Adaptive Control of Differentially Private Linear Quadratic Systems

Sayak Ray Chowdhury*
Indian Institute of Science
Bangalore, India
Email: sayak@iisc.ac.in

Xingyu Zhou*
Wayne State University
Detroit, USA
Email: xingyu.zhou@wayne.edu

Ness Shroff
The Ohio State University
Columbus, USA
Email: shroff.11@osu.edu

***Abstract*—In this paper we study the problem of regret minimization in reinforcement learning (RL) under differential privacy constraints. This work is motivated by the wide range of RL applications for providing personalized service, where privacy concerns are becoming paramount. In contrast to previous works, we take the first step towards *non-tabular* RL settings, while providing a rigorous privacy guarantee. In particular, we consider the adaptive control of differentially private linear quadratic (LQ) systems. We develop the first private RL algorithm,** Private-OFU-RL **which is able to attain a sub-linear regret while guaranteeing privacy protection. More importantly, the additional cost due to privacy is only on the order of $\frac{\ln(1/\delta)^{1/4}}{\varepsilon^{1/2}}$ given privacy parameters $\varepsilon, \delta > 0$. Through this process, we also provide a general procedure for adaptive control of LQ systems under *changing regularizers*, which not only generalizes previous non-private controls, but also serves as the basis for general private controls, and may be of independent interest.**

I. Introduction

Reinforcement learning (RL) is a control-theoretic problem, which adaptively learns to make sequential decisions in an unknown environment through trial and error. RL has shown to have significant success for delivering a wide variety of personalized services, including online news and advertisement recommendation [1], medical treatment design [2], natural language processing [3], and social robot [4]. In these applications, an RL agent improves its personalization algorithm by interacting with users to maximize the reward. In particular, in each round, the RL agent offers an action based on the user's state, and then receives the feedback from the user (i.e., state information, state transition, reward, etc.). This feedback is used by the agent to learn the unknown environment and improve its action selection strategy.

However, in most practical scenarios, the feedback from the users often encodes their sensitive information. For example, in a personalized healthcare setting, the states of a patient include personal information such as age, gender, height, weight, state of the treatment etc. Similarly, the states of a virtual keyboard user (e.g., Google GBoard users) are the words and sentences she typed in, which inevitably contain private information about the user. Another intriguing example is the social robot for second language education of children. The states include facial expressions, and the rewards contain whether they have passed the quiz. Users may not want any of this information to be inferred by others. This directly results in an increasing concern about privacy protection in personalized services. To be more specific, although a user might be willing to share her own information to the agent to obtain a better tailored service, she would not like to allow third parties to infer her private information from the output of the learning algorithm. For example, in the healthcare application, we would like to ensure that an adversary with arbitrary side knowledge cannot infer a particular patient's state from the treatments prescribed to her.

Differential privacy (DP) [5] has become a standard mechanism for designing interactive learning algorithms under a rigorous privacy guarantee for individual data. Most of the previous works on differentially private learning under partial feedback focus on the simpler bandit setting (i.e., no state transition) [6]–[10]. For the general RL problem, there are only a few works that consider differential privacy [11]–[13]. More importantly, only the *tabula-rasa* discrete-state discrete-action environments are considered in these works. However, in real-world applications mentioned above, the number of states and actions are often very large and can even be infinite. Over the years, for various non-tabular environments, efficient and provably optimal algorithms for *reward maximization* or, equivalently, *regret minimization* have been developed (see, e.g., [14]–[18]). This directly motivates the following question: *Is it possible to obtain the optimal reward while providing individual privacy guarantees in the non-tabular RL scenario?*

In this paper, we take the first step to answer the aforementioned question by considering a particular non-tabular RL problem – adaptive control of linear quadratic (LQ) systems, in which the state transition is a linear function and the immediate reward (cost) is a quadratic

*Equal contribution

function of the current state and action. In particular, our main contributions can be summarized as follows.

- First, we provide a general framework for adaptive control of LQ systems under *changing regularizers* using the optimism in the face of uncertainty (OFU) principle, which covers both the extreme cases – non-private and fully private LQ control.
- We then develop the first private RL algorithm, namely Private-OFU-RL, for regret minimization in LQ systems by adapting the *binary counting mechanism* to ensure differencial privacy.
- In particular, we show that Private-OFU-RL satisfies *joint differential privacy* (JDP), which, informally, implies that sensitive information about a given user is protected even if an adversary has access to the actions prescribed to all other users.
- Finally, we prove that Private-OFU-RL achieves a sub-linear regret guarantee, where the regret due to privacy only grows as $\frac{\ln(1/\delta)^{1/4}}{\varepsilon^{1/2}}$ with privacy levels $\varepsilon, \delta > 0$ implying that a high amount of privacy (low ε, δ) comes at a high cost and vice-versa.

II. Preliminaries

A. Stochastic Linear Quadratic Control

We consider the discrete-time episodic linear quadratic (LQ) control problem with H time steps at every episode. Let $x_h \in \mathbb{R}^n$ be the state of the system, $u_h \in \mathbb{R}^d$ be the control and $c_h \in \mathbb{R}$ be the cost at time h. An LQ problem is characterized by linear dynamics and a quadratic cost function

$$x_{h+1} = Ax_h + Bu_h + w_h, \quad c_h = x_h^\top Q x_h + u_h^\top R u_h \ , \quad (1)$$

where A, B are *unknown* matrices, and Q, R are known positive definite (p.d.) matrices. The starting state x_1 is fixed (can possibly be chosen by an adversary) and the system noise $w_h \in \mathbb{R}^n$ is zero-mean. We summarize the unknown parameters in $\Theta = [A, B]^\top \in \mathbb{R}^{(n+d)\times n}$.

The goal of the agent is to design a closed-loop control policy $\pi : [H] \times \mathbb{R}^n \to \mathbb{R}^d$ mapping states to controls that minimizes the expected cost

$$J_h^\pi(\Theta, x) := \mathbb{E}_\pi \Big[\sum\nolimits_{h'=h}^{H} c_{h'} \ \big| \ x_h = x \Big], \quad (2)$$

for all $h \in [H]$ and $x \in \mathbb{R}^n$. Here the expectation is over the random trajectory induced by the policy π starting from state x at time h. From the standard theory for LQ control (e.g., [19]), the optimal policy π^* has the form

$$\pi_h^*(x) = K_h(\Theta)x \ , \quad \forall h \in [H] \ ,$$

where the gain matrices $K_h(\Theta)$ are given by

$$K_h(\Theta) = -(R + B^\top P_h(\Theta) B)^{-1} B^\top P_h(\Theta) A. \quad (3)$$

Here the symmetric positive semidefinite matrices $P_h(\Theta)$ are defined recursively by the *Riccati iteration*

$$P_h(\Theta) = Q + A^\top P_{h+1}(\Theta) A \quad (4)$$
$$- A^\top P_{h+1}(\Theta) B (R + B^\top P_{h+1}(\Theta) B)^{-1} B^\top P_{h+1}(\Theta) A.$$

with $P_{H+1}(\Theta) := 0$. The optimal cost is given by

$$J_h^*(\Theta, x) = x^\top P_h(\Theta) x + \sum\nolimits_{h'=h}^{H} \mathbb{E} \left[w_{h'}^\top P_{h'+1}(\Theta) w_{h'} \right]. \quad (5)$$

We let the agent play K episodes and measure the performance by cumulative regret.[1] In particular, if the true system dynamics are $\Theta_* = [A_*, B_*]^\top$, the cumulative regret of the first K episodes is given by

$$\mathcal{R}(K) := \sum\nolimits_{k=1}^{K} \left(J_1^{\pi_k}(\Theta_*, x_{k,1}) - J_1^*(\Theta_*, x_{k,1}) \right), \quad (6)$$

where $J_1^*(\Theta_*, x_{k,1})$ is the (expected) cost under an optimal policy for episode k (computed via (5)), and $J_1^{\pi_k}(\Theta_*, x_{k,1})$ is the (expected) cost under the chosen policy π_k at the start of episode k (computed via (2)). We seek to attain a sublinear regret $\mathcal{R}(K) = o(K)$, which ensures that the agent finds the optimal policy as $K \to \infty$. We end this section by presenting our assumptions on the LQ system (1), which are common in the LQ control literature [17].

