recent advances in regret guarantees for policy optimization (Cai et al. 2020; Efroni et al. 2020), we are not aware of any existing work on private policy optimization. Thus, our work takes the first step towards a unified framework for private policy-based RL algorithms in tabular MDPs with general privacy and regret guarantees.

2 Problem formulation

In this section, we recall the basics of episodic Markov Decision Processes and introduce the notion of differential privacy in reinforcement learning.

2.1 Learning model and regret in episodic MDPs

We consider episodic reinforcement learning (RL) in a finite horizon stochastic Markov decision process (MDP) (Puterman 1994; Sutton 1988) given by a tuple $(\mathcal{S}, \mathcal{A}, H, (P_h)_{h=1}^H, (c_h)_{h=1}^H)$, where $\mathcal{S}$ and $\mathcal{A}$ are state and action spaces with cardinalities S and A, respectively, $H \in \mathbb{N}$ is the episode length, $P_h(s'|s, a)$ is the probability of transitioning to state s' from state s provided action a is taken at step h and $c_h(s, a)$ is the mean of the cost distribution at step h supported on $[0, 1]$. The actions are chosen following some policy $\pi = (\pi_h)_{h=1}^H$, where each π_h is a mapping from the state space $\mathcal{S}$ into a probability distribution over the action space $\mathcal{A}$, i.e. $\pi_h(a|s) \geq 0$ and $\sum_{a\in\mathcal{A}} \pi_h(a|s) = 1$ for

all $s \in \mathcal{S}$. The agent would like to find a policy π that minimizes the long term expected cost starting from every state $s \in \mathcal{S}$ and every step $h \in [H]$, defined as

$$V_h^\pi(s) := \mathbb{E}\left[\sum\nolimits_{h'=h}^{H} c_{h'}(s_{h'}, a'_h) \mid s_h = s, \pi\right],$$

where the expectation is with respect to the randomness of the transition kernel and the policy. We call V_h^π the value function of policy π at step h. Now, defining the Q-function of policy π at step h as

$$Q_h^\pi(s,a) := \mathbb{E}\left[\sum\nolimits_{h'=h}^{H} c_{h'}(s_{h'}, a'_h)) \mid s_h = s, a_h = a, \pi\right],$$

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

A policy π^* is said to be optimal if it minimizes the value for all states s and step h simultaneously, and the corresponding optimal value function is denoted by $V_h^*(s) = \min_{\pi \in \Pi} V_h^\pi(s)$ for all $h \in [H]$, where Π is the set of all non-stationary policies. The agent interacts with the environment for K episodes to learn the unknown transition probabilities $P_h(s'|s,a)$ and mean costs $c_h(s,a)$, and thus, in turn, the optimal policy π^*. At each episode k, the agent chooses a policy $\pi^k = (\pi_h^k)_{h=1}^H$ and samples a trajectory $\{s_1^k, a_1^k, c_1^k, \ldots, s_H^k, a_H^k, c_H^k, s_{H+1}^k\}$ by interacting with the MDP using this policy. Here, at a given step h, s_h^k denotes the state of the MDP, $a_h^k \sim \pi_h^k(\cdot|s_h^k)$ denotes the action taken by the agent, $c_h^k \in [0,1]$ denotes the (random) cost suffered by the agent with the mean value $c_h(s_h^k, a_h^k)$ and $s_{h+1}^k \sim P_h(\cdot|s_h^k, a_h^k)$ denotes the next state. The initial state s_1^k is assumed to be fixed and history independent. We measure performance of the agent by the cumulative (pseudo) regret accumulated over K episodes, defined as

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

where $T = KH$ denotes the total number of steps. We seek algorithms with regret that is sublinear in T, which demonstrates the agent's ability to act near optimally.

2.2 Differential privacy in episodic RL

In the episodic RL setting described above, it is natural to view each episode $k \in [K]$ as a trajectory associated to a specific user. To this end, we let $U_K = (u_1, \ldots, u_K) \in \mathcal{U}^K$ to denote a sequence of K unique[3] users participating in the private RL protocol with an RL agent $\mathcal{M}$, where $\mathcal{U}$ is the set of all users. Each user u_k is identified by the cost and state responses $(c_h^k, s_{h+1}^k)_{h=1}^H$ she gives to the actions $(a_h^k)_{h=1}^H$ chosen by the agent. We let $\mathcal{M}(U_k) = (a_1^1, \ldots, a_H^K) \in \mathcal{A}^{KH}$ to denote the set of all actions chosen by the agent $\mathcal{M}$ when interacting with the user sequence U_k. Informally, we will be interested in (centralized) randomized mechanisms (in this case, RL agents) $\mathcal{M}$ so that the knowledge of the output $\mathcal{M}(U_k)$ and all but the k-th user u_k does not reveal 'much' information about u_k. We formalize this in the following definition.

[3]Uniqueness is assumed wlog, as for a returning user one can group her with her previous occurrences.

Definition 1 (Differential Privacy (DP)). *For any $\varepsilon \geq 0$ and $\delta \in [0,1]$, a mechanism $\mathcal{M} : \mathcal{U}^K \to \mathcal{A}^{KH}$ is (ε, δ)-differentially private if for all $U_K, U'_K \in \mathcal{U}^K$ differing on a single user and for all subset of actions $\mathcal{A}_0 \subset \mathcal{A}^{KH}$,*

$$\mathbb{P}\left[\mathcal{M}(U_K) \in \mathcal{A}_0\right] \leq \exp(\varepsilon)\mathbb{P}\left[\mathcal{M}(U'_K) \in \mathcal{A}_0\right] + \delta\,.$$

If $\delta = 0$, we call the mechanism $\mathcal{M}$ to be ε-differentially private (ε-DP).

