Rényi Differentially Private Bandits

Achraf Azize,[1] **Debabrota Basu** [1]

[1] Équipe Scool, Univ. Lille, Inria,
CNRS, Centrale Lille, UMR 9189- CRIStAL
F-59000 Lille, France
achraf.azize@inria.fr, debabrota.basu@inria.fr

Abstract

Rényi Differential Privacy (RDP) is a popular and practical relaxation of pure Differential Privacy (DP) studied in distributed and deep learning settings. In this paper, we study algorithms to ensure Rényi DP in different bandit problems. Bandits serve as a theoretical foundation of sequential learning, and also as an algorithmic foundation of modern recommender systems. As recommenders often involve the private data of users, it motivates the study of private bandits. Specifically, we study three settings, i.e. finite-armed bandits, linear bandits, and linear contextual bandits, where we render the sequence of observed rewards Rényi DP with a centralised decision-maker. For each of these settings, we propose a Rényi DP bandit algorithm sharing similar algorithmic ingredients, namely the Gaussian mechanism and adaptive episodes. Further, we analyse the *regret* of the three algorithms. Our analysis shows that the prices of imposing Rényi DP in all of these settings are negligible in comparison with the regrets incurred oblivious to privacy. Specifically, for a horizon T and RDP budget (α, ϵ), the regret due to RDP is $\tilde{\mathcal{O}}(\sqrt{\frac{\alpha}{\epsilon}}\log(T))$, while the regret independent of privacy is $\tilde{\mathcal{O}}(\sqrt{T})$.

1 Introduction

Multi-armed bandit (in brief, bandits) (Lattimore and Szepesvári 2018) is the archetypal setting of reinforcement learning consisting of K actions and an unknown underlying state. Here, each action $a \in [K]$ corresponds to a distribution over rewards $r \in \mathbb{R}$ with mean reward μ_a. A bandit policy π must choose, at each time-step t, an action (or arm) a_t and receives a reward r_t from the reward distribution corresponding to a_t. The goal of the policy is to maximize the cumulative reward $\sum_{t=1}^{T} r_t$. This is the simplest setting of sequential decision-making that encounters the exploration–exploitation dilemma. The policy has to choose between exploring arms about which it knows little, and exploiting the arms that currently appears to have maximal mean reward. Bandits are widely used to address a wide range of sequential decision-making tasks under uncertainty, such as recommender systems (Silva et al. 2022), strategic pricing (Bergemann and Välimäki 1996) or clinical trials (Thompson 1933) to name a few. These applications often involve individuals' sensitive data, such as health conditions, personal preferences, financial situation, and thus, naturally invoke data privacy concerns.

Algorithm 1: Interaction Protocol in Bandits

```
Input: A policy $\pi$ and Users $u_1, \dots, u_T$
Output: Actions satisfying $(\alpha,\epsilon)$-global RDP
for $t = 1, \dots, T$ do
    $u_t$ sends context $c_t$ to $\pi$ (if available)
    $\pi$ recommends action $a_t$
    $u_t$ sends the sensitive reward $r_t$ to $\pi$
end for
```

Example 1. *As a motivating example, we consider the problem of vaccine recommendation. Each day, a new patient u_t arrives, to whom a vaccine a_t is recommended. While recommending a vaccine a_t, the bandit policy might either consider the specific medical conditions (or context) c_t corresponding to u_t, or ignore it. Then, a reaction to the vaccine is observed. If the vaccine works, the observed reward $r_t = 1$, otherwise $r_t = 0$. This reward information can reveal sensitive information about the health condition of patient u_t. Thus, the goal of the bandit policy is to recommend a sequence of vaccines (actions) that cures the maximum number of patients while protecting the privacy of these patients.*

In this paper, we adhere to Differential Privacy (DP) framework to ensure data privacy of users. DP (Dwork, Roth et al. 2014) is the gold standard of privacy-preserving data analysis in both academia and industry, requiring that an algorithm's output have a limited dependency on the presence of any single user. Rényi DP (Mironov 2017) is a popularly deployed relaxation of DP that shares similar properties as those of DP while allowing tighter analysis of composite mechanisms.

Related Works. Privacy issues have been studied for bandits under different settings, such as stochastic bandits (Mishra and Thakurta 2015; Tossou and Dimitrakakis 2016; Sajed and Sheffet 2019; Azize and Basu 2022; Hu and Hegde 2022), adversarial bandits (Tossou and Dimitrakakis 2017), and linear contextual bandits (Shariff and Sheffet 2018; Neel and Roth 2018; Hanna et al. 2022). Also, multiple formulations of DP, namely *local* and *global*, are extended to bandits (Basu, Dimitrakakis, and Tossou 2019). *Local DP* aims to preserve the privacy of a sequence of observed rewards by sending noisy rewards to the bandit policy (Duchi, Jordan, and Wainwright 2013). Though local DP provides stronger privacy as the data curator has no access to the original reward stream, it injects more noise leading to higher regret. Also, the fundamental hardness of ensuring local DP in bandits in terms of regret lower bound and also the corresponding optimal algorithms are well-understood (Zheng et al. 2020). *Global DP* allows the bandit policy to access rewards without noise. In global DP, one aims to keep the sequence of observed rewards private while the sequence of actions taken by the policy is publicly revealed (Basu, Dimitrakakis, and Tossou 2019). Here, *we focus on the global DP setting.*

The existing works on private bandits consider either pure ϵ-DP or (ϵ, δ)-DP as the privacy framework. Only (Chowdhury and Zhou 2022) studies Rényi Differential Privacy in a distributed bandit setting by adding a Skellam noise. In this paper, *we investigate even more fundamental bandit settings and aim to ensure (α, ϵ)-global RDP*. The main question that we aim to address is:

What is the additional cost in the regret due to imposing Rényi Differential Privacy in (α, ϵ)-global RDP bandits?

Our Contributions. First, we formally define (α, ϵ)-global RDP bandits. Following that, we study the cost of RDP in terms of regret for three different settings, namely stochastic bandits with finitely many actions (Section 4), stochastic linear bandits with (fixed) finitely many actions (Section 5) and linear contextual bandits with context-dependent feasible actions (Section 6).

For each setting, we propose an algorithm that achieves (α, ϵ)-global RDP, almost for free. These three algorithms share the same blueprint. First, they add a calibrated *Gaussian noise* to the reward statistics each time they interact with the private reward sequence. Second, they run in *adaptive episodes*, with the number of episodes logarithmic in T. This means that the algorithm only accesses the private reward sequence in $\log(T)$ time-steps, rather than accessing it at each step. A lower number of interactions leads to a less sensitive estimate of reward statistics, and thus lower addition of Gaussian noise.

We further show that imposing (α, ϵ)-global RDP is almost free in terms of regret of the bandit algorithms. Specifically, the cost of (α, ϵ)-global RDP in the regret of these algorithms is shown to be $\tilde{\mathcal{O}}(\sqrt{\frac{\alpha}{\epsilon}} \log(T))$, which is significantly lower than the regret oblivious to privacy, i.e. $\tilde{\mathcal{O}}(\sqrt{T})$. In Table 1, we summarise the corresponding regret upper-bounds. Another interesting consequence of our analysis is that the gap-dependent regret of AdaR-UCB incurs an additive term $\mathcal{O}\left(\sqrt{\frac{\alpha}{\epsilon}} \log(T)\right)$ in regret due to (α, ϵ)-global RDP. In contrast, the price of ϵ-global DP for the same algorithms is $\Omega\left(\frac{\log(T)}{\epsilon}\right)$. This result indicates a fundamental difference between ϵ-global DP and (α, ϵ)-global RDP that we aim to understand in this work.

2 Background: Rényi Differential Privacy

Differential Privacy (DP) renders an individual corresponding to a datapoint indistinguishable by constraining the output of an algorithm to remain almost the same under a change in one input datapoint.

Definition 1 ((ϵ, δ)-DP (Dwork, Roth et al. 2014) and (α, ϵ)-RDP (Mironov 2017)). *A mechanism $\mathcal{M}$, that assigns to each dataset d a probability distribution $\mathcal{M}_d$ on some measurable space $(\mathbb{X}, \mathcal{F})$, is*

- (ϵ, δ)*-DP for a given $\delta \in [0, 1)$ if*

$$\sup_{A \in \mathcal{F}, d \sim d'} \mathcal{M}_d(A) - e^{\epsilon} \mathcal{M}_{d'}(A) \leq \delta. \quad (1)$$

- (α, ϵ)*-RDP for a given $\alpha > 1$, if*

$$\sup_{d \sim d'} D_{\alpha}(\mathcal{M}_d \| \mathcal{M}_{d'}) \leq \epsilon. \quad (2)$$

Here, two datasets d and d' are said to be neighbouring (denoted by $d \sim d'$) if their Hamming distance is one. $D_{\alpha}(P \| Q) \triangleq \frac{1}{\alpha - 1} \log \mathbb{E}_Q \left[\left(\frac{dP}{dQ} \right)^{\alpha} \right]$ denotes the Rényi divergence of order α between P and Q.

Rényi DP (RDP) was initiated by (Abadi et al. 2016) in the moments' accountant method for the Gaussian mechanism, and then extensively studied in (Mironov 2017; Dwork and Rothblum 2016; Bun and Steinke 2016; Bun et al. 2018).

The Gaussian mechanism (Dwork, Roth et al. 2014; Mironov 2017) ensures (α,ϵ)-RDP by injecting a random noise to the output of the algorithm that is sampled from a calibrated Gaussian distribution (as specified in Theorem 2).

Theorem 2 ((α,ϵ)-RDP of The Gaussian Mechanism (Corollary 3, (Mironov 2017))). *Let f be a mechanism in $\mathbb{R}^d$ with L_2 sensitivity $s(f) \triangleq \max_{d \sim d'} \| f(d) - f(d') \|_2$. Then $f + Z$ is (α, ϵ)-RDP where $Z \sim \mathcal{N}(0, \frac{\alpha s(f)^2}{2\epsilon} I_d)$. We use $\mathcal{N}(\mu, \Sigma)$ to denote the Gaussian distribution with mean μ and covariance matrix Σ, and $\| \cdot \|_2$ to denote the L_2 norm on $\mathbb{R}^d$.*

Gaussian mechanism ensures RDP when the input database is static. In a sequential setting like bandits, a mechanism must update the published statistics as new data arrives (Dwork et al. 2010a; Basu, Dimitrakakis, and Tossou 2019), and thus, we extend the RDP definitions accordingly.

3 Bandits with (α, ϵ)-Global RDP

A bandit algorithm (or policy) interacts with an environment ν consisting of K arms with reward distributions $\{\nu_a\}_{a=1}^{K}$ for a given horizon $T \in \mathbb{N}$ and produces a history $\mathcal{H}_T \triangleq \{(A_t, R_t)\}_{t=1}^{T}$. At each step t, the choice of the arm depends on the previous history $\mathcal{H}_{t-1}$. The reward R_t is sampled from the reward distribution ν_{A_t} and is conditionally independent of the previous history $\mathcal{H}_{t-1}$. π can be represented by a sequence $(\pi_t)_{t=1}^{T}$, where $\pi_t : \mathcal{H}_{t-1} \to [K]$ is a probability kernel. Thus, if we denote a sequence of actions as $A \triangleq [a_1, \dots, a_T]$, a sequence of rewards as $R \triangleq [r_1, \dots, r_T]$, then the policy could be seen as a (randomized) mechanism that takes as input the sensitive reward dataset R and outputs a sequence of actions A with probability $\mathcal{M}_R^{\pi}(A) \triangleq \prod_{t=1}^{T} \pi_t(a_t \mid a_1, r_1, \dots, a_{t-1}, r_{t-1})$. To define DP in bandits, we extend the event-level privacy under continuous observations framework (Dwork et al. 2010a).

Table 1: Regret upper bounds for Rényi DP bandits. Terms in blue correspond to the cost of Rényi DP

Setting	Regret Upper Bound	Reference
Finite-arm Rényi DP Bandits	$\mathcal{O}\left(\sqrt{KT\log(T)}\right) + \mathcal{O}\left(K\sqrt{\frac{\alpha}{\epsilon}}\log(T)\right)$	Corollary 6
Linear Rényi DP Bandits	$\mathcal{O}\left(\sqrt{dT\log(KT)}\right) + \mathcal{O}\left(\sqrt{\frac{\alpha}{\epsilon}}d^2(\log(KT))^{\frac{3}{2}}\right)$	Theorem 8
Linear Contextual Rényi DP Bandits	$\mathcal{O}\left(d\log(T)\sqrt{T}\right) + \mathcal{O}\left(\sqrt{\frac{\alpha}{\epsilon}}d^2\log(T)^2\right)$	Theorem 10

Algorithm 2: AdaR-UCB

```
1: Input: Privacy budget $(\alpha,\epsilon)$, an environment $\nu$ with $K$ arms, optimism parameter $\beta > 3$
2: Output: Actions satisfying $(\alpha,\epsilon)$-global RDP
3: Initialisation: Choose each arm once and let $t = K$
4: for $\ell = 1, 2, \dots$ do
5:     Let $t_\ell = t + 1$
6:     Compute $A_\ell = \mathrm{argmax}_a \, \mathrm{I}_a^{\alpha,\epsilon}(t_\ell - 1, \beta)$ (Eq. (6))
7:     Choose arm $A_\ell$ until round t such that $N_{A_\ell}(t) = 2N_{A_\ell}(t_\ell - 1)$
8: end for
```

Definition 3. *A policy* $\pi = (\pi_t)_{t=1}^T$ *is*

- *ϵ-global DP if*

$$\sup_{A\in[K]^T, R\sim R'} \mathcal{M}_R^\pi(A) - e^\epsilon \mathcal{M}_{R'}^\pi(A) \leq 0. \quad (3)$$

- *(α, ϵ)-global RDP for a given $\alpha > 1$, if*

$$\sup_{R\sim R'} D_\alpha(\mathcal{M}_R^\pi \| \mathcal{M}_{R'}^\pi) \leq \epsilon. \quad (4)$$

Here, $R \sim R'$ denotes that the reward sequences are neighbouring, i.e they only differ on a single time-step t.