Assumption 1 (Boundedness). *(a) The true system dynamics Θ_* is a member of a set $\mathcal{S} := \{\Theta = [A, B]^\top : \|\Theta\|_{\mathrm{F}} \leqslant 1$ and $[A, B]$ is controllable$\}$. (b) There exist constants C, C_A, C_B such that $\|A_*\| \leqslant C_A < 1$, $\|B_*\| \leqslant C_B < 1$, and $\|Q\| \leqslant C$, $\|R\| \leqslant C$, (c) For all $k \geqslant 1$, $\|x_{k,1}\| \leqslant 1$. (d) The noise $w_{k,h}$ at any $k \geqslant 1$ and $h \in [H]$, is (i) independent of all other randomness, (ii) $\mathbb{E}[w_{k,h}] = 0$, and (iii) $\|w_{k,h}\|_2 \leqslant C_w < 1$. (e) There exists a constant γ such that $C_A + \gamma C_B + C_w \leqslant 1$.*

B. Differential Privacy

We now formally define the notion of differential privacy in the context of episodic LQ control. We write $v = (v_1, \ldots, v_K) \in \mathcal{V}^K$ to denote a sequence of K unique users participating in the private RL protocol with an RL agent $\mathcal{M}$, where $\mathcal{V}$ is the set of all users. Each user v_k is identified by the state responses $\{x_{k,h+1}\}_{h\in[H]}$ she gives to the controls $\{u_{k,h}\}_{h\in[H]}$ chosen by the agent. We write $\mathcal{M}(v) = \{u_{k,h}\}_{k\in[K],h\in[H]} \in (\mathbb{R}^d)^{KH}$ to denote the privatized controls chosen by the agent $\mathcal{M}$ when interacting with the users v. Informally, we will be interested in randomized algorithms $\mathcal{M}$ so that the knowledge of the output $\mathcal{M}(v)$ and all but the k-th user v_k does not reveal 'much' about v_k. We formalize in the following definition, which is adapted from [20].

Definition 1 (Differential Privacy (DP)). *For any $\varepsilon \geqslant 0$ and $\delta \in [0, 1]$, an algorithm $\mathcal{M} : \mathcal{V}^K \to (\mathbb{R}^d)^{KH}$ is (ε, δ)-differentially private if for all $v, v' \in \mathcal{V}^K$ differing on a single user and all subset of controls $\mathcal{U} \subset \big(\mathbb{R}^d\big)^{KH}$,*

$$\mathbb{P}\left[\mathcal{M}(v) \in \mathcal{U}\right] \leqslant \exp(\varepsilon) \mathbb{P}\left[\mathcal{M}(v') \in \mathcal{U}\right] + \delta \ .$$

We now relax this definition motivated by the fact that the controls recommended to a given user v_k is only

[1] In the following, we add subscript k to denote the variables for the k-th episode – state $x_{k,h}$, control $u_{k,h}$, noise $w_{k,h}$ and cost $c_{k,h}$.

observed by her. We consider *joint differential privacy* [21], which requires that simultaneously for all k, the joint distribution on controls sent to users other than v_k will not change substantially upon changing the state responses of the user v_k. To this end, we let $\mathcal{M}_{-k}(v) := \mathcal{M}(v)\backslash\{u_{k,h}\}_{h\in[H]}$ to denote all the controls chosen by the agent $\mathcal{M}$ excluding those recommended to v_k.

Definition 2 (Joint Differential Privacy (JDP))**.** *For any $\varepsilon \geqslant 0$ and $\delta\in[0,1]$, an algorithm $\mathcal{M}:\mathcal{V}^K\to(\mathbb{R}^d)^{KH}$ is (ε,δ)-jointly differentially private if for all $k\in[K]$, all $v,v'\in\mathcal{V}$ differing on the k-th user and all subset of controls $\mathcal{U}_{-k}\subset\left(\mathbb{R}^d\right)^{(K-1)H}$ given to all but the k-th user,*

$$\mathbb{P}\left[\mathcal{M}_{-k}(v)\in\mathcal{U}_{-k}\right]\leqslant\exp(\varepsilon)\mathbb{P}\left[\mathcal{M}_{-k}(v')\in\mathcal{U}_{-k}\right]+\delta\ .$$

This relaxation is necessary in our setting since knowledge of the controls recommended to the user v_k can reveal a lot of information about her state responses. It weakens the constraint of DP only in that the controls given specifically to v_k may be sensitive in her state responses. However, it is still a very strong definition since it protects v_k from any arbitrary collusion of other users against her, so long as she does not herself make the controls reported to her public.

In this work, we look for algorithms that are (ε,δ)-JDP. But, we will build our algorithm upon standard DP mechanisms. Furthermore, to establish privacy, we will use a different relaxation called *concentrated differential privacy* (CDP) [22]. Roughly, a mechanism is CDP if the privacy loss has Gaussian tails. To this end, we let $\mathcal{M}$ to be a mechanism taking as input a data-stream $x\in\mathcal{X}^n$ and releasing output from some range $\mathcal{Y}$.

Definition 3 (Concentrated Differential Privacy (CDP))**.** *For any $\rho\geqslant 0$, an algorithm $\mathcal{M}:\mathcal{X}^n\to\mathcal{Y}$ is ρ-zero-concentrated differentially private if for all $x,x'\in\mathcal{X}^n$ differing on a single entry and all $\alpha\in(1,\infty)$,*

$$\mathrm{D}_\alpha\left(\mathcal{M}(x)||\mathcal{M}(x')\right)\leqslant\rho\ \alpha\ ,$$

where $\mathrm{D}_\alpha\left(\mathcal{M}(x)||\mathcal{M}(x')\right)$ is the α-Renyi divergence between the distributions of $\mathcal{M}(x)$ and $\mathcal{M}(x')$.[2]

III. OFU-Based Control

Our proposed private RL algorithm implements the optimism in the face of uncertainty (OFU) principle in LQ systems. As in [14], the key to implementing the OFU-based control is a high-probability confidence set for the unknown parameter matrix Θ_*.

A. Adaptive Control with Changing Regularizers

We start with the adaptive LQ control with *changing regularizers*. This not only allows us to generalize previous results for non-private control, but more importantly serves as a basis for the analysis of private control in the next section. We first define the following compact notations. For a state and control pair at step h in episode k, i.e., $x_{k,h}$ and $u_{k,h}$, we write $z_{k,h}=[x_{k,h}^\top,u_{k,h}^\top]^\top$. For any $k\geqslant 1$, we define the following matrices: $Z_k := [z_{k',h'}^\top]_{k'\in[k-1],h'\in[H]}$, $X_k^{\text{next}} := [x_{k',h'+1}^\top]_{k'\in[k-1],h'\in[H]}$ and $W_k := [w_{k',h'}^\top]_{k'\in[k-1],h'\in[H]}$. For two matrices A and B, we also define $\|A\|_B^2 := \mathrm{trace}(A^\top BA)$. Now, at every episode k, we consider the following ridge regression estimate w.r.t. a regularizing p.d. matrix $H_k\in\mathbb{R}^{(n+d)\times(n+d)}$:

$$\begin{aligned}\Theta_k &:= \operatorname*{arg\,min}_{\Theta\in\mathbb{R}^{(n+d)\times n}}\left\|X_k^{\text{next}}-Z_k\Theta\right\|_{\mathrm{F}}^2+\|\Theta\|_{H_k}^2\\ &=(Z_k^\top Z_k+H_k)^{-1}Z_k^\top X_k^{\text{next}},\end{aligned}$$

In contrast to the standard online LQ control [14], here the sequence of matrices $\{Z_k^\top Z_k\}_{k\geqslant 1}$ is perturbed by a sequence of regularizers $\{H_k\}_{k\geqslant 1}$. In particular, when $H_k=\lambda I$, we get back the standard estimate of [14]. In addition, we also allow $Z_k^\top X_k^{\text{next}}$ to be perturbed by a matrix L_k at every episode k. Setting $V_k := Z_k^\top Z_k + H_k$ and $U_k := Z_k^\top X_k^{\text{next}} + L_k$, we now define the estimate under changing regularizers $\{H_k\}_{k\geq 1}$ and $\{L_k\}_{k\geq 1}$ as

$$\widehat{\Theta}_k = V_k^{-1}U_k\ . \tag{7}$$

We make the following assumptions on the sequence of regularizers $\{H_k\}_{k\geq 1}$ and $\{L_k\}_{k\geq 1}$.

Assumption 2 (Regularity)**.** *For any $\alpha\in(0,1]$, there exist constants $\lambda_{\max}$, $\lambda_{\min}$ and ν depending on α such that, with probability at least $1-\alpha$, for all $k\in[K]$,*

$$\|H_k\|\leqslant\lambda_{\max},\quad\left\|H_k^{-1}\right\|\leqslant 1/\lambda_{\min}\ \ \textit{and}\ \ \|L_k\|_{H_k^{-1}}\leqslant\nu.$$

Lemma 1 (Concentration under changing regularizers)**.** *Under assumptions 1 and 2, the following holds:*

$$\forall\alpha\in(0,1],\quad\mathbb{P}\left[\exists k\in\mathbb{N}\colon\ \left\|\Theta_*-\widehat{\Theta}_k\right\|_{V_k}\geqslant\beta_k(\alpha)\right]\leqslant\alpha\ ,$$

where $\beta_k(\alpha) := C_w\sqrt{2\ln\left(\frac{2}{\alpha}\right)+n\ln\det\left(I+\lambda_{\min}^{-1}Z_k^\top Z_k\right)}+\sqrt{\lambda_{\max}}+\nu$.