This is a direct adaptation of the classic notion of differential privacy (Dwork, Roth et al. 2014). However, we need to relax this definition for our purpose, because although the actions recommended to the user u_k have only a small effect on the types (i.e., state and cost responses) of other users participating in the RL protocol, those can reveal a lot of information about the type of the user u_k. Thus, it becomes hard to privately recommend the actions to user u_k while protecting the privacy of it's type, i.e., it's state and cost responses to the suggested actions. Hence, to preserve the privacy of individual users, we consider the notion of *joint differential privacy* (JDP) (Kearns et al. 2014), which requires that simultaneously for all user u_k, the joint distribution of the actions recommended to all users other than u_k be differentially private in the type of the user u_k. It weakens the constraint of DP only in that the actions suggested specifically to u_k may be sensitive in her type (state and cost responses). However, JDP is still a very strong definition since it protects u_k from any arbitrary collusion of other users against her, so long as she does not herself make the actions suggested to her public. To this end, we let $\mathcal{M}_{-k}(U_k) := \mathcal{M}(U_k) \setminus (a_h^k)_{h=1}^H$ to denote all the actions chosen by the agent $\mathcal{M}$ excluding those recommended to u_k and formally define JDP as follows.

Definition 2 (Joint Differential Privacy (JDP)). *For any $\varepsilon \geq 0$, a mechanism $\mathcal{M} : \mathcal{U}^K \to \mathcal{A}^{KH}$ is ε-joint differentially private if for all $k \in [K]$, for all user sequences $U_K, U'_K \in \mathcal{U}^K$ differing only on the k-th user and for all set of actions $\mathcal{A}_{-k} \subset \mathcal{A}^{(K-1)H}$ given to all but the k-th user,*

$$\mathbb{P}\left[\mathcal{M}_{-k}(U_K) \in \mathcal{A}_{-k}\right] \leq \exp(\varepsilon)\mathbb{P}\left[\mathcal{M}_{-k}(U'_K) \in \mathcal{A}_{-k}\right].$$

JDP has been used extensively in private mechanism design (Kearns et al. 2014), in private matching and allocation problems (Hsu et al. 2016), in designing privacy-preserving algorithms for linear contextual bandits (Shariff and Sheffet 2018), and it has been introduced in private tabular RL by Vietri et al. (2020).

JDP allows the agent to observe the data (i.e., the entire trajectory of state-action-cost sequence) associated with each user and the privacy burden lies on the agent itself. In some scenarios, however, the users may not even be willing to share it's data with the agent directly. This motivates a stringer notion of privacy protection, called the *local differential privacy* (LDP) (Duchi, Jordan, and Wainwright 2013). In this setting, each user is assumed to have her own privacy mechanism that can do randomized mapping on its data to guarantee privacy. To this end, we denote by X a trajectory $(s_h, a_h, c_h, s_{h+1})_{h=1}^H$ and by $\mathcal{X}$ the set of all possible trajectories. We write $\mathcal{M}'(X)$ to denote the privatized trajectory generated by a (local) randomized mechanism $\mathcal{M}'$. With this notation, we now formally define LDP for our RL protocol.

Algorithm 1: PRIVATE-UCB-PO

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

Initialize policy $\pi_h^1(a|s) = 1/A$ for all (s, a, h)

Initialize private counts $\widetilde{C}_h^1(s, a) = 0$, $\widetilde{N}_h^1(s, a) = 0$ and $\widetilde{N}_h^1(s, a, s') = 0$ for all (s, a, s', h)

Set precision levels $E_{\varepsilon,\delta,1}, E_{\varepsilon,\delta,2}$ of the PRIVATIZER

for $k = 1, 2, 3, \ldots, K$ **do**

  Initialize private value estimates: $\widetilde{V}_{H+1}^k(s) = 0$

  for $h = H, H-1, \ldots, 1$ **do**

    Compute $\widetilde{c}_h^k(s, a)$ and $\widetilde{P}_h^k(s, a)$ $\forall(s, a)$ as in (1) using the private counts

    Set exploration bonus using Lemma 1: $\beta_h^k(s, a) = \beta_h^{k,c}(s, a) + H\beta_h^{k,p}(s, a)$ $\forall(s, a)$

    Compute: $\forall(s, a)$, $\widetilde{Q}_h^k(s, a) = \min\{H - h + 1, \max\{0, \widetilde{c}_h^k(s, a) + \sum_{s' \in \mathcal{S}} \widetilde{V}_{h+1}^k(s')\widetilde{P}_h^k(s'|s, a) - \beta_h^k(s, a)\}\}$

    Compute private value estimates: $\forall s$, $\widetilde{V}_h^k(s) = \sum_{a \in \mathcal{A}} \widetilde{Q}_h^k(s, a)\pi_h^k(a|s)$

  Roll out a trajectory $(s_1^k, a_1^k, c_1^k, \ldots, s_{H+1}^k)$ by acting the policy $\pi^k = (\pi_h^k)_{h=1}^H$

  Receive private counts $\widetilde{C}_h^{k+1}(s, a)$, $\widetilde{N}_h^{k+1}(s, a)$, $\widetilde{N}_h^{k+1}(s, a, s')$ from the PRIVATIZER

  Update policy: $\forall(s, a, h)$, $\pi_h^{k+1}(a|s) = \frac{\pi_h^k(a|s)\exp(-\eta\widetilde{Q}_h^k(s,a))}{\sum_{a \in \mathcal{A}} \pi_h^k(a|s)\exp(-\eta\widetilde{Q}_h^k(s,a))}$

Definition 3 (Local Differential Privacy (LDP))**.** *For any $\varepsilon \geq 0$, a mechanism $\mathcal{M}'$ is ε-local differentially private if for all trajectories $X, X' \in \mathcal{X}$ and for all possible subsets $\mathcal{E}_0 \subset \{\mathcal{M}'(X)|X \in \mathcal{X}\}$,*

$$\mathbb{P}\left[\mathcal{M}'(X) \in \mathcal{E}_0\right] \leq \exp(\varepsilon)\mathbb{P}\left[\mathcal{M}'(X') \in \mathcal{E}_0\right].$$

LDP ensures that if any adversary (can be the RL agent itself) observes the output of the privacy mechanism $\mathcal{M}'$ for two different trajectories, then it is statistically difficult for it to guess which output is from which trajectory. This has been used extensively in multi-armed bandits (Zheng et al. 2020; Ren et al. 2020), and introduced in private tabular RL by Garcelon et al. (2020).

