Concentrated Differential Privacy for Bandits

Achraf Azize
Équipe Scool
Univ. Lille, Inria, CNRS, Centrale Lille, CRIStAL
Lille, France
achraf.azize@inria.fr

Debabrota Basu
Équipe Scool
Univ. Lille, Inria, CNRS, Centrale Lille, CRIStAL
Lille, France
debabrota.basu@inria.fr

***Abstract*—Bandits serve as the theoretical foundation of sequential learning and an algorithmic foundation of modern recommender systems. However, recommender systems often rely on user-sensitive data, making privacy a critical concern. This paper contributes to the understanding of Differential Privacy (DP) in bandits with a trusted centralised decision-maker, and especially the implications of ensuring *zero Concentrated Differential Privacy* (zCDP). First, we formalise and compare different adaptations of DP to bandits, depending on the considered input and the interaction protocol. Then, we propose three private algorithms, namely AdaC-UCB, AdaC-GOPE and AdaC-OFUL, for three bandit settings, namely finite-armed bandits, linear bandits, and linear contextual bandits. The three algorithms share a generic algorithmic blueprint, i.e. the Gaussian mechanism and adaptive episodes, to ensure a good privacy-utility trade-off. We analyse and upper bound the *regret* of these three algorithms. Our analysis shows that in all of these settings, the prices of imposing zCDP are (asymptotically) negligible in comparison with the regrets incurred oblivious to privacy. Next, we complement our regret upper bounds with the first *minimax lower bounds* on the regret of bandits with zCDP. To prove the lower bounds, we elaborate a new proof technique based on couplings and optimal transport. We conclude by experimentally validating our theoretical results for the three different settings of bandits.**

***Index Terms*—Differential Privacy, Multi-armed Bandits, Regret Analysis, Lower bounds**

I. Introduction

For almost a century, Multi-armed bandits (in brief, *bandits*) are studied to understand the cost of partial information and feedback in reinforcement learning, and sequential decision making [1], [2]. In a bandit problem, an agent aims to maximise its accumulated utility by choosing a sequence of actions (or decisions), while the utility of each action is unknown and can be estimated only by choosing it. A Bandit consists of K actions corresponding to K unknown reward distributions $\{\nu_a\}_{a\in[K]}$. We call $\boldsymbol{\nu} \triangleq \{\nu_a\}_{a\in[K]}$ an *environment* or a bandit instance. For T time steps, a bandit algorithm (or policy) π chooses an action (or arm) $a_t \in [K]$ and receives a reward r_t from the reward distribution ν_{a_t}. The goal of the policy is to maximise the cumulative reward $\sum_{t=1}^{T} r_t$ or equivalently minimise the regret, i.e. the cumulative reward that π cannot achieve since it does not know the optimal reward distribution a *priori*.

Bandits constitute the theoretical basis of modern Reinforcement Learning (RL) theory [2]. They are also increasingly used in a wide range of sequential decision-making tasks under uncertainty, such as recommender systems [3], strategic pricing [4], clinical trials [1] to name a few. These applications often involve individuals' sensitive data, such as personal preferences, financial situation, and health conditions, and thus, naturally, invoke data privacy concerns in bandits.

Example 1 (DoctorBandit)**.** *Let us consider a bandit algorithm recommending one of K medicines with distributions of outcomes $\{\nu_a\}_{a\in[K]}$. Specifically, on the t-th day, a new patient u_t arrives, and medicine $a_t \in [K]$ is recommended to her by a policy π. To recommend a medicine a_t, the policy might either consider the specific medical conditions (or context) of patient u_t, or ignore it. Then, the patient's reaction to the medicine is observed. If the medicine cures the patient, the observed reward $r_t = 1$, otherwise $r_t = 0$. This observed reward can reveal sensitive information about the health condition of patient u_t. Thus,* the goal of a privacy-preserving bandit algorithm is to recommend a sequence of medicines (actions) that cures the maximum number of patients while protecting the privacy of these patients. *We present this interactive process in Algorithm 1.*

Algorithm 1 Sequential interaction between a policy and users

Input: A policy $\pi = \{\pi_t\}_{t=1}^{T}$ and Users $\{u_t\}_{t=1}^{T}$
Output: A sequence of actions $a_1, \dots, a_T$
for $t = 1, \dots, T$ **do**
  π recommends $a_t \sim \pi_t(. \mid a_1, r_1, \dots, a_{t-1}, r_{t-1})$
  u_t sends the **sensitive** reward r_t to π
end for

Motivated by such data-sensitive scenarios, privacy issues are widely studied for bandits in different settings, such as finite-armed bandits [5]–[9], adversarial bandits [10], linear contextual bandits [11]–[13], and best-arm identification [14]. All these works adhere to Differential Privacy (DP) [15] as the framework to ensure the data privacy of users, which is presently the gold standard of privacy-preserving data analysis. DP dictates that an algorithm's output has a limited dependency on the presence of any single user. Also, multiple formulations of DP, namely *local* and *global*, are extended to bandits [16]. Here, *we focus on the global DP formulation*, where *users trust the centralised decision-maker*, i.e. the policy, and provide it access to the raw sensitive rewards. The goal of the policy is to reveal the sequence of actions while

protecting the privacy of the users and achieving minimal regret.

The complexity of pure global DP is widely studied for different settings of bandits. In the literature, the lower bound on the regret achievable by any reasonable policy is used to quantify the hardness of imposing privacy in the corresponding bandit setting. In tandem, the goal of the algorithm design is to construct an algorithm whose upper bound on the achievable regret matches the lower bound as much as possible. Recently, lower bounds on regret for finite-armed and linear bandits preserving pure global DP, and algorithm design techniques to match the lower bounds are proposed [8]. This leaves open the question of what will be the minimal cost of preserving the relaxations of pure DP in bandits, as stated in [8], [11]. *Our goal is to provide a complete picture of regret's lower and upper bounds for a relaxation of pure DP*.

In private bandits, proving regret lower bounds often rely on coupling arguments where group privacy is a central property [8]. Since zCDP scales well under group privacy, we adopt zCDP as the relaxation of pure DP. In this work, we investigate zCDP in three settings of bandits: *finite-armed bandits*, *stochastic linear bandits with (fixed) finitely many arms*, and *contextual linear bandits*. To our knowledge, we are the first to study the complexity of zCDP for bandits with global DP.

Contributions. Specifically, our contributions are as follows:

1) **Privacy Definitions for Bandits.** We compare different ways of adopting relaxations of DP for bandits. We observe that, though for pure DP some of these definitions are equivalent, more care is needed for approximate and zero Concentrated DP. We explicate two main distinctions in the definitions. The first is dealing with the bandit feedback when defining the private input dataset. The second is whether to consider or not the interactive nature of the policy as a mechanism. Formalising and linking these definitions is a crucial step that was missing in the private bandits literature. Our first contribution is to fill this gap.
2) **Algorithm Design.** Following the study of privacy definitions for bandits, we adhere to ρ-Interactive zCDP as the main privacy definition. We propose three algorithms, namely AdaC-UCB, AdaC-GOPE, and AdaC-OFUL, that achieve ρ-Interactive zCDP, *almost for free*, for three bandit settings, namely finite-armed bandits, stochastic linear bandits with (fixed) finitely many actions and linear contextual bandits with context-dependent feasible actions. *These three algorithms share the same blueprint.* First, they add a calibrated *Gaussian noise* to reward statistics. Second, they run in *adaptive episodes*, with the number of episodes logarithmic in T. This means that the algorithm only accesses the private reward dataset 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, less injection of Gaussian noise.
3) **Regret Analysis.** We analyse the regrets of the proposed algorithms and show that ρ-Interactive zCDP can be preserved almost for free in terms of the regrets. Specifically, for a fixed privacy budget ρ, and asymptotically in the horizon T, the cost of ρ-Interactive zCDP in the regret of these algorithms exhibits an additional $\tilde{\mathcal{O}}(\rho^{-1/2}\log(T))$, which is significantly lower than the privacy oblivious regret, i.e. $\tilde{\mathcal{O}}(\sqrt{T})$. In Table I, we summarise the regret upper bounds corresponding to the three proposed algorithms. We also numerically validate the performance of the three algorithms and the corresponding theoretical results in different settings.
4) **Hardness of Preserving Privacy in Bandits as Lower Bounds.** Addressing the open problem of [8], [11], we prove minimax lower bounds for finite-armed bandits and linear bandits with ρ-Interactive zCDP, that quantify the cost to ensure ρ-Interactive zCDP in these settings. To prove the lower bound, we develop a new proof technique that relates minimax lower bounds to a transport problem. The minimax lower bounds show the existence of two privacy regimes depending on the privacy budget ρ and the horizon T. Specifically, for $\rho = \Omega(T^{-1})$, *an optimal algorithm does not have to pay any cost to ensure privacy* in both settings. The regret lower bounds show that AdaC-UCB, AdaC-GOPE, and AdaC-OFUL are optimal, up to poly-logarithmic factors. In Table I, we summarise the corresponding regret lower bounds.

Outline. The outline of the paper is as follows. First, we discuss privacy definitions for bandits in Section III. In Section IV, we propose AdaC-GOPE and AdaC-OFUL, for linear and contextual bandits. We provide a privacy and regret analysis of these two algorithms in Section V. We discuss lower bounds for zCDP in Section VI. The analysis of the complexity of zCDP in finite-armed bandits is deferred to Appendix C. Finally, we experimentally validate the theoretical insights in Section VII before concluding. Before diving into the technical details, we discuss the relevant literature of differentially private bandits in Section II.

II. Related Works

In this section, we discuss the relevant literature of differentially private bandits, and posit our contributions in the light of them.

a) Privacy Definitions for Bandits: In this paper, we first aim to clarify different definitions of Differential Privacy (DP) considered in the context of bandits. In the presence of a trusted centralised decision-maker, the two formulations of DP considered for bandits are Table DP and View DP. Interestingly, existing DP bandit literature has considered as a "folklore" result that View DP and Table DP are equivalent, e.g. footnote 1 in [17] and Section 3 of [16]. To the best of our knowledge, *we provide the first formal proof of the equivalence between View DP and Table DP in the case of pure ϵ-DP and falsify the equivalence for the relaxations of DP, such as (ϵ, δ)-DP*. This difference is not clear if we look into an atomic sequence of actions (e.g. probability of $\{a_1, \ldots, a_T\}$) but they

TABLE I: Regret bounds for bandits with ρ-Interactive zCDP. Terms in blue correspond to the cost of ρ-Interactive zCDP.

Bandit Setting	Regret Upper Bound	Regret Lower Bound
Finite-armed bandits	$\mathcal{O}\left(\sqrt{KT\log(T)}\right) + \mathcal{O}\left(\frac{K}{\sqrt{\rho}}\sqrt{\log(T)}\right)$ (Thm 11)	$\Omega\left(\max\left(\sqrt{KT}, \sqrt{\frac{K}{\rho}}\right)\right)$ (Thm 17)
Linear bandits	$\mathcal{O}\left(\sqrt{dT\log(KT)}\right) + \mathcal{O}\left(\frac{d}{\sqrt{\rho}}\log^{\frac{3}{2}}(KT)\right)$ (Thm 4)	$\Omega\left(\max\left(d\sqrt{T}, \frac{d}{\sqrt{\rho}}\right)\right)$ (Thm 9) [a]
Linear Contextual bandits	$\mathcal{O}\left(d\log(T)\sqrt{T}\right) + \mathcal{O}\left(\frac{d^2}{\sqrt{\rho}}\log(T)^2\right)$ (Thm 5)	

[a] The non-private lower bound of $\Omega(d\sqrt{T})$ does not contradict the $\mathcal{O}\left(\sqrt{dT\log(KT)}\right)$ of linear bandits with K arms. As explained in Sec 24.1. of [2], the size of the action set in the proof of the lower bound corresponds to $K = 2^d$, and thus, the dependence on d is tight.

differ while considering composite events (e.g. probability of $\{(a_1, \ldots, a_T), (a'_1, \ldots, a'_T)\}$). Control of such composite events becomes important under the relaxations of DP. We discuss this in detail in Section III and Appendix B.