Lemma 1 helps us to introduce the following high probability confidence set

$$\mathcal{C}_k(\alpha) := \left\{\Theta : \left\|\Theta-\widehat{\Theta}_k\right\|_{V_k}\leqslant\beta_k(\alpha)\right\}. \tag{8}$$

We then search for an optimistic estimate $\widetilde{\Theta}_k$ within this confidence region $\mathcal{C}_k(\alpha)$, such that

$$\widetilde{\Theta}_k\in\operatorname*{arg\,min}_{\Theta\in\mathcal{C}_k(\alpha)\cap\mathcal{S}} J_1^*(\Theta,x_{k,1}), \tag{9}$$

where $J_1^*(\Theta,x_{k,1})$ is the optimal cost when system dynamics are Θ (can be computed from (5)). With the estimate $\widetilde{\Theta}_k$, the agent then chooses policy π_k and selects the controls recommended by this policy

$$u_{k,h} := \pi_{k,h}(x_{k,h}) = K_h(\widetilde{\Theta}_k)x_{k,h}\ , \tag{10}$$

where $K_h(\widetilde{\Theta}_k)$ can be computed from (3). We call this procedure OFU-RL and bound its regret as follows.

Theorem 1 (Regret under changing regularizers)**.** *Under*

[2]For two probability distributions P and Q on Ω, the α-Renyi divergence $\mathrm{D}_\alpha(P||Q) := \frac{1}{\alpha-1}\ln\left(\int_\Omega P(x)^\alpha Q(x)^{1-\alpha}\,\mathrm{d}x\right)$.

Assumptions 1 and 2, for any $\alpha \in (0,1]$, *with probability at least* $1-\alpha$, *the cumulative regret of* OFU-RL *satisfies*

$$\mathcal{R}(K) = O\left(H\sqrt{K}\left(\sqrt{H} + n(n+d)\psi_{\lambda_{\min}} + \ln(1/\alpha)\right)\right) + O\left(H\sqrt{K}\left(\sqrt{\lambda_{\max}} + \nu\right)\sqrt{n(n+d)\psi_{\lambda_{\min}}}\right),$$

where $\psi_{\lambda_{\min}} := \ln(1+HK/(n+d)\lambda_{\min})$.

Proof sketch. Inspired by [17], we first decompose the regret under the following 'good' event: $\mathcal{E}_K(\alpha) := \{\Theta_* \in \mathcal{C}_k(\alpha) \cap \mathcal{S}, \forall k \in [K]\}$, which, by Assumption 1 and Lemma 1, holds w.p. at least $1-\alpha$. Then, under the 'good' event, the cumulative regret (6) can be written as $\mathcal{R}(K) \leqslant \sum_{k=1}^{K}\sum_{h=1}^{H}(\Delta_{k,h} + \Delta'_{k,h} + \Delta''_{k,h})$, where

$$\Delta_{k,h} := \mathbb{E}\left[J_{h+1}^{\pi_k}(\Theta_*, x_{k,h+1}) | \mathcal{F}_{k,h}\right] - J_{h+1}^{\pi_k}(\Theta_*, x_{k,h+1}),$$

$$\Delta'_{k,h} := \|x_{k,h+1}\|_{\widetilde{P}_{k,h+1}} - \mathbb{E}\left[\|x_{k,h+1}\|_{\widetilde{P}_{k,h+1}} | \mathcal{F}_{k,h}\right] \text{ and}$$

$$\Delta''_{k,h} := \left\|\Theta_*^\top z_{k,h}\right\|_{\widetilde{P}_{k,h+1}} - \left\|\widetilde{\Theta}_k^\top z_{k,h}\right\|_{\widetilde{P}_{k,h+1}},$$

in which $\widetilde{P}_{k,h} := P_h(\widetilde{\Theta}_k)$ is given by (4) and $\mathcal{F}_{k,h}$ denotes all the randomness present before time (k,h).

Now, we are going to bound each term, respectively. For the first two terms, we can show that both of them are bounded martingale difference sequences. Therefore, by Azuma–Hoeffding inequality, we have $\sum_{k,h}\Delta_{k,h} = O(\sqrt{KH^3})$ and $\sum_{k,h}\Delta'_{k,h} = O(\sqrt{KH})$ with high probability. We use Lemma 1 and the OFU principle (9) to bound the third term as $\sum_{k,h}\Delta''_{k,h} = O\big(H\sqrt{K}\beta_k(\alpha)\sqrt{\ln\det\left(I+\lambda_{\min}^{-1}Z_k^\top Z_k\right)}\big)$. To put everything together, first note from Assumption 1 that

$$\ln\det\left(I+\lambda_{\min}^{-1}Z_k^\top Z_k\right) \leqslant (n+d)\ln\left(1+\frac{HK(1+\gamma)^2}{(n+d)\lambda_{\min}}\right).$$

Plugging this into $\beta_k(\alpha)$ given in Lemma 1 and the third term above, yields the final result. □

We end the section with a proof sketch of Lemma 1.

Proof sketch (Lemma 1). Under Assumptions 1 and 2, with some basic algebra, we first have

$$\left\|\Theta_* - \widehat{\Theta}_k\right\|_{V_k} = \left\|H_k\Theta_* - Z_k^\top W_k - L_k\right\|_{V_k^{-1}} \leqslant \underbrace{\left\|Z_k^\top W_k\right\|_{(Z_k^\top Z_k + \lambda_{\min} I)^{-1}}}_{\mathcal{T}_1} + \underbrace{\left\|H_k^{\frac{1}{2}}\right\|_2 + \|L_k\|_{H_k^{-1}}}_{\mathcal{T}_2}.$$

By Assumption 2, we have w.p. at least $1-\alpha$, $\mathcal{T}_2 \leqslant \sqrt{\lambda_{\max}} + \nu$. To bound $\mathcal{T}_1$, by the boundedness of $w_{k,h}$ in Assumption 1, we first note that each row of the matrix W_k is a sub-Gaussian random vector with parameter C_w. We then generalize the self-normalized concentration inequality of vector-valued martingales [23, Theorem 1] to the setting of matrix-valued martingales. In particular, we show that w.p. at least $1-\alpha$,

$$\mathcal{T}_1 \leqslant C_w\sqrt{2\ln(1/\alpha) + n\ln\det(I+\lambda_{\min}^{-1}Z_k^\top Z_k)}.$$

Combining the bounds on $\mathcal{T}_1$ and $\mathcal{T}_2$ using a union bound argument, yields the final result. □

B. Private Control

In this section, we introduce the Private-OFU-RL algorithm (Alg. 1). At every episode k, we keep track of the history via private version of the matrices $Z_k^\top Z_k$ and $Z_k^\top X_k^{\text{next}}$. To do so, we first initialize two private counter mechanisms $\mathcal{B}_1$ and $\mathcal{B}_2$, which take as parameters the privacy levels ε, δ, number of episodes K, horizon H and a problem-specific constant γ (see Assumption 1). The counter $\mathcal{B}_1$ (resp. $\mathcal{B}_2$) take as input an event stream of matrices $\{\sum_{h=1}^{H} z_{k,h}z_{k,h}^\top\}_{k\in[K]}$ (resp. $\{\sum_{h=1}^{H} z_{k,h}x_{k,h+1}^\top\}_{k\in[K]}$), and at the start of each episode k, release the private version of the matrix $Z_k^\top Z_k$ (resp. $Z_k^\top X_k^{\text{next}}$), which itself is a matrix of the same dimension. Let $T_{1,k}$ and $T_{2,k}$ denote the privatized versions for $Z_k^\top Z_k$ and $Z_k^\top X_k^{\text{next}}$, respectively. For some $\eta > 0$ (will be determined later), we define $V_k := T_{1,k} + \eta I$ and $U_k := T_{2,k}$. We now instantiate the general OFU-RL procedure under changing regularizers (Section III-A) with these private statistics. First, we compute the point estimate $\widehat{\Theta}_k$ from (7) and build the confidence set $\mathcal{C}_k(\alpha)$ from (8). Then, we choose the most optimisitic policy π_k w.r.t. the entire set $\mathcal{C}_k(\alpha)$ from (9) and (10). Finally, we execute the policy for the entire episode and update the counters with observed trajectory.