3 Private Policy Optimization

In this section, we introduce a policy-optimization based private RL algorithm PRIVATE-UCB-PO (Algorithm 1) that can be instantiated with any private mechanism (henceforth, referred as a PRIVATIZER) satisfying a general condition. We derive a generic regret bound for PRIVATE-UCB-PO, which can be applied to obtain bounds under JDP and LDP requirements by instantiating PRIVATE-UCB-PO with a CENTRAL-PRIVATIZER and a LOCAL-PRIVATIZER, respectively. All the proofs are deferred to the appendix.

Let us first introduce some notations. We denote by $N_h^k(s, a) := \sum_{k'=1}^{k-1} \mathbb{I}\{s_h^{k'} = s, a_h^{k'} = a\}$, the number of times that the agent has visited state-action pair (s, a) at step h *before* episode k. Similarly, $N_h^k(s, a, s') := \sum_{k'=1}^{k-1} \mathbb{I}\{s_h^{k'} = s, a_h^{k'} = a, s_{h+1}^{k'} = s'\}$ denotes the count of going to state s' from s upon playing action a at step h *before* episode k. Finally, $C_h^k(s, a) := \sum_{k'=1}^{k-1} \mathbb{I}\{s_h^{k'} = s, a_h^{k'} = a\}c_h^{k'}$ denotes the total cost suffered by taking action a on state s and step h *before* episode k. In non-private learning, these counters are sufficient to find estimates of the transition kernels $(P_h)_h$ and mean cost functions $(c_h)_h$ to design the policy $(\pi_h^k)_h$ for episode k. However, in private learning, the challenge is that the counters depend on users' state and cost responses to suggested actions, which is considered sensitive information. Therefore, the PRIVATIZER must release the counts in a privacy-preserving way on which the learning agent would rely. To this end, we let $\widetilde{N}_h^k(s, a)$, $\widetilde{C}_h^k(s, a)$, and $\widetilde{N}_h^k(s, a, s')$ to denote the privatized versions of $N_h^k(s, a)$, $C_h^k(s, a)$, and $N_h^k(s, a, s')$, respectively. Now, we make a general assumption on the counts released by the PRIVATIZER (both LOCAL and CENTRAL), which roughly means that with high probability the private counts are close to the actual ones.

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

In the following, we assume Assumption 1 holds. Then, we define, for all (s, a, h, k), the *private* mean empirical costs and *private* empirical transition probabilities as

$$\widetilde{c}_h^k(s, a) := \frac{\widetilde{C}_h^k(s, a)}{\max\{1, \widetilde{N}_h^k(s, a) + E_{\varepsilon,\delta,1}\}}$$
$$\widetilde{P}_h^k(s'|s, a) := \frac{\widetilde{N}_h^k(s, a, s')}{\max\{1, \widetilde{N}_h^k(s, a) + E_{\varepsilon,\delta,1}\}}. \quad (1)$$

The following concentration bounds on the private estimates will be the key to our algorithm design.

Lemma 1 (Concentration of private estimates)**.** *Fix any $\varepsilon > 0$ and $\delta \in (0, 1]$. Then, under Assumption 1, with probability at least $1 - 2\delta$, uniformly over all (s, a, h, k),*

$$|c_h(s, a) - \widetilde{c}_h^k(s, a)| \leq \beta_h^{k,c}(s, a), \quad \textit{and}$$
$$\left\|P_h(\cdot|s, a) - \widetilde{P}_h^k(\cdot|s, a)\right\|_1 \leq \beta_h^{k,p}(s, a),$$

where $\beta_h^{k,c}(s, a) := \frac{L_c(\delta)}{\sqrt{\max\{1, \widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1}\}}} + \frac{3E_{\varepsilon,\delta,1}}{\max\{1, \widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1}\}}$, $L_c(\delta) := \sqrt{2\ln\frac{4SAT}{\delta}}$, $\beta_h^{k,p}(s, a) := \frac{L_p(\delta)}{\sqrt{\max\{1, \widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1}\}}} + \frac{SE_{\varepsilon,\delta,2}+2E_{\varepsilon,\delta,1}}{\max\{1, \widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1}\}}$, *and* $L_p(\delta) := \sqrt{4S\ln\frac{6SAT}{\delta}}$.

PRIVATE-UCB-PO algorithm. PRIVATE-UCB-PO (Algorithm 1) is a private policy optimization (PO) algorithm based on the celebrated upper confidence bound (UCB) philosophy (Auer, Cesa-Bianchi, and Fischer 2002; Jaksch, Ortner, and Auer 2010). Similar to the non-private setting (Efroni et al. 2020), it basically has two stages at each episode k: *policy evaluation* and *policy improvement*. In the policy evaluation stage, it evaluates the policy π^k based on $k-1$ historical trajectories. In contrast to the non-private case, PRIVATE-UCB-PO relies only on the private counts (returned by the PRIVATIZER) to calculate the private mean empirical costs and private empirical transitions. These two along with a UCB exploration bonus term (which also depends only on private counts) are used to compute Q-function estimates. The Q-estimates are then truncated and corresponding value estimates are computed by taking their expectation with respect to the policy. Next, a new trajectory is rolled out by acting the policy π^k and the PRIVATIZER translates all non-private counts into the private ones to be used for the policy evaluation in the next episode. Finally, in the policy improvement stage, PRIVATE-UCB-PO employs a 'soft' update of the current policy π^k by following a standard mirror-descent step together with a Kullback–Leibler (KL) divergence proximity term (Beck and Teboulle 2003; Cai et al. 2020; Efroni et al. 2020). The following theorem presents a general regret bound of PRIVATE-UCB-PO (Algorithm 1) when instantiated with any PRIVATIZER (LOCAL or CENTRAL) that satisfies Assumption 1.