On the other hand, we discuss why considering an interactive adversary is important in a sequential setting like bandits. We develop an Interactive DP definition for bandits (Definition 4) based on the framework of [18], [19]. Recently, a similar definition of Interactive DP has been proposed by [20] for the continual observation setting under adaptively chosen queries (Section 5.1, [20]). Our Interactive DP definition can be perceived as an adaptation of the Interactive DP definition of [20] to the "partial information setting" of bandits. Detailed discussion is deferred to Remark 2.

b) Algorithm Design: [8] proposes a generic framework to make any index-based algorithms achieve ϵ-pure global DP, in the stochastic finite-armed bandit setting. This framework has three main ingredients: *per-arm doubling*, *forgetting*, and *adding Laplace noise*. AdaC-UCB is an extension of this framework to zCDP. On the other hand, the design choices for AdaC-GOPE and AdaC-OFUL are quite different from the framework in [8]. AdaC-GOPE runs in phases. However, these phases are *not* arm-dependent and *not* necessarily doubling. On the other hand, one can perceive that AdaC-OFUL deploys a *generalisation of per-arm doubling* to contextual linear bandits, using the *doubling of the determinant of the design matrix* trick. However, AdaC-OFUL does not forget the samples from the previous phases (Line 8, Algorithm 3). For linear bandits with a finite number of arms, [13], [21] also propose two private variants of GOPE algorithm [2]. In Section IV, we show that AdaC-GOPE achieves lower regret than both [13], [21]. [11], [12] also propose two differentially private variants of OFUL [22] for linear contextual bandits. In Section IV-B, we propose a differentially private variant of OFUL, namely AdaC-OFUL, that achieves lower regret than the existing algorithms (Theorem 5). In this work, we consider rewards to be the private information and contexts to be public [12], whereas one can consider both of them to be jointly private [11], which we do not consider in this paper.

c) Comparison with Regret Bounds under Pure DP: Every ϵ-DP algorithm is ρ-zCDP with $\rho = \frac{1}{2}\epsilon^2$ (Proposition 1.4, [23]). Due to this observation, it is possible to provide zCDP regret upper bounds from the ϵ-DP bandit literature, by replacing ϵ with $\sqrt{2\rho}$ in those results. Our zCDP upper bounds improve on these "converted" upper bounds on logarithmic terms in T, K, and d. This improvement is due to the use of the Gaussian Mechanism rather than the Laplace mechanism. Table II summarises the comparison.

d) Hardness of Preserving Privacy in Bandits as Lower Bounds: To prove regret lower bounds in bandits, we leverage the generic proof ideas in [2]. The main technical challenge in these proofs is to quantify the extra cost of "indistinguishability" due to DP. This cost is expressed in terms of an upper bound on KL-divergence of observations induced by two 'confusing' bandit environments. For pure DP [8], the upper bound on the KL-divergence (Theorem 10 in [8]) is proved by adapting the Karwa-Vadhan lemma [24] to the bandit sequential setting. To our knowledge, there is no zCDP version of the Karwa-Vadhan lemma. Thus, we first provide a general result in Theorem 6, which could be seen as a generalisation of the Karwa-Vadhan lemma to zCDP. To prove this result, we derive a new maximal coupling argument relating the KL upper bound to an optimal transport problem, which can be of parallel interest. Then, we adapt it to the bandit setting in Theorem 7. The regret lower bounds are retrieved by plugging in these upper bounds on the KL-divergence in the generic lower bound proof of bandits.

III. Privacy definitions for bandits

We first recall the definition of Differential Privacy (DP) and the bandit canonical model. Then, we compare different adaptations of DP to bandits under the centralised model. These adaptations differ in the nature of the input considered and the nature of the interaction protocol.

A. Background: Differential Privacy and Bandits

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

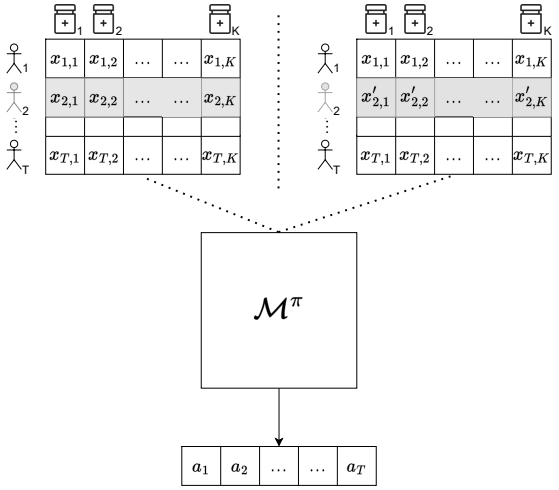


Fig. 1: Table DP

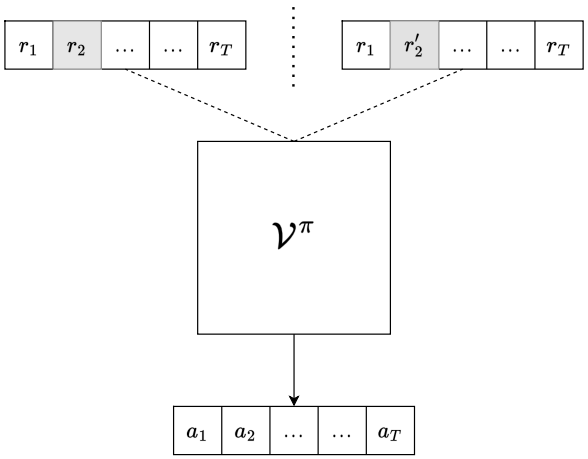


Fig. 2: View DP

Definition 1 ((ϵ, δ)-DP [15] and ρ-zCDP [23])**.** *A mechanism $\mathcal{M}$, which assigns to each dataset d a probability distribution $\mathcal{M}_d$ on some measurable space $(\mathbb{X}, \mathcal{F})$, satisfies*

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

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

- *ρ-zCDP if, for all $\alpha > 1$,*

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

Here, two datasets d and d' are said to be neighbouring, and are 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.

Now, we recall the **canonical model of bandits** (Sec 4.6., [2]).

Definition 2. *A bandit algorithm (or **policy**) π is a sequence of rules $(\pi_t)_{t=1}^T$, where $\pi_t : \mathcal{H}_{t-1} \to \Delta_K$ is a probability kernel that assigns to a history $\mathcal{H}_{t-1}$ a distribution over arms, and Δ_K is the simplex over $[K]$.*

A bandit algorithm (or policy) π interacts with an environment ν consisting of K arms (or actions) with reward distributions $\{\nu_a\}_{a=1}^K$ for a given horizon T, and produces a history $\mathcal{H}_T \triangleq \{(A_t, R_t)\}_{t=1}^T$. At each step t, the choice of the arm A_t depends on the previous history $\mathcal{H}_{t-1}$, i.e. $A_t \sim \pi_t(. \,|\mathcal{H}_{t-1})$. The reward R_t is sampled from the reward distribution ν_{A_t} and is conditionally independent of the previous history $\mathcal{H}_{t-1}$.

In order to rigorously adapt DP to bandits, it is important to specify: (a) *the mechanism* in question, (b) its *input dataset*, (c) the *neighbouring relationship between the input datasets* and (d) the *output* of the mechanism.

B. Challenges in Adapting DP for Bandits

In the DoctorBandit (Example 1), privacy concerns emerge from the sensitivity of the reward information, i.e. the reaction of a patient to a medicine could disclose private information about their health condition. The published output is the sequence of recommended medicines, i.e. $(a_1, \ldots, a_T)$. Thus, the mechanism to be made private is induced by the policy π.

As privacy is a worst-case constraint, any definition of privacy in bandits should only depend on the policy π, and be independent of any (stochastic) environment considerations. Rather, a privacy definition should be perceived as a constraint on the class of policies to be considered.

The *first challenge* in defining DP for bandits is *to determine the private input dataset, due to the bandit feedback*. Specifically, each patient u_t can be represented by the vector of their potential reactions $x_t \triangleq (x_{t,1}, \ldots, x_{t,K}) \in \{0, 1\}^K$. If the policy π recommends an action a_t for user u_t, only the reward $r_t \triangleq x_{t,a_t}$ is observed. There are two possible ways to deal with the partial information in adapting DP.

i. Consider that the private input is the table of all potential rewards $d \triangleq (x_1, \ldots, x_T) \in (\{0, 1\}^K)^T$, which we call **Table DP**.
ii. Consider the input as a list of "fixed in advance" observed rewards $\mathbf{r} \triangleq \{r_1, \ldots, r_T\}$, which we call **View DP**.

The *second challenge* in defining DP is *to determine the composition protocol*. The sequence of the published actions can be seen as the answer to T adaptively chosen queries, on adaptively gathered data. A policy π can be seen as a mechanism that interactively produces a sequence of actions, answering T adaptively chosen queries, by a potentially adversarial analyst. It is thus natural to induce an interactive mechanism from the policy π and adapt to it the **Interactive DP** definition as studied in [19], [25].

C. Table DP vs. View DP

We denote the mechanism induced by the interaction of a policy π and a table of rewards $d \triangleq \{(x_{t,i})_{i \in [K]}\}_{t \in [T]} \in (\mathbb{R}^K)^T$ as the mechanism $\mathcal{M}^\pi$, such that

$$\begin{aligned} \mathcal{M}^\pi : \quad (\mathbb{R}^K)^T &\to \mathcal{P}([K]^T) \\ d \; &\to \mathcal{M}_d^\pi \,. \end{aligned}$$

Here, $\mathcal{M}_d^\pi$ is a distribution over the sequence of actions, and $\mathcal{M}_d^\pi(a_1, \dots, a_T) = \prod_{t=1}^T \pi_t\left(a_t | a_1, x_{1,a_1}, \dots a_{t-1}, x_{t-1,a_{t-1}}\right)$. The hamming distance between two table of rewards $d, d' \in (\mathbb{R}^K)^T$ is the number of different *rows* in d and d', i.e. $d_{\text{Ham}}(d, d') \triangleq \sum_{t=1}^T \mathbb{1}\{x_t \neq x'_t\} = \sum_{t=1}^T \mathbb{1}\left\{\exists i \in [K],\ x_{t,i} \neq x'_{t,i}\right\}$.

The mechanism induced by the interaction of π and a list of rewards $\mathbf{r} \triangleq (r_t)_{t\in[T]} \in \mathbb{R}^T$ is denoted by $\mathcal{V}^\pi$, such that

$$\begin{aligned} \mathcal{V}^\pi :\ \mathbb{R}^T &\to \mathcal{P}([K]^T) \\ \mathbf{r} &\to \mathcal{V}^\pi_{\mathbf{r}}\,. \end{aligned}$$

Here, $\mathcal{V}_d^\pi$ is a distribution over the sequence of actions, and $\mathcal{V}_{\mathbf{r}}^\pi(a_1, \dots, a_T) = \prod_{t=1}^T \pi_t(a_t | a_1, r_1, \dots a_{t-1}, r_{t-1})$. The Hamming distance between two lists of rewards $r, r' \in \mathbb{R}^T$ is the number of different elements in r and r', i.e. $d_{\text{Ham}}(r, r') \triangleq \sum_{t=1}^T \mathbb{1}\{r_t \neq r'_t\}$

Remark 1. *The expressions of $\mathcal{V}_{\mathbf{r}}^\pi(a_1, \dots, a_T)$ and $\mathcal{M}_r^\pi(a_1, \dots, a_T)$ as products capture the sequential nature of producing the sequence of actions $(a_1, \dots, a_T)$. At first glance, the two expressions look very similar. However, the differences arise when $\mathcal{V}_{\mathbf{r}}^\pi$ and $\mathcal{M}_d^\pi$ are applied to non-atomic event $E \in \mathcal{P}([K]^T)$. For example, if we define an event $E \triangleq \{(a_1, \dots, a_T), (b_1, \dots, b_T)\}$, then $\mathcal{V}_{\mathbf{r}}^\pi(E) = \prod_{t=1}^T \pi_t(a_t | a_1, r_1, \dots a_{t-1}, r_{t-1}) + \prod_{t=1}^T \pi_t(b_t | b_1, r_1, \dots b_{t-1}, r_{t-1})$, while $\mathcal{M}_d^\pi(E) = \prod_{t=1}^T \pi_t(a_t | a_1, x_{1,a_1}, \dots a_{t-1}, x_{t-1,a_{t-1}}) + \prod_{t=1}^T \pi_t(b_t | a_1, x_{1,b_1}, \dots b_{t-1}, x_{t-1,b_{t-1}})$. In the expression of $\mathcal{V}_{\mathbf{r}}^\pi(E)$, the same rewards appear in the elements of the sum. In contrast, in the expression of $\mathcal{M}_d^\pi(E)$, each sequence of actions generates different trajectories of reward in the table. As we show later, this subtle difference is the source of the difference between Table DP and View DP.*

Now that the mechanisms are explicit, the corresponding definitions of DP follow naturally.