Remark. For a given R, $\mathcal{M}_R^\pi$ is a probability distribution on $([K]^T, \mathcal{P}([K]^T))$, i.e $\sum_{A\in[K]^T} \mathcal{M}_R^\pi(A) = 1$.

In the following three sections, we consider three bandit settings, namely stochastic bandits with finitely many arms, stochastic linear bandits with (fixed) finitely many arms and linear contextual bandits with context-dependent feasible arms. For all of these bandit settings, we impose (α, ϵ)-global RDP (Definition 3) as the privacy constraint, where the rewards are the private data to protect.

4 Stochastic Bandits with (α, ϵ)-global RDP

In this section, we first specify the stochastic bandits with (α, ϵ)-global RDP. Then, we provide an (α, ϵ)-global RDP algorithm, namely AdaR-UCB, and analyse its performance (regret) to quantify the cost of (α, ϵ)-global RDP.

Formulation. Let $\nu = (P_a : a \in [K])$ a bandit instance with K arms and means $(\mu_a)_{a\in[K]}$. The goal is to design a (α, ϵ)-global RDP policy that maximizes the cumulative reward or equivalently minimizes regret over a horizon T:

$$\mathrm{Reg}_T(\pi, \nu) \triangleq T\mu^\star - \mathbb{E}\left[\sum_{t=1}^T R_t\right] = \sum_{a=1}^K \Delta_a \mathbb{E}\left[N_a(T)\right]. \quad (5)$$

Here, $\mu^\star \triangleq \max_{a\in[K]} \mu_a$ is the mean of the optimal arm $a^\star$. $\Delta_a \triangleq \mu^\star - \mu_a$ is the sub-optimality gap of the arm a. $N_a(T) \triangleq \sum_{t=1}^T \mathbb{1}\left\{A_t = a\right\}$ is the number of times the arm a is played till T, where the expectation is taken both on the randomness of the environment ν and the policy π.

Algorithm: AdaR-UCB. Now, we propose AdaR-UCB algorithm. To design AdaR-UCB, we first define the private index to select the arms (Line 6 of Algorithm 2):

$$\mathrm{I}_a(t_\ell - 1, \beta) \triangleq \hat{\mu}_a^\ell + \mathcal{N}\left(0, \sigma_{a,\ell}^2\right) + B_a(t_\ell - 1, \beta). \quad (6)$$

Here, $\hat{\mu}_a^\ell$ is the empirical mean of rewards computed using only samples from the last episode arm a was played. $\sigma_{a,\ell}^2 \triangleq \frac{\alpha}{2\epsilon\times\left(\frac{1}{2}N_a(t_\ell-1)\right)^2}$ is the noise calibrated using Theorem 2 to make the empirical mean (α, ϵ)-RDP and finally the exploration bonus is defined as $B_a(t_\ell - 1, \beta) \triangleq \sqrt{\left(\frac{1}{2\times\frac{1}{2}N_a(t_\ell-1)} + \frac{\alpha}{\epsilon\times\left(\frac{1}{2}N_a(t_\ell-1)\right)^2}\right)\beta\log(t_\ell)}$.

AdaR-UCB is an extension of the generic algorithmic wrapper proposed in (Azize and Basu 2022), which turns any index-based bandit algorithm ϵ-global DP, to the (α, ϵ)-global RDP setting. Following (Azize and Basu 2022), AdaR-UCB relies on three ingredients: arm-dependent doubling, forgetting, and adding calibrated Gaussian noise. First, the algorithm runs in episodes. In each episode, the same arm is played for double the number of times it was last played. Second, at the beginning of a new episode, the index of an arm a, as defined in Eq. (6), is computed only using samples from the last episode arm a was played and forgetting all the other samples. Due to these two ingredients, each empirical mean computed in the index of Eq. (6) only needs to be (α, ϵ)-RDP so that the whole sequence of released empirical means is (α, ϵ)-RDP. We formalise this intuition in Lemma 11 of Appendix A.

Theorem 4 (Privacy of AdaR-UCB)**.** *For rewards in* $[0, 1]$, AdaR-UCB *satisfies* (α, ϵ)*-global RDP.*

Proof Sketch. The main idea is that a change in reward *only affects* the empirical mean calculated in one episode, which is made private using the Gaussian Mechanism and Lemma 11. Since the actions are only calculated using the private empirical means, AdaR-UCB is (α, ϵ)-global RDP following the post-processing lemma. We refer to Appendix A for the complete proof.

Regret Analysis. We derive both gap-dependent and gap-independent upper bounds on regret of AdaR-UCB and discuss the additive cost due to (α, ϵ)-global RDP.

Theorem 5 (Gap-dependent Regret)**.** *For rewards in* $[0,1]$ *and* $\beta > 3$*,* AdaR-UCB *yields a regret upper bound of*

$$\sum_{a:\Delta_a>0}\left(\frac{8\beta}{\Delta_a}\log(T)+8\sqrt{\frac{\beta\alpha}{\epsilon}}\sqrt{\log(T)}+\frac{2\beta}{\beta-3}\right).$$

Corollary 6 (Gap-independent Regret)**.** *For rewards in* $[0,1]$ *and* $\beta > 3$*,* AdaR-UCB *yields a regret upper bound of*

$$\mathcal{O}\left(\sqrt{KT\log(T)}\right)+\mathcal{O}\left(K\sqrt{\frac{\alpha}{\epsilon}\log(T)}\right).$$

Proof Sketch. The proof is a direct application of Theorem 12 in (Azize and Basu 2022), which is a generic regret decomposition for adaptive phased index-based bandit algorithms, with forgetting. The difference comes from the form of the index, with a different privacy bonus, that comes from the concentrated Gaussian Mechanism specific to Rényi DP. We refer to Appendix B for the complete proof.

Discussion: RDP for Almost-free. The bound of Theorem 5 shows that, in the gap-dependent regret, the price of (α, ϵ)-global RDP is the additive term $\mathcal{O}\left(\sqrt{\frac{\alpha}{\epsilon}\log(T)}\right)$. For a fixed budget (α, ϵ), this additive term is negligible in comparison to the non private part $\mathcal{O}\left(\frac{\log(T)}{\Delta}\right)$ as the horizon $T \to \infty$. This is in sharp contrast with the ϵ-global DP setting, where the price of DP is $\Omega\left(\frac{\log(T)}{\epsilon}\right)$ (Shariff and Sheffet 2018).

5 Linear Bandits with (α, ϵ)-global RDP

In this section, we study (α, ϵ)-global RDP for linear bandits with finite number of arms.

Formulation. Similar to the formulation of Section 4, here, we consider that a fixed set of actions $\mathcal{A} \subset \mathbb{R}^d$ is available at each round such that $|\mathcal{A}| = K$. In addition, the reward is not just a sample from distribution but is generated by a linear structural equation (Lattimore and Szepesvári 2018). Specifically, at step t, the observed reward

$$R_t \triangleq \langle \theta^\star, A_t \rangle + \eta_t, \tag{7}$$

where η_t is a conditionally 1-subgaussian noise, i.e. $\mathbb{E}\left[\exp\left(\lambda\eta_t\right) \mid A_1, \eta_1, \dots, A_{t-1}\right] \leq \exp\left(\lambda^2/2\right)$ almost surely for all $\lambda \in \mathbb{R}$.

Assumption 1 (Boundedness)**.** *We assume that all the quantities of interest are bounded.*

1. *Actions are bounded:* $\|a\|_2 \leq 1$ *for all* a *in* $\mathcal{A}$*.*
2. *Rewards are bounded:* $|R_t| \leq 1$*.*
3. *The unknown parameter is bounded:* $\|\theta^\star\|_2 \leq 1$

Similar to Section 4, the goal is to design a (α, ϵ)-global RDP policy that minimizes regret (Eq. (5)). However, the particularity of this setup is that an optimal policy should take advantage of the structure existing between arms to get rid of the polynomial dependence of regret on K.

Algorithm. In order to design a near-optimal algorithm with RDP, we propose an (α, ϵ)-global RDP extension of the G-Optimal design-based Phased Elimination (GOPE) algorithm (Algorithm 12 (Lattimore and Szepesvári 2018)), namely AdaR-GOPE. We state the pseudocode of AdaR-GOPE in Algorithm 3.

Algorithm 3: AdaR-GOPE

```
Input: Privacy budget $(\alpha, \epsilon)$, $\mathcal{A} \subset \mathbb{R}^d$ and $\delta$
Output: Actions satisfying $(\alpha,\epsilon)$-global RDP
Initialisation: Set $\ell = 1$, $t_1 = 1$ and $\mathcal{A}_1 = \mathcal{A}$
for $\ell = 1, 2, \dots$ do
    $\beta_\ell \leftarrow 2^{-\ell}$
    Step 1: Find the $G$-optimal design $\pi_\ell$ for $\mathcal{A}_\ell$:
        $\max_{\substack{\pi \in \mathcal{P}(\mathcal{A}_\ell) \\ \sum_{a \in \mathcal{A}_\ell} \pi(a)=1,\ |\mathrm{Supp}(\pi)| \leq d(d+1)/2}} \log\det V(\pi).$
    Step 2: Choose each action $a \in \mathcal{A}_\ell$ for $T_\ell(a)$ times
        where $T_\ell(a)$ is defined by Eq 8.
    Observe rewards $\{R_t\}_{t=t_\ell}^{t_\ell + \sum_a T_\ell(a)}$
    $T_\ell \leftarrow \sum_{a \in \mathcal{A}_\ell} T_\ell(a)$ and $t_{\ell+1} \leftarrow t_\ell + T_\ell + 1$
    Step 3: Estimate the parameter as
        $\hat{\theta}_\ell = V_\ell^{-1} \sum_{t=t_\ell}^{t_{\ell+1}-1} A_t R_t$ with $V_\ell = \sum_{a \in \mathrm{Supp}(\pi_\ell)} T_\ell(a) a a^\top$
    Step 4: Make the parameter estimate private
        $\tilde{\theta}_\ell = \hat{\theta}_\ell + V_\ell^{-1} \sum_{a \in \mathrm{Supp}(\pi_\ell)} a N_a$
        where $(N_a)_{a \in \mathrm{Supp}(\pi_\ell)} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \frac{2\alpha}{\epsilon}\right)$.
    Step 4: Eliminate low rewarding arms:
        $\mathcal{A}_{\ell+1} = \left\{a \in \mathcal{A}_\ell : \max_{b \in \mathcal{A}_\ell} \left\langle \tilde{\theta}_\ell, b - a \right\rangle \leq 2\beta_\ell \right\}.$
end for
```

At the end of each phase of AdaR-GOPE, arms that are likely to be sub-optimal, i.e. the ones with a gap exceeding the current target (β_ℓ), are eliminated. The elimination criterion only depends on the samples collected in the current phase. In addition, the actions to be played during a phase are chosen based on the solution of an optimal design problem to minimise the number of required samples to eliminate arms that are sub-optimal.

In particular, if π_ℓ is the G-optimal solution for $\mathcal{A}_\ell$ at phase ℓ, then each action $a \in \mathcal{A}_\ell$ is played $T_\ell(a)$ times, where

$$T_\ell(a) \triangleq \left\lceil \frac{8d\pi_\ell(a)}{\beta_\ell^2}\log\left(\frac{4K\ell(\ell+1)}{\delta}\right) + \frac{2d\pi_\ell(a)}{\beta_\ell}\sqrt{\frac{2\alpha}{\epsilon}d(d+1)\log\left(\frac{4K\ell(\ell+1)}{\delta}\right)}\right\rceil \tag{8}$$

The samples collected in the present phase do not influence which actions are played in it. This decoupling has two advantages: (a) It allows us to make use of the tighter confidence bounds available in the fixed design setting (Appendix C.2), and (b) use Lemma 11 to make the algorithm private.

Theorem 7 (Privacy of Algorithm 3)**.** *Under Assumption 1, AdaR-GOPE (Algorithm 3) satisfies (α, ϵ)-global RDP.*

Proof Sketch. A change in reward at any time-step only affects the estimate $\hat{\theta}_\ell$ in the corresponding phase. By making each $\hat{\theta}_\ell$ (α, ϵ)-RDP with respect to the sequence of rewards observed in the corresponding phase, the whole sequence of released estimates $(\hat{\theta}_\ell)_\ell$ becomes (α, ϵ)-RDP, due to Lemma 11. Since the action selection only depends on $(\hat{\theta}_\ell)_\ell$, the algorithm is (α, ϵ)-global RDP by the post-processing lemma. We refer to Appendix A for the complete proof.

Regret Analysis. Now, we quantify the additional cost incurred by AdaR-GOPE due to (α, ϵ)-global RDP.