We now describe the private counters $\mathcal{B}_1$ adapting the *Binary counting mechanism* of [24]. First, we write $\Sigma_1[i,j] = \sum_{k=i}^{j}\sum_{h=1}^{H} z_{k,h}z_{k,h}^\top$ to denote a partial sum (P-sum) involving the state-control pairs in episodes i through j. Next, we consider a binary interval tree, where each leaf node represents an episode (i.e., the tree has $k-1$ leaf nodes at the start of episode k), and each interior node represents the range of episodes covered by its children. At the start of episode k, we first release a noisy P-sum $\widehat{\Sigma}_1[i,j]$ corresponding to each node in the tree. Here $\widehat{\Sigma}_1[i,j]$ is obtained by perturbing both (p,q)-th and (q,p)-th, $1 \leqslant p \leqslant q \leqslant (n+d)$, entries of $\Sigma_1[i,j]$ with i.i.d. Gaussian noise $\zeta_{p,q} \sim \mathcal{N}(0,\sigma_1^2)$.[3] Then $T_{1,k}$ is computed by summing up the noisy P-sums released by the set of nodes that uniquely cover the range $[1,k-1]$. Observe that, at the end each episode, the mechanism only needs to store noisy P-sums required for computing private statistics at future episodes, and can safely discard P-sums that are no longer needed. For the private counter $\mathcal{B}_2$, we maintain P-sums $\Sigma_2[i,j] = \sum_{k=i}^{j}\sum_{h=1}^{H} z_{k,h}z_{k,h}^\top$ with i.i.d. noise $\mathcal{N}(0,\sigma_2^2)$ and compute the private statistics $T_{2,k}$ using a similar procedure. The noise levels σ_1 and σ_2 depends on the problem intrinsics (K, H, γ) and privacy parameters (ε, δ). These, in turn, govern the constants $\lambda_{\max}$, $\lambda_{\min}$, ν appearing in the confidence set

[3]This will ensure symmetry of the P-sums even after adding noise.

Algorithm 1: Private-OFU-RL

Input: Number of episodes K, horizon H, privacy level $\varepsilon > 0$, $\delta \in (0,1]$, constants γ, C_w, confidence level $\alpha \in (0,1]$

initialize private counters $\mathcal{B}_1$ and $\mathcal{B}_2$ with parameters $K, H, \varepsilon, \delta, \gamma$

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

  compute private statistics $T_{1,k}$ and $T_{2,k}$

  construct confidence set $\mathcal{C}_k(\alpha)$

  find $\widetilde{\Theta}_k \in \arg\min_{\Theta \in \mathcal{C}_k(\alpha)\cap\mathcal{S}} J_1^*(\Theta, x_{k,1})$

  for *each step* $h = 1,2,\ldots,H$ **do**

    execute control $u_{k,h} = K_h(\widetilde{\Theta}_k)x_{k,h}$

    observe cost $c_{k,h}$ and next state $x_{k,h+1}$

  send $\sum_{h=1}^{H} z_{k,h}z_{k,h}^\top$ and $\sum_{h=1}^{H} z_{k,h}x_{k,h+1}^\top$ to the counters $\mathcal{B}_1$ and $\mathcal{B}_2$, respectively

$\mathcal{C}_k(\alpha)$ and the regularizer η. The details will be specified in the next Section as needed.

IV. Privacy and Regret Guarantees

In this section, we show that Private-OFU-RL is a JDP algorithm with sublinear regret guarantee.

A. Privacy Guarantee

Theorem 2 (Privacy)**.** *Under Assumption 1, for any* $\varepsilon > 0$ *and* $\delta \in (0,1]$, Private-OFU-RL *is* (ε,δ)*-JDP.*

Proof sketch. We first show that both the counters $\mathcal{B}_1$ and $\mathcal{B}_2$ are $(\varepsilon/2, \delta/2)$-DP. We begin with the counter $\mathcal{B}_1$. To this end, we need to determine a global upper bound Δ_1 over the L_2-sensitivity of all the P-sums $\Sigma_1[i,j]$. Informally, Δ_1 encodes the maximum change in the Frobenious norm of each P-sum if the trajectory of a single episode is changed. By Assumption 1, we have $\|z_{k,h}\| \leqslant 1+\gamma$, and hence $\Delta_1 = H(1+\gamma)^2$. Since the noisy P-sums $\widehat{\Sigma}_1[i,j]$ are obtained via Gaussian mechanism, we have that each $\widehat{\Sigma}_1[i,j]$ is $(\Delta_1^2/2\sigma_1^2)$-CDP [22, Proposition 1.6]. We now see that every episode appears only in at most $m := \lceil \log_2 K \rceil$ P-sums $\Sigma_1[i,j]$. Therefore, by the composition property, the whole counter $\mathcal{B}_1$ is $(m\Delta_1^2/2\sigma_1^2)$-CDP, and thus, in turn, $\left(\frac{m\Delta_1^2}{2\sigma_1^2} + 2\sqrt{\frac{m\Delta_1^2}{2\sigma_1^2}\ln\left(\frac{2}{\delta}\right)}, \frac{\delta}{2}\right)$-DP for any $\delta > 0$ [22, Lemma 3.5]. Now, setting $\sigma_1^2 \approx 8m\Delta_1^2\ln(2/\delta)/\varepsilon^2$, we can ensure that $\mathcal{B}_1$ is $(\varepsilon/2, \delta/2)$-DP. A similar analysis yields that counter $\mathcal{B}_2$ is $(\varepsilon/2, \delta/2)$-DP if we set $\sigma_2^2 \approx 8m\Delta_2^2\ln(2/\delta)/\varepsilon^2$, where $\Delta_2 := H(1+\gamma)$.

To prove Theorem 2, we now use the *billboard lemma* [25, Lemma 9] which, informally, states that an algorithm is JDP under continual observation if the output sent to each user is a function of the user's private data and a common quantity computed using standard differential privacy. Note that at each episode k, Private-OFU-RL computes private statistics $T_{1,k}$ and $T_{2,k}$ for all users using the counters $\mathcal{B}_1$ and $\mathcal{B}_2$. These statistics are then used to compute the policy π_k. By composition and post-processing properties of DP, we can argue that the sequence of policies $\{\pi_k\}_{k\in[K]}$ are computed using an (ε,δ)-DP mechanism. Now, the controls $\{u_{k,h}\}_{h\in[H]}$ during episode k are generated using the policy π_k and the user's private data $x_{k,h}$ as $u_{k,h} = \pi_{k,h}(x_{k,h})$. Then, by the billboard lemma, the composition of the controls $\{u_{k,h}\}_{k\in[K],h\in[H]}$ sent to all the users is (ε,δ)-JDP. □

B. Regret Guarantee

Theorem 3 (Private regret)**.** *Under Assumption 1, for any privacy parameters* $\varepsilon > 0$ *and* $\delta \in (0,1]$*, and for any* $\alpha \in (0,1]$*, with probability at least* $1-\alpha$*,* Private-OFU-RL *enjoys the regret bound*

$$\mathcal{R}(K) = O\Big(H^{3/2}\sqrt{K}\big(n(n+d)\ln K + \ln(1/\alpha)\big)\Big)$$
$$+ O\left(H^{3/2}\sqrt{K}\ln K\left(n(n+d) + \sqrt{\ln K/\alpha}\right)\frac{\ln(1/\delta)^{1/4}}{\varepsilon^{1/2}}\right).$$

Theorems 2 and 3 together imply that Private-OFU-RL can achieve a sub-linear regret under (ε,δ)-JDP privacy guarantee. Furthermore, comparing Theorem 3 with Theorem 1, we see that the first term in the regret bound corresponds to the non-private regret, and the second term is the cost of privacy. More importantly, the cost due to privacy grows only as $\frac{\ln(1/\delta)^{1/4}}{\varepsilon^{1/2}}$ with ε, δ.