Definition 3 (Table DP and View DP)**.** *A policy π ensures*

- *(ϵ, δ)-Table DP if and only if $\mathcal{M}^\pi$ is (ϵ, δ)-DP,*
- *(ϵ, δ)-View DP if and only if $\mathcal{V}^\pi$ is (ϵ, δ)-DP,*
- *ρ-Table zCDP if and only if $\mathcal{M}^\pi$ is ρ-zCDP,*
- *ρ-View zCDP if and only if $\mathcal{V}^\pi$ is ρ-zCDP.*

Table DP is a formalisation of the privacy definition adopted in [5], [12], while View DP is a formalisation of the definition adopted in [7], [8], [13], [26].

We summarise the relations between Table DP and View DP in the following proposition. For brevity, the proofs are deferred to Appendix B.

Proposition 1 (Relation between Table DP and View DP)**.** *For any policy π, we have that*

(a) $\mathcal{M}^\pi$ is ϵ-DP $\Leftrightarrow$ $\mathcal{V}^\pi$ is ϵ-DP.
(b) $\mathcal{M}^\pi$ is (ϵ, δ)-DP $\Rightarrow$ $\mathcal{V}^\pi$ is (ϵ, δ)-DP.
(c) $\mathcal{M}^\pi$ is ρ-zCDP $\Rightarrow$ $\mathcal{V}^\pi$ ρ-zCDP.
(d) $\mathcal{V}^\pi$ is (ϵ, δ)-DP $\Rightarrow$ $\mathcal{M}^\pi$ is $(\epsilon, K^T\delta)$-DP.
(e) $\Pi_{\text{Table}}^{(\epsilon,\delta)} \subsetneq \Pi_{\text{View}}^{(\epsilon,\delta)}$,

where $\Pi_{\text{Table}}^{(\epsilon,\delta)}$ and $\Pi_{\text{View}}^{(\epsilon,\delta)}$ are the class of all policies verifying (ϵ, δ)-Table DP and (ϵ, δ)-View DP, respectively.

The Consequences of Proposition 1. Proposition 1 establishes that Table DP is a "stronger" notion of privacy than View DP. Table DP protects all the **potential** responses of an individual rather than just the **observed** one.

Specifically, Proposition 1(a) shows that Table DP and View DP are equivalent for pure DP, i.e. $(\epsilon, 0)$-DP. For relaxations of pure DP, i.e. for (ϵ, δ)-DP and ρ-zCDP, Proposition 1(b) and 1(c) show that Table DP always implies View DP with the same privacy budget.

However, the converse from View DP to Table DP happens with a loss in the privacy budget. Proposition 1(e) states that the class of policies verifying (ϵ, δ)-Table DP is *strictly* included in the class of policies verifying (ϵ, δ)-View DP. To prove this, we build a policy that verifies some (ϵ_1, δ_1)-View DP but is shown to be never (ϵ_1, δ_1)-Table DP. This validates that going from View DP to Table DP must happen with a *loss* in the privacy budget. Proposition 1(d) yields a simple quantification of the loss. We leave it as an open problem to quantify the best privacy loss conversion from View DP to Table DP. It would be an interesting question to investigate if the equivalence between View DP and Table DP is still valid for (ϵ, δ)-DP, for some very small δ regime.

An Intuition. We observe that under bandit feedback, pure DP and relaxations of DP behave differently for Table DP and View DP. To provide an intuition behind this phenomenon, we would like to revise Remark 1. For pure DP, it is enough tp bound the change in the probability for "atomic" sequences of actions $(a_1, \dots, a_T)$. For such "atomic" event, it is easy to go from $\mathcal{V}^\pi$ to $\mathcal{M}^\pi$ (Reduction 1) and back (Reduction 2). For relaxations of DP, this does not hold true anymore. The details of the proof of Proposition 1 are available in Appendix B.

D. Interactive DP

Bandits inherently operate through an interactive process (Algorithm 1). It is possible to induce an interactive mechanism from a policy π, viewed as a party in an interactive protocol, interacting with a possibly adversarial analyst.

The interaction protocol has three elements: (i) the policy $\pi = \{\pi_t\}_{t=1}^T$, (ii) a private input dataset which we consider to be the table of potential rewards[1] $d \triangleq (x_1, \dots, x_T) \in (\mathbb{R}^K)^T$, and (iii) an adversary $B \triangleq \{B_t\}_{t=1}^T$.

The interaction protocol is the following:

For $t = 1, \dots, T$

1) The bandit algorithm selects an action
$$o_t \sim \pi_t(\cdot \mid q_1, x_{1,q_1}, \dots, q_{t-1}, x_{t-1,q_{t-1}}),$$
2) The adversary returns a query action
$$q_t = B_t(o_1, o_2, \dots, o_t).$$
3) The bandit algorithm observes the reward corresponding to q_t for user u_t, i.e. x_{t,q_t}.

[1]It is also possible to consider a "View" definition of Interactive DP.

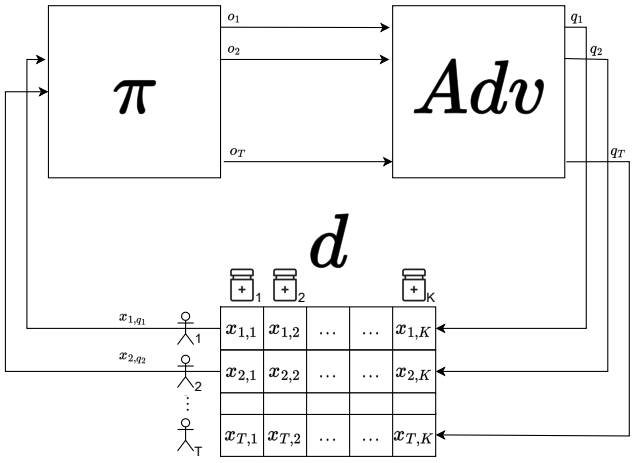


Fig. 3: Sequential interaction between the policy, an adversary, and a table of rewards.

We represent this interaction by $\pi \leftrightarrow^d B$, and illustrate it in Figure 3.

The main difference between this interaction protocol and Algorithm 1 is that the reward revealed to the bandit algorithm, at step t, is not the reward corresponding to the action recommended by the policy, i.e. o_t, but from a query action chosen by the adversary, i.e. q_t. The query action q_t is chosen by the adversary depending on its current view, i.e. the sequence of recommended actions $(o_1, \ldots, o_t)$.

Following the Interactive DP framework [19], the policy π is a differentially private interactive mechanism if the view of adversary B, i.e.

$$\text{View}_{B,\pi,d} \triangleq \text{View}_B(\pi \leftrightarrow^d B) \triangleq (o_1, \ldots, o_T),$$

is indistinguishable when the interaction is run on two neighbouring tables of rewards d and d'.

Definition 4 (Interactive DP)**.** *A policy π satisfies*

- *(ϵ, δ)-Interactive DP for a given $\epsilon \geq 0$ and $\delta \in [0, 1)$, if for all adversaries B and all subset of views $\mathcal{S} \subseteq [K]^T$,*

$$\sup_{d \sim d'} \Pr[\text{View}_{B,\pi,d} \in \mathcal{S}] - e^{\epsilon} \Pr[\text{View}_{B,\pi,d'} \in \mathcal{S}] \leq \delta.$$

- *ρ-Interactive zCDP policy for a given $\rho \geq 0$, if for every $\alpha > 1$, and every adversary B,*

$$\sup_{d \sim d'} D_\alpha(\text{View}_{B,\pi,d} \| \text{View}_{B,\pi,d'}) \leq \rho\alpha.$$

Remark 2. *Interactive DP in Definition 4 can be perceived as an adaptation of the "adaptive" privacy definition in Section 5.1 of [20] to the bandit setting. At each step t, the adversary of [20] chooses a query r_t to send to the policy π, depending on the history of the interaction between the policy and the adversary. The adversary in Definition 4 also adaptively chooses a query $r_t = x_{t,q_t}$ to send to the policy, but by shuffling over a "fixed-in-advance" table of rewards d. Thus, the adversary in [20] might be stronger than the one in Definition 4. However, it is an interesting question to see if the two definitions are equivalent, i.e. can a "fully" adaptive adversary for bandits be simulated as a shuffling adversary over a fixed table of rewards?*

We elaborate on three interesting implications of the interactive definition of privacy in bandits.

(a) Interactive DP defends against a more realistic sequential adversary, who can "manipulate" the rewards observed by the policy at every and any step.

(b) Interactive DP protects the privacy of the users even if the users are non-compliant [27], [28], i.e. the users decide to ignore the recommendations of the policy and choose a different arm.

(c) Interactive DP inherently provides robustness against online reward poisoning attacks [29].

We relate Interactive DP and Table DP in the following proposition. The proofs are detailed in Appendix B.

Proposition 2. *For any policy π, we have that*

(a) π is ρ-Interactive zCDP $\Rightarrow$ π is ρ-Table zCDP

(b) π is ρ-Interactive zCDP if and only if, for every deterministic adversary $B = \{B_t\}_{t=1}^T$, π^B is ρ-Table zCDP. Here, $\pi^B \triangleq \{\pi_t^B\}_{t=1}^T$ is a post-processing of the policy π induced by the adversary B such that

$$\pi_t^B(a \mid a_1, r_1, .., a_{t-1}, r_{t-1}) \triangleq$$
$$\pi_t\Big(a \mid B_1(a_1), r_1, B_2(a_1, a_2), r_2, .., B_{t-1}(a_1, .., a_{t-1}), r_{t-1}\Big).$$

Proposition 2 shows that, for policies that are "closed" under interactive post-processing, ρ-Interactive zCDP and ρ-Table zCDP are equivalent. Algorithms in private bandits literature, which are based on the binary tree mechanism [30], [31] and our non-overlapping adaptive episode mechanism (Lemma 1) verify both Table and Interactive DP[2].

The motivation for proposing Interactive DP as a privacy definition for bandits is that the class of Interactive DP policies gives a better representation of the algorithms already developed in the private bandits' literature, has interesting implications, and provides better "group privacy" decomposition, which plays a crucial role when deriving lower bounds in Section VI.

Theorem 1 (Group Privacy for ρ-Interactive DP)**.** *If π is a ρ-Interactive zCDP policy then, for any sequence of actions $(a_1, \ldots, a_T)$ and any two sequence of rewards $\boldsymbol{r} \triangleq \{r_1, \ldots, r_T\}$ and $\boldsymbol{r'} \triangleq \{r'_1, \ldots, r'_T\}$, we have that*

$$\sum_{t=1}^{T} \text{KL}\left(\pi_t(. \mid \mathcal{H}_{t-1}) \;\middle\|\; \pi_t(. \mid \mathcal{H}'_{t-1})\right) \leq \rho d_{Ham}(\boldsymbol{r}, \boldsymbol{r'})^2$$

where $\mathcal{H}_t \triangleq (a_1, r_1, \ldots, a_t, r_t)$, $\mathcal{H}'_t \triangleq (a_1, r'_1, \ldots, a_t, r'_t)$ and $d_{Ham}(\boldsymbol{r}, \boldsymbol{r'}) = \sum_{t=1}^T \mathbb{1}\{r_t \neq r'_t\}$.

The proof is provided in Appendix B, and uses the decoupling induced by a constant adversary.

Hereafter, we adhere to ρ-Interactive zCDP as the definition of privacy for bandits. We refer to the class of policies

[2] [20] studies the binary tree mechanism under Interactive DP.

verifying ρ-Interactive zCDP as Π^{ρ}_{Int}. The goal is to design a policy $\pi \in \Pi^{\rho}_{\text{Int}}$ that maximises the expected sum of rewards, or equivalently minimises the expected regret when interacting with a class of environments, *using the bandit canonical model*. In Section IV, we define the exact regret for each setting under study. Note that, for the contextual bandit setting, contexts can assumed to be either public or private depending on the application of interest. In Appendix B-D, we discuss how to extend the privacy definitions when the contexts are assumed to be private, and also the limitations of the "public contexts" assumption, which is considered here.

Remark 3. *ρ-Interactive DP can be perceived as a constraint on the class of policies to be considered. To express this constraint, the interactive protocol between the policy, an adversary, and a table of rewards is defined (Fig 3). An interactive private policy is constrained to show a similar view to any "privacy" adversary when interacting with two neighbouring reward tables. On the other hand, to measure the quality of a policy in the class of ρ-Interactive DP policies, we compute the regret of the policy when interacting with a class of environments using the canonical bandit protocol, i.e. the rewards are stochastically generated from an arm-dependent distribution, and there is no "privacy" adversary changing the arms chosen by the policy. In other words, the interaction protocols to analyse privacy and regret are different.*

IV. Algorithm Design

In this section, we propose AdaC-GOPE and AdaC-OFUL, two algorithms that satisfy ρ-Interactive zCDP for linear bandits and contextual linear bandits respectively. The two algorithms share a similar blueprint: adding Gaussian noise and having adaptive episodes. AdaC-UCB share similar ingredients for finite armed bandits. AdaC-UCB is presented and analysed in the appendix.