Theorem 8. *Under Assumption 1 and for $\delta \in (0, 1)$, with probability at least $1 - \delta$, the regret R_T of AdaR-GOPE (Algorithm 3) is upper-bounded by*

$$C_1\sqrt{dT\log\left(\frac{K\log(T)}{\delta}\right)}+C_2d^2\sqrt{\frac{\alpha}{\epsilon}\log\left(\frac{K\log(T)}{\delta}\right)}\log(T)$$

where C_1 and C_2 are universal constants. If $\delta = \frac{1}{T}$, then $\mathbb{E}(R_T) \leq \mathcal{O}\left(\sqrt{dT\log(KT)}\right) + \mathcal{O}\left(\sqrt{\frac{\alpha}{\epsilon}}d^2(\log(KT))^{\frac{3}{2}}\right)$

Proof Sketch. Under the "good event" that all the private parameters $\tilde{\theta}_\ell$ are well estimated, we show that the optimal action never gets eliminated. But the sub-optimal arms get eliminated as soon as the elimination threshold is smaller than their sub-optimality gaps. The regret upper bound follows directly. We refer to Appendix C for complete proof.

Discussions. Here, we discuss briefly about the regret bound.

1. *RDP for Almost-free:* Theorem 8 shows that the price of (α, ϵ)-global RDP is the additive term $\tilde{\mathcal{O}}\left(\sqrt{\frac{\alpha}{\epsilon}}d^2\right)$. For a fixed RDP budget (α, ϵ), the regret due to privacy becomes negligible in comparison with the privacy-oblivious term in regret, i.e. $\tilde{\mathcal{O}}\left(\sqrt{dT}\right)$, as $T \to \infty$.
2. *Extension to ϵ-global DP:* An ϵ-global DP version of Algorithm 3 could easily be given, by changing the Gaussian noise with a calibrated Laplace noise. The regret bound will be similar to the one in Theorem 8, up to added multiplicative $\log(T)$ terms in the private part, due to the concentration of the sum of Laplace noise.
3. *Related Regret Bounds.* A similar algorithm achieving ϵ-global DP is proposed in (Algorithm 1, (Hanna et al. 2022)). However, their algorithm has two shortcomings. First, the algorithm does not retrieve the optimal non-private regret bound for this setting. The non-private regret bound in their analysis is $\tilde{\mathcal{O}}\left(d\sqrt{T}\right)$ (ref. Eq.18 (Hanna et al. 2022)), which is $\sqrt{d}$ loose compared to the regret bound $\tilde{\mathcal{O}}\left(\sqrt{dT}\right)$ achievable by a non-private algorithm. As explained in Section 22 of (Lattimore and Szepesvári 2018), the main reason for proposing a G-optimal elimination algorithm is to obtain a $\sqrt{d}$ dependency in the regret, rather than a d dependency of typical optimism-based strategies.

   This leads to the second shortcoming: an added fine-tuned parameter, q, as in Line 3 of Algorithm 1 (Hanna et al. 2022). q depends on the horizon T and dictates the length of the episodes. This makes the algorithm not anytime, in contrast to AdaR-GOPE, which is anytime. Their algorithm encounters this issue due to a non-tight regret analysis, that depends on the fine-tuned parameter q.

6 Linear Contextual Bandits with (α, ϵ)-global RDP

Now, we consider an even more general setting of bandits, where the feasible arms at each step depend on a context observed at that set. Thus, the set of feasible arms change over steps. We study (α, ϵ)-global RDP in this problem, called linear contextual bandits.

Formulation. Contextual bandits generalise the finite-armed bandits by allowing the learner to use side-information. At each step t, the policy observes a context $c_t \in \mathcal{C}$, which might be random or not. Having observed the context, the policy chooses an action $A_t \in [K]$ and observes a reward R_t. For the linear contextual bandits, the reward R_t depends on both the arm a_t and the context c_t in terms of a linear structural equation:

$$R_t = \langle \theta^\star, \psi(a_t, c_t) \rangle + \eta_t. \tag{9}$$

Here, $\psi : [K] \times \mathcal{C} \to \mathbb{R}^d$ is the feature map, $\theta^\star \in \mathbb{R}^d$ is the unknown parameter, and η_t is the noise, which we assume to be conditionally 1-subgaussian.

Under Eq. (9), all that matters is the feature vector that results in choosing a given action rather than the identical the action itself. This justifies studying a reduced model: in round t, the policy is served with the decision set $\mathcal{A}_t \subset \mathbb{R}^d$, from which it chooses an action $a_t \in \mathcal{A}_t$ and receives a reward

$$R_t = \langle \theta^\star, a_t \rangle + \eta_t$$

where η_t is 1-subgaussian given $\mathcal{A}_1, A_1, R_1, \ldots, \mathcal{A}_{t-1}, A_{t-1}, R_{t-1}, \mathcal{A}_t$, and A_t.

Different choices of $\mathcal{A}_t$ lead to different settings. If $\mathcal{A}_t = \{\psi(c_t, a) : a \in [K]\}$, then we have a contextual linear bandit. On the other hand, if $\mathcal{A}_t = \{e_1, \ldots, e_d\}$ where $(e_i)_i$ are the unit vectors of $\mathbb{R}^d$ then the resulting bandit problem reduces to the stochastic finite-armed bandit of Section 4.

The goal is to design a (α, ϵ)-global RDP policy that minimizes the regret, which is defined as

$$\hat{R}_T \triangleq \sum_{t=1}^{T} \max_{a \in \mathcal{A}_t} \langle \theta^\star, a - a_t \rangle, \quad R_T \triangleq \mathbb{E}[\hat{R}_T]$$

We suppose that Assumption 1 also holds in this setting.

Remark 1. *In this section, we suppose that c_t is **public** information, and thus $\mathcal{A}_t$ is too. Rewards are the only private statistics to protect. The main difference compared to Section 5 is that the set of actions $\mathcal{A}_t$ is allowed to change at each time-step t. Thus, the action-elimination based strategies, as used in Section 5, are not useful.*

Stochastic contexts. In this section, we have an additional assumption on context generation. Specifically, we adopt the same assumption as in (Gentile, Li, and Zappella 2014), i.e. the contexts are stochastically generated.

Assumption 2 (Stochastic Contexts)**.** *At each step t, the context set $\mathcal{A}_t \triangleq \{a_1^t, \dots, a_{k_t}^t\}$ is generated conditionally i.i.d (conditioned on k_t and the history $H_t \triangleq \{\mathcal{A}_1, A_1, X_1, \dots, \mathcal{A}_{t-1}, A_{t-1}, X_{t-1}, \mathcal{A}_t, A_t\}$) from a random process A such that*

- $\|A\|_2 = 1$
- $\mathbb{E}[AA^T]$ *is full rank, with minimum eigenvalue* $\lambda_0 > 0$
- $\forall z \in \mathbb{R}^d$, $\|z\|_2 = 1$, *the random variable* $(z^T A)^2$ *is conditionally subgaussian, with variance*

$$\nu_t^2 \triangleq \mathbb{V}\left[(z^T A)^2 \mid k_t, H_t\right] \leq \frac{\lambda_0^2}{8\log(4k_t)}$$

This additional assumption gives us explicit control on the minimum eigenvalue of the design matrix $V_t \triangleq \sum_{t'=1}^{t} A_{t'} A_{t'}^T$. Using Lemma 27 on the minimum eigenvalue, we quantify more precisely the effect of the added noise due to (α, ϵ)-global RDP and derive tighter confidence bounds.

Algorithm. One of the popular and well-analysed algorithm for non-private contextual bandits is the Rarely Switching OFUL (Optimism in Face of Uncertainty- Linear) algorithm (Abbasi-Yadkori, Pál, and Szepesvári 2011). We propose an (α, ϵ)-global RDP extension of Rarely Switching OFUL, namely AdaR-OFUL.

The OFUL algorithm applies the "optimism in face of uncertianty principle" to the contextual linear bandit setting, which is to act in each round as if the environment is as nice as plausibly possible. In finite-action stochastic bandits, this means choosing the action with the largest upper confidence bound. In the case of linear contextual bandits, the idea remains the same, but the form of the confidence bound is more complicated. This is because the observed rewards yield side-information about more than just the arm played. OFUL algorithm computes a regularized least square estimate of the parameter θ and an ellipsoid confidence set around the estimated parameter. For each new observed action set $\mathcal{A}_t$, OFUL chooses the action with the largest upper confidence bound in the confidence ellipsoid.

The Rarely Switching OFUL Algorithm (RS-OFUL) can be seen as an "adaptively" phased version of the OFUL algorithm. RS-OFUL runs in episodes. At the beginning of each episode, the least square estimate and the confidence ellipsoid are updated. For the whole episode, the same estimate and confidence ellipsoid are used to choose the optimistic action. The condition to update the estimates (Line 6 of Algorithm 4) is to accumulate enough "useful information" in terms of the design matrix, which makes an update worth enough. RS-OFUL only updates the estimates $\log(T)$ times, while OFUL updates at each time-step of OFUL. RS-OFUL achieves similar regret as OFUL, up to a $\sqrt{1+C}$ multiplicative constant.

AdaR-OFUL (Algorithm 4) extends RS-OFUL by privately estimating the least-square estimate, while adapting the confidence ellipsoid accordingly. By Lemma 11, AdaR-OFUL only needs to make the estimate at every episode (α, ϵ)-RDP.

Theorem 9 (Privacy of Algorithm 4)**.** *Under Assumptions 1 and 2, AdaR-OFUL (Algorithm 4) satisfies* (α, ϵ)*-global RDP.*

Algorithm 4: AdaR-OFUL

```
 1: Input: Privacy budget $(\alpha, \epsilon)$, Horizon $T$, Regularizer $\lambda$, Dimension $d$, Doubling Schedule $C$
 2: Output: A sequence of $T$-actions satisfying $(\alpha, \epsilon)$-global RDP
 3: Initialisation: $V_0 = \lambda I_d$, $\tilde{\theta} = 0_d$, $\tau = 0$, $\ell = 1$
 4: for $t = 1, 2, \dots$ do
 5:     Observe $\mathcal{A}_t$
 6:     if $\det(V_{t-1}) > (1 + C)\det(V_\tau)$ then
 7:         Sample $Y_\ell \sim \mathcal{N}(0, \frac{2\alpha}{\epsilon} I_d)$
 8:         Compute $\tilde{\theta}_{t-1} = (V_{t-1})^{-1}(\sum_{s=1}^{t-1} A_s R_s + \sum_{m=1}^{\ell} Y_m)$
 9:         $\ell \leftarrow \ell + 1$ and $\tau \leftarrow t - 1$
10:     end if
11:     Compute $A_t = \mathrm{argmax}_{a \in \mathcal{A}_t} \left\langle \tilde{\theta}_\tau, a \right\rangle + \tilde{\beta}_\tau \|a\|_{(V_\tau)^{-1}}$
12:     Play arm $A_t$, Observe reward $R_t$
13:     $V_t \leftarrow V_{t-1} + A_t A_t^T$
14: end for
```

Proof Sketch. Similar to the previous algorithms, a change in the reward only affects the estimate of θ in the corresponding episode. Since the context is public, there is no privacy cost in estimating the design matrix V. Thus, making each estimate (α, ϵ)-RDP concludes the proof. We refer to Appendix A for the complete proof.

Remark 2. *When* $C = 1$*, Algorithm 4 can be seen as a direct generalisation of* AdaR-UCB *to the linear contextual bandits. Then, Line 6 of Algorithm 4 reduces to the arm-dependent doubling for the finite-arm stochastic bandits.*

Regret Analysis. Now, we upper bound the additional regret incurred by AdaR-OFUL due to global RDP.

Theorem 10. *Under Assumptions 1 and 2, and for* $\delta \in (0, 1]$*, with probability at least* $1 - \delta$*, the regret* R_T *of AdaR-OFUL (Algorithm 4) is upper bounded by*

$$R_T \leq \mathcal{O}\left(d\log(T)\sqrt{T}\right) + \mathcal{O}\left(\sqrt{\frac{\alpha}{\epsilon}} d^2 \log(T)^2\right)$$

Proof Sketch. As for the non-private proofs of OFUL and RS-OFUL, the main challenge in regret analysis of AdaR-OFUL is to design tight ellipsoid confidence sets around the private estimate $\tilde{\theta}_t$. To do so, we rely on the self-normalized bound for vector-valued martingales theorem of (Abbasi-Yadkori, Pál, and Szepesvári 2011). The regret us yielded by adjoining this theorem with the assumption of stochastic contexts controlling $\lambda_{\min}(G_t)$ and the concentration of χ^2 distribution. The rest of the proof is adapted from the analysis of RS-OFUL (Abbasi-Yadkori, Pál, and Szepesvári 2011). The only difference is that $\tilde{\beta}$ that controls the optimism width is not increasing anymore, but could be decomposed into the sum of an increasing and decreasing part. We also show that the number of updates of the estimated parameters is in $\mathcal{O}(\log(T))$. We refer to Appendix D for the complete proof.

Discussion: RDP for Almost-free. The upper bound of Theorem 10 shows that the price of (α, ϵ)-global RDP for linear contextual bandits is the additive term $\tilde{\mathcal{O}}\left(\sqrt{\frac{\alpha}{\epsilon}} d^2\right)$. For a fixed budget (α, ϵ), the regret dependent on RDP turns negligible in comparison with the privacy-oblivious regret term of $\tilde{\mathcal{O}}\left(d\sqrt{T}\right)$, as $T \to \infty$.