Theorem 1 (Regret bound of PRIVATE-UCB-PO)**.** *Fix any $\varepsilon > 0$ and $\delta \in (0,1]$ and set $\eta = \sqrt{2\log A/(H^2K)}$. Then, under Assumption 1, with probability at least $1-\delta$, the cumulative regret of* PRIVATE-UCB-PO *is*

$$\mathcal{R}(T) = \widetilde{O}\left(\sqrt{S^2AH^3T} + \sqrt{S^3A^2H^4}\right) + \widetilde{O}\left(E_{\varepsilon,\delta,2}S^2AH^2 + E_{\varepsilon,\delta,1}SAH^2\right).$$

Remark 1 (Cost of privacy)**.** *Theorem 1 shows that regret of* PRIVATE-UCB-PO *is lower bounded by the regret in non-private setting (Efroni et al. 2020, Theorem 1), and depends directly on the privacy parameter ε through the permitted precision levels $E_{\varepsilon,\delta,1}$ and $E_{\varepsilon,\delta,2}$ of the* PRIVATIZER*. Thus, choosing $E_{\varepsilon,\delta,1}, E_{\varepsilon,\delta,2}$ appropriately to guarantee JDP or LDP, we can obtain regret bounds under both forms of privacy. The cost of privacy, as we shall see in Section 5, is lower order than the non-private regret under JDP, and is of the same order under the stringer requirement of LDP.*[4]

Proof sketch. We first decompose the regret as sum of three terms: $\mathcal{T}_1 = \sum_k (V_1^{\pi^k}(s_1^k) - \widetilde{V}_1^k(s_1^k))$, $\mathcal{T}_2 = \sum_{k,h} \mathbb{E}\left[\langle \widetilde{Q}_h^k(s_h,\cdot), \pi_h^k(\cdot|s_h) - \pi_h^*(\cdot|s_h)\rangle | s_1^k, \pi^*\right]$ and $\mathcal{T}_3 = \sum_{k,h} \mathbb{E}\left[\widetilde{Q}_h^k(s_h,a_h) - c_h(s_h,a_h) - P_h(\cdot|s_h,a_h)\widetilde{V}_{h+1}^k | s_1^k, \pi^*\right]$. We then bound each of the three terms. First, by setting $\eta = \sqrt{2\log A/(H^2K)}$, we show that $\mathcal{T}_2 \le \sqrt{2H^4K\log A}$ via a standard online mirror descent analysis. Then, from Lemma 1 and our choice of bonus terms, we get $\mathcal{T}_3 \le 0$. Next, we bound $\mathcal{T}_1$ by the sum of expected bonus terms, i.e., $\mathcal{T}_1 \le \sum_{k,h} \mathbb{E}\left[2\beta_h^{k,c}(s_h,a_h) + 2H\beta_h^{k,p}(s_h,a_h) | s_1^k, \pi^k\right]$. Now, by Assumption 1, the expected bonuses can be controlled using $\mathbb{E}\left[\frac{L_c(\delta)}{\sqrt{\max\{N_h^k(s_h,a_h),1\}}} + \frac{3E_{\varepsilon,\delta,1}}{\max\{N_h^k(s_h,a_h),1\}} | s_1^k, \pi^k\right]$, and $\mathbb{E}\left[\frac{L_p(\delta)}{\sqrt{\max\{N_h^k(s_h,a_h),1\}}} + \frac{SE_{\varepsilon,\delta,2}+2E_{\varepsilon,\delta,1}}{\max\{N_h^k(s_h,a_h),1\}} | s_1^k, \pi^k\right]$, respectively. We can now complete the proof by showing that[5] $\sum_{k,h} \mathbb{E}\left[\frac{1}{\max\{1,N_h^k(s_h,a_h)\}}\right] \approx \widetilde{O}(SAH)$ and $\sum_{k,h} \mathbb{E}\left[\frac{1}{\sqrt{\max\{1,N_h^k(s_h,a_h)\}}}\right] \approx \widetilde{O}\left(\sqrt{SAHT} + SAH\right)$.

[4] The lower order terms scale with S^2, which is quite common for optimistic tabular RL algorithms (Azar, Osband, and Munos 2017; Dann, Lattimore, and Brunskill 2017).

4 Private UCB-VI Revisited

In this section, we turn to investigate value-iteration based private RL algorithms. It is worth noting that private valued-based RL algorithms have been studied under both JDP and LDP requirements (Vietri et al. 2020; Garcelon et al. 2020). However, to the best of our understanding, the regret analysis of the JDP algorithm presented in Vietri et al. (2020) has gaps and does not support the claimed result.[6] Under LDP, the regret bound presented in Garcelon et al. (2020) is sub-optimal in the cardinality of the state space and as the authors have remarked, it is possible to achieve the optimal scaling using a refined analysis. Motivated by this, we revisit private value iteration by designing an optimistic algorithm PRIVATE-UCB-VI (Algorithm 2) that can be instantiated with a PRIVATIZER (CENTRAL and LOCAL) to achieve both JDP and LDP.

PRIVATE-UCB-VI algorithm. Our algorithm design principle is again based on the UCB philosophy, the private estimates defined in (1) and a value-aware concentration result for the estimates stated in Lemma 2 below. Similar to the non-private setting (Azar, Osband, and Munos 2017), PRIVATE-UCB-VI (Algorithm 2) follows the procedure of optimistic value iteration. Specifically, at each episode k, using the private counts and a private UCB bonus term, it first compute private Q-estimates and value estimates using optimistic Bellman recursion. Next, a greedy policy π^k is obtained directly from the estimated Q-function. Finally, a trajectory is rolled out by acting the policy π^k and then PRIVATIZER translates all non-private statistics into private ones

[5] These are generalization of results proved under stationary transition model (Efroni et al. 2019; Zanette and Brunskill 2019) to our non-stationary setting (similar results appear in Efroni, Mannor, and Pirotta (2020); Efroni et al. (2020), but without proofs).