A. Stochastic Linear Bandits

Here, we study ρ-Interactive zCDP for stochastic linear bandits with a finite number of arms.

1) Setting: We consider that a fixed set of actions $\mathcal{A} \subset \mathbb{R}^d$ is available at each round, such that $|\mathcal{A}| = K$. The rewards are generated by a linear structural equation. Specifically, at step t, the observed reward is $r_t \triangleq \langle \theta^\star, a_t \rangle + \eta_t$, where $\theta^\star \in \mathbb{R}^d$ is the unknown parameter, and η_t is a conditionally 1-subgaussian noise, i.e. $\mathbb{E}\left[\exp\left(\lambda \eta_t\right) \mid a_1, \eta_1, \ldots, a_{t-1}\right] \leq \exp\left(\lambda^2/2\right)$ almost surely for all $\lambda \in \mathbb{R}$.

For any horizon $T > 0$, the regret of a policy π is

$$\mathrm{Reg}_T(\pi, \mathcal{A}, \theta^\star) \triangleq \mathbb{E}_{\theta^\star}\left[\sum_{t=1}^{T} \Delta_{A_t}\right], \tag{3}$$

where suboptimality gap $\Delta_a \triangleq \max_{a' \in \mathcal{A}} \langle a' - a, \theta^\star \rangle$. $\mathbb{E}_{\theta^\star}[\cdot]$ is the expectation with respect to the measure of outcomes induced by the interaction of π and the linear bandit environment $(\mathcal{A}, \theta^\star)$.

2) Algorithm: We propose AdaC-GOPE (Algorithm 2), which is a ρ-Interactive zCDP extension of the G-Optimal design-based Phased Elimination (GOPE) algorithm [2, Algorithm 12]. AdaC-GOPE is a phased elimination algorithm. At the end of each episode ℓ, AdaC-GOPE eliminates the arms that are likely to be sub-optimal, i.e. the ones with an empirical gap exceeding the current threshold ($\beta_\ell = 2^{-\ell}$). The elimination criterion only depends on the samples collected in the current episode. In addition, the actions to be played during an episode are chosen based on the solution of an optimal design problem (Equation (5)) that helps to exploit the structure of arms and to minimise the number of samples needed to eliminate a sub-optimal arm.

In particular, if π_ℓ is the G-optimal solution (Definition 5) for $\mathcal{A}_\ell$ at phase ℓ, then each action $a \in \mathcal{A}_\ell$ is played $T_\ell(a) \triangleq \lceil c_\ell \pi_\ell(a) \rceil$ times, where for $\delta_{K,\ell} \triangleq \frac{\delta}{K\ell(\ell+1)}$ and $f(d,\delta) \triangleq d + 2\sqrt{d \log\left(\frac{2}{\delta}\right)} + 2\log\left(\frac{2}{\delta}\right)$,

$$c_\ell \triangleq \frac{8d}{\beta_\ell^2} \log\left(\frac{4}{\delta_{K,\ell}}\right) + \frac{2d}{\beta_\ell}\sqrt{\frac{2}{\rho} f\left(d, \delta_{K,\ell}\right)} \tag{4}$$

The term in blue is the additional length of the episode to compensate for the noisy statistics used to ensure privacy. The samples collected in the current episode do not influence which actions are played in it. This decoupling allows: (a) the use of the tighter confidence bounds available in the fixed design setting (Appendix E-A), and (b) avoiding privacy composition theorems and using, therefore, Lemma 1 to make the algorithm private. Note that AdaC-GOPE can be seen as a generalisation of DP-SE [7] to the linear bandit setting.

Here, we present the definitions of optimal design and a classic equivalence result required to state Algorithm 2.

Definition 5 (Optimal design [32])**.** *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*

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

- *π is called a **design**.*
- *The set $\mathrm{Supp}(\pi) \triangleq \{a \in \pi : \pi(a) \neq 0\}$ is called the **core set** of $\mathcal{A}$.*
- *A design that maximises f is called a **D-optimal design**.*
- *A design that minimises g is called a **G-optimal design**.*

Theorem 2 (Kiefer–Wolfowitz theorem [33])**.** *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, and*
- *$g(\pi^\star) = d$.*

Also, there exists a minimiser $\pi^\star$ of g such that $|\mathrm{Supp}(\pi^\star)| \leq \frac{d(d+1)}{2}$.

Algorithm 2 AdaC-GOPE

1: **Input:** Privacy budget ρ, $\mathcal{A} \subset \mathbb{R}^d$ and δ
2: **Output:** Actions satisfying ρ-Interactive zCDP
3: **Initialisation:** Set $\ell = 1$, $t_1 = 1$ and $\mathcal{A}_1 = \mathcal{A}$
4: **for** $\ell = 1, 2, \dots$ **do**
5: $\quad \beta_\ell \leftarrow 2^{-\ell}$
6: $\quad$ **Step 1:** Find the G-optimal design π_ℓ for $\mathcal{A}_\ell$:

$$\max_{\substack{\pi \in \mathcal{P}(\mathcal{A}_\ell) \\ |\mathrm{Supp}(\pi)| \leq d(d+1)/2}} \log\det V(\pi). \tag{5}$$

7: $\quad$ **Step 2:** $\mathcal{S}_\ell \leftarrow \mathrm{Supp}\left(\pi_\ell\right)$
8: $\quad$ Choose each action $a \in \mathcal{S}_\ell$ for $T_\ell(a) \triangleq \lceil c_\ell \pi_\ell(a) \rceil$ times where c_ℓ is defined by Eq (4).
9: $\quad$ Observe rewards $\{r_t\}_{t=t_\ell}^{t_\ell + \sum_a T_\ell(a)}$
10: $\quad T_\ell \leftarrow \sum_{a \in \mathcal{S}_\ell} T_\ell(a)$ and $t_{\ell+1} \leftarrow t_\ell + T_\ell + 1$
11: $\quad$ **Step 3:** Estimate the parameter as

$$\hat{\theta}_\ell = V_\ell^{-1} \sum_{t=t_\ell}^{t_{\ell+1}-1} a_t r_t \quad \text{with} \quad V_\ell = \sum_{a \in \mathcal{S}_\ell} T_\ell(a) a a^\top$$

12: $\quad$ **Step 4:** Make the parameter estimate private

$$\tilde{\theta}_\ell = \hat{\theta}_\ell + V_\ell^{-\frac{1}{2}} N_\ell,$$

$\quad$ where $N_\ell \sim \mathcal{N}\left(0, \frac{2d}{\rho c_\ell} I_d\right)$.
13: $\quad$ **Step 4:** Eliminate low rewarding arms:

$$\mathcal{A}_{\ell+1} = \left\{ a \in \mathcal{A}_\ell : \max_{b \in \mathcal{A}_\ell} \left\langle \tilde{\theta}_\ell, b - a \right\rangle \leq 2\beta_\ell \right\}.$$

14: **end for**

B. Contextual Linear Bandits

Now, we consider an even more general setting of bandits, where the feasible arms at each step may vary and depend on some contextual information.

1) Setting: 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:

$$r_t = \langle \theta^\star, \psi(a_t, c_t) \rangle + \eta_t. \tag{6}$$

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 Equation (6), all that matters is the feature vector that results in choosing a given action rather than the identity of 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 $\mathcal{A}_1, a_1, R_1, \dots, \mathcal{A}_{t-1}, a_{t-1}, R_{t-1}, \mathcal{A}_t$, 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 $\mathcal{A}_t = \{e_1, \dots, e_d\}$, where $(e_i)_i$ are the unit vectors of $\mathbb{R}^d$ then the resulting bandit problem reduces to the stochastic finite-armed bandit.

The goal is to design a ρ-Interactive zCDP policy that minimises the regret, which is defined as

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

Remark 4. *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 IV-A 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 IV-A, are not useful.*

2) Algorithm: We propose AdaC-OFUL, a ρ-Interactive zCDP extension of the Rarely Switching OFUL algorithm [22]. The OFUL algorithm applies the "optimism in the face of uncertainty principle" to the contextual linear bandit setting, which is to act in each round as if the environment is as nice as plausibly possible. 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 3) 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 the estimates at each time step. RS-OFUL achieves similar regret as OFUL, up to a $\sqrt{1+C}$ multiplicative constant.

AdaC-OFUL (Algorithm 3) extends RS-OFUL by privately estimating the least-square estimate (Line 8 of Algorithm 3) while adapting the confidence ellipsoid accordingly. Specifically, we set $\tilde{\beta}_t = \beta_t + \frac{\gamma_t}{\sqrt{t}}$, where $\beta_t = \mathcal{O}\left(\sqrt{d \log(t)}\right)$ and $\gamma_t = \mathcal{O}\left(\sqrt{\frac{1}{\rho} d \log(t)}\right)$. Further details are in App. F.

Remark 5 (A Generic Blueprint for AdaC-GOPE and AdaC-OFUL)**.** AdaC-GOPE *and* AdaC-OFUL *share two main ingredients. First, they add calibrated noise using the Gaussian Mechanism (Theorem 18), i.e. Line 12 in* AdaC-GOPE *and Line 8 in* AdaC-OFUL*). Second, both of them run in adaptive episodes.* AdaC-GOPE *runs in phases (Line 4 in* AdaC-GOPE*), where arms that are likely sub-optimal are eliminated.* AdaC-OFUL *only updates the parameter estimates $\tilde{\theta}$ when the determinant of the design matrix increases enough, i.e. Line 6 in* AdaC-OFUL*. Both algorithms do not access the private rewards at each step t of the interaction, but only at the beginning of the corresponding phases. As it is explained in the Parallel Composition lemma (Lemma 1), and detailed in the generic privacy proof in Appendix D, we leverage this*

Algorithm 3 AdaC-OFUL

Input: Privacy budget ρ, Horizon T, Regulariser λ, Dimension d, Doubling Schedule C
Output: A sequence of T-actions satisfying ρ-Interactive zCDP
Initialisation: $V_0 = \lambda I_d$, $\tilde{\theta} = 0_d$, $\tau = 0$, $\ell = 1$
for $t = 1, 2, \dots$ **do**
  Observe $\mathcal{A}_t$
  if $\det(V_{t-1}) > (1 + C)\det(V_\tau)$ **then**
    Sample $Y_\ell \sim \mathcal{N}(0, \frac{2}{\rho} I_d)$
    Compute $\tilde{\theta}_{t-1} = (V_{t-1})^{-1}(\sum_{s=1}^{t-1} a_s r_s + \sum_{m=1}^{\ell} Y_m)$
    $\ell \leftarrow \ell + 1$ and $\tau \leftarrow t - 1$
  end if
  Compute $a_t = \text{argmax}_{a \in \mathcal{A}_t} \left\langle \tilde{\theta}_\tau, a \right\rangle + \tilde{\beta}_\tau \|a\|_{(V_\tau)^{-1}}$
  Play arm a_t, Observe reward r_t
  $V_t \leftarrow V_{t-1} + a_t a_t^T$
end for

"sparser" access to the private input to add less-noise, and thus, circumvent the need to use composition theorems of DP.

V. Privacy and Regret Analysis

In this section, we provide a privacy and regret analysis of AdaC-GOPE and AdaC-OFUL. Under boundness assumptions, we show that both algorithms are ρ-Interactive zCDP. We also upper bound the regrets of both algorithms and quantify the cost of privacy in the regret.

A. Privacy Analysis

We formalise the intuition behind the blueprint of the algorithm design in Lemma 1. The Privacy Lemma shows that when a mechanism $\mathcal{M}$ is applied to non-overlapping subsets of an input dataset, there is no need to use the composition theorems. Plus, there is no additional cost in the privacy budget.

Lemma 1 (Parallel Composition)**.** *Let $\mathcal{M}$ be a mechanism that takes a **set** as input. Let $\ell < T$ and $t_1, \dots t_\ell, t_{\ell+1}$ be in $[1, T]$ such that $1 = t_1 < \dots < t_\ell < t_{\ell+1} - 1 = T$.*
Let's define the following mechanism

$$\mathcal{G} : \{x_1, \dots, x_T\} \to \bigotimes_{i=1}^{\ell} \mathcal{M}_{\{x_{t_i}, \dots, x_{t_{i+1}-1}\}} \tag{7}$$

$\mathcal{G}$ is the mechanism we get by applying $\mathcal{M}$ to the partition of the input dataset $\{x_1, \dots, x_T\}$ according to $t_1 < \dots < t_\ell < t_{\ell+1}$, i.e.

$$(x_1, x_2, \dots, x_T) \xrightarrow{\mathcal{G}} (o_1, o_2, \dots, o_T),$$

where $o_i \sim \mathcal{M}_{\{x_{t_i}, \dots, x_{t_{i+1}-1}\}}$.
We have that

(a) If $\mathcal{M}$ is (ϵ, δ)-DP then $\mathcal{G}$ is (ϵ, δ)-DP
(b) If $\mathcal{M}$ is ρ-zCDP then $\mathcal{G}$ is ρ-zCDP

The proof is deferred to Appendix D. The main idea is that a change in one element of the input dataset only affects one entry of the output, which already verifies DP. Now, we state some classic assumptions that bound the quantities of interest.