Limitations. Here, we discuss the limitations of our analysis.

- *The context c_t is public:* This assumption gets rid of the additional task of estimating privately both the design matrix for computing $\tilde{\theta}$, and the determinant of the design matrix to decide whether or not to update the estimates (Line 6 of Algorithm 4). Since the last condition should be checked at each time-step, a possible way to adapt AdaR-OFUL would be to estimate privately the design matrix at each step using the tree-based mechanism (Dwork et al. 2010b; Chan, Shi, and Song 2011) as in (Shariff and Sheffet 2018). However, with this change, the current analysis will not hold. With the added noise in the design matrix, the private estimate of the design matrix at step $t+1$ is no longer a rank 1 update of the private estimate at step t. The trick of Lemma 28 cannot be used anymore, which is the basis of the current analysis.
- *The context c_t is stochastic:* Our regret upper bound relies on Assumption 2, i.e. stochastic generation of the contexts. Without this assumption, there would not be a $\frac{1}{\sqrt{T}}$ in the confidence width $\tilde{\beta}$ that gives an additional cost of $\log(T)$ due to (α, ϵ)-global RDP rather than $\sqrt{T}$. The (ϵ, δ)-Joint DP algorithm (Shariff and Sheffet 2018), proposed for *private and adversarial contexts*, has an additional regret of $\frac{1}{\epsilon}\sqrt{T}$ due to privacy. It is still an open problem whether it is possible to design a private algorithm for linear contextual bandits with private and adversarially chosen contexts, such that the additional regret due to privacy is only in $\log(T)$.

Remark 3. *Related Problem Setups: (Neel and Roth 2018) proposes LinPriv, which is a reward-private extension of Linear UCB aimed for ensuring ϵ-global DP in linear contextual bandit. The context is assumed to be public but adversely chosen. Theorem 5 in (Neel and Roth 2018) claims that the regret of LinPriv is of order $\tilde{\mathcal{O}}\left(d\sqrt{T} + \frac{1}{\epsilon}Kd\log T\right)$. We believe that there is an error in this regret analysis. It should be $\tilde{\mathcal{O}}\left(d\sqrt{T} + \frac{1}{\epsilon}Kd\sqrt{T}\right)$. We refer to Appendix D.3 for details. Thus, the problem of achieving lower than $\sqrt{T}$ cost due to imposing privacy, while also considering adversarial contexts, still remain open.*

7 Conclusion and Future Work

We study bandits with (α, ϵ)-global RDP under three settings. We design an (α, ϵ)-global RDP bandit algorithm for each setting, and show that the additional cost in the regret incurred due to Rényi DP is negligible compared to the regret incurred oblivious to privacy. The three algorithms share the same algorithmic blueprint. They add calibrated *Gaussian noise* and they run in *adaptive episodes*. This revelation allows us to devise a generic and simple algorithmic approach to make any index-based bandit algorithm (α, ϵ)-global RDP.

One limitation of our analysis in the linear contextual bandit setting is the assumption that the contexts are public and stochastic. In recommender systems, the context may contain sensitive information about individuals. Also, in practice, we might not have information about the generation process of the observed contexts. *Designing and analysing an algorithm that does not rely on these two assumptions, and still achieves (α, ϵ)-global RDP almost for free, is an interesting open question.*

Another future direction is to derive regret lower bounds for bandits with (α, ϵ)-global RDP. Regret lower bounds for bandits with ϵ-global DP have been studied in the literature. (Shariff and Sheffet 2018) first showed that Bernoulli stochastic bandits with ϵ-global DP should incur an additional regret of $\Omega\left(\frac{\log(T)}{\epsilon}\right)$. (Azize and Basu 2022) refine this result and generalize it to any probability distributions. (Azize and Basu 2022) also give regret lower bounds for linear bandit with ϵ-global DP. However, *deriving regret lower bounds for bandits with (α, ϵ)-global RDP or (ϵ, δ)-global DP is still an open problem.* The problem-dependent regret upper bound for stochastic bandits with (α, ϵ)-global RDP (Theorem 5) suggests that the lower bounds for bandits with (α, ϵ)-global RDP may be of order $\Omega\left(\sqrt{\frac{\alpha}{\epsilon}\log(T)}\right)$. Proving this conjecture would be an interesting technical challenge to explore.

Acknowledgments

This work is supported by the AI_PhD@Lille grant. D. Basu acknowledges the funding of the ANR JCJC project "REPUBLIC". We also thank Philippe Preux for his support.

References

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep Learning with Differential Privacy. In *Proc. of CCS*, 308–318. ISBN 978-1-4503-4139-4.

Abbasi-Yadkori, Y.; Pál, D.; and Szepesvári, C. 2011. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24.

Azize, A.; and Basu, D. 2022. When Privacy Meets Partial Information: A Refined Analysis of Differentially Private Bandits. *arXiv preprint arXiv:2209.02570*.

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

Bergemann, D.; and Välimäki, J. 1996. Learning and strategic pricing. *Econometrica: Journal of the Econometric Society*, 1125–1149.

Bun, M.; Dwork, C.; Rothblum, G. N.; and Steinke, T. 2018. Composable and Versatile Privacy via Truncated CDP. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2018, 74–86. New York, NY, USA: ACM. ISBN 978-1-4503-5559-9.

Bun, M.; and Steinke, T. 2016. Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds. In *Theory of Cryptography*, 635–658. Berlin, Heidelberg: Springer Berlin Heidelberg.

Chan, T.-H. H.; Shi, E.; and Song, D. 2011. Private and Continual Release of Statistics. *ACM Trans. Inf. Syst. Secur.*, 14(3).

Chowdhury, S. R.; and Zhou, X. 2022. Distributed Differential Privacy in Multi-Armed Bandits. *arXiv preprint arXiv:2206.05772*.

Duchi, J. C.; Jordan, M. I.; and Wainwright, M. J. 2013. Local privacy, data processing inequalities, and statistical minimax rates. *arXiv preprint arXiv:1302.3203*.

Dwork, C.; Naor, M.; Pitassi, T.; and Rothblum, G. N. 2010a. Differential privacy under continual observation. In *Proceedings of the forty-second ACM symposium on Theory of computing*, 715–724. ACM.

Dwork, C.; Naor, M.; Pitassi, T.; and Rothblum, G. N. 2010b. Differential Privacy under Continual Observation. In *Proceedings of the Forty-Second ACM Symposium on Theory of Computing*, STOC '10, 715–724. New York, NY, USA: Association for Computing Machinery. ISBN 9781450300506.

Dwork, C.; Roth, A.; et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4): 211–407.

Dwork, C.; and Rothblum, G. N. 2016. Concentrated Differential Privacy. *ArXiv*, abs/1603.01887.

Gentile, C.; Li, S.; and Zappella, G. 2014. Online clustering of bandits. In *International Conference on Machine Learning*, 757–765. PMLR.

Hanna, O. A.; Girgis, A. M.; Fragouli, C.; and Diggavi, S. 2022. Differentially Private Stochastic Linear Bandits:(Almost) for Free. *arXiv preprint arXiv:2207.03445*.

Hu, B.; and Hegde, N. 2022. Near-optimal Thompson sampling-based algorithms for differentially private stochastic bandits. In *Uncertainty in Artificial Intelligence*, 844–852. PMLR.

Lattimore, T.; and Szepesvári, C. 2018. Bandit algorithms. *preprint*.

Mironov, I. 2017. Rényi Differential Privacy. In *Proceedings of 30th IEEE Computer Security Foundations Symposium (CSF)*, 263–275.

Mishra, N.; and Thakurta, A. 2015. (Nearly) Optimal Differentially Private Stochastic Multi-Arm Bandits. In *UAI*.

Neel, S.; and Roth, A. 2018. Mitigating bias in adaptive data gathering via differential privacy. In *International Conference on Machine Learning*, 3720–3729. PMLR.

Sajed, T.; and Sheffet, O. 2019. An Optimal Private Stochastic-MAB Algorithm Based on an Optimal Private Stopping Rule.

Shariff, R.; and Sheffet, O. 2018. Differentially Private Contextual Linear Bandits. In *Advances in Neural Information Processing Systems*, 4296–4306.

Silva, N.; Werneck, H.; Silva, T.; Pereira, A. C.; and Rocha, L. 2022. Multi-Armed Bandits in Recommendation Systems: A survey of the state-of-the-art and future directions. *Expert Systems with Applications*, 197: 116669.

Thompson, W. R. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4): 285–294.

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

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

Zhang, F. 2011. *Matrix theory: basic results and techniques*. Springer.

Zheng, K.; Cai, T.; Huang, W.; Li, Z.; and Wang, L. 2020. Locally Differentially Private (Contextual) Bandits Learning.

Appendix

A Privacy Proofs

Lemma 11 (Privacy Lemma)**.** *Let $\mathcal{M}$ be a mechanism that takes as input a set $\{r_1, \dots, r_k\}$ for every $k \in \mathbb{N}^*$ and outputs a distribution.*
Let $\ell < T$ and $t_1, \dots t_\ell, t_{\ell+1}$ be in $[1, T]$ such that $1 = t_1 < \cdots < t_\ell < t_{\ell+1} - 1 = T$.
Let's define the following mechanism

$$\mathcal{G} : \{r_1, \dots, r_T\} \to \bigotimes_{i=1}^{\ell} \mathcal{M}_{\{r_{t_i}, \dots, r_{t_{i+1}-1}\}} \tag{10}$$

If $\mathcal{M}$ is (α, ϵ)-RDP then $\mathcal{G}$ is (α, ϵ)-RDP

Proof. Suppose that $\mathcal{M}$ is (α, ϵ)-RDP, and let $r \triangleq \{r_1, \dots, r_T\}$ and $r' \triangleq \{r'_1, \dots, r'_T\}$ be two neighboring rewards. This implies that $\exists j \in [1, T]$ such that $r_j \neq r'_j$ and $\forall t \neq j$, $r_t = r'_t$.

Let ℓ' be such that $t_{\ell'} \leq j \leq t_{\ell'+1} - 1$.

We have that

$$D_\alpha(\mathcal{G}_r \| \mathcal{G}_{r'}) = \frac{1}{\alpha - 1} \log \left(\int_{o=(o_1, \dots, o_\ell)} \mathcal{G}_{r'}(o) \left(\frac{\mathcal{G}_r(o)}{\mathcal{G}_{r'}(o)} \right)^\alpha \right)$$

Since

$$\mathcal{G}_r(o) = \prod_{i=1}^{\ell} \mathcal{M}_{\{r_{t_i}, \dots, r_{t_{i+1}-1}\}}(o_i)$$

and

$$\mathcal{G}_{r'}(o) = \prod_{i=1}^{\ell} \mathcal{M}_{\{r'_{t_i}, \dots, r'_{t_{i+1}-1}\}}(o_i)$$

we get

$$\frac{\mathcal{G}_r(o)}{\mathcal{G}_{r'}(o)} = \frac{\mathcal{M}_{\{r_{t_{\ell'}}, \dots, r_{t_j}, \dots, r_{t_{\ell'+1}-1}\}}(o_i)}{\mathcal{M}_{\{r_{t_{\ell'}}, \dots, r'_{t_j}, \dots, r_{t_{\ell'+1}-1}\}}(o_i)}$$

Thus,

$$D_\alpha(\mathcal{G}_r \| \mathcal{G}_{r'}) = D_\alpha(\mathcal{M}_{\{r_{t_{\ell'}}, \dots, r_{t_j}, \dots, r_{t_{\ell'+1}-1}\}} \| \mathcal{M}_{\{r_{t_{\ell'}}, \dots, r'_{t_j}, \dots, r_{t_{\ell'+1}-1}\}}) \leq \epsilon$$

□

Theorem 4 (Privacy of AdaR-UCB)**.** *For rewards in $[0, 1]$,* AdaR-UCB *satisfies (α, ϵ)-global RDP.*

Proof. Fix two neighboring reward streams $r^T = \{r_1, \dots, r_T\}$ and $r'^T = \{r'_1, \dots, r'_T\}$.
This implies that $\exists j \in [1, T]$ such that $r_j \neq r'_j$ and $\forall t \neq j$, $r_t = r'_t$.

Let $\mathcal{M}_{\{r_{t_i}, \dots, r_{t_{i+1}-1}\}} \triangleq \frac{1}{t_{i+1} - t_i} \sum_{s=t_i}^{t_{i+1}-1} r_s + Z_i$ where $Z_i \sim \mathcal{N}\left(0, \frac{\alpha}{2\epsilon(t_{i+1} - t_i)^2}\right)$.

For rewards in $[0, 1]$, the L2 sensitivity of $f : \{r_{t_i}, \dots, r_{t_{i+1}-1}\} \to \frac{1}{t_{i+1} - t_i} \sum_{s=t_i}^{t_{i+1}-1} r_s$ is $\frac{1}{t_{i+1} - t_i}$. Using Theorem 2, $\mathcal{M}$ is (α, ϵ)-RDP.

For fixed episodes, we apply Lemma 11 to show that $\mathcal{G}$, as defined in Eq. 10 is also (α, ϵ)-RDP.