[6] The gap lies in Vietri et al. (2020, Lemma 18) in which the private estimates were incorrectly used as the true cost and transition functions. This lead to a simpler but incorrect regret decomposition since it omits the 'error' term between the private estimates and true values. Moreover, the error term cannot be simply upper bounded by its current bonus term ($\widetilde{\text{conf}}_t$ in Vietri et al. (2020, Algorithm 3)) since one cannot directly use Hoeffding's inequality due to the fact that the value function is not fixed in this term (please refer to Appendix for more detailed discussions).

to be used in the next episode.

Lemma 2 (Refined concentration of private estimates)**.** *Fix any $\varepsilon > 0$ and $\delta \in (0,1)$. Then, under Assumption 1, with probability at least $1-3\delta$, uniformly over all (s,a,s',h,k),*

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

$$|P_h(s'|s,a) - \widetilde{P}_h^k(s'|s,a)| \le C\sqrt{\frac{L'(\delta)P_h(s'|s,a)}{\max\{1,\widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1}\}}} + \frac{CL'(\delta)+2E_{\varepsilon,\delta,1}+E_{\varepsilon,\delta,2}}{\max\{1,\widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1}\}},$$

where $\beta_h^{k,c}(s,a)$ and $L_c(\delta)$ is as defined in Lemma 1, $(PV_{h+1})(s,a) := \sum_{s'} P(s'|s,a)V_{h+1}(s')$, $\beta_h^{k,pv}(s,a) := \frac{HL_c(\delta)}{\sqrt{\max\{1,\widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1}\}}} + \frac{H(SE_{\varepsilon,\delta,2}+2E_{\varepsilon,\delta,1})}{\max\{1,\widetilde{N}_h^k(s,a)+E_{\varepsilon,\delta,1}\}}$, $C > 0$ is some constant, and $L'(\delta) := \log\left(\frac{6SAT}{\delta}\right)$.

The bonus term $\beta_h^{k,pv}$ in PRIVATE-UCB-VI does not have the factor $\sqrt{S}$ in the leading term compared to $\beta_h^{k,p}$ in PRIVATE-UCB-PO. This is achieved by following a similar idea in UCB-VI (Azar, Osband, and Munos 2017). That is, instead of bounding the transition dynamics as in Lemma 1, we maintain a confidence bound directly over the optimal value function (the second result in Lemma 2). Due to this, we have an extra term in the regret bound, which can be carefully bounded by using a Bernstein-type inequality (the third result in Lemma 2). These two steps enable us to obtain an improved dependence on S in the regret bound compared to existing private value-based algorithms (Vietri et al. 2020; Garcelon et al. 2020) under both JDP and LDP. This is stated formally in the next theorem, which presents a general regret bound of PRIVATE-UCB-VI (Algorithm 2) when instantiated with any PRIVATIZER (LOCAL or CENTRAL).

Theorem 2 (Regret bound for PRIVATE-UCB-VI)**.** *Fix any $\varepsilon > 0$ and $\delta \in (0,1]$. Then, under Assumption 1, with probability $\ge 1-\delta$, the regret of* PRIVATE-UCB-VI *is*

$$\mathcal{R}(T) = \widetilde{O}\left(\sqrt{SAH^3T} + S^2AH^3\right) + \widetilde{O}\left(S^2AH^2E_{\varepsilon,\delta,1} + S^2AH^2E_{\varepsilon,\delta,2}\right).$$

Remark 2 (Cost of privacy)**.** *Similar to* PRIVATE-UCB-PO*, the regret of* PRIVATE-UCB-VI *is lower bounded by the regret in non-private setting (see Azar, Osband, and Munos (2017, Theorem 1)),*[7] *and the privacy parameter appear only in the lower order terms.*

Remark 3 (VI vs. PO)**.** *The regret bound of* PRIVATE-UCB-VI *is a $\sqrt{S}$ factor better in the leading privacy-independent term compared to* PRIVATE-UCB-PO*. This follows the same pattern as in the non-private case, i.e.,* UCB-VI *(Azar, Osband, and Munos 2017) vs.* OPPO *(Efroni et al. 2020).*

[7] In the non-private setting, Azar, Osband, and Munos (2017) assume stationary transition kernels $P_h = P$ for all h. We consider non-stationary kernels, which adds a multiplicative $\sqrt{H}$ factor in our non-private regret.

Algorithm 2: PRIVATE-UCB-VI

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

Initialize private counts $\widetilde{C}_h^1(s,a) = 0$, $\widetilde{N}_h^1(s,a) = 0$ and $\widetilde{N}_h^1(s,a,s') = 0$ for all (s,a,s',h)

Set precision levels $E_{\varepsilon,\delta,1}, E_{\varepsilon,\delta,2}$ of the PRIVATIZER

for $k = 1,\dots,K$ **do**

  Initialize private value estimates: $\widetilde{V}_{H+1}^k(s) = 0$

  for $h = H, H-1, \dots, 1$ **do**

    Compute $\widetilde{c}_h^k(s,a)$ and $\widetilde{P}_h^k(s'|s,a)\ \forall(s,a,s')$ as in (1) using the private counts

    Set exploration bonus using Lemma 2: $\beta_h^k(s,a) = \beta_h^{k,c}(s,a) + \beta_h^{k,pv}(s,a)\ \forall(s,a)$

    Compute: $\forall(s,a)$, $\widetilde{Q}_h^k(s,a) = \min\{H-h+1, \max\{0, \widetilde{c}_h^k(s,a) + \sum_{s'\in\mathcal{S}} \widetilde{V}_{h+1}^k(s')\widetilde{P}_h^k(s'|s,a) - \beta_h^k(s,a)\}\}$