Assumption 1 (Boundedness)**.** *We assume that:*
(1) actions are bounded: $\forall a \in \mathcal{A}$, $\|a\|_2 \leq 1$ in linear bandits, and $\forall t \in [1, T], \forall a \in \mathcal{A}_t$, $\|a\|_2 \leq 1$ in contextual bandits
(2) rewards are bounded: $|r_t| \leq 1$, and
(3) the unknown parameter is bounded: $\|\theta^\star\|_2 \leq 1$.

Theorem 3. *Under Assumption 1, both* AdaC-GOPE *and* AdaC-OFUL *satisfy ρ-Interactive zCDP.*

In appendix D, we provide a generic proof for both AdaC-GOPE and AdaC-OFUL, which combines Lemma 1 and the Gaussian Mechanism (Theorem 18) to show that the sequence of private parameter estimates $\{\tilde{\theta}_\ell\}_\ell$ are ρ-zCDP. We note that since the episodes are adaptive, i.e. the steps corresponding to the start and end of an episode depend on the private input dataset, more care is needed to adapt Lemma 1. Finally, since the actions only depend on the estimates $\{\tilde{\theta}_\ell\}_\ell$, the algorithms are ρ-Interactive zCDP by the post-processing lemma (Lemma 3).

B. Regret Analysis

1) Stochastic Linear Bandits with Finite Number of Arms:

Theorem 4 (Regret Analysis of AdaC-GOPE)**.** *Under Assumption 1 and for $\delta \in (0, 1)$, with probability at least $1 - \delta$, the regret R_T of* AdaC-GOPE *is upper-bounded by*

$$A\sqrt{dT \log\left(\frac{K \log(T)}{\delta}\right)} + \frac{Bd}{\sqrt{\rho}}\sqrt{\log\left(\frac{K \log(T)}{\delta}\right)}\log(T),$$

where A and B are universal constants. If $\delta = \frac{1}{T}$, then

$$\mathbb{E}(R_T) \leq \mathcal{O}\left(\sqrt{dT \log(KT)}\right) + \mathcal{O}\left(\frac{d}{\sqrt{\rho}}(\log(KT))^{\frac{3}{2}}\right).$$

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 β_ℓ is smaller than their sub-optimality gaps. The regret upper bound follows directly. We refer to Appendix E for complete proof.

We discuss the implications of our regret upper bound:

1. Achieving ρ-Interactive zCDP 'almost for free': Theorem 4 shows that the price of ρ-Interactive zCDP is the additive term $\tilde{\mathcal{O}}\left(\frac{d}{\sqrt{\rho}}\right)$[3]. For a fixed RDP budget ρ and as $T \to \infty$, the regret due to privacy becomes negligible in comparison with the privacy-oblivious term in regret, i.e. $\tilde{\mathcal{O}}\left(\sqrt{dT}\right)$.

2. Optimality of AdaC-GOPE. In Section VI, we prove a $\Omega(\frac{d}{\sqrt{\rho}})$ minimax private regret lower bound that matches the regret upper bound of AdaC-GOPE up to an extra $(\log KT)^{\frac{3}{2}}$ factor. If K is exponential in d, then there is a mismatch between the regret upper and lower bounds, in their dependence on the dimension d. This gap could be improved with a better mechanism to make $\hat{\theta}$ private (Step 4 in Algorithm

[3] $\tilde{\mathcal{O}}$ hides poly-logarithmic factors in the horizon T.

2). In Appendix E-C, we discuss in detail how different ways of adding noise at Step 4 impact the dependence of the regret upper bound on d.

Related Algorithms and Bounds. Concurrently to our work, both [13] and [21] study private variants of the GOPE algorithm for pure ϵ-global DP and (ϵ, δ)-global DP, respectively. However, both algorithms differ in how they make private the estimated parameter $\hat{\theta}$ compared to AdaC-GOPE. Both [13] and [21] add noise to each sum of rewards $\sum_{t=t_\ell}^{t_{\ell+1}-1} r_t$ (Line 11, Alg. 2), whereas AdaC-GOPE add noise in $\hat{\theta}_l$ (Line 12, Alg. 2). As a result, though AdaC-GOPE achieves linear dependence on the dimension d as suggested by the lower bound, others do not (d^2 for [13] and $d^{3/2}$ for [21]).

In Appendix E-C, we analyse in detail the impact of adding noise at different steps of GOPE, both theoretically and experimentally.

2) Contextual Linear Bandits:

To analyse the regret of AdaC-OFUL, we impose a stochastic assumption on the context generation. Specifically, we adopt the same assumption that is often used in on-policy [34], [35] and off-policy [36], [37] linear contextual bandits.

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*

1. $\|A\|_2 = 1$

2. $\mathbb{E}[AA^T]$ is full rank, with minimum eigenvalue $\lambda_0 > 0$

3. $\forall z \in \mathbb{R}^d$, $\|z\|_2 = 1$, 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 helps control the minimum eigenvalue of the design matrix $V_t \triangleq \sum_{s=1}^t a_s a_s^T$. Using Lemma 12 on the minimum eigenvalue, we quantify more precisely the effect of the added noise due to ρ-Interactive zCDP and derive tighter confidence bounds.

Theorem 5. *Under Assumptions 1 and 2, and for $\delta \in (0, 1]$, with probability at least $1 - \delta$, the regret R_T of* AdaC-OFUL *is upper bounded by*

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

Proof Sketch. The main challenge in the regret analysis is to design tight ellipsoid confidence sets around the private estimate $\tilde{\theta}_t$, since the regret can be shown to be the sum of the confidence widths. To design the non-private part of the ellipsoid confidence sets, we rely on the self-normalised bound for vector-valued martingales theorem of [22]. For the private part, we rely on the assumption of stochastic contexts controlling $\lambda_{\min}(G_t)$ and the concentration of χ^2 distribution to control the introduced Gaussian noise. The rest of the proof is adapted from the analysis of RS-OFUL [22]. We also show that the number of episodes, i.e. updates of the estimated parameters, is in $\mathcal{O}(\log(T))$. We refer to Appendix F for the complete proof.

We discuss the implications of our regret upper bound:

1. Achieving ρ-Interactive zCDP 'almost for free': The upper bound of Theorem 5 shows that the price of ρ-Interactive zCDP for linear contextual bandits is the additive term $\tilde{\mathcal{O}}\left(\frac{d^2}{\sqrt{\rho}}\right)$. For a fixed budget ρ and as $T \to \infty$, the regret due to zCDP turns negligible in comparison with the privacy-oblivious regret term of $\tilde{\mathcal{O}}\left(d\sqrt{T}\right)$.

2. Adapting AdaC-OFUL *for private contexts:* To make AdaC-OFUL achieve Joint-DP [11], the estimate $\tilde{\theta}$ at line 8 should be made private with respect to both rewards and context. A straightforward way to do so is by estimating the design matrix V_t privately, e.g. as it is done in [11]. A first regret analysis of this adaptation shows that the price of privacy in the regret will become not negligible, i.e. the regret is $O\left(\sqrt{T} + \sqrt{T/\rho}\right)$. This shows that the bottleneck in the problem is the private estimation of the design matrix.

3. Connecting Related Settings. [12] proposes LinPriv, which is an ϵ-global DP extension of OFUL. The context is assumed to be public but *adversely chosen*. Theorem 5 in [12] states that the regret of LinPriv is $\tilde{\mathcal{O}}\left(d\sqrt{T} + \frac{1}{\epsilon}Kd\log T\right)$. We revisit their regret analysis and show that the bound should be $\tilde{\mathcal{O}}\left(d\sqrt{T} + \frac{1}{\epsilon}Kd\sqrt{T}\right)$ instead (Appendix F-C). Also, [11] proposes an (ϵ, δ)-Joint DP algorithm for *private and adversarial contexts*. The algorithm is based on OFUL and privately estimates $\hat{\theta}_t$ at each step using the tree-based mechanism [31], [38]. However, this algorithm has an additional regret of $\frac{1}{\epsilon}\sqrt{T}$ due to privacy.

Open Problem. It is still an open problem *whether it is possible* to design a private algorithm for linear contextual bandits with *private and/or adversarially chosen contexts*, such that the additional regret due to privacy in $\mathcal{O}(\log(T))$.

VI. Lower Bounds on Regret

In this section, we quantify the cost of ρ-Interactive zCDP for bandits by providing regret lower bounds for any ρ-Interactive zCDP policy. These lower bounds on regret provide valuable insight into the inherent hardness of the problem and establish a target for optimal algorithm design. We first derive a ρ-Interactive zCDP version of the KL decomposition Lemma using a sequential coupling argument. The regret lower bounds are then retrieved by plugging the KL upper bound in classic regret lower bound proofs. A summary of the lower bounds is in Table I, while the proof details are deferred to Appendix G.

A. KL Decomposition Lemma under ρ-zCDP

In order to proceed with the lower bounds, first, we are interested in controlling the Kullback-Leibler (KL) divergence between marginal distributions induced by a ρ-zCDP mechanism $\mathcal{M}$ when the datasets are generated using two different distributions. This type of information-theoretic bounds is generally the main step for many standard methods for obtaining minimax lower bounds.

In particular, if $\mathcal{P}_1$ and $\mathcal{P}_2$ are two data-generating distributions over $\mathcal{X}^n$, we define the marginals M_1 and M_2 over the output of mechanism $\mathcal{M}$ as

$$M_\nu(A) \triangleq \int_{d\in\mathcal{X}^n} \mathcal{M}_d(A)\, \mathrm{d}\mathcal{P}_\nu(d), \tag{8}$$

when the inputs are generated from $\mathcal{P}_1$ and $\mathcal{P}_2$ respectively, i.e. for $\nu \in \{1,2\}$ and $A \in \mathcal{F}$.

Define $\mathcal{C}$ as a coupling of $(\mathcal{P}_1, \mathcal{P}_2)$, i.e. the marginals of $\mathcal{C}$ are $\mathcal{P}_1$ and $\mathcal{P}_2$. We denote by $\Pi(\mathcal{P}_1, \mathcal{P}_2)$ the set of all the couplings between $\mathcal{P}_1$ and $\mathcal{P}_2$.

Theorem 6 (KL Upper Bound as a Transport Problem)**.** *If $\mathcal{M}$ is ρ-zCDP, then*

$$\mathrm{KL}(M_1 \parallel M_2) \le \rho \inf_{\mathcal{C}\in\Pi(\mathcal{P}_1,\mathcal{P}_2)} \mathbb{E}_{(d,d')\sim\mathcal{C}}[d_{Ham}(d,d')^2].$$

Deriving the sharpest upper bound for the KL would require solving the transport problem

$$\inf_{\mathcal{C}\in\Pi(\mathcal{P}_1,\mathcal{P}_2)} \mathbb{E}_{(d,d')\sim\mathcal{C}}[d_{\mathrm{Ham}}(d,d')^2]. \tag{9}$$

As a proxy, we will use maximal couplings.

Proposition 3. *Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be two probability distributions that share the same σ-algebra. There exists a coupling $c_\infty(\mathcal{P}_1,\mathcal{P}_2) \in \Pi(\mathcal{P}_1,\mathcal{P}_2)$ called a maximal coupling, such that*

$$\mathbb{E}_{(X_1,X_2)\sim c_\infty(\mathcal{P}_1,\mathcal{P}_2)}[\mathbb{1}\{X_1 \neq X_2\}] = \mathrm{TV}(\mathcal{P}_1 \parallel \mathcal{P}_2)$$

Using maximal coupling for data-generating distributions that are product distributions yields the following bound.