Since for the two neighbouring rewards r^T and r'^T, we have that $r^{j-1} = r'^{j-1}$, and $\{r_{j+1}, \dots, r_T\} = \{r'_{j+1}, \dots, r'_T\}$, the time-steps t_ℓ corresponding to the beginning of each adaptive episodes are random variables that have the same law under r^T and r'^T.

Taking the expectation over the adaptive episodes shows that the whole sequence of released noisy empirical means $\mathcal{G}$ is (α, ϵ)-RDP.

The released actions of AdaR-UCB only depend on the sequence of released noisy empirical means $\mathcal{G}$. By post-processing, AdaR-UCB is (α, ϵ)-global RDP.

□

We use the same proof structure for the privacy of Algorithm 3 and Algorithm 4. The only difference is to specify the actual mechanism $\mathcal{M}$ used for each episode and show that it is an (α, ϵ)-global RDP mechanism.

Theorem 7 (Privacy of AdaR-GOPE)**.** *For rewards in [0,1], AdaR-GOPE (Algorithm 3) satisfies (α, ϵ)-global RDP.*

Proof. Fix two neighboring reward streams $r^T = \{r_1, \dots, r_T\}$ and $r'^T = \{r'_1, \dots, r'_T\}$.
This implies that $\exists j \in [1, T]$ such that $r_j \neq r'_j$ and $\forall t \neq j$, $r_t = r'_t$.

Let $\hat{\theta}$ be the mechanism that estimates the parameter θ using the least square estimate, i.e $\hat{\theta} : \{R_t\}_{t=t_\ell}^{t_{\ell+1}-1} \to V_\ell^{-1} \sum_{t=t_\ell}^{t_{\ell+1}-1} A_t R_t$ such that $V_\ell = \sum_{a \in \mathrm{Supp}(\pi_\ell)} T_\ell(a) a a^\top$.

We can write that

$$\sum_{t=t_\ell}^{t_{\ell+1}-1} A_t R_t = \sum_{a \in \mathrm{Supp}(\pi_\ell)} a \sum_{A_t = a, t \in [t_\ell, t_{\ell+1}-1]} R_t.$$

For rewards in $[-1, 1]$, the L2 sensitivity of $\sum_{A_t = a, t \in [t_\ell, t_{\ell+1}-1]} R_t$ is 1.

Let $(N_a)_{a \in \mathrm{Supp}(\pi_\ell)} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \frac{2\alpha}{\epsilon}\right)$.

Using Theorem 2 we have that $\sum_{a \in \mathrm{Supp}(\pi_\ell)} a \left(\sum_{A_t = a, t \in [t_\ell, t_{\ell+1}-1]} R_t + N_a\right)$ is (α, ϵ)-RDP.

Which gives that $\mathcal{M} : \{R_t\}_{t=t_\ell}^{t_{\ell+1}-1} \to \hat{\theta}\left(\{R_t\}_{t=t_\ell}^{t_{\ell+1}-1}\right) + V_\ell^{-1} \sum_{a \in \mathrm{Supp}(\pi_\ell)} a N_a$ is (α, ϵ)-RDP.

For fixed episodes, we apply Lemma 11 to show that $\mathcal{G}$, as defined in Eq. 10 is also (α, ϵ)-RDP.

Since for the two neighboring rewards r^T and r'^T, we have that $r^{j-1} = r'^{j-1}$, and $\{r_{j+1}, \dots, r_T\} = \{r'_{j+1}, \dots, r'_T\}$, the time-steps t_ℓ corresponding to the beginning of each adaptive episodes are random variables that have the same law under r^T and r'^T.

Taking the expectation over the adaptive episodes shows that the whole sequence of released noisy empirical means $\mathcal{G}$ is (α, ϵ)-RDP.

The released actions of Algorithm 3 only depend on the sequence of released noisy empirical means $\mathcal{G}$. By post-processing, Algorithm 3 is (α, ϵ)-global RDP.

□

Theorem 9 (Privacy of AdaR-OFUL)**.** *For rewards in [0,1], AdaR-OFUL (Algorithm 4) satisfies (α, ϵ)-global RDP.*

Proof. Fix two neighboring reward streams $r^T = \{r_1, \dots, r_T\}$ and $r'^T = \{r'_1, \dots, r'_T\}$.
This implies that $\exists j \in [1, T]$ such that $r_j \neq r'_j$ and $\forall t \neq j$, $r_t = r'_t$.

Let $\hat{\theta}$ be the mechanism that estimates the parameter θ using the least square estimate on the whole sequence, i.e $\hat{\theta} : \{R_t\}_{s=1}^{t} \to V_t^{-1} \sum_{s=1}^{t} A_s R_s$ such that $V_t = \sum_{s=1}^{t} a_s a_s^\top$.

We can write that

$$\sum_{s=1}^{t} A_s R_s = \sum_{\ell=1}^{\ell(t)} \sum_{t \in [t_\ell, t_{\ell+1}-1]} A_t R_t.$$

For rewards in $[-1, 1]$ and actions A_t such that $\|A_t\| \leq 1$, the L2 sensitivity of the sum $\sum_{t \in [t_\ell, t_{\ell+1}-1]} A_t R_t$ is 2.

Let $Y_\ell \sim \mathcal{N}(0, \frac{2\alpha}{\epsilon} I_d)$.

Using Theorem 2 we have that $\sum_{\ell=1}^{\ell(t)} \sum_{t \in [t_\ell, t_{\ell+1}-1]} A_t R_t + Y_\ell$ is (α, ϵ)-RDP.

Which gives that $\mathcal{M} : \{R_t\}_{s=1}^{t} \to \hat{\theta}\left(\{R_t\}_{s=1}^{t} + \sum_{m=1}^{\ell} Y_m\right)$ is (α, ϵ)-RDP.

For fixed episodes, we apply Lemma 11 to show that $\mathcal{G}$, as defined in Eq. 10 is also (α, ϵ)-RDP.

Since for the two neighboring rewards r^T and r'^T, we have that $r^{j-1} = r'^{j-1}$, and $\{r_{j+1}, \dots, r_T\} = \{r'_{j+1}, \dots, r'_T\}$, the time-steps t_ℓ corresponding to the beginning of each adaptive episodes are random variables that have the same law under r^T and r'^T.

Taking the expectation over the adaptive episodes shows that the whole sequence of released noisy empirical means $\mathcal{G}$ is (α, ϵ)-RDP.

The released actions of Algorithm 3 only depend on the sequence of released noisy empirical means $\mathcal{G}$. By post-processing, Algorithm 3 is (α, ϵ)-global RDP.

□

B Finite-arm Rényi DP Bandits

B.1 Concentration Inequalities

Lemma 12. *Assume that $(X_i)_{1\leq i\leq n}$ are iid random variables in $[0,1]$, with $\mathbb{E}(X_i)=\mu$. Then, for any $\delta \geq 0$,*

$$\mathbb{P}\left(\hat{\mu}_n + Z_n - \sqrt{\left(\frac{1}{2n}+\frac{\alpha}{\epsilon n^2}\right)\log\left(\frac{1}{\delta}\right)} \geq \mu\right) \leq \delta, \tag{11}$$

and

$$\mathbb{P}\left(\hat{\mu}_n + Z_n + \sqrt{\left(\frac{1}{2n}+\frac{\alpha}{\epsilon n^2}\right)\log\left(\frac{1}{\delta}\right)} \leq \mu\right) \leq \delta, \tag{12}$$

where $\hat{\mu}_n = \frac{1}{n}\sum_{t=1}^{n} X_t$ and $Z_n \sim \mathcal{N}\left(0,\, \frac{\alpha}{2\epsilon n^2}\right)$.

Proof. Let $Y = (\hat{\mu}_n + Z_n - \mu)$.
Using Properties 2. and 3. of Lemma 22, we get that Y is $\sqrt{\frac{1}{4n}+\frac{\alpha}{2\epsilon n^2}}$-subgaussian.
We conclude using the concentration on subgaussian random variables, i.e. Lemma 21.

□

B.2 Regret Analysis

Theorem 5. *For rewards in $[0,1]$,* AdaR-UCB *satisfies (α,ϵ)-global RDP, and for $\beta > 3$, it yields a regret upper bound of*

$$\sum_{a:\Delta_a>0}\left(\frac{8\beta}{\Delta_a}\log(T) + 8\sqrt{\frac{\beta\alpha}{\epsilon}}\sqrt{\log(T)} + \frac{2\beta}{\beta-3}\right)$$

Proof. By the generic regret decomposition of Theorem 11 in (Azize and Basu 2022), for every suboptimal arm a, we have that

$$\mathbb{E}[N_a(T)] \leq 2^{\ell+1} + \mathbb{P}\left(G^c_{a,\ell,T}\right)T + \frac{\beta}{\beta-3},$$

where

$$G_{a,\ell,T} = \left\{\hat{\mu}_{a,2^\ell} + Z_\ell + \sqrt{\left(\frac{1}{2\times 2^\ell}+\frac{\alpha}{\epsilon\times(2^\ell)^2}\right)\beta\log(T)} < \mu_1\right\}.$$

such that $Z_\ell \sim \mathcal{N}\left(0,\, \frac{\alpha}{2\epsilon\times(2^\ell)^2}\right)$

Step 1: Choosing an ℓ. Now, we observe that

$$\begin{aligned}\mathbb{P}(G^c_{a,\ell,T}) &= \mathbb{P}\left(\hat{\mu}_{a,2^\ell} + Z_\ell + \sqrt{\left(\frac{1}{2\times 2^\ell}+\frac{\alpha}{\epsilon\times(2^\ell)^2}\right)\beta\log(T)} \geq \mu_1\right)\\ &= \mathbb{P}\left(\hat{\mu}_{a,2^\ell} + Z_\ell - \sqrt{\left(\frac{1}{2\times 2^\ell}+\frac{\alpha}{\epsilon\times(2^\ell)^2}\right)\beta\log(T)} \geq \mu_a + \epsilon\right)\end{aligned}$$

for $\epsilon = \left(\Delta_a - 2\sqrt{\left(\frac{1}{2\times 2^\ell}+\frac{\alpha}{\epsilon\times(2^\ell)^2}\right)\beta\log(T)}\right)$.

The idea is to choose ℓ big enough so that $\epsilon \geq 0$.
Let us consider the contrary, i.e.

$$\begin{aligned}\epsilon < 0 &\Rightarrow 2^\ell < \frac{2\beta\log(T)}{\Delta_a^2}\left(1+\Delta_a\sqrt{\frac{\alpha}{\epsilon\beta\log(T)}}\right)\\ &\Rightarrow 2^\ell < \frac{2\beta}{\Delta_a^2}\log(T) + 2\sqrt{\frac{\beta\alpha}{\epsilon\Delta_a^2}}\sqrt{\log(T)}\end{aligned}$$

Thus, by choosing

$$\ell = \left\lceil \frac{1}{\log(2)}\log\left(\frac{2\beta}{\Delta_a^2}\log(T) + 2\sqrt{\frac{\beta\alpha}{\epsilon\Delta_a^2}}\sqrt{\log(T)}\right)\right\rceil$$

we ensure $\epsilon > 0$. This also implies that

$$\mathbb{P}(G^c_{a,\ell,T}) \leq \mathbb{P}\left(\hat{\mu}_{a,2^\ell} + Z_\ell - \sqrt{\left(\frac{1}{2 \times 2^\ell} + \frac{\alpha}{\epsilon \times (2^\ell)^2}\right)\beta \log(T)} \geq \mu_a\right) \leq \frac{1}{T^\beta}$$

The last inequality is due to Equation 11 of Lemma 12.

Step 2: The Regret Bound. Combining Steps 1 and 2, we get that

$$\begin{aligned}\mathbb{E}[N_a(T)] &\leq \frac{\beta}{\beta - 3} + 2^{\ell+1} + T \times \frac{1}{T^\beta} \\ &\leq \frac{8\beta}{\Delta_a^2}\log(T) + 8\sqrt{\frac{\beta\alpha}{\epsilon\Delta_a^2}}\sqrt{\log(T)} + \frac{2\beta}{\beta - 3}.\end{aligned}$$

Plugging this upper bound back in the definition of problem-dependent regret

$$\mathrm{Reg}_T(\mathsf{AdaR\text{-}UCB}, \nu) \leq \sum_{a:\Delta_a > 0}\left(\frac{8\beta}{\Delta_a}\log(T) + 8\sqrt{\frac{\beta\alpha}{\epsilon}}\sqrt{\log(T)} + \frac{2\beta}{\beta - 3}\right).$$

□

C Linear Rényi DP Bandits

C.1 Basic Definitions of Optimal Design

Definition 13 (Optimal Design)**.** *Let $\mathcal{A} \subset \mathbb{R}^d$ and $\pi : \mathcal{A} \to [0,1]$ be a distribution on $\mathcal{A}$ so that $\sum_{a\in\mathcal{A}} \pi(a) = 1$. Let $V(\pi) \in \mathbb{R}^{d\times d}$ and $f(\pi), g(\pi) \in \mathbb{R}$ be given by*

$$V(\pi) = \sum_{a\in\mathcal{A}} \pi(a) a a^T, \qquad f(\pi) = \log\det V(\pi), \qquad g(\pi) = \max_{a\in\mathcal{A}} \|a\|_{V(\pi)^{-1}}$$

- *π is called a design*
- *The set $\text{Supp}\,(\pi) \triangleq \{a \in \pi : \pi(a) \neq 0\}$ is called the core set of $\mathcal{A}$*
- *A design that maximizes f is known as a **D-optimal design***
- *A design that minimizes g is known as **G-optimal design***

Theorem 14 (Kiefer–Wolfowitz Theorem)**.** *Assume that $\mathcal{A}$ is compact and span$(\mathcal{A}) = \mathbb{R}^d$. The following are equivalent:*

- *$\pi^\star$ is a minimiser of g.*
- *$\pi^\star$ is a maximiser of f.*
- *$g(\pi^\star) = d$*

Furthermore, there exists a minimised $\pi^\star$ of g such that $|\text{Supp}\,(\pi)| \leq \frac{d(d+1)}{2}$

C.2 Concentration Inequalities

Let $A_1, \dots, A_t$ be deterministically chosen without the knowledge of $XR1, \dots, R_t$ and π be an optimal design for $\mathcal{A}$.
Let $V_t \triangleq \sum_{s=1}^t A_s A_s^T = \sum_{a\in\mathcal{A}} N_a(t) a a^T$ be the design matrix, $\hat{\theta}_t = V_t^{-1} \sum_{s=1}^t A_s R_s$ be the least square estimate and $\tilde{\theta}_t = \hat{\theta}_t + V_t^{-1} \left(\sum_{b\in\text{Supp}(\pi)} b N_b \right)$ where $N_b \sim \mathcal{N}\left(0, \frac{2\alpha}{\epsilon}\right)$

Theorem 15. *For every $a \in \mathcal{A}$ and $\delta \in [0,1]$, we have that*

$$\mathbb{P}\left(\left| \left\langle \tilde{\theta}_t - \theta^\star, a \right\rangle \right| \geq g_t \sqrt{2 \log\left(\frac{4}{\delta}\right)} + g_t^2 \sqrt{\frac{2\alpha d(d+1)}{\epsilon} \log\left(\frac{4}{\delta}\right)} \right) \leq \delta$$

where $g_t = \max_{b\in\mathcal{A}} \|b\|_{V_t^{-1}}$.