    Compute private value function: $\forall s$, $\widetilde{V}_h^k(s) = \min_{a\in\mathcal{A}} \widetilde{Q}_h^k(s,a)$

  Compute policy: $\forall(s,h)$, $\pi_h^k(s) = \operatorname{argmin}_{a\in\mathcal{A}} \widetilde{Q}_h^k(s,a)$ (with breaking ties arbitrarily)

  Roll out a trajectory $(s_1^k, a_1^k, c_1^k, \dots, s_{H+1}^k)$ by acting the policy $\pi^k = (\pi_h^k)_{h=1}^H$

  Receive private counts $\widetilde{C}_h^{k+1}(s,a)$, $\widetilde{N}_h^{k+1}(s,a)$, $\widetilde{N}_h^{k+1}(s,a,s')$ from the PRIVATIZER

5 Privacy and regret guarantees

In this section, we instantiate PRIVATE-UCB-PO and PRIVATE-UCB-VI using a CENTRAL-PRIVATIZER and a LOCAL-PRIVATIZER, and derive corresponding privacy and regret guarantees.

5.1 Achieving JDP using CENTRAL-PRIVATIZER

The CENTRAL-PRIVATIZER runs a private K-bounded binary-tree mechanism (counter) (Chan, Shi, and Song 2010) for each count $N_h^k(s,a), C_h^k(s,a), N_h^k(s,a,s')$, i.e. it uses $2SAH + S^2AH$ counters in total. Let us focus on the counters – there are SAH many of thems – for the number of visited states $N_h^k(s,a)$. Each counter takes as input the data stream $\sigma_h(s,a) \in \{0,1\}^K$, where the j-th bit $\sigma_h^j(s,a) := \mathbb{I}\{s_h^j = s, a_h^j = a\}$ denotes whether the pair (s,a) is encountered or not at step h of episode j, and at the start of each episode k, release a private version $\widetilde{N}_h^k(s,a)$ of the count $N_h^k(s,a) := \sum_{j=1}^{k-1} \sigma_h^j(s,a)$. Let us now discuss how private counts are computed. To this end, we let $N_h^{i,j}(s,a) = \sum_{k=i}^j \sigma_h^k(s,a)$ to denote a partial sum (P-sum) of the counts in episodes i through j, and consider a binary interval tree, each leaf node of which represents an episode (i.e., the tree has $k-1$ leaf nodes at the start of episode k). Each interior node of the tree represents the range of episodes covered by its children. At the start of

episode k, first a noisy P-sum corresponding to each node in the tree is released by perturbing it with an independent Laplace noise Lap($\frac{1}{\varepsilon'}$), where $\varepsilon' > 0$ is a given privacy parameter.[8] Then, the private count $\widetilde{N}_h^k(s,a)$ is computed by summing up the noisy P-sums released by the set of nodes – which has cardinality at most $O(\log k)$ – that uniquely cover the range $[1, k-1]$. Observe that, at the end of episode k, the counter only needs to store noisy P-sums required for computing private counts at future episodes, and can safely discard P-sums those are no longer needed.

The counters corresponding to empirical rewards $C_h^k(s,a)$ and state transitions $N_h^k(s,a,s')$ follow the same underlying principle to release the respective private counts $\widetilde{C}_h^k(s,a)$ and $\widetilde{N}_h^k(s,a,s')$. The next lemma sums up the properties of the CENTRAL-PRIVATIZER.

Lemma 3 (Properties of CENTRAL-PRIVATIZER)**.** *For any $\varepsilon > 0$, CENTRAL-PRIVATIZER with parameter $\varepsilon' = \frac{3H \log K}{\varepsilon}$ is ε-DP. Furthermore, for any $\delta \in (0,1]$, it satisfies Assumption 1 with $E_{\varepsilon,\delta,1} = \frac{3H}{\varepsilon}\sqrt{8 \log^3 K \log(6SAT/\delta)}$, and $E_{\varepsilon,\delta,2} = \frac{3H}{\varepsilon}\sqrt{8 \log^3 K \log(6S^2AT/\delta)}$.*

Lemma 3 follows from the privacy guarantee of the Laplace mechanism, and the concentration bound on the sum of i.i.d. Laplace random variables (Dwork, Roth et al. 2014). Using Lemma 3, as corollaries of Theorem 1 and Theorem 2, we obtain the regret and privacy guarantees for PRIVATE-UCB-PO and PRIVATE-UCB-VI with the CENTRAL-PRIVATIZER.

Corollary 1 (Regret under JDP)**.** *For any $\varepsilon > 0$ and $\delta \in (0,1]$, both PRIVATE-UCB-PO and PRIVATE-UCB-VI, if instantiated using CENTRAL-PRIVATIZER with parameter $\varepsilon' = \frac{3H \log K}{\varepsilon}$, satisfy ε-JDP. Furthermore, with probability at least $1-\delta$, we obtain the regret bounds:*

$$\mathcal{R}^{\text{PRIVATE-UCB-PO}}(T) = \widetilde{O}\left(\sqrt{S^2AH^3T} + S^2AH^3/\varepsilon\right),$$

$$\mathcal{R}^{\text{PRIVATE-UCB-VI}}(T) = \widetilde{O}\left(\sqrt{SAH^3T} + S^2AH^3/\varepsilon\right).$$

We prove the JDP guarantees using the *billboard model* (Hsu et al. 2016, Lemma 9) which, informally, states that an algorithm is JDP if the output sent to each user is a function of the user's private data and a common quantity computed using a standard DP mechanism. Note that by Lemma 3 and the post-processing property of DP (Dwork, Roth et al. 2014), the sequence of policies $(\pi^k)_k$ are ε-DP. Therefore, by the billboard model, the actions $(a_h^k)_{h,k}$ suggested to all the users are ε-JDP.