Theorem 7 (KL Decomposition for Product Distributions)**.** *Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be two product distributions over $\mathcal{X}^n$, i.e. $\mathcal{P}_1 = \bigotimes_{i=1}^n p_{1,i}$ and $\mathcal{P}_2 = \bigotimes_{i=1}^n p_{2,i}$, where $p_{\nu,i}$ for $\nu \in \{1,2\}, i \in [1,n]$ are distributions over $\mathcal{X}$. Let $t_i \triangleq \mathrm{TV}(p_{1,i} \parallel p_{2,i})$. If $\mathcal{M}$ is ρ-zCDP, then*

$$\mathrm{KL}(M_1 \parallel M_2) \le \rho\left(\sum_{i=1}^n t_i\right)^2 + \rho\sum_{i=1}^n t_i(1-t_i) \tag{10}$$

This is a centralised ρ-zCDP version of the KL-decomposition lemma under local DP [39, Theorem 1], and a ρ-zCDP version of the Karwa-Vadhan lemma [24]. Similar coupling ideas have been developed in [40] to derive ρ-zCDP variants of LeCam and Fano inequalities.

B. Lower Bound on Regret for Linear Bandits

Now, we adapt Theorem 7 for the bandit marginals. Let $\nu = \{P_a, a \in [K]\}$ and $\nu' = \{P'_a, a \in [K]\}$ be two bandit instances. When the policy π interacts with the bandit instance ν, it induces a marginal distribution $m_{\nu\pi}$ over the sequence of actions, i.e.

$$m_{\nu\pi}(a_1,\dots,a_T) \triangleq \int_{r_1,\dots,r_T} \prod_{t=1}^T \pi_t(a_t \mid \mathcal{H}_{t-1}) p_{a_t}(r_t)\, \mathrm{d}r_t.$$

We define $m_{\nu'\pi}$ similarly.

Theorem 8 (KL Decomposition for ρ-Interactive zCDP)**.** *If π is ρ-Interactive zCDP, then*

$$\mathrm{KL}(m_{\nu\pi} \parallel m_{\nu'\pi}) \le \rho\Bigg(\left[\mathbb{E}_{\nu\pi}\left(\sum_{t=1}^T t_{a_t}\right)\right]^2 + \mathbb{E}_{\nu\pi}\left(\sum_{t=1}^T t_{a_t}(1-t_{a_t})\right) + \mathbb{V}_{\nu\pi}\left(\sum_{t=1}^T t_{a_t}\right)\Bigg),$$

where $t_{a_t} \triangleq \mathrm{TV}\left(P_{a_t} \parallel P'_{a_t}\right)$ and $\mathbb{E}_{\nu\pi}$ and $\mathbb{V}_{\nu\pi}$ are the expectation and variance under $m_{\nu\pi}$ respectively.

The proof of Theorem 8 combines the ρ-Interactive DP group privacy property (Theorem 1) and the maximal coupling ideas developed in Theorem 7.

Leveraging this decomposition, we derive the *minimax* regret lower bound, i.e. the best regret achievable by a policy on the corresponding worst-case environment.

Definition 6 (Minimax Regret)**.** *The minimax regret lower bound is defined as*

$$\mathrm{Reg}^{minimax}_{T,\rho}(\mathcal{A},\Theta) \triangleq \inf_{\pi\in\Pi^\rho_{Int}} \sup_{\theta\in\Theta} \mathrm{Reg}_T(\pi,\mathcal{A},\theta).$$

Theorem 9 (Minimax Lower Bounds for Linear Bandits)**.** *Let $\mathcal{A} = [-1,1]^d$ and $\Theta = \mathbb{R}^d$. Then, for any ρ-Interactive zCDP policy, we have that*

$$\mathrm{Reg}^{minimax}_{T,\rho}(\mathcal{A},\Theta) \ge \max\left\{\underbrace{\frac{e^{-2}}{8}d\sqrt{T}}_{\text{without } \rho\text{-zCDP}},\ \underbrace{\frac{e^{-2.25}}{4}\frac{d}{\sqrt{\rho}}}_{\text{with } \rho\text{-zCDP}}\right\}$$

In order to prove the lower bounds, we deploy the KL upper bound of Theorem 7 in the classic proof scheme of regret lower bounds [2]. The high-level idea of proving bandit lower bounds is selecting two *hard* environments, which are hard to statistically distinguish but are conflicting, i.e. actions that may be optimal in one are sub-optimal in other. The KL upper bound of Theorem 8 allows us to quantify the extra-hardness to statistically distinguish environments due to the additional "blurriness" created by the ρ-zCDP constraint.

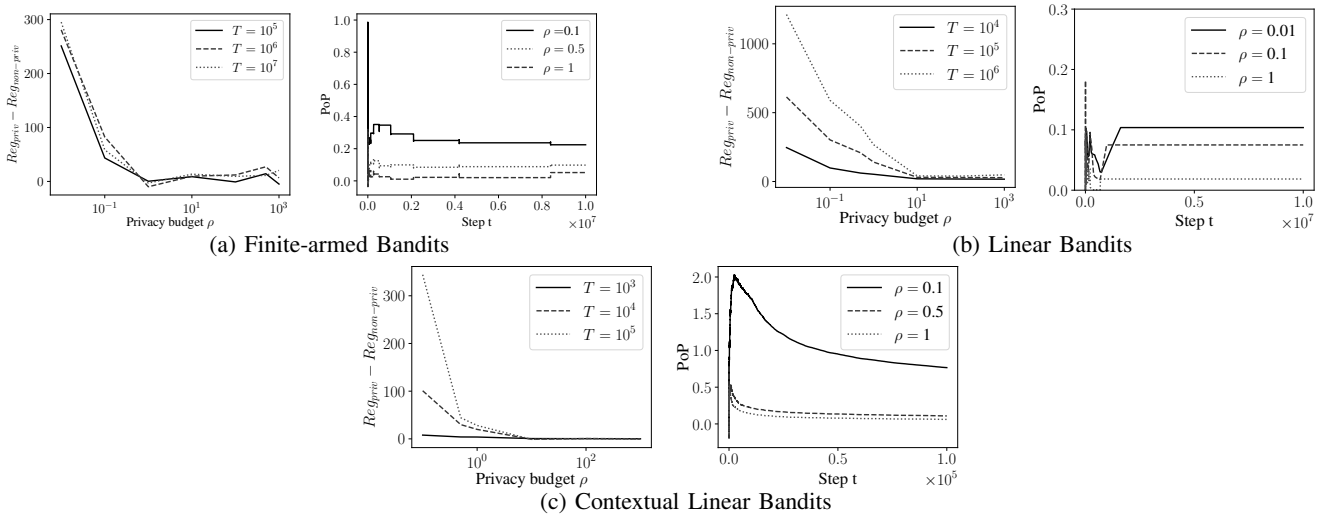
The minimax regret lower bound suggests the existence of two hardness regimes depending on ρ, T and d. When $\rho < 4e^{-0.5}/T$, i.e. the **high-privacy regime**, the lower bound becomes $\Omega\left(d/\sqrt{\rho}\right)$, and ρ-Interactive zCDP bandits incur more regret than non-private ones. When $\rho \ge 4e^{-0.5}/T$, i.e. in the **low-privacy regime**, the lower bound retrieves the non-private lower bound, i.e. $\Omega(d\sqrt{T})$, and privacy can be for free.

VII. Experimental Analysis

We empirically verify whether AdaC-UCB, AdaC-GOPE and AdaC-OFUL can achieve privacy for free.

A. Experimental Setup

For finite-armed bandits, we test AdaC-UCB with $\beta = 1$ and compare it to its non-private counterpart, i.e. a UCB algorithm with adaptive episodes and forgetting. We test



(a) Finite-armed Bandits

(b) Linear Bandits

(c) Contextual Linear Bandits

Fig. 4: For each bandit setting, the left figure represents the evolution of the difference between the private and non-private regret w.r.t. the privacy budget ρ. The right figure represents the evolution of the price of privacy (PoP) w.r.t. the time step.

the algorithms for Bernoulli bandits with 5-arms and means $\{0.75, 0.625, 0.5, 0.375, 0.25\}$ (as in [7]).

For linear bandits with finitely many arms, we implement AdaC-GOPE and compare it to GOPE. We set the failure probability to $\delta = 0.001$ and the noise to be $\rho_t = \mathcal{N}(0, 1)$. We use the Frank-Wolfe algorithm to solve the G-optimal design problem [2]. We chose $K = 10$ actions randomly on the unit tri-dimensional sphere ($d = 3$). The true parameter $\theta^\star$ is also chosen randomly on the tri-dimensional sphere.

For linear contextual bandits, we implement AdaC-OFUL and compare it to RS-OFUL. We set $C = 1$, the regularisation constant $\lambda = 0.1$, the failure probability to $\delta = 0.001$ and the noise $\rho_t = \mathcal{N}(0, 1)$. We set $K = 10$ and $d = 3$. To generate the contexts, at each time step, we sample from a new set of actions $\mathcal{A}_t$ which is 10 dimensional multivariate Gaussian $\mathcal{N}\left(\left(\frac{1}{\sqrt{d}}, \ldots, \frac{1}{\sqrt{d}}\right), \frac{1}{10}\mathbf{I}_d\right)$. This way, we sample the contexts near the unit sphere, while having a sub-Gaussian generation process corresponding to the context-generation Assumption 2. The true parameter $\theta^\star$ is chosen randomly on the tri-dimensional sphere.

For the three settings, we run the private and non-private algorithms 100 times for a horizon $T = 10^7$, and compare their average regrets (Figure 4).

B. Results and Analysis

From the experimental results illustrated in Figure 4, we reach to two conclusions for all three settings.

1. *Free-privacy in low-privacy regime.* For a fixed horizon T, the difference between the private and non-private regret, $Reg_{priv} - Reg_{non-priv}$, converges to zero as the privacy budget $\rho \to \infty$. Thus, our algorithms achieve the same regret as their non-private counterparts in the low-privacy regime.

2. *Asymptotic no price of privacy.* For a fixed privacy budget ρ, the Price of Privacy (PoP), i.e. $\text{PoP} \triangleq \frac{\text{Reg}_{priv} - \text{Reg}_{non-priv}}{\text{Reg}_{non-priv}}$ converges to zero as the horizon T increases. This observation resonates with both the theoretical regret upper bounds of the algorithms and the hardness suggested by the lower bounds, where cost due to privacy appears as lower-order terms.

VIII. Conclusion and Future Works

We study bandits with ρ-zCDP and a centralised decision-maker for three settings: stochastic, linear and contextual bandits. First, we compare different ways of adapting DP to bandits. We adhere to the ρ-Interactive zCDP as the DP framework, as it encapsulates the other definitions. Then, for each bandit setting, we design a ρ-Interactive zCDP policy and show that the additional cost in the regret due to ρ-Interactive zCDP is negligible in comparison to the regret incurred oblivious to privacy. The three algorithms share similar algorithmic blueprint. They add calibrated *Gaussian noise* and they run in *adaptive episodes*. These ingredients allow devising a generic and simple algorithmic approach to make index-based bandit algorithms achieving privacy with minimal cost. We derive minimax regret lower bounds for finite-armed and linear bandits, showing the existence of two hardness regimes and privacy can be achieved for free in low-privacy regime.

One future direction for the linear contextual bandit is to lift the assumptions that the contexts are public and stochastic. For example, in personalised recommender systems, the context may contain sensitive information of individuals. *Designing and analysing an algorithm that does not rely on these assumptions, and achieves ρ-Interactive zCDP almost for free in linear contextual bandits, is an interesting open question.*

Another future direction is to derive regret lower bounds for bandits with (ϵ, δ)-DP. Both pure ϵ-DP and ρ-zCDP enjoy a ('tight') group privacy property that gives meaningful lower bounds for bandits when applied with coupling arguments. These arguments fail to adapt to (ϵ, δ)-DP. An interesting technical challenge would be to adapt, for bandits, the fingerprinting lemma, which is a technique used for proving (ϵ, δ)-DP lower bounds [41], [42]. For the algorithm design, it would be also interesting to see how to close the multiplicative gaps.

ACKNOWLEDGMENT

This work is supported by the AI_PhD@Lille grant. D. Basu acknowledges the Inria-Kyoto University Associate Team "RELIANT" for supporting the project, and the ANR JCJC for the REPUBLIC project (ANR-22-CE23-0003-01). We also thank Philippe Preux for his support.

REFERENCES

[1] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3-4, pp. 285–294, 1933.

[2] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.

[3] N. Silva, H. Werneck, T. Silva, A. C. Pereira, and L. Rocha, "Multi-armed bandits in recommendation systems: A survey of the state-of-the-art and future directions," *Expert Systems with Applications*, vol. 197, p. 116669, 2022.

[4] D. Bergemann and J. Välimäki, "Learning and strategic pricing," *Econometrica: Journal of the Econometric Society*, pp. 1125–1149, 1996.

[5] N. Mishra and A. Thakurta, "(Nearly) optimal differentially private stochastic multi-arm bandits," in *UAI*, 2015.