Proof sketch (Theorem 3). First note that the private statistics $T_{1,k}$ can be computed by summing at most $m = \lceil \log_2 K \rceil$ noisy P-sums $\widehat{\Sigma}_1[i,j]$. We then have that the total noise N_k in each $T_{1,k}$ is a symmetric matrix with it's (p,q)-th entry, $1 \leqslant p \leqslant q \leqslant (n+d)$, being i.i.d. $\mathcal{N}(0, m\sigma_1^2)$. Therefore, by an adaptation of [26, Corollary 4.4.8], we have w.p. at least $1-\alpha/2K$,

$$\|N_k\| \leqslant \Lambda := \sigma_1\sqrt{m}\left(4\sqrt{n+d} + \sqrt{8\ln(4K/\alpha)}\right).$$

Similarly, the total noise L_k in each $T_{2,k}$ is an $(n+d)\times n$ matrix, whose each entry is i.i.d. $\mathcal{N}(0, m\sigma_2^2)$. Hence $\|L_k\|_{\mathrm{F}}^2/m\sigma_2^2$ is a χ^2-statistic with $n(n+d)$ degrees of freedom, and therefore, by [27, Lemma 1], we have w.p. at least $1-\alpha/2K$,

$$\|L_k\|_{\mathrm{F}} \leqslant \sigma_2\sqrt{m}\left(\sqrt{2n(n+d)} + \sqrt{4\ln(2K/\alpha)}\right).$$

By construction, we have the regularizer $H_k = N_k + \eta I$. Setting $\eta = 2\Lambda$, we ensure that H_k is p.d., and hence $\|L_k\|_{H_k^{-1}} \leqslant \Lambda^{-1/2}\|L_k\|_{\mathrm{F}}$. Then, by a union bound argument, Assumption 2 holds for $\lambda_{\min} = \Lambda$, $\lambda_{\max} = 3\Lambda$ and $\nu = \sigma_2\sqrt{m/\Lambda}\left(\sqrt{2n(n+d)} + \sqrt{4\ln(2K/\alpha)}\right)$. Appropriating noise levels σ_1, σ_2 from Section IV-A, the regret bound now follows from Theorem 1. □

V. Conclusion

We develop the first DP algorithm, Private-OFU-RL, for episodic LQ control. Through the notion of JDP, we show that it can protect private user information from being inferred by observing the control policy without losing much on its regret performance. We leave as future work private control of non-linear systems [16].

REFERENCES

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

[2] Y. Zhao, M. R. Kosorok, and D. Zeng, "Reinforcement learning design for cancer clinical trials," *Statistics in medicine*, vol. 28, no. 26, pp. 3294–3315, 2009.

[3] A. R. Sharma and P. Kaushik, "Literature survey of statistical, deep and reinforcement learning in natural language processing," in *2017 International Conference on Computing, Communication and Automation (ICCCA)*. IEEE, 2017, pp. 350–354.

[4] G. Gordon, S. Spaulding, J. K. Westlund, J. J. Lee, L. Plummer, M. Martinez, M. Das, and C. Breazeal, "Affective personalization of a social robot tutor for children's second language skills," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.

[5] C. Dwork, "Differential privacy: A survey of results," in *International conference on theory and applications of models of computation*. Springer, 2008, pp. 1–19.

[6] A. Tossou and C. Dimitrakakis, "Algorithms for differentially private multi-armed bandits," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.

[7] ——, "Achieving privacy in the adversarial multi-armed bandit," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[8] D. Basu, C. Dimitrakakis, and A. Tossou, "Differential privacy for multi-armed bandits: What is it and what is its cost?" *arXiv preprint arXiv:1905.12298*, 2019.

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

[10] X. Zhou and J. Tan, "Local differential privacy for bayesian optimization," *arXiv preprint arXiv:2010.06709*, 2020.

[11] B. Balle, M. Gomrokchi, and D. Precup, "Differentially private policy evaluation," in *International Conference on Machine Learning*. PMLR, 2016, pp. 2130–2138.

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

[13] E. Garcelon, V. Perchet, C. Pike-Burke, and M. Pirotta, "Local differentially private regret minimization in reinforcement learning," *arXiv preprint arXiv:2010.07778*, 2020.

[14] Y. Abbasi-Yadkori and C. Szepesvári, "Regret bounds for the adaptive control of linear quadratic systems," in *Proceedings of the 24th Annual Conference on Learning Theory*, 2011, pp. 1–26.

[15] I. Osband and B. V. Roy, "Model-based reinforcement learning and the eluder dimension," in *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 1*, 2014, pp. 1466–1474.

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

[17] T. Wang and L. F. Yang, "Episodic linear quadratic regulators with low-rank transitions," *arXiv preprint arXiv:2011.01568*, 2020.

[18] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan, "Provably efficient reinforcement learning with linear function approximation," in *Conference on Learning Theory*, 2020, pp. 2137–2143.

[19] D. Bertsekas, *Dynamic programming and optimal control*. Athena scientific Belmont, MA, 3 edition, 2004.

[20] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy." 2014.

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

[22] M. Bun and T. Steinke, "Concentrated differential privacy: Simplifications, extensions, and lower bounds," in *Theory of Cryptography Conference*. Springer, 2016, pp. 635–658.

[23] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *Advances in Neural Information Processing Systems*, 2011, pp. 2312–2320.

[24] T.-H. H. Chan, E. Shi, and D. Song, "Private and continual release of statistics," *ACM Transactions on Information and System Security (TISSEC)*, vol. 14, no. 3, pp. 1–24, 2011.

[25] J. Hsu, Z. Huang, A. Roth, T. Roughgarden, and Z. S. Wu, "Private matchings and allocations," *SIAM Journal on Computing*, vol. 45, no. 6, pp. 1953–1984, 2016.

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

[27] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Annals of Statistics*, pp. 1302–1338, 2000.

[28] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *arXiv preprint arXiv:1011.3027*, 2010.

[29] M. Ledoux, *The concentration of measure phenomenon*. American Mathematical Soc., 2001, no. 89.

APPENDIX

A. Important Facts

Under Assumption 1, we have that the states and controls are bounded, which directly follows from Proposition 8 in [17].

Lemma 2. *Under Assumption 1, our algorithm satisfies*

$$\|x_{k,h}\|_2 \le 1, \|u_{k,h}\|_2 \le \gamma, \|z_{k,h}\|_2 \le (1+\gamma)$$

for all $k \ge 1$ and $h \in [H]$.

We also have an upper bound on the matrices $P_h(\Theta)$, which follows from Proposition 6 in [17].

Lemma 3. *Under Assumption 1, there exists a constant D such that for all $\Theta \in \mathcal{C}_k(\delta) \cap \mathcal{S}$,*

$$\|P_{k,h}(\Theta)\|_2 \le D$$

holds for all $k \ge 1$ and $h \in [H]$.

Lemma 4 (Operator norm of symmetric Gaussian random matrices)**.** *Let A be an $n \times n$ random symmetric matrix whose entries A_{ij} on and above the diagonal are $i.i.d$ Gaussian random variables $\mathcal{N}(0, \sigma^2)$. Then, for any $t > 0$, we have*

$$\|A\| \le 4\sigma\sqrt{n} + 2t,$$

with probability at least $1 - 2\exp(-\frac{t^2}{2\sigma^2})$.

Lemma 5 (Concentration of chi-square; Corollary to Lemma 1 of [27])**.** *Let U be a χ^2 statistic with D degrees of freedom. Then, for any positive x,*

$$\mathbb{P}\left\{U \ge D + 2\sqrt{Dx} + 2x\right\} \le \exp(-x).$$

B. Proof of Theorem ??

Proof. For any $k \ge 1$, we have

$$\begin{aligned}
&\Theta_* - \Theta_k \\
=&\Theta_* - V_k^{-1} U_k \\
=&\Theta_* - V_k^{-1}\left(Z_k^T Z_k \Theta_* + Z_k^T W_k + L_k\right) \\
=&\Theta_* - V_k^{-1}\left(V_k \Theta_* - H_k \Theta_* + Z_k^T W_k + L_k\right) \\
=&\widehat{V}_k^{-1}\left(H_k \Theta_* - Z_k^T W_k - L_k\right).
\end{aligned}$$

Hence, we have

$$\begin{aligned}
&\|\Theta_k - \Theta_*\|_{V_k} \\
=&\left\|H_k\Theta_* - Z_k^T W_k - L_k\right\|_{V_k^{-1}} \\
\overset{(a)}{\le}&\left\|Z_k^T W\right\|_{V_k^{-1}} + \|H_k\Theta_*\|_{V_k^{-1}} + \|L_k\|_{V_k^{-1}} \\
\overset{(b)}{\le}&\left\|Z_k^T W_k\right\|_{V_k^{-1}} + \|H_k\Theta_*\|_{H_k^{-1}} + \|L_k\|_{H_k^{-1}} \\
\overset{(c)}{\le}&\left\|Z_k^T W_k\right\|_{(G_k+\lambda_{min}I)^{-1}} + \left\|H_k^{\frac{1}{2}}\Theta_*\right\|_F + \|L_k\|_{H_k^{-1}} \\
\overset{(d)}{\le}&\underbrace{\left\|Z_k^T W_k\right\|_{(G_k+\lambda_{min}I)^{-1}}}_{\mathcal{T}_1} + \underbrace{\left\|H_k^{\frac{1}{2}}\right\|_2 + \|L_k\|_{H_k^{-1}}}_{\mathcal{T}_2},
\end{aligned}$$

where (a) follows from triangle inequality; (b) holds by $V_k \succeq H_k$; (c) is true since $V_k \succeq (G_k + \lambda_{min} I)$ under Assumption 2; (d) follows from the fact for any two matrices A and B, $\|AB\|_F \leq \|A\|_2 \|B\|_F$, and the assumption $\|\Theta_*\|_F \leq 1$ in Assumption 1.