Proof. For every $a \in \mathcal{A}$

$$\begin{aligned} \left\langle \tilde{\theta}_t - \theta^\star, a \right\rangle &= \left\langle \hat{\theta}_t - \theta^\star, a \right\rangle + \sum_{b\in\text{Supp}(\pi)} \left(a^T V_t^{-1} b\right) N_b \\ &= \left\langle \hat{\theta}_t - \theta^\star, a \right\rangle + Z_t \end{aligned}$$

where $Z_t \triangleq \sum_{b\in\text{Supp}(\pi)} \left(a^T V_t^{-1} b\right) N_b$.

<u>**Step 1: Concentration of the least square estimate.**</u> Using Eq.(20.2) from Chapter 20 of (Lattimore and Szepesvári 2018), we have that

$$\mathbb{P}\left(\left| \left\langle \hat{\theta}_t - \theta^\star, a \right\rangle \right| \geq g_t \sqrt{2 \log\left(\frac{4}{\delta}\right)} \right) \leq \frac{\delta}{2}$$

<u>**Step 2: Concentration of the additional Gaussian noise.**</u> On the other hand, we have that

$$\left|a^T V_t^{-1} b\right| \leq \ |V_t^{-\frac{1}{2}} a\| \, |V_t^{-\frac{1}{2}} b\| \leq \max_{b\in\text{Supp}(\pi)} |b\|^2_{V_t^{-1}} = g_t^2$$

using that $|\text{Supp}\,(\pi_\ell)| \leq \frac{d(d+1)}{2}$, $N_b \sim \mathcal{N}(0, \frac{2\alpha}{\epsilon})$ and Properties 1 and 2 of Lemma 22, we get that

$$Z_t \text{ is } \sigma_t\text{-subgaussian}, \ \text{ where } \sigma_t = g_t^2 \sqrt{\frac{\alpha d(d+1)}{\epsilon}}$$

Using Lemma 21, we get that

$$\mathbb{P}\left(|Z_t| \geq g_t^2 \sqrt{\frac{2\alpha d(d+1)}{\epsilon} \log\left(\frac{4}{\delta}\right)} \right) \leq \frac{\delta}{2}$$

Steps 1 and 2 together conclude the proof. □

Corollary 16. *Let β be a fixed confidence level. If we choose each action $a \in \mathcal{A}$*

$$N_a(t) = \left\lceil \frac{8d\pi(a)}{\beta^2} \log\left(\frac{4}{\delta}\right) + \frac{2d\pi(a)}{\beta} \sqrt{\frac{2\alpha d(d+1)}{\epsilon} \log\left(\frac{4}{\delta}\right)} \right\rceil$$

then, for $t = \sum_{a \in \operatorname{Supp}(\pi)} N_a(t)$, we get that

$$\mathbb{P}\left(\left|\left\langle \tilde{\theta}_t - \theta^\star, a \right\rangle\right| \geq \beta\right) \leq \delta$$

Proof. We have that

$$V_t = \sum_{a \in \operatorname{Supp}(\pi)} N_a(t) a a^T \geq c V(\pi)$$

where $c \triangleq \frac{8d}{\beta^2} \log\left(\frac{4}{\delta}\right) + \frac{2d}{\beta}\sqrt{\frac{2\alpha d(d+1)}{\epsilon} \log\left(\frac{4}{\delta}\right)}$.
Which means that

$$g_t^2 = \max_{b \in \mathcal{A}} \|b\|^2_{V_t^{-1}} \leq \frac{1}{c} \max_{b \in \mathcal{A}} \|b\|^2_{V(\pi)^{-1}} = \frac{g(\pi)}{c} = \frac{d}{c}$$

Which gives that

$$\begin{aligned} g_t \sqrt{2 \log\left(\frac{4}{\delta}\right)} + g_t^2 \sqrt{\frac{2\alpha d(d+1)}{\epsilon} \log\left(\frac{4}{\delta}\right)} &\leq \frac{\sqrt{2d \log\left(\frac{4}{\delta}\right)}}{\sqrt{\frac{8d}{\beta^2} \log\left(\frac{4}{\delta}\right)}} + \frac{d \sqrt{\frac{2\alpha d(d+1)}{\epsilon} \log\left(\frac{4}{\delta}\right)}}{\frac{2d}{\beta} \sqrt{\frac{2\alpha d(d+1)}{\epsilon} \log\left(\frac{4}{\delta}\right)}} \\ &\leq \frac{\beta}{2} + \frac{\beta}{2} = \beta \end{aligned}$$

and conclude using Theorem 15. □

C.3 Regret Analysis

Theorem 8. *If rewards are in $[0, 1]$, Algorithm 3 is (α, ϵ)-global RDP and with probability at least $1 - \delta$, the regret R_T of Algorithm 3 is upper bounded by*

$$R_T \leq C_1 \sqrt{dT \log\left(\frac{k \log(T)}{\delta}\right)} + C_2 d^2 \sqrt{\frac{\alpha}{\epsilon} \log\left(\frac{k \log(T)}{\delta}\right)} \log(T)$$

where C_1 and C_2 are positive-valued universal constants.
If $\delta = \frac{1}{T}$, then $\mathbb{E}(R_T) \leq C_1 \sqrt{dT \log(kT)} + C_2 \sqrt{\frac{\alpha}{\epsilon}} d^2 \log(kT)^{\frac{3}{2}}$

Proof. **Step 1: Defining the good event E.** Let

$$E \triangleq \bigcap_{\ell=1}^{\infty} \bigcap_{a \in \mathcal{A}_\ell} \left\{\left|\left\langle \tilde{\theta}_\ell - \theta_*, a \right\rangle\right| \leq \beta_\ell\right\}.$$

Using Corollary 16, we get that

$$\begin{aligned} \mathbb{P}(\neg E) &\leq \sum_{\ell=1}^{\infty} \sum_{a \in \mathcal{A}_\ell} \mathbb{P}\left(\left|\left\langle \tilde{\theta}_\ell - \theta_*, a \right\rangle\right| > \beta_\ell\right) \\ &\leq \sum_{\ell=1}^{\infty} \sum_{a \in \mathcal{A}_\ell} \frac{\delta}{k\ell(\ell+1)} \leq \delta \end{aligned}$$

Step 2: Good properties under the event E. We have that under E

- The optimal arm $a^\star \in \arg\max_{a \in \mathcal{A}} \langle \theta^*, a \rangle$ is never eliminated.

  Proof. for every episode ℓ and $b \in \mathcal{A}_\ell$, we have that

$$\begin{aligned} \left\langle \tilde{\theta}_\ell, b - a^\star \right\rangle = \left\langle \tilde{\theta}_\ell - \theta^\star, b - a^\star \right\rangle + \langle \theta^\star, b - a^\star \rangle &\leq \left\langle \tilde{\theta}_\ell - \theta^\star, b - a^\star \right\rangle \\ &\leq \left|\left\langle \tilde{\theta}_\ell - \theta_*, a^\star \right\rangle\right| + \left|\left\langle \tilde{\theta}_\ell - \theta_*, b \right\rangle\right| \leq 2\beta_\ell \end{aligned}$$

□

- Each sub-optimal arm a will be removed after ℓ_a rounds where $\ell_a \triangleq \min\{\ell : 4\beta_\ell < \Delta_a\}$.

  Proof. We have that

$$\begin{aligned}\left\langle \tilde{\theta}_{\ell_a}, a^\star - a \right\rangle &\geq \langle \theta^\star, a^\star \rangle - \beta_{\ell_a} - \langle \theta^\star, a \rangle - \beta_{\ell_a} \\ &= \Delta_a - 2\beta_{\ell_a} > 2\beta_{\ell_a}\end{aligned}$$

□

- for $a \in \mathcal{A}_{\ell+1}$, we have that $\Delta_a \leq 4\beta_\ell$.

  Proof. If $\Delta_a > 4\beta_\ell$, then $\ell \geq \ell_a$ and arm a is already eliminated, i.e. $a \notin \mathcal{A}_{\ell+1}$ □

Step 3: Regret decomposition under E.
Fix Δ to be optimized later.
Under E, each sub-optimal action a such that $\Delta_a > \Delta$ will only be played for the first ℓ_Δ rounds where

$$\ell_\Delta \triangleq \min\{\ell : 4\beta_\ell < \Delta\} = \left\lceil \log_2\left(\frac{4}{\Delta}\right) \right\rceil$$

We have that

$$\begin{aligned}R_T &= \sum_{a \in \mathcal{A}} \Delta_a N_a(T) \\ &= \sum_{a : \Delta_a > \Delta} \Delta_a N_a(T) + \sum_{a : \Delta_a \leq \Delta} \Delta_a N_a(T) \\ &= \sum_{\ell=1}^{\ell_\Delta \wedge \ell(T)} \sum_{a \in \mathcal{A}_\ell} \Delta_a T_\ell(a) + T\Delta \\ &\leq \sum_{\ell=1}^{\ell_\Delta \wedge \ell(T)} 4\beta_{\ell-1} T_\ell + T\Delta\end{aligned}$$

where $\ell(T)$ is the total number of episodes played until timestep T.
Step 4: Upper-bounding T_ℓ and $\ell(T)$ under E. We have that

$$\begin{aligned}T_\ell &= \sum_{a \in Supp(\pi_\ell)} T_\ell(a) \\ &= \sum_{a \in Supp(\pi_\ell)} \left\lceil \frac{8d\pi_\ell(a)}{\beta_\ell^2} \log\left(\frac{4k\ell(\ell+1)}{\delta}\right) + \frac{2d\pi_\ell(a)}{\beta_\ell} \sqrt{\frac{2\alpha}{\epsilon} d(d+1) \log\left(\frac{4k\ell(\ell+1)}{\delta}\right)} \right\rceil \\ &\leq \frac{d(d+1)}{2} + \frac{8d}{\beta_\ell^2} \log\left(\frac{4k\ell(\ell+1)}{\delta}\right) + \frac{2d}{\beta_\ell} \sqrt{\frac{2\alpha}{\epsilon} d(d+1) \log\left(\frac{4k\ell(\ell+1)}{\delta}\right)}\end{aligned}$$

since $\beta_{\ell+1} = \frac{1}{2}\beta_\ell$ and $\sum_{\ell=1}^{\ell(T)} T_\ell = T$, there exists a constant C such that $\ell(T) \leq C\log(T)$.
Which means that, for $\ell \leq \ell(T)$, there exists a constant C' such that

$$\log\left(\frac{4k\ell(\ell+1)}{\delta}\right) \leq C' \log\left(\frac{k\log(T)}{\delta}\right)$$

hence

$$T_\ell \leq \frac{d(d+1)}{2} + \frac{8d}{\beta_\ell^2} C' \log\left(\frac{k\log(T)}{\delta}\right) + \frac{2d}{\beta_\ell} \sqrt{\frac{2\alpha}{\epsilon} d(d+1) C' \log\left(\frac{k\log(T)}{\delta}\right)}$$

Step 5: Upper-bounding the regret under E.