[6] A. C. Tossou and C. Dimitrakakis, "Algorithms for differentially private multi-armed bandits," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[7] T. Sajed and O. Sheffet, "An optimal private stochastic-mab algorithm based on optimal private stopping rule," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5579–5588.

[8] A. Azize and D. Basu, "When privacy meets partial information: A refined analysis of differentially private bandits," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 199–32 210, 2022.

[9] B. Hu and N. Hegde, "Near-optimal thompson sampling-based algorithms for differentially private stochastic bandits," in *Uncertainty in Artificial Intelligence*. PMLR, 2022, pp. 844–852.

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

[11] R. Shariff and O. Sheffet, "Differentially private contextual linear bandits," in *Advances in Neural Information Processing Systems*, 2018, pp. 4296–4306.

[12] S. Neel and A. Roth, "Mitigating bias in adaptive data gathering via differential privacy," in *International Conference on Machine Learning*. PMLR, 2018, pp. 3720–3729.

[13] O. A. Hanna, A. M. Girgis, C. Fragouli, and S. Diggavi, "Differentially private stochastic linear bandits:(almost) for free," *arXiv preprint arXiv:2207.03445*, 2022.

[14] A. Azize, M. Jourdan, A. A. Marjani, and D. Basu, "On the complexity of differentially private best-arm identification with fixed confidence," *arXiv preprint arXiv:2309.02202*, 2023.

[15] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[16] 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.

[17] A. G. Thakurta and A. Smith, "(Nearly) optimal algorithms for private online learning in full-information and bandit settings," *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[18] S. Vadhan and T. Wang, "Concurrent composition of differential privacy," in *Theory of Cryptography: 19th International Conference, TCC 2021, Raleigh, NC, USA, November 8–11, 2021, Proceedings, Part II 19*. Springer, 2021, pp. 582–604.

[19] S. Vadhan and W. Zhang, "Concurrent composition theorems for all standard variants of differential privacy," *arXiv preprint arXiv:2207.08335*, 2022.

[20] P. Jain, S. Raskhodnikova, S. Sivakumar, and A. Smith, "The price of differential privacy under continual observation," in *International Conference on Machine Learning*. PMLR, 2023, pp. 14 654–14 678.

[21] F. Li, X. Zhou, and B. Ji, "Differentially private linear bandits with partial distributed feedback," in *2022 20th International Symposium on Modeling and Optimization in Mobile, Ad hoc, and Wireless Networks (WiOpt)*. IEEE, 2022, pp. 41–48.

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

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

[24] V. Karwa and S. Vadhan, "Finite sample differentially private confidence intervals," 2017. [Online]. Available: https://arxiv.org/abs/1711.03908

[25] X. Lyu, "Composition theorems for interactive differential privacy," in *Advances in Neural Information Processing Systems*, 2022.

[26] B. Hu, Z. Huang, and N. A. Mehta, "Optimal algorithms for private online learning in a stochastic environment," 2021. [Online]. Available: https://arxiv.org/abs/2102.07929

[27] N. Kallus, "Instrument-armed bandits," in *Algorithmic Learning Theory*. PMLR, 2018, pp. 529–546.

[28] A. Stirn and T. Jebara, "Thompson sampling for noncompliant bandits," *arXiv preprint arXiv:1812.00856*, 2018.

[29] F. Liu and N. Shroff, "Data poisoning attacks on stochastic bandits," in *International Conference on Machine Learning*. PMLR, 2019, pp. 4042–4050.

[30] C. Dwork, M. Naor, T. Pitassi, G. N. Rothblum, and S. Yekhanin, "Pan-private streaming algorithms." in *ICS*, 2010, pp. 66–80.

[31] T.-H. H. Chan, E. Shi, and D. Song, "Private and continual release of statistics," *ACM Trans. Inf. Syst. Secur.*, vol. 14, no. 3, nov 2011. [Online]. Available: https://doi.org/10.1145/2043621.2043626

[32] J. López-Fidalgo, *Optimal Experimental Design: A Concise Introduction for Researchers*. Springer Nature, 2023, vol. 226.

[33] J. Kiefer and J. Wolfowitz, "The equivalence of two extremum problems," *Canadian Journal of Mathematics*, vol. 12, pp. 363–366, 1960.

[34] C. Gentile, S. Li, and G. Zappella, "Online clustering of bandits," in *International Conference on Machine Learning*. PMLR, 2014, pp. 757–765.

[35] Z. Li, L. Ratliff, K. G. Jamieson, L. Jain *et al.*, "Instance-optimal pac algorithms for contextual bandits," *Advances in Neural Information Processing Systems*, vol. 35, pp. 37 590–37 603, 2022.

[36] A. Zanette, K. Dong, J. N. Lee, and E. Brunskill, "Design of experiments for stochastic contextual linear bandits," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 720–22 731, 2021.

[37] M. Jörke, J. Lee, and E. Brunskill, "Simple regret minimization for contextual bandits using bayesian optimal experimental design," in *ICML2022 Workshop on Adaptive Experimental Design and Active Learning in the Real World*, 2022.

[38] C. Dwork, M. Naor, T. Pitassi, and G. N. Rothblum, "Differential privacy under continual observation," in *Proceedings of the Forty-Second ACM Symposium on Theory of Computing*, ser. STOC '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 715–724. [Online]. Available: https://doi.org/10.1145/1806689.1806787

[39] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *Proc. of IEEE Foundations of Computer Science (FOCS)*, 2013.

[40] C. Lalanne, A. Garivier, and R. Gribonval, "On the statistical complexity of estimation and testing under privacy constraints," *arXiv preprint arXiv:2210.02215*, 2022.

[41] M. Bun, J. Ullman, and S. Vadhan, "Fingerprinting codes and the price of approximate differential privacy," in *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, 2014, pp. 1–10.

[42] G. Kamath, A. Mouzakis, and V. Singhal, "New lower bounds for private estimation and a generalized fingerprinting lemma," *arXiv preprint arXiv:2205.08532*, 2022.

[43] I. Mironov, "Rényi differential privacy," in *Proceedings of 30th IEEE Computer Security Foundations Symposium (CSF)*, 2017, pp. 263–275.

APPENDIX A
OUTLINE

The appendices are organised as follows:

- The relations between privacy definitions are detailed in Appendix B.
- AdaC-UCB is proposed and analysed in Appendix C.
- TABLE II compares the regret upper bounds of our algorithms compared to the "converted" regret upper bounds from the pure DP bandit literature.
- A generic privacy proof and its specification for AdaC-UCB, AdaC-GOPE and AdaC-OFUL is presented in Appendix D.
- The regret analysis of AdaC-GOPE alongside the concentration inequalities under optimal design are presented in Appendix E.
- The regret analysis of AdaC-OFUL alongside the concentration inequalities for private least square estimator are presented in Appendix F.
- A new proof to generate lower bounds for ρ-zCDP is developed in Appendix G and adapted to bandits.
- Extended experiments are presented in Appendix H.
- Existing technical results and definitions are summarised in Appendix I

APPENDIX B
PRIVACY DEFINITIONS FOR BANDITS

In this section, we present the missing proofs of Section III and discuss privacy definitions for the contextual bandit setting.

A. Proof of Proposition 1

Proposition 1 (Relation between Table DP and View DP)**.** *For any policy π, we have that*

(a) $\mathcal{M}^\pi$ is ϵ-DP $\Leftrightarrow$ $\mathcal{V}^\pi$ is ϵ-DP.
(b) $\mathcal{M}^\pi$ is (ϵ,δ)-DP $\Rightarrow$ $\mathcal{V}^\pi$ is (ϵ,δ)-DP.
(c) $\mathcal{M}^\pi$ is ρ-zCDP $\Rightarrow$ $\mathcal{V}^\pi$ ρ-zCDP.
(d) $\mathcal{V}^\pi$ is (ϵ,δ)-DP $\Rightarrow$ $\mathcal{M}^\pi$ is $(\epsilon, K^T\delta)$-DP.
(e) $\Pi^{(\epsilon,\delta)}_{\textit{Table}} \subsetneq \Pi^{(\epsilon,\delta)}_{\textit{View}}$

where $\Pi^{(\epsilon,\delta)}_{\textit{Table}}$ and $\Pi^{(\epsilon,\delta)}_{\textit{View}}$ are the class of all policies verifying (ϵ,δ)-Table DP and (ϵ,δ)-View DP respectively.

Before proving the proposition, we define two handy reductions, to go from list to table of rewards and vice-versa.

Reduction 1 (From list to table of rewards)**.** *For every $r \in \mathbb{R}^T$ a list of rewards, we define $d(r)$ to be the table such that $d(r)_{t,i} = r_t$ for all $i \in [K]$ and all $t \in [T]$.*

In other words, $d(r)$ is the table of rewards where r is concatenated colon-wise K times.

This transformation has two interesting consequences:

- *for every $E \in \mathcal{P}([K]^T)$, $\mathcal{V}^\pi_r(E) = \mathcal{M}^\pi_{d(r)}(E)$*
- *for every $\alpha > 1$, $D_\alpha(\mathcal{V}^\pi_r \| \mathcal{V}^\pi_{r'}) = D_\alpha(\mathcal{M}^\pi_{d(r)} \| \mathcal{M}^\pi_{d(r')})$*
- *If $r \sim r'$ are neighbouring list of rewards, then $d(r) \sim d(r')$ are neighbouring table of rewards*

Reduction 2 (From table of rewards to lists)**.** *For every atomic event $a^T \triangleq (a_1, \dots, a_T)$ and a table of reward $d \in (\mathbb{R}^K)^T$, we define $r(d, a^T) \in \mathbb{R}^T$ to be the list of rewards such that $r(d, a^T)_t = d_{t,a_t}$.*

In other words, $r(d, a^T)$ is the list of rewards corresponding to the trajectory of a^T in d.

This transformation has two interesting consequences:

- *for every a^T, $\mathcal{M}^\pi_d(a^T) = \mathcal{V}^\pi_{r(d,a^T)}(a^T)$*
- *If $d \sim d'$ are neighbouring table of rewards, then for every a^T, $r(d, a^T) \sim r(d', a^T)$ are neighbouring list of rewards.*

Proof (Proposition 1).

(b): Suppose that $\mathcal{M}^\pi$ is (ϵ,δ)-DP.

Let $r \sim r'$ two neighbouring lists of rewards. For every event $E \in \mathcal{P}([K]^T)$, we have that

$$\mathcal{V}^\pi_{\mathbf{r}}(E) - e^\epsilon \mathcal{V}^\pi_{r'}(E) = \mathcal{M}^\pi_{d(r)}(E) - e^\epsilon \mathcal{M}^\pi_{d(r')}(E) \leq \delta$$

where the last inequality is because $\mathcal{M}^\pi$ is (ϵ,δ)-DP and $d(r) \sim d(r')$.

We conclude that $\mathcal{V}^\pi$ is (ϵ,δ)-DP.

(c): Suppose that $\mathcal{M}^\pi$ is ρ-zCDP.

Let $r \sim r'$ two neighbouring lists of rewards. For every $\alpha > 1$, we have that

$$D_\alpha(\mathcal{V}^\pi_r \| \mathcal{V}^\pi_{r'}) = D_\alpha(\mathcal{M}^\pi_{d(r)} \| \mathcal{M}^\pi_{d(r')}) \leq \rho\alpha$$

where the last inequality is because $\mathcal{M}^\pi$ is ρ-zCDP and $d(r) \sim d(r')$.

We conclude that $\mathcal{V}^\pi$ is ρ-zCDP.

(a) $\Rightarrow$) Is a direct consequence of (b) for $\delta = 0$.

$\Leftarrow$) Suppose that $\mathcal{V}^\pi$ is ϵ-DP.

Let $d \sim d'$ be two tables of rewards in $(\mathbb{R}^K)^T$.

For ϵ-DP, it is enough to consider atomic events $a^T \triangleq (a_1, \dots, a_T)$.

For any atomic event a^T, we have that

$$\mathcal{M}^\pi_d(a^T) = \mathcal{V}^\pi_{r(d,a^T)}(a^T) \leq e^\epsilon \mathcal{V}^\pi_{r(d',a^T)}(a^T) = e^\epsilon \mathcal{M}^\pi_{d'}(a^T)$$

where the first inequality is because $\mathcal{V}^\pi$ is ϵ-DP and $r(d, a^T) \sim r(d', a^T)$.

We conclude that $\mathcal{M}^\pi$ is ϵ-DP.

(d) Suppose that $\mathcal{V}^\pi$ is (ϵ,δ)-DP.

Let $d \sim d'$ be two tables of rewards in $(\mathbb{R}^K)^T$.