Now, we are left to bound $\mathcal{T}_1$ and $\mathcal{T}_2$, respectively. First, under Assumption 2, by a union bound, we have w.p. at least $1-\delta/2$,

$$\mathcal{T}_2 \leq \sqrt{\lambda_{max}} + \nu.$$

For $\mathcal{T}_1$, we can apply Lemma **??** since noise $w_{k,h}$ is a sub-Gaussian vector with parameter C_w under the assumption (b) in Assumption 1. Thus, we have w.p. at least $1-\delta/2$,

$$\mathcal{T}_1 \leq C_w \sqrt{2\ln\left(\frac{2}{\delta}\right) + n \ln \frac{\det(G_k + \lambda_{min} I)}{\det(\lambda_{min} I)}}.$$

Finally, putting the bounds on $\mathcal{T}_1$ and $\mathcal{T}_2$ together, yields the required result. □

C. Proof of Theorem 1

Proof. We choose an error probability $\delta > 0$. Given this, we define the following 'good' event, i.e., the confidence set holds.

$$\mathcal{E}_K(\delta) := \{\Theta_* \in \mathcal{C}_k(\delta) \cap \mathcal{S}, \forall k = 1, 2 \ldots, K\},$$

in which $\mathcal{C}_k(\delta)$ is given by (8). As a direct result of Theorem **??** and (a) in Assumption 1, we know with probability at least $1-\delta$, the event $\mathcal{E}_K(\delta)$ is true.

In the following, we will bound the regret under event $\mathcal{E}_K(\delta)$. First, we can decompose the regret as in the following lemma.

Lemma 6. *Let $\widetilde{P}_{k,h} := P_h(\widetilde{\Theta}_k)$ given by* (4) *and $\mathcal{F}_{k,h}$ is is all randomness before time (k,h). Under event $\mathcal{E}_K(\delta)$, we have*

$$\mathcal{R}(K) \leq \sum_{k=1}^{K} \sum_{h=1}^{H} (\Delta_{k,h} + \Delta'_{k,h} + \Delta''_{k,h}),$$

where

$$\Delta_{k,h} := \mathbb{E}\left[J_{h+1}^{\pi_k}(\Theta_*, x_{k,h+1}) \mid \mathcal{F}_{k,h}\right] - J_{h+1}^{\pi_k}(\Theta_*, x_{k,h+1}),$$
$$\Delta'_{k,h} := \|x_{k,h+1}\|_{\widetilde{P}_{k,h+1}} - \mathbb{E}\left[\|x_{k,h+1}\|_{\widetilde{P}_{k,h+1}} \mid \mathcal{F}_{k,h}\right],$$
$$\Delta''_{k,h} := \left\|\Theta_*^\top z_{k,h}\right\|_{\widetilde{P}_{k,h+1}} - \left\|\widetilde{\Theta}_k^\top z_{k,h}\right\|_{\widetilde{P}_{k,h+1}}.$$

Now, we will turn to bound each of the three terms above. For the first two terms, we can bound them by using the following lemma.

Lemma 7. *Under Assumption 1 and the event $\mathcal{E}_K(\delta)$, with probability at least $1-2\delta$, we have both*

$$\left|\sum_{k=1}^{K}\sum_{h=1}^{H} \Delta_{k,h}\right| \leq O\left(\sqrt{KH^3 \ln\frac{2}{\delta}}\right)$$

and

$$\left|\sum_{k=1}^{K}\sum_{h=1}^{H} \Delta'_{k,h}\right| \leq O\left(\sqrt{KH \ln\frac{2}{\delta}}\right)$$

For the third term, we can bound it as follows.

Lemma 8. *Under Assumption 1 and the event $\mathcal{E}_K(\delta)$, we have*

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \Delta''_{k,h} = O\left(H\sqrt{K}\beta_k(\delta)\left(\ln\frac{\det(G_k + \lambda_{min}I)}{\det(\lambda_{min}I)}\right)^{1/2}\right).$$

Finally, we are ready to put everything together. We first note that by the boundness result in Lemma 2

$$\ln\det(G_k + \lambda_{min}I) \leq (n+d)\ln\left(\lambda_{min} + \frac{HK(1+\gamma)^2}{n+d}\right).$$

Based on this and $\widehat{\gamma} := (1+\gamma)^2$, we have from (**??**)

$$\beta_k(\delta) = C_w\sqrt{2\ln\frac{2}{\delta} + (n^2+nd)\ln\left(1 + \frac{HK\widehat{\gamma}}{(n+d)\lambda_{min}}\right)} + \sqrt{\lambda_{max}} + \nu.$$

Thus, we finally obtain that

$$\begin{aligned}
&\mathcal{R}(K)\\
=&O(H^{3/2}\sqrt{K})\\
+&O\left(H\sqrt{K}(\sqrt{\lambda_{max}}+\nu)\sqrt{(n+d)\psi(\lambda_{min}, n, d, H, K)}\right)\\
+&O\left(H\sqrt{K}\left(\ln\frac{1}{\delta} + n(n+d)\psi(\lambda_{min}, n, d, H, K)\right)\right),
\end{aligned}$$

in which

$$\psi(\lambda_{min}, n, d, H, K) := \ln\left(1 + \frac{HK}{(n+d)\lambda_{min}}\right).$$

□

D. Proofs of Lemmas for Theorem 1

Proof of Lemma 6. Let $\Gamma_{k,h} := J_h^{\pi_k}(\Theta_*, x_{k,h}) - J_h^*(\widetilde{\Theta}_k, x_{k,h})$.

$$\begin{aligned}
R(K) &= \sum_{k=1}^{K} J_1^{\pi_k}(\Theta_*, x_{k,1}) - J_1^*(\Theta_*, x_{k,1})\\
&\overset{(a)}{\leq} \sum_{k=1}^{K} J_1^{\pi_k}(\Theta_*, x_{k,1}) - J_1^*(\widetilde{\Theta}_k, x_{k,1})\\
&= \sum_{k=1}^{K} \Gamma_{k,1},
\end{aligned} \tag{11}$$

where (a) holds by the optimistic algorithm (i.e., (9)) under the event $\mathcal{E}_K(\delta)$.

To bound this, we first investigate $\Gamma_{k,h}$. Note that the action $u_{k,h}$ under π_k is the same as that under an optimal policy when the true dynamics is $\widetilde{\Theta}$, and hence

$$\begin{aligned}
\Gamma_{k,h} =& \|x_{k,h}\|_{Q_h} + \|u_{k,h}\|_{R_h} + \mathbb{E}\left[J_{h+1}^{\pi_k}(\Theta_*, x_{k,h+1}) \mid \mathcal{F}_{k,h}\right]\\
&- \|x_{k,h}\|_{Q_h} - \|u_{k,h}\|_{R_h} - \sum_{h'=h+1}^{H} \mathbb{E}\left[w_{h'}^\top P_{h'+1}(\widetilde{\Theta}_k) w_{h'}\right]\\
&- \mathbb{E}\left[\|x_{k,h+1}\|_{\widetilde{P}_{k,h+1}} \mid \mathcal{F}_{k,h}\right].
\end{aligned}$$

Let $\widetilde{\psi}_{k,h+1} := \sum_{h'=h+1}^{H} \mathbb{E}\left[w_{h'}^\top P_{h'+1}(\widetilde{\Theta}_k) w_{h'}\right]$, we have

$$\begin{aligned}\Gamma_{k,h} =& \mathbb{E}\left[J_{h+1}^{\pi_k}(\Theta_*, x_{k,h+1}) \mid \mathcal{F}_{k,h}\right] - \widetilde{\psi}_{k,h+1}\\ &- \mathbb{E}\left[\left\|\widetilde{\Theta}_k^\top z_{k,h} + w_{k,h}\right\|_{\widetilde{P}_{k,h+1}} \mid \mathcal{F}_{k,h}\right]\\ =& \Delta_{k,h} + J_{h+1}^{\pi_k}(\Theta_*, x_{k,h+1}) - \widetilde{\psi}_{k,h+1}\\ &- \mathbb{E}\left[\left\|\widetilde{\Theta}_k^\top z_{k,h} + w_{k,h}\right\|_{\widetilde{P}_{k,h+1}} \mid \mathcal{F}_{k,h}\right]\end{aligned}$$

where

$\Delta_{k,h} := \mathbb{E}\left[J_{h+1}^{\pi_k}(\Theta_*, x_{k,h+1}) \mid \mathcal{F}_{k,h}\right] - J_{h+1}^{\pi_k}(\Theta_*, x_{k,h+1})$.