Under E

$$\begin{aligned}\sum_{\ell=1}^{\ell_\Delta \wedge \ell(T)} 4\beta_{\ell-1} T_\ell &\leq \sum_{\ell=1}^{\ell_\Delta \wedge \ell(T)} 8\beta_\ell \left(\frac{d(d+1)}{2} + \frac{8d}{\beta_\ell^2} C' \log\left(\frac{k\log(T)}{\delta}\right) + \frac{2d}{\beta_\ell}\sqrt{\frac{2\alpha}{\epsilon} d(d+1) C' \log\left(\frac{k\log(T)}{\delta}\right)}\right) \\ &\leq 4d(d+1) + 64dC' \log\left(\frac{k\log(T)}{\delta}\right)\left(\sum_{\ell=1}^{\ell_\Delta} 2^\ell\right) + 16(d+1)^2\sqrt{\frac{2\alpha}{\epsilon} C' \log\left(\frac{k\log(T)}{\delta}\right)}\ell(T) \\ &\leq 4d(d+1) + 16dC' \log\left(\frac{k\log(T)}{\delta}\right)\left(\frac{16}{\Delta}\right) + 16(d+1)^2\sqrt{\frac{2\alpha}{\epsilon} C' \log\left(\frac{k\log(T)}{\delta}\right)}\ell(T) \\ &\leq 4d(d+1) + C_1 d \log\left(\frac{k\log(T)}{\delta}\right)\frac{1}{\Delta} + C_2 d^2\sqrt{\frac{\alpha}{\epsilon} \log\left(\frac{k\log(T)}{\delta}\right)}\log(T)\end{aligned}$$

All in all, we have that

$$R_T \leq 4d(d+1) + C_2 d^2 \sqrt{\frac{\alpha}{\epsilon}\log\left(\frac{k\log(T)}{\delta}\right)}\log(T) + C_1 d\log\left(\frac{k\log(T)}{\delta}\right)\frac{1}{\Delta} + T\Delta$$

Step 6: Optimizing for Δ. Taking $\Delta = \sqrt{\frac{C_1 d}{T}\log\left(\frac{k\log(T)}{\delta}\right)}$, we get that

$$R_T \leq C_1\sqrt{dT\log\left(\frac{k\log(T)}{\delta}\right)} + C_2 d^2\sqrt{\frac{\alpha}{\epsilon}\log\left(\frac{k\log(T)}{\delta}\right)}\log(T)$$

Step 7: Upper bounding the expected regret. For $\delta = \frac{1}{T}$, we get that

$$\begin{aligned}\mathbb{E}(R_T) &\leq (1-\delta)R_T(\delta) + \delta T \\ &\leq R_T(\delta) + 1 \\ &\leq C_1'\sqrt{dT\log(kT)} + C_2'\sqrt{\frac{\alpha}{\epsilon}} d^2 \log(kT)^{\frac{3}{2}}\end{aligned}$$

□

D Reward-Private Linear Contextual Rényi DP Bandits

D.1 Confidence bound for the private least square estimator

Theorem 17. *Let $\delta \in (0,1)$. Then, with probability $1-\mathcal{O}(\delta)$, it holds that, for all $t \in [1,T]$,*

$$\|\tilde{\theta}_t - \theta^\star\|_{V_t} \leq \tilde{\beta}_t$$

where

$$\tilde{\beta}_t = \beta_t + \frac{\gamma_t}{\sqrt{t}}$$

such that

$$\beta_t = \mathcal{O}\left(\sqrt{d\log(t)}\right) \text{ and } \gamma_t = \mathcal{O}\left(\sqrt{\frac{\alpha}{\epsilon}} d\log(t)\right)$$

and β_t and γ_t are increasing in t.

Proof. **Step 1: Decomposing $\tilde{\theta}_t - \theta^\star$.** We have that

$$\begin{aligned}
\tilde{\theta}_t - \theta^\star &= V_t^{-1}\left(\sum_{s=1}^{t} A_s R_s + \sum_{m=1}^{\ell(t)} Y_m\right) - \theta^\star \\
&= V_t^{-1}\left(\sum_{s=1}^{t} A_s(A_s^T\theta^\star + \eta_s) + \sum_{m=1}^{\ell(t)} Y_m\right) - \theta^\star \\
&= V_t^{-1}\left((V_t - \lambda I_d)\theta^\star + \sum_{s=1}^{t} A_s\eta_s + \sum_{m=1}^{\ell(t)} Y_m\right) - \theta^\star \\
&= V_t^{-1}\left(S_t + \sum_{m=1}^{\ell(t)} Y_m - \lambda\theta^\star\right) \\
\|\tilde{\theta}_t - \theta^\star\|_{V_t} &= \|S_t + N_t - \lambda\theta^\star\|_{V_t^{-1}}
\end{aligned}$$

where $S_t \triangleq \sum_{s=1}^{t} A_s\eta_s$, $N_t = \sum_{m=1}^{\ell(t)} Y_m \sim \mathcal{N}\left(0, \frac{2\alpha\ell(t)}{\epsilon} I_d\right)$ and $\ell(t)$ is the number of episodes until time-step t (number of updates of $\tilde{\theta}$.

Step 2: Defining the good event E. Let

$$\begin{aligned}
E_1 &= \left\{\forall t \in [1,T] : \|S_t\|_{V_t^{-1}} \leq \sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(V_t)}{\lambda^d}\right)}\right\}, \\
E_2 &= \left\{\forall t \in [1,T] : \lambda_{\text{min}}(G_t) \geq \frac{\lambda_0 t}{4} - 8\log\left(\frac{t+3}{\delta/d}\right) - 2\sqrt{t\log\left(\frac{t+3}{\delta/d}\right)}\right\}, \\
E_3 &= \left\{\forall t \in [1,T] : \|N_t\| \leq \sqrt{\frac{2\alpha\ell(t)}{\epsilon}\left(d + 2\sqrt{d\log\left(\frac{T}{\delta}\right)} + 2\log\left(\frac{T}{\delta}\right)\right)}\right\}
\end{aligned}$$

where $G_t \triangleq \sum_{s=1}^{t} A_s A_s^T$, and let

$$E = E_1 \cap E_2 \cap E_3 \tag{13}$$

Step 3: Showing that E happens with high probability.

For event E_1:

By a direct application of Lemma 26, we get that

$$\mathbb{P}(\neg E_1) \leq \delta.$$

For event E_2:

By a direct application of Lemma 27, we get that

$$\mathbb{P}(\neg E_2) \leq \delta.$$

For event E_3:

Since $N_t \sim \mathcal{N}\left(0, \frac{2\alpha\ell(t)}{\epsilon} I_d\right)$, a direct application of Lemma 24 gives that

$$\mathbb{P}(\neg E_3) \leq \delta.$$

All in all, we get that $\mathbb{P}(E) \geq 1 - 3\delta$.

Step 4: Upper bounding $\|\tilde{\theta}_t - \theta^\star\|_{V_t}$ under E. We have that,

$$\|\tilde{\theta}_t - \theta^\star\|_{V_t} \leq \|S_t\|_{V_t^{-1}} + \|N_t\|_{V_t^{-1}} + \|\lambda\theta^\star\|_{V_t^{-1}}$$

Under E, $V_t \geq (\lambda + \lambda_{\text{min}}(G_t))I_d \geq \lambda I_d$.
Which gives that, under E,

$$\begin{aligned}\|N_t\|_{V_t^{-1}} &\leq \frac{1}{\sqrt{\lambda + \lambda_{\text{min}}(G_t)}}\|N_t\| \\ &\leq \sqrt{\frac{\frac{2\alpha\ell(t)}{\epsilon}\left(d + 2\sqrt{d\log\left(\frac{1}{\delta}\right)} + 2\log\left(\frac{T}{\delta}\right)\right)}{\lambda + \frac{\lambda_0 t}{4} - 8\log\left(\frac{t+3}{\delta/d}\right) - 2\sqrt{t\log\left(\frac{t+3}{\delta/d}\right)}}} \triangleq \frac{\gamma_t}{\sqrt{t}}\end{aligned}$$

and

$$\begin{aligned}\|S_t\|_{V_t^{-1}} + \|\lambda\theta^\star\|_{V_t^{-1}} &\leq \sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(V_t)}{\lambda^d}\right)} + \frac{\lambda}{\sqrt{\lambda}}\|\theta^\star\| \\ &= \sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(V_t)}{\lambda^d}\right)} + \sqrt{\lambda}\|\theta^\star\| \triangleq \beta_t\end{aligned}$$

So, under E, we have that

$$\|\tilde{\theta}_t - \theta^\star\|_{V_t} \leq \tilde{\beta}_t$$

where

$$\tilde{\beta}_t = \beta_t + \frac{\gamma_t}{\sqrt{t}}$$

Step 5: Upper bounding $\det(V_t)$ and $\ell(t)$.

Under E, using the determinant trace inequality, we have that

$$\det(V_t) \leq \left(\frac{1}{d}\text{trace}(V_t)\right)^d \leq \left(\frac{d\lambda + t}{d}\right)^d$$

which gives that

$$\beta_t = \sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(1 + \frac{t}{\lambda d}\right)} + \sqrt{\lambda}\|\theta^\star\|$$

We can say that $\beta_t = \mathcal{O}(\sqrt{d\log(t)})$.

On the other hand, after each episode, the $\det(V_t)$ is, at least, increased multiplicatively by $(1 + C)$, which means that under E, we have that

$$(1 + C)^{\ell(t)}\det(V_0) \leq \det(V_t) \leq \left(\lambda + \frac{t}{d}\right)^d$$

which gives that

$$\ell(t) \leq \frac{d}{\log(1+C)} \log\left(1 + \frac{t}{\lambda d}\right)$$

so $\ell(t) = \mathcal{O}(d \log(t))$ and $\gamma_t = \mathcal{O}\left(\sqrt{\frac{\alpha}{\epsilon}} d \log(t)\right)$

Step 6: Putting everything together.

Under event E, we have that $\|\tilde{\theta}_t - \theta^\star\|_{V_t} \leq \tilde{\beta}_t$ where $\tilde{\beta}_t = \beta_t + \frac{\gamma_t}{\sqrt{t}}$, $\beta_t = \mathcal{O}(\sqrt{d \log(t)})$ and $\gamma_t = \mathcal{O}\left(\sqrt{\frac{\alpha}{\epsilon}} d \log(t)\right)$ such that β_t and γ_t are increasing.

□

D.2 Regret Analysis

Theorem 10. *For rewards in [0,1], Algorithm 4 is (α, ϵ)-global RDP and with probability at least $1 - \mathcal{O}(\delta)$, the regret of Algorithm 4 is upper-bounded by*

$$\mathcal{O}\left(d \log(T) \sqrt{T}\right) + \mathcal{O}\left(\sqrt{\frac{\alpha}{\epsilon}} d^2 \log(T)^2\right)$$

Proof. Let E be the event defined in equation 13.

Step 1: Regret decomposition.

Let $A_t^\star = \arg\max_{a \in \mathcal{A}_t} \langle \theta^\star, a \rangle$.
We have that

$$R_T = \sum_{t=1}^{T} r_t, \quad \text{where} \;\; r_t = \langle \theta^\star, A_t^\star - A_t \rangle$$

Step 2: Instantaneous regret upper bound, under E.

At step t, let τ_t be the last step where $\tilde{\theta}$ was updated.
Let $\mathcal{C}_t = \{\theta \in \mathbb{R}^d : \|\theta - \tilde{\theta}_{t-1}\|_{V_{t-1}} \leq \tilde{\beta}_{t-1}\}$ and $\mathrm{UCB}_t(a) = \max_{\theta \in \mathcal{C}_t} \langle \theta, a \rangle$.
Also, define $\breve{\theta}_{\tau_t} = \arg\max_{\theta \in \mathcal{C}_{\tau_t}} \langle \theta, A_t \rangle$ so that $\mathrm{UCB}_{\tau_t}(A_t) = \left\langle \breve{\theta}_{\tau_t}, A_t \right\rangle$.
Finally, Line 11 of Algorithm 4 could be re-written as $A_t = \arg\max_{a \in \mathcal{A}_t} \mathrm{UCB}_{\tau_t}(a)$.
Under E, we have that

$$\begin{aligned}
r_t &= \langle \theta^\star, A_t^\star - A_t \rangle \\
&\overset{(a)}{\leq} \left\langle \breve{\theta}_{\tau_t} - \theta^\star, A_t \right\rangle \\
&\overset{(b}{\leq} \|\breve{\theta}_{\tau_t} - \theta^\star\|_{V_{t-1}} \|A_t\|_{V_{t-1}^{-1}} \\
&\overset{(c)}{\leq} \sqrt{\frac{\det(V_{t-1})}{\det(V_{\tau_t})}} \|\breve{\theta}_{\tau_t} - \theta^\star\|_{V_{\tau_t}} \|A_t\|_{V_{t-1}^{-1}} \\
&\overset{(d)}{\leq} \sqrt{1+C}(2\tilde{\beta}_{\tau_t}) \|A_t\|_{V_{t-1}^{-1}}
\end{aligned}$$

where:
(a) Under E, $\theta^\star \in \mathcal{C}_{\tau_t}$ and $\langle \theta^\star, A_t^\star \rangle \leq \max_{\theta \in \mathcal{C}_{\tau_t}} \langle \theta, A_t^\star \rangle = \mathrm{UCB}_{\tau_t}(A_t^\star) \leq \mathrm{UCB}_{\tau_t}(A_t) = \left\langle \breve{\theta}_{\tau_t}, A_t \right\rangle$.
(b) By the Cauchy-Schwartz inequality.
(c) By Lemma 28.
(d) By definition of τ_t and Line 6 of Algorithm 4 , we have that $\det(V_{t-1}) \leq (1+C)\det(V_{\tau_t})$ and under E, $\theta^\star \in \mathcal{C}_{\tau_t}$, so $\|\breve{\theta}_{\tau_t} - \theta^\star\|_{V_{\tau_t}} \leq 2\tilde{\beta}_{\tau_t}$.