Remark 4. *Corollary 1, to the best of our understanding, provides the first regret bound for private PO, and a correct regret bound for private VI as compared to Vietri et al. (2020), under the requirement of JDP.*

5.2 Achieving LDP using LOCAL-PRIVATIZER

The LOCAL-PRIVATIZER, at each episode k, release the private counts by injecting Laplace noise into the aggregated statistics computed from the trajectory generated in that episode. Let us discuss how private counts for the number of visited states are computed. At each episode j, given privacy parameter $\varepsilon' > 0$, LOCAL-PRIVATIZER perturbs $\sigma_h^j(s,a)$ with an independent Laplace noise Lap($\frac{1}{\varepsilon'}$), i.e. it makes SAH noisy perturbations in total. The private counts for the k-th episode are computed as $\widetilde{N}_h^k(s,a) = \sum_{j=1}^{k-1} \widetilde{\sigma}_h^j(s,a)$, where $\widetilde{\sigma}_h^j(s,a)$ denotes the noisy perturbations. The private counts corresponding to empirical rewards $C_h^k(s,a)$ and state transitions $N_h^k(s,a,s')$ are computed similarly. The next lemma sums up the properties of the LOCAL-PRIVATIZER.

Lemma 4 (Properties of LOCAL-PRIVATIZER)**.** *For any $\varepsilon > 0$, LOCAL-PRIVATIZER with parameter $\varepsilon' = \frac{3H}{\varepsilon}$ is ε-LDP. Furthermore, for any $\delta \in (0,1]$, it satisfies Assumption 1 with $E_{\varepsilon,\delta,1} = \frac{3H}{\varepsilon}\sqrt{8K \log(6SAT/\delta)}$ and $E_{\varepsilon,\delta,2} = \frac{3H}{\varepsilon}\sqrt{8K \log(6S^2AT/\delta)}$.*

Corollary 2 (Regret under LDP)**.** *For any $\varepsilon > 0$ and $\delta \in (0,1]$, instantiating PRIVATE-UCB-PO and PRIVATE-UCB-VI using LOCAL-PRIVATIZER with parameter $\varepsilon' = \frac{3H}{\varepsilon}$, we obtain, with probability $\geq 1-\delta$, the regret bounds:*

$$\mathcal{R}^{\text{PRIVATE-UCB-PO}}(T) = \widetilde{O}\left(\sqrt{S^2AH^3T} + S^2A\sqrt{H^5T}/\varepsilon\right),$$

$$\mathcal{R}^{\text{PRIVATE-UCB-VI}}(T) = O\left(\sqrt{SAH^3T} + S^2A\sqrt{H^5T}/\varepsilon\right).$$

Remark 5. *Corollary 2, to the best of our knowledge, provides the first regret guarantee for private PO, and an improved regret bound for private VI as compared to Garcelon et al. (2020), under the requirement of LDP.*

Remark 6 (JDP vs. LDP)**.** *The noise level in the private counts is $O(\log k)$ under JDP and $O(k)$ under LDP. Due to this, the privacy cost for LDP is $\tilde{O}(\sqrt{T}/\varepsilon)$, whereas for JDP it is only $\tilde{O}(1/\varepsilon)$.*

Remark 7 (Alternative LDP mechanisms)**.** *Other than the Laplace noise, one can also use Bernoulli and Gaussian noise in the LOCAL-PRIVATIZER to achieve LDP (Kairouz, Bonawitz, and Ramage 2016; Wang et al. 2019b). Thanks to Theorem 1 and Theorem 2, the regret bounds are readily obtained by plugging in the corresponding $E_{\varepsilon,\delta,1}$ and $E_{\varepsilon,\delta,2}$.*

6 Conclusions

In this work, we presented the first private policy-optimization algorithm in tabular MDPs with regret guarantees under both JDP and LDP requirements. We also revisited private value-iteration algorithms by improving the regret bounds of existing results. These are achieved by developing a general framework for algorithm design and regret analysis in private tabular RL settings. Though we focus on statistical guarantees of private RL algorithms, it will be helpful to understand these from a practitioner's perspective. We leave this as a possible future direction. Another important direction is to apply our general framework to MDPs with function approximation, e.g., linear MDPs (Jin et al. 2019), kernelized MDPs (Chowdhury and Gopalan 2019) and generic MDPs (Ayoub et al. 2020).

[8] A random variable $X \sim \text{Lap}(b)$, with scale parameter $b > 0$, if $\forall x \in \mathbb{R}$, it's p.d.f. is given by $f_X(x) = \frac{1}{2b}\exp(-|x|/b)$.

References

Agarwal, N.; and Singh, K. 2017. The price of differential privacy for online learning. In *International Conference on Machine Learning*, 32–40. PMLR.

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3): 235–256.

Ayoub, A.; Jia, Z.; Szepesvari, C.; Wang, M.; and Yang, L. F. 2020. Model-Based Reinforcement Learning with Value-Targeted Regression. *arXiv preprint arXiv:2006.01107*.

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

Balle, B.; Gomrokchi, M.; and Precup, D. 2016. Differentially private policy evaluation. In *International Conference on Machine Learning*, 2130–2138. PMLR.

Beck, A.; and Teboulle, M. 2003. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3): 167–175.

Cai, Q.; Yang, Z.; Jin, C.; and Wang, Z. 2020. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, 1283–1294. PMLR.

Chan, T. H.; Shi, E.; and Song, D. 2010. Private and continual release of statistics. In *International Colloquium on Automata, Languages, and Programming*, 405–417. Springer.

Chen, X.; Zheng, K.; Zhou, Z.; Yang, Y.; Chen, W.; and Wang, L. 2020. (Locally) Differentially Private Combinatorial Semi-Bandits. In *International Conference on Machine Learning*, 1757–1767. PMLR.

Chowdhury, S. R.; and Gopalan, A. 2019. Online learning in kernelized markov decision processes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 3197–3205. PMLR.