Now, by the assumptions on the noise (i.e., (b) in Assumption 1), we can write $\Gamma_{k,h}$ as follows.

$$\begin{aligned}&\Gamma_{k,h}\\ &\overset{(a)}{=}\Delta_{k,h} + J_{h+1}^{\pi_k}(\Theta_*, x_{k,h+1}) - \widetilde{\psi}_{k,h+1} - \left\|\widetilde{\Theta}_k^\top z_{k,h}\right\|_{\widetilde{P}_{k,h+1}}\\ &\quad - \mathbb{E}\left[\left\|w_{k,h}\right\|_{\widetilde{P}_{k,h+1}} \mid \mathcal{F}_{k,h}\right]\\ &=\Delta_{k,h} + J_{h+1}^{\pi_k}(\Theta_*, x_{k,h+1}) - \widetilde{\psi}_{k,h+1} - \left\|\widetilde{\Theta}_k^\top z_{k,h}\right\|_{\widetilde{P}_{k,h+1}}\\ &\quad - \mathbb{E}\left[\left\|x_{k,h+1} - \Theta_*^\top z_{k,h}\right\|_{\widetilde{P}_{k,h+1}} \mid \mathcal{F}_{k,h}\right]\\ &\overset{(b)}{=}\Delta_{k,h} + J_{h+1}^{\pi_k}(\Theta_*, x_{k,h+1}) - \widetilde{\psi}_{k,h+1} - \left\|\widetilde{\Theta}_k^\top z_{k,h}\right\|_{\widetilde{P}_{k,h+1}}\\ &\quad - \mathbb{E}\left[\left\|x_{k,h+1}\right\|_{\widetilde{P}_{k,h+1}} \mid \mathcal{F}_{k,h}\right] + \left\|\Theta_*^\top z_{k,h}\right\|_{\widetilde{P}_{k,h+1}},\end{aligned}$$

where in (a) and (b), we have used the independence and mean zero of $w_{k,h}$. In order to include the optimal cost term, we observe from (5) that

$$\widetilde{\psi}_{k,h+1} = J_h^*(\widetilde{M}_k, x_{k,h+1}) - \|x_{k,h+1}\|_{\widetilde{P}_{k,h+1}}.$$

Based on this, we can further rewrite $\Gamma_{k,h}$ as

$$\begin{aligned}&\Gamma_{k,h}\\ &=\Delta_{k,h} + J_{h+1}^{\pi_k}(\Theta_*, x_{k,h+1}) - J_{h+1}^*(\widetilde{\Theta}_k, x_{k,h+1})\\ &\quad + \|x_{k,h+1}\|_{\widetilde{P}_{k,h+1}} - \mathbb{E}\left[\|x_{k,h+1}\|_{\widetilde{P}_{k,h+1}} \mid \mathcal{F}_{k,h}\right] \qquad (12)\\ &\quad + \left\|\Theta_*^\top z_{k,h}\right\|_{\widetilde{P}_{k,h+1}} - \left\|\widetilde{\Theta}_k^\top z_{k,h}\right\|_{\widetilde{P}_{k,h+1}} \qquad (13)\\ &=\Delta_{k,h} + \Gamma_{k,h+1} + \Delta'_{k,h} + \Delta''_{k,h}, \qquad (14)\end{aligned}$$

in which (12) is denoted by $\Delta'_{k,h}$ and (13) is denoted by $\Delta''_{k,h}$, respectively.

Finally, due to the fact that the cost for $H+1$ and beyond are zero, we can combine (11) and (14) to obtain that

$$R(K) \le \sum_{k_1}^{K} \sum_{h=1}^{H-1} (\Delta_{k,h} + \Delta'_{k,h} + \Delta''_{k,h}),$$

which completes the proof. $\square$

Proof of Lemma 7. We will bound both terms by using Azuma–Hoeffding inequality. Recall that $\mathcal{F}_{k,h}$ is is all randomness *before* time (k,h), we have

$$\mathbb{E}\left[\Delta_{k,h} \mid \mathcal{F}_{k,h}\right] = 0 \text{ and } \mathbb{E}\left[\Delta'_{k,h} \mid \mathcal{F}_{k,h}\right] = 0.$$

Thus, all we need to show is that both of them are bounded.

$$\begin{aligned}|\Delta'_{k,h}| &= \left|\|x_{k,h+1}\|_{\widetilde{P}_{k,h+1}} - \mathbb{E}\left[\|x_{k,h+1}\|_{\widetilde{P}_{k,h+1}} \mid \mathcal{F}_{k,h}\right]\right|\\ &\le 2C,\end{aligned}$$

which follows from the boundness results in Lemmas 2 and 3. To bound $|\Delta_{k,h}|$, we can bound it backwards by using the assumptions that $\|Q_h\|_2 \le C$ and $\|R_h\|_2 \le C$. First notice that

$$\begin{aligned}|J_H^{\pi_k}(\Theta_*, x_{k,H})| &= \|x_{k,H}\|_{Q_h} + \|u_{k,H}\|_{R_h}\\ &\le (1+\gamma^2)C,\end{aligned}$$

which again uses the boundness result in Lemma 2. Thus, for $h \in [H]$, we have

$$\begin{aligned}|J_h^{\pi_k}(\Theta_*, x_{k,H})| \le& \|x_{k,h}\|_{Q_h} + \|u_{k,H}\|_{R_h}\\ &+ |\mathbb{E}\left[J_{h+1}^{\pi_k}(\Theta_*, x_{k,h+1}) \mid \mathcal{F}_{k,h}\right]|\\ \le& H(1+\gamma^2)C.\end{aligned}$$

Finally, a direct application of Azuma–Hoeffding inequality yields the result. $\square$

Proof of Lemma 8. Note that

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{h=1}^{H} \Delta''_{k,h} \le \sum_{k=1}^{K}\sum_{h=1}^{H} |\Delta''_{k,h}|\\ &= \sum_{k=1}^{K}\sum_{h=1}^{H} \left|\left\|\widetilde{P}_{k,h+1}^{1/2}\Theta_*^\top z_{k,h}\right\|_2^2 - \left\|\widetilde{P}_{k,h+1}^{1/2}\widetilde{\Theta}_k^\top z_{k,h}\right\|_2^2\right|\\ &\le \left[\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\left\|\widetilde{P}_{k,h+1}^{1/2}\Theta_*^\top z_{k,h}\right\|_2 - \left\|\widetilde{P}_{k,h+1}^{1/2}\widetilde{\Theta}_k^\top z_{k,h}\right\|_2\right)^2\right]^{1/2}\\ &\quad \cdot \left[\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\left\|\widetilde{P}_{k,h+1}^{1/2}\Theta_*^\top z_{k,h}\right\|_2 + \left\|\widetilde{P}_{k,h+1}^{1/2}\widetilde{\Theta}_k^\top z_{k,h}\right\|_2\right)^2\right]^{1/2}.\end{aligned}$$

We will bound the two terms by using the following two claims, respectively.

Claim 1. *Under Assumption 1 and the event $\mathcal{E}_K(\delta)$, we have for some constant D*

$$\begin{aligned}&\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\left\|\widetilde{P}_{k,h+1}^{1/2}\Theta_*^\top z_{k,h}\right\|_2 + \left\|\widetilde{P}_{k,h+1}^{1/2}\widetilde{\Theta}_k^\top z_{k,h}\right\|_2\right)^2\right]^{1/2}\\ &\le 2D(1+\gamma)\sqrt{HK}.\end{aligned}$$

Claim 2. *Under Assumption 1 and the event $\mathcal{E}_K(\delta)$, we have for some constant D,*

$$\begin{aligned}&\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\left\|\widetilde{P}_{k,h+1}^{1/2}\Theta_*^\top z_{k,h}\right\|_2 - \left\|\widetilde{P}_{k,h+1}^{1/2}\widetilde{\Theta}_k^\top z_{k,h}\right\|_2\right)^2\right]^{1/2}\\ &\le 2\sqrt{H}D(1+\gamma)\beta_k(\delta)\left(\ln\frac{\det(G_k + \lambda_{min} I)}{\det(\lambda_{min} I)}\right)^{1/2}.\end{aligned}$$