We also have that $r_t \leq 2$ and $\tilde{\beta}_{\tau_t} \leq \beta_T + \frac{\gamma_T}{\sqrt{\tau_t}}$, which gives

$$r_t \leq 2\sqrt{1+C}\beta_T \left(1 \wedge \|A_t\|_{V_{t-1}^{-1}}\right) + 2\sqrt{1+C}\frac{\gamma_T}{\sqrt{\tau_t}} \left(1 \wedge \|A_t\|_{V_{t-1}^{-1}}\right)$$

Step 3: Upper-bounding the regret, under E.
Under E, we have that

$$
\begin{aligned}
R_T &= \sum_{t=1}^{T} r_t \\
&\le 2\sqrt{1+C}\beta_T \sum_{t=1}^{T} \left(1 \wedge \|A_t\|_{V_{t-1}^{-1}}\right) + 2\sqrt{1+C}\gamma_T \sum_{t=1}^{T} \frac{1}{\sqrt{\tau_t}} \left(1 \wedge \|A_t\|_{V_{t-1}^{-1}}\right) \\
&\le 2\sqrt{1+C}\beta_T \sqrt{T \sum_{t=1}^{T} 1 \wedge \|A_t\|^2_{V_{t-1}^{-1}}} + 2\sqrt{1+C}\gamma_T \sqrt{\left(\sum_{t=1}^{T} \frac{1}{\tau_t}\right)\left(\sum_{t=1}^{T} 1 \wedge \|A_t\|^2_{V_{t-1}^{-1}}\right)} \qquad (14)
\end{aligned}
$$

where the last inequality is due to the Cauchy-Schwartz inequality.

Step 4: The elliptical potential lemma.
We use that $1 \wedge x \le \log(1+x)$ and $\det(V_t) = \det(V_{t-1})\left(1 + \|A_t\|^2_{G_{t-1}(\lambda)^{-1}}\right)$ to have that

$$
\begin{aligned}
\sum_{t=1}^{T} \left(1 \wedge \|A_t\|^2_{V_{t-1}^{-1}}\right) &\le 2\sum_{t=1}^{T} \log\left(1 + \|A_t\|^2_{V_{t-1}^{-1}}\right) \\
&= 2\log\left(\frac{\det(V_T)}{\det(V_0)}\right) \\
&\le 2d\log\left(1 + \frac{T}{\lambda d}\right) \qquad (15)
\end{aligned}
$$

often known as the elliptical potential lemma.

Step 5: Upper-bounding the length of every episode under E.
Episode ℓ starts at t_ℓ and ends at $t_{\ell+1}-1$, so we have that

$$
\frac{\det(V_{t_{\ell+1}-1})}{\det(V_{t_\ell})} \le 1 + C \qquad (16)
$$

On the other hand,

$$
\frac{\det(V_{t_{\ell+1}-1})}{\det(V_{t_\ell})} = \prod_{t=t_\ell+1}^{t_{\ell+1}-1} \left(1 + \|A_t\|^2_{V_{t-1}^{-1}}\right) \qquad (17)
$$

Under E, we use that

$$
V_{t-1} \le (\lambda + \lambda_{\max}(G_{t-1}))\, I_d \le (\lambda + t - 1)\, I_d
$$

since $\lambda_{\max}(G_{t-1}) \le \operatorname{trace}(G_{t-1}) \le t-1$.

which gives that

$$
\|A_t\|^2_{V_{t-1}^{-1}} \ge \frac{1}{\lambda + t - 1}
$$

Plugging in Equation 17, we get that

$$
\begin{aligned}
\frac{\det(V_{t_{\ell+1}-1})}{\det(V_{t_\ell})} &\ge \prod_{t=t_\ell+1}^{t_{\ell+1}-1} \left(1 + \frac{1}{\lambda + t - 1}\right) \\
&= \prod_{t=t_\ell+1}^{t_{\ell+1}-1} \left(\frac{\lambda + t}{\lambda + t - 1}\right) = \frac{\lambda + t_{\ell+1} - 1}{\lambda + t_\ell} \\
&\ge \frac{1}{\lambda + 1} \frac{t_{\ell+1}}{t_\ell}
\end{aligned}
$$

where the last inequality uses that $t_\ell \ge 1$ and $\lambda \ge 1$.

Finally using the upper bound of Equation 16, we get that

$$
\frac{t_{\ell+1}}{t_\ell} \le (1+C)(1+\lambda)
$$

Which gives that

$$\sum_{t=1}^{T} \frac{1}{\tau_t} = \sum_{\ell=1}^{\ell(T)} \sum_{t=t_\ell}^{t_{\ell+1}-1} \frac{1}{t_\ell} = \sum_{\ell=1}^{\ell(T)} \frac{t_{\ell+1} - t_\ell}{t_\ell} \leq (1+C)(1+\lambda)\ell(T) \tag{18}$$

Step 6: Putting everything together.

Plugging the upper bounds of Equation 15 and 18 in the regret upper bound of Equation 14, we get that

$$R_T \leq 2\sqrt{1+C}\sqrt{2d\log\left(1+\frac{T}{\lambda d}\right)}\left(\beta_T\sqrt{T} + \gamma_T\sqrt{(1+C)(1+\lambda)\ell(T)}\right)$$

We finalize by using that

$$\beta_T = \mathcal{O}\left(\sqrt{d\log(T)}\right),\ \ \gamma_T = \mathcal{O}\left(\sqrt{\frac{\alpha}{\epsilon}}d\log(T)\right) \quad \text{and} \ \ \ell(T) = \mathcal{O}\left(d\log(T)\right)$$

We get that

$$R_T \leq \mathcal{O}\left(d\log(T)\sqrt{T}\right) + \mathcal{O}\left(\sqrt{\frac{\alpha}{\epsilon}}d^2\log(T)^2\right)$$

□

D.3 Rectifying LinPriv Regret Analysis

(Neel and Roth 2018) propose LinPriv: Reward-Private Linear UCB, an ϵ-global DP linear contextual bandit algorithm. The context is assumed to be public but adversely chosen. The algorithm is an ϵ-global DP extension of OFUL, where the reward statistics are estimated, at each time-step and for every arm, using a tree-based mechanism (Dwork et al. 2010b; Chan, Shi, and Song 2011).

Theorem 5 in (Neel and Roth 2018) claims that the regret of LinPriv is of order

$$\tilde{\mathcal{O}}\left(d\sqrt{T} + \frac{1}{\epsilon}Kd\log T\right)$$

.

We believe there is a mistake in their regret analysis. In the proof of Theorem 5, they say that

"The crux of their analysis is actually the bound $\sum_{t=1}^{n} \|x_{i,t}\|_{V_{i,t}^{-1}} \leq 2d\log\left(1+\frac{n}{\lambda d}\right)$. "

However, we believe that the result they are citing from (Abbasi-Yadkori, Pál, and Szepesvári 2011) is erroneous. The correct one is

$$\sum_{t=1}^{n} \|x_{i,t}\|^2_{V_{i,t}^{-1}} \leq 2d\log\left(1+\frac{n}{\lambda d}\right),$$

which is known as the elliptical potential lemma (Eq. (15)).

To get the sum, a Cauchy-Schwartz inequality is generally used which leads to

$$\sum_{t=1}^{n} \|x_{i,t}\|_{V_{i,t}^{-1}} \leq \sqrt{n\sum_{t=1}^{n} \|x_{i,t}\|^2_{V_{i,t}^{-1}}} \leq \sqrt{2nd\log\left(1+\frac{n}{\lambda d}\right)}$$

After n is replaced by $\frac{T}{K}$, an additional multiplicative $\sqrt{T}$ should appear in the private regret.

Thus, the rectified regret should be $\tilde{\mathcal{O}}\left(d\sqrt{T} + \frac{1}{\epsilon}Kd\sqrt{T}\right)$.

Remark 4. *In the proof of Theorem 5 (Neel and Roth 2018), to bound the sum $\sum w_{i,t} \leq \mathcal{O}(\sqrt{\log T})\sum_{t=1}^{n} \|x_{i,t}\|_{V_{i,t}^{-1}}$, they use the correct bound on the sum $\sum_{t=1}^{n} \|x_{i,t}\|_{V_{i,t}^{-1}}$ with the $\sqrt{T}$ appearing. However, they misuse it for the private part.*

E Existing Technical Results and Definitions

In this section, we summarise the existing technical results and definitions required to establish our proofs.

Lemma 18 (Post-processing Lemma (Proposition 2.1, (Dwork, Roth et al. 2014)))**.** *If a randomised algorithm $\mathcal{A}$ satisfies (ϵ, δ)-Differential Privacy and f is an arbitrary randomised mapping defined on $\mathcal{A}$'s output, then $f \circ \mathcal{A}$ satisfies (ϵ, δ)-DP.*

Lemma 19 (Markov's Inequality)**.** *For any random variable X and $\varepsilon > 0$,*

$$\mathbb{P}(|X| \geq \varepsilon) \leq \frac{\mathbb{E}[|X|]}{\varepsilon}.$$

Definition 20 (Subgaussianity)**.** *A random variable X is σ-subgaussian if for all $\lambda \in \mathbb{R}$, it holds that*

$$\mathbb{E}[\exp(\lambda X)] \leq \exp(\lambda^2 \sigma^2 / 2)$$

Lemma 21 (Concentration of Subgaussian random variables)**.** *If X is σ-subgaussian, then for any $\epsilon \geq 0$,*

$$\mathbb{P}\left(X \geq \epsilon\right) \leq \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$$

Lemma 22 (Properties of Subgaussian random variables)**.** *Suppose that X_1 and X_2 are independent and σ_1 and σ_2-subgaussian, respectively, then*

1. *cX is $|c|\,\sigma$-subgaussian for all $c \in \mathbb{R}$.*
2. *$X_1 + X_2$ is $\sqrt{\sigma_1^2 + \sigma_2^2}$-subgaussian.*
3. *If X has mean zero and $X \in [a, b]$ almost surely, then X is $\frac{b-a}{2}$-subgaussian.*

Lemma 23 (Theorem 7.8 of (Zhang 2011))**.** *If $A \geq B \geq 0$, then*

- $\det(A) \geq \det(B)$
- *$A^{-1} \leq B^{-1}$ if A and B are non-singular.*

Lemma 24 (Concentration of the χ^2-distribution, Claim 17 of (Shariff and Sheffet 2018))**.** *If $X \sim \mathcal{N}(0, I_d)$ and $\delta \in (0, 1)$, then*

$$\mathbb{P}\left(\|X\|^2 \geq d + 2\sqrt{d \log\left(\frac{1}{\delta}\right)} + 2\log\left(\frac{1}{\delta}\right)\right) \leq \delta$$

Lemma 25 (Concentration of top singular value, Section 4.2 of (Shariff and Sheffet 2018))**.** *If $M \in \mathbb{R}^{d \times d}$ such that $M_{i,j} \overset{iid}{\sim} \mathcal{N}(0, 1)$, $\|M\| \triangleq$ top singular value of M and $\delta \in (0, 1)$, then*

$$\mathbb{P}\left(\|M\| > 4\sqrt{d+1} + 2\log\left(\frac{1}{\delta}\right)\right) \leq \delta$$

Lemma 26 (Theorem 20.4 of (Lattimore and Szepesvári 2018))**.** *Let the noise ρ_t be conditionally 1-subgaussian (conditioned on $A_1, X_1, \ldots, A_{t-1}, X_{t-1}, A_t$), $S_t = \sum_{s=1}^{t} A_s \rho_s$ and $V_t(\lambda) = \lambda I_d + \sum_{s=1}^{t} A_s A_s^T$. Then, for all $\lambda > 0$ and $\delta \in (0, 1)$,*

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \|S_t\|^2_{V_t(\lambda)^{-1}} \geq 2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(V_t(\lambda))}{\lambda^d}\right)\right) \leq \delta$$

Lemma 27 (Lemma 2, Equation (6) of (Gentile, Li, and Zappella 2014))**.** *Let, at each round, $\mathcal{A}_t = \{a_1^t, \ldots, a_{k_t}^t\}$ be generated i.i.d (conditioned on k_t and the history H_t) from a random process A such that*

- $\|A\| = 1$
- *$\mathbb{E}[AA^T]$ is full rank, with minimum eigenvalue $\lambda_0 > 0$*
- *$\forall z \in \mathbb{R}^d$, $\|z\| = 1$, the random variable $(z^T A)^2$ is conditionally subgaussian, with variance*

$$\nu_t^2 = \mathbb{V}\left[(z^T A)^2 \mid k_t, H_t\right] \leq \frac{\lambda_0^2}{8\log(4k_t)}$$

Then

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \lambda_{min}\left(\sum_{s=1}^{t} A_s A_s^T\right) \leq \frac{\lambda_0 t}{4} - 8\log\left(\frac{t+3}{\delta/d}\right) - 2\sqrt{t\log\left(\frac{t+3}{\delta/d}\right)}\right) \leq \delta$$

Lemma 28 (Lemma 12 in (Abbasi-Yadkori, Pál, and Szepesvári 2011))**.** *Let A, B and C be positive semi-definite matrices such that $A = B + C$. Then, we have that*

$$\sup_{x \neq 0} \frac{x^T A x}{x^T B x} \leq \frac{\det(A)}{\det(B)}$$