Chowdhury, S. R.; Zhou, X.; and Shroff, N. 2021. Adaptive Control of Differentially Private Linear Quadratic Systems. In *2021 IEEE International Symposium on Information Theory (ISIT)*, 485–490. IEEE.

Dann, C.; Lattimore, T.; and Brunskill, E. 2017. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. *arXiv preprint arXiv:1703.07710*.

Duan, Y.; Chen, X.; Houthooft, R.; Schulman, J.; and Abbeel, P. 2016. Benchmarking deep reinforcement learning for continuous control. In *International conference on machine learning*, 1329–1338. PMLR.

Dubey, A. 2021. No-Regret Algorithms for Private Gaussian Process Bandit Optimization. In *International Conference on Artificial Intelligence and Statistics*, 2062–2070. PMLR.

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

Dwork, C. 2008. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, 1–19. Springer.

Dwork, C.; Roth, A.; et al. 2014. The algorithmic foundations of differential privacy.

Efroni, Y.; Mannor, S.; and Pirotta, M. 2020. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*.

Efroni, Y.; Merlis, N.; Ghavamzadeh, M.; and Mannor, S. 2019. Tight regret bounds for model-based reinforcement learning with greedy policies. *arXiv preprint arXiv:1905.11527*.

Efroni, Y.; Shani, L.; Rosenberg, A.; and Mannor, S. 2020. Optimistic policy optimization with bandit feedback. *arXiv preprint arXiv:2002.08243*.

Gajane, P.; Urvoy, T.; and Kaufmann, E. 2018. Corrupt bandits for preserving local privacy. In *Algorithmic Learning Theory*, 387–412. PMLR.

Garcelon, E.; Perchet, V.; Pike-Burke, C.; and Pirotta, M. 2020. Local Differentially Private Regret Minimization in Reinforcement Learning. *arXiv preprint arXiv:2010.07778*.

Gordon, G.; Spaulding, S.; Westlund, J. K.; Lee, J. J.; Plummer, L.; Martinez, M.; Das, M.; and Breazeal, C. 2016. Affective personalization of a social robot tutor for children's second language skills. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

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

Guha Thakurta, A.; and Smith, A. 2013. (Nearly) optimal algorithms for private online learning in full-information and bandit settings. *Advances in Neural Information Processing Systems*, 26: 2733–2741.

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

Hu, B.; Huang, Z.; and Mehta, N. A. 2021. Optimal Algorithms for Private Online Learning in a Stochastic Environment. *arXiv preprint arXiv:2102.07929*.

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

Jaksch, T.; Ortner, R.; and Auer, P. 2010. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr): 1563–1600.

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

Jin, C.; Yang, Z.; Wang, Z.; and Jordan, M. I. 2019. Provably efficient reinforcement learning with linear function approximation. *arXiv preprint arXiv:1907.05388*.

Kairouz, P.; Bonawitz, K.; and Ramage, D. 2016. Discrete distribution estimation under local privacy. In *International Conference on Machine Learning*, 2436–2444. PMLR.

Kakade, S. M. 2001. A natural policy gradient. *Advances in neural information processing systems*, 14.

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

Konda, V. R.; and Tsitsiklis, J. N. 2000. Actor-critic algorithms. In *Advances in neural information processing systems*, 1008–1014. Citeseer.

Li, J.; Monroe, W.; Ritter, A.; Galley, M.; Gao, J.; and Jurafsky, D. 2016. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*.

Li, L.; Chu, W.; Langford, J.; and Schapire, R. E. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, 661–670.

Liu, B.; Cai, Q.; Yang, Z.; and Wang, Z. 2019. Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv preprint arXiv:1906.10306*.

Maurer, A.; and Pontil, M. 2009. Empirical Bernstein bounds and sample variance penalization. In *22nd Conference on Learning Theory (COLT)*.

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

Ono, H.; and Takahashi, T. 2020. Locally private distributed reinforcement learning. *arXiv preprint arXiv:2001.11718*.

Puterman, M. L. 1994. *Markov Decision Processes — Discrete Stochastic Dynamic Programming*. New York, NY: John Wiley & Sons, Inc.

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

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

Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.; and Moritz, P. 2015. Trust region policy optimization. In *International conference on machine learning*, 1889–1897. PMLR.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shariff, R.; and Sheffet, O. 2018. Differentially private contextual linear bandits. *arXiv preprint arXiv:1810.00068*.

Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. 2017. Mastering the game of go without human knowledge. *nature*, 550(7676): 354–359.

Sutton, R. S. 1988. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1): 9–44.

Tossou, A.; and Dimitrakakis, C. 2016. Algorithms for differentially private multi-armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Tossou, A.; and Dimitrakakis, C. 2017. Achieving privacy in the adversarial multi-armed bandit. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

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

Wang, B.; and Hegde, N. 2019. Privacy-preserving Q-Learning with Functional Noise in Continuous State Spaces. *arXiv preprint arXiv:1901.10634*.

Wang, L.; Cai, Q.; Yang, Z.; and Wang, Z. 2019a. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*.

Wang, T.; Zhao, J.; Yang, X.; and Ren, X. 2019b. Locally differentially private data collection and analysis. *arXiv preprint arXiv:1906.01777*.

Wang, W. Y.; Li, J.; and He, X. 2018. Deep reinforcement learning for NLP. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, 19–21.

Weissman, T.; Ordentlich, E.; Seroussi, G.; Verdu, S.; and Weinberger, M. J. 2003. Inequalities for the L1 deviation of the empirical distribution. *Hewlett-Packard Labs, Technical Report*.

Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4): 229–256.

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

Zheng, K.; Cai, T.; Huang, W.; Li, Z.; and Wang, L. 2020. Locally differentially private (contextual) bandits learning. *arXiv preprint arXiv:2006.00701*.

Zhou, X.; and Tan, J. 2020. Local Differential Privacy for Bayesian Optimization. *arXiv preprint arXiv:2010.06709*.