$\square$

E. Proofs for Claims

Proof of Claim 2. Under Assumption 1 and the event $\mathcal{E}_K(\delta)$, we have

$$\begin{aligned}
&\left(\left\|\widetilde{P}_{k,h+1}^{1/2}\Theta_*^\top z_{k,h}\right\|_2 - \left\|\widetilde{P}_{k,h+1}^{1/2}\widetilde{\Theta}_k^\top z_{k,h}\right\|_2\right)^2 \\
&\leq \left\|\widetilde{P}_{k,h+1}^{1/2}(\widetilde{\Theta}_k - \Theta_*)^\top z_{k,h}\right\|_2^2 \\
&\overset{(a)}{\leq} D^2 \left\|(\widetilde{\Theta}_k - \Theta_*)^\top z_{k,h}\right\|_2^2 \\
&\leq D^2 \left\|(\widetilde{\Theta}_k - \Theta_*)^\top \widehat{V}_k^{1/2}\right\|_2^2 \left\|\widehat{V}_k^{-1/2} z_{k,h}\right\|_2^2 \\
&\leq 2D^2\beta_k^2(\delta) \left\|\widehat{V}_k^{-1/2} z_{k,h}\right\|_2^2.
\end{aligned}$$

Combining the result above with the boundeness of $\widetilde{P}_{k,h+1}$, Θ_*, $\widetilde{\Theta}_k$ and $z_{k,h}$, yields

$$\begin{aligned}
&\left\|\widetilde{P}_{k,h+1}^{1/2}(\widetilde{\Theta}_k - \Theta_*)^\top z_{k,h}\right\|_2^2 \\
&\overset{(a)}{\leq} 2D^2(1+\gamma)^2\beta_k^2(\delta) \min\left\{1, \left\|\widehat{V}_k^{-1/2} z_{k,h}\right\|_2^2\right\} \\
&\overset{(b)}{\leq} 2D^2(1+\gamma)^2\beta_k^2(\delta) \min\left\{1, \|z_{k,h}\|^2_{(G_k+\lambda_{min})^{-1}}\right\} \\
&\overset{(c)}{\leq} 4D^2(1+\gamma)^2\beta_k^2(\delta) \ln\left(1 + \|z_{k,h}\|^2_{(G_k+\lambda_{min})^{-1}}\right).
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\sum_{k=1}^K\sum_{h=1}^H \left(\left\|\widetilde{P}_{k,h+1}^{1/2}\Theta_*^\top z_{k,h}\right\|_2 - \left\|\widetilde{P}_{k,h+1}^{1/2}\widetilde{\Theta}_k^\top z_{k,h}\right\|_2\right)^2 \\
&\leq \sum_{k=1}^K\sum_{h=1}^H 8D^2(1+\gamma)^2\beta_k^2(\delta) \ln\left(1 + \|z_{k,h}\|^2_{(G_k+\lambda_{min})^{-1}}\right) \\
&\overset{(a)}{\leq} 4(H)D^2(1+\gamma)^2\beta_k^2(\delta) \ln\frac{\det(G_k+\lambda_{min}I)}{\det(\lambda_{min}I)}.
\end{aligned}$$

Taking the square root complemtes the proof. $\square$

F. Lemmas for Private Control

Lemma 9 (Slepian's inequality; Lemma 5.33 in [28])**.** *Consider two Gaussian processes $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ whose increments satisfy the inequality $\mathbb{E}\left[|X_s - X_t|^2\right] \leq \mathbb{E}\left[|Y_s - Y_t|^2\right]$ for all $s, t \in T$, then, $\mathbb{E}\left[\sup_{t\in T} X_t\right] \leq \mathbb{E}\left[\sup_{t\in T} Y_t\right]$.*

Lemma 10 (Concentration in the Gauss space)**.** *Let f be a real valued 1-Lipschitz function on $\mathbb{R}^n$. Let X be a Gaussian random vector in $\mathbb{R}^n$ whose entries X_i is $i.i.d$ $\mathcal{N}(0,\sigma^2)$. Then, for every $t \geq 0$,*

$$\mathbb{P}\left\{f(X) - \mathbb{E}\left[f(X)\right] \geq t\right\} \leq \exp\left(-\frac{t^2}{2\sigma^2}\right)$$

Proof. This result directly follows from Proposition 2.18 in [29] by considering γ as the Gaussian measure on $\mathbb{R}^n$ with density of a multivariate normal distribution, i.e., $\mathcal{N}(0,\sigma^2 I)$ and hence c in Proposition 2.18 of [29] is equal to $\frac{1}{\sigma^2}$. $\square$

Lemma 11 (Operator norm of Gaussian random matrices)**.** *Let A be an $m \times n$ random matrix whose entries A_{ij} are $i.i.d$ Gaussian random variables $\mathcal{N}(0,\sigma^2)$. Then, for any $t > 0$, we have*

$$\mathbb{P}\{\|A\| \geq \sigma(\sqrt{n} + \sqrt{m}) + t\} \leq \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Proof. We will first show that

$$\mathbb{E}\left[\|A\|\right] \leq \sigma(\sqrt{m} + \sqrt{n})$$

by using Lemma 9. To this end, note that the operator norm of A can be computed as follows.

$$\|A\| = \max_{u\in S^{n-1}, v\in S^{m-1}} \langle Au, v\rangle.$$

where S^{n-1} and s^{m-1} are unit spheres. Thus, $\|A\|$ can be regarded as the supremum of the Gaussian process $X_{u,v} := \langle Au, v\rangle$ indexed by the pair of vectors $(u,v) \in S^{n-1} \times S^{m-1}$. Let us define $Y_{u,v} := \langle g, u\rangle + \langle h, v\rangle$ where $g \in \mathbb{R}^n$ and $h \in \mathbb{R}^m$ are independent Gaussian random vectors whose entries are $i.i.d$ Gaussian random variables $\mathcal{N}(0,\sigma^2)$. Now, we compare the increments of these two Gaussian processes for every $(u,v), (u',v') \in S^{n-1} \times S^{m-1}$.

$$\begin{aligned}
&\mathbb{E}\left[|X_{u,v} - X_{u',v'}|^2\right] \\
&= \sigma^2 \sum_{i=1}^n\sum_{j=1}^m |u_i v_j - u'_i v'_j|^2 \\
&= \sigma^2 \left\|uv^T - u'v'^T\right\|_F^2 \\
&= \sigma^2 \left\|(u-u')v^T + u'(v-v')^T\right\|_F^2 \\
&= \sigma^2 \left(\|u-u'\|_2^2 + \|v-v'\|_2^2\right) \\
&\quad + 2\sigma^2 \text{trace}\left(v(u-u')^T u'(v-v')^T\right) \\
&= \sigma^2 \left(\|u-u'\|_2^2 + \|v-v'\|_2^2 + 2(u^T u' - 1)(1 - v^T v')\right) \\
&\leq \sigma^2 \left(\|u-u'\|_2^2 + \|v-v'\|_2^2\right) \\
&= \mathbb{E}\left[|Y_{u,v} - Y_{u',v'}|^2\right].
\end{aligned}$$

Therefore, Lemma 9 applies here, and hence

$$\begin{aligned}
\mathbb{E}\left[\|A\|\right] &= \mathbb{E}\left[\max_{(u,v)} X_{u,v}\right] \leq \mathbb{E}\left[\max_{(u,v)} Y_{u,v}\right] \\
&= \mathbb{E}\left[\|g\|_2\right] + \mathbb{E}\left[\|h\|_2\right] \leq \sigma(\sqrt{m} + \sqrt{n}).
\end{aligned}$$

Finally, note that $\|A\|$ is a 1-Lipschitz function of A when considered as a vector in $\mathbb{R}^{mn}$. Then, consider the function f in Lemma 10 as the operator norm, we directly obtain the result. $\square$

Lemma 12 (Operator norm of symmetric Gaussian random matrices)**.** *Let A be an $n \times n$ random symmetric matrix whose entries A_{ij} on and above the diagonal are $i.i.d$ Gaussian random variables $\mathcal{N}(0,\sigma^2)$. Then, for any $t > 0$, we have*

$$\|A\| \leq 4\sigma\sqrt{n} + 2t,$$

with probability at least $1 - 2\exp(-\frac{t^2}{2\sigma^2})$.

Proof. Decompose the matrix A into the upper triangular matrix A^+ and the lower triangular matrix A^- such that $A = A^+ + A^-$. Note that without loss of generality, the entries on the diagonal are in A^+. Then, apply Lemma 11 to both A^+ and A^-, and by a union bound, we have that the following inequalities hold simultaneously

$$\left\|A^+\right\| \leq 2\sigma\sqrt{n} + t \quad \text{and} \quad \left\|A^-\right\| \leq 2\sigma\sqrt{n} + t$$

with probability at least $1 - 2\exp(-\frac{t^2}{2\sigma^2})$ for any $t > 0$. Finally, by triangle inequality $\|A\| \leq \|A^+\| + \|A^-\|$, we prove the result. $\square$

Lemma 13 (Concentration of chi-square; Corollary to Lemma 1 of [27]). *Let U be a χ^2 statistic with D degrees of freedom. Then, for any positive x,*

$$\mathbb{P}\left\{U \geq D + 2\sqrt{Dx} + 2x\right\} \leq \exp(-x).$$