Let $E \in \mathcal{P}([K]^T)$ be an event, i.e. a set of sequences. We have that

$$\begin{aligned}
\mathcal{M}^\pi_d(E) = \sum_{a^T \in E} \mathcal{M}^\pi_d(a^T) &= \sum_{a^T \in E} \mathcal{V}^\pi_{r(d,a^T)}(a^T) \\
&\underset{(a)}{\leq} \sum_{a^T \in E} (e^\epsilon \mathcal{V}^\pi_{r(d',a^T)}(a^T) + \delta) \\
&\underset{(b)}{\leq} e^\epsilon \mathcal{M}^\pi_{d'}(E) + K^T\delta,
\end{aligned}$$

where (a) holds true because $\mathcal{V}^\pi$ is (ϵ,δ)-DP, and (b) is true because $\mathrm{card}(E) \leq K^T$.

We conclude that $\mathcal{M}^\pi$ is $(\epsilon, K^T\delta)$-DP.

(e) To prove the strict inclusion, we build a policy π for $T = 3$, $K = 2$ with action 0 and action 1, and rewards in $\{0, 1\}$.

TABLE II: Comparaison between our regret upper bounds, and converted upper bounds from the pure-DP litterature

Bandit Setting	Our Regret Upper Bound	"Converted" Regret Upper Bounds
Finite-armed bandits	$\sum_{a:\Delta_a>0}\left(\frac{8\beta}{\Delta_a}\log(T)+8\sqrt{\frac{\beta}{\rho}}\sqrt{\log(T)}\right)$ (Thm 10)	$\sum_{a:\Delta_a>0}\left(\frac{8\beta}{\Delta_a}\log(T)+8\sqrt{\frac{\beta}{\rho}}\log(T)\right)$ (Thm 7 in [8])
	$\mathcal{O}\left(\sqrt{KT\log(T)}\right)+\mathcal{O}\left(\frac{K}{\sqrt{\rho}}\sqrt{\log(T)}\right)$ (Thm 11)	$\mathcal{O}\left(\sqrt{KT\log(T)}\right)+\mathcal{O}\left(\frac{K}{\sqrt{\rho}}\log(T)\right)$ (Thm 12 in [8])
Linear bandits	$\mathcal{O}\left(\sqrt{dT\log(KT)}\right)+\mathcal{O}\left(\frac{d}{\sqrt{\rho}}\log^{\frac{3}{2}}(KT)\right)$ (Thm 4)	$\mathcal{O}\left(\sqrt{dT\log(KT)}\right)+\mathcal{O}\left(\frac{d^2}{\sqrt{\rho}}\log^2(KT)\right)$ (Eq.17 in [13])
Linear Contextual bandits	$\mathcal{O}\left(d\log(T)\sqrt{T}\right)+\mathcal{O}\left(\frac{d^2}{\sqrt{\rho}}\log(T)^2\right)$ (Thm 5)	-

A policy here is a sequence of three decision rules

$$\pi = \{\pi_1, \pi_2, \pi_3\},$$

where each decision rule is a function from the history. Since the possible histories at each step are finite, specifying a decision rule is just specifying the probability weights of choosing action 0 and action 1 for every possible history.

We consider the following decision rules

$$\pi_1 = \begin{bmatrix} 2/3 & 1/3 \end{bmatrix}$$

$$\pi_2 = \begin{bmatrix} 1/2 & 1/2 \\ 1/3 & 2/3 \\ 1/4 & 3/4 \\ 1/3 & 2/3 \end{bmatrix}$$

$$\pi_3 = \begin{bmatrix} 1/2 & 1/2 \\ 1/3 & 2/3 \\ 1/4 & 3/4 \\ 1/5 & 4/5 \\ 1/2 & 1/2 \\ 2/3 & 1/3 \\ 1/4 & 3/4 \\ 0 & 1 \\ 1/3 & 2/3 \\ 1/7 & 6/7 \\ 3/4 & 1/4 \\ 2/5 & 3/5 \\ 1/2 & 1/2 \\ 1 & 0 \\ 1/4 & 3/4 \\ 2/3 & 1/3 \end{bmatrix}$$

The history is first represented as a binary string, and then converted to decimals. Finally, the index in the decision rule corresponding to this decimal value is chosen. We elaborate this procedure in the two examples below.

Example 1. If the policy observed the history $\{1, 0\}$, i.e. action 1 was played in the first round and the reward 0 was observed, this leads to index 2 in π_2, so the policy plays arm 0 with probability $1/4$ and arm 1 with probability $3/4$.

Example 2. If the policy observed the history $\{0, 1, 1, 1\}$, i.e. action 0 was played in the first round, the reward 1 was observed, then action 1 was played in the second round and the reward 1 was observed. This corresponds to index 7 in π_3. Thus, the policy plays arm 0 with probability 0 and arm 1 with probability 1.

Since the events and the neighbouring datasets are finite (and have a small number), it is easy to build the following two sets:

$$A = \left\{\left(\mathcal{V}^{\pi}_{\mathbf{r}}(E), \mathcal{V}^{\pi}_{\mathbf{r}'}(E)\right), \forall E \in \mathcal{P}([2]^3), \text{ and } \forall \mathbf{r} \sim \mathbf{r}'\right\}$$
$$B = \left\{(\mathcal{M}^{\pi}_{d}(E), \mathcal{M}^{\pi}_{d'}(E)), \forall E \in \mathcal{P}([2]^3), \text{ and } \forall d \sim d'\right\}$$

A and B represent all the probability tuples (p, q) computed on all neighbouring lists and tables of rewards, respectively, for all possible events on the sequence of actions.

Then, by checking over all the elements of A and B, it is possible to show that π is (ϵ_1, δ_1)-View DP but never (ϵ_1, δ_1)-Table DP for $\epsilon_1 = 0.95$ and $\delta_1 = 0.17$. Specifically, we mean that for $\epsilon_1 = 0.95$ and $\delta_1 = 0.17$, we obtain that $\forall (p, q) \in A,\ p \leq e^{\epsilon_1} q + \delta_1$, while $\exists (p', q') \in B,\ p' > e^{\epsilon_1} q' + \delta_1$. In fact, we can show that the smallest ϵ_0, for which π is (ϵ_0, δ_1)-Table DP, is $\epsilon_0 = 0.98$.

Thus, we conclude our proof with this construction. □

B. Proof of Proposition 2

Remark 6. *We recall that to check the interactive DP condition, it is enough to only consider deterministic adversaries (Lemma 2.2 in [18]).*

Proposition 2. *For any policy π, we have that*

(a) π is ρ-Interactive zCDP $\Rightarrow$ π is ρ-Table zCDP

(b) π is ρ-Interactive ADP if and only if, for every deterministic adversary $B = \{B_t\}_{t=1}^T$, π^B is ρ-Table zCDP. Here, $\pi^B \triangleq \{\pi_t^B\}_{t=1}^T$ is a post-processing of the policy π induced by the adversary B such that

$$\pi_t^B(a \mid a_1, r_1, .., a_{t-1}, r_{t-1}) \triangleq$$
$$\pi_t\Big(a \mid B_1(a_1), r_1, B_2(a_1, a_2), r_2, .., B_{t-1}(a_1, .., a_{t-1}), r_{t-1}\Big).$$

Proof. (a) is direct by taking the identity-adversary B^{id} defined by $B_t^{\text{id}}(o_1, \ldots, o_t) = o_t$.

(b) is direct by observing that for every deterministic adversary B, the view of adversary B reduces to $\text{View}(B \leftrightarrow \mathcal{M}(d)) = \mathcal{M}^{\pi^B}$. □

C. Proof of Theorem 1

Theorem 1 (Group privacy for ρ-Interactive DP)**.** *If π is a ρ-Interactive zCDP policy then, for any sequence of ac-*

tions $(a_1, \dots, a_T)$ *and any two sequence of rewards* $\boldsymbol{r} \triangleq \{r_1, \dots, r_T\}$ *and* $\boldsymbol{r'} \triangleq \{r'_1, \dots, r'_T\}$*, we have that*

$$\sum_{t=1}^{T} \mathrm{KL}\left(\pi_t(. \mid \mathcal{H}_{t-1}) \;\middle\|\; \pi_t(. \mid \mathcal{H}'_{t-1})\right) \leq \rho d_{Ham}(\boldsymbol{r}, \boldsymbol{r'})^2$$

where $\mathcal{H}_t \triangleq (a_1, r_1, \dots, a_t, r_t)$*,* $\mathcal{H}'_t \triangleq (a_1, r'_1, \dots, a_t, r'_t)$ *and* $d_{Ham}(\boldsymbol{r}, \boldsymbol{r'}) = \sum_{t=1}^{T} \mathbb{1}\{r_t \neq r'_t\}$*.*

Proof. Let $\mathbf{a} \triangleq (a_1, \dots, a_T)$ be a fixed sequence of actions. Let $\mathbf{r} \triangleq \{r_1, \dots, r_T\}$ and $\mathbf{r'} \triangleq \{r'_1, \dots, r'_T\}$ be two sequences of rewards.

Step 1: The constant adversary. We consider the constant adversary $B_{\mathbf{a}}$ defined as

$$B_{\mathbf{a}}(o_1, \dots, o_t) \triangleq a_t$$

i.e. $B_{\mathbf{a}}$ is the adversary that always queries at step t the action a_t, independently of the actions recommended by the policy. Let $\pi_{\mathbf{a}} \triangleq \pi^{B_{\mathbf{a}}}$ be the policy corresponding to the post-processing B_a.

Since π is ρ-Interactive zCDP, using Proposition 2, (b), then $\mathcal{M}^{\pi_{\mathbf{a}}}$ is ρ-zCDP. And Proposition 1, (c) gives that $\mathcal{V}^{\pi_{\mathbf{a}}}$ is ρ-zCDP.

Step 2: Group privacy of zCDP. Using the group privacy property of ρ-zCDP i.e. Theorem 16 with $\alpha = 1$, we get that

$$\mathrm{KL}\left(\mathcal{V}^{\pi_{\mathbf{a}}}_{\mathbf{r}} \parallel \mathcal{V}^{\pi_{\mathbf{a}}}_{\mathbf{r'}}\right)) \leq \rho \; d_{\mathrm{Ham}}(\mathbf{r}, \mathbf{r'})^2. \tag{11}$$

Step 3: Decomposing the view of the constant adversary. On the other hand, we have that

$$\mathcal{V}^{\pi_{\mathbf{a}}}_{\mathbf{r}}(o_1, \dots, o_T) = \prod_{t=1}^{T} \pi_t(o_t \mid a_1, r_1, \dots, a_{t-1}, r_{t-1}).$$

In other words $\mathcal{V}^{\pi_{\mathbf{a}}}_{\mathbf{r}} = \bigotimes_{t=1}^{T} \pi_t(. \mid a_1, r_1, \dots, a_{t-1}, r_{t-1})$.
Similarly, $\mathcal{V}^{\pi_{\mathbf{a}}}_{\mathbf{r'}} = \bigotimes_{t=1}^{T} \pi_t(. \mid a_1, r'_1, \dots, a_{t-1}, r'_{t-1})$.
Hence, we get

$$\mathrm{KL}\left(\mathcal{V}^{\pi_{\mathbf{a}}}_{\mathbf{r}} \parallel \mathcal{V}^{\pi_{\mathbf{a}}}_{\mathbf{r'}}\right)) = \sum_{t=1}^{T} \mathrm{KL}\left(\pi_t(. \mid \mathcal{H}_{t-1}) \;\middle\|\; \pi_t(. \mid \mathcal{H}'_{t-1})\right) \tag{12}$$

Plugging Equaion (12) in Inequality (11) concludes the proof. □

D. Privacy definitions for contextual bandits and Joint DP.

Joint DP is a definition of privacy proposed by [11] for linear contextual bandits when both contexts and rewards contain sensitive information. First, we recall their definition adapted to our notations and terminology.

Definition 7 (Joint "View" DP [11])**.** *We say two sequences* $S \triangleq \{(\mathcal{A}_1, r_1), (\mathcal{A}_2, r_2), \dots, (\mathcal{A}_n, r_n)\}$ *and* $S' \triangleq \{(\mathcal{A}'_1, r'_1), \dots, (\mathcal{A}'_n, r'_n)\}$ *are* t-neighbors *if for all* $s \neq t$ *it holds that* $(\mathcal{A}_s, r_s) = (\mathcal{A}'_s, r'_s)$*.*

A randomised algorithm π *for the contextual bandit problem is* (ϵ, δ)-Jointly "View" Differentially Private *(View JDP) if for any* t *and any pair of* t*-neighbouring sequences* S *and* S'*, and any subset* $E_{>t} \subset \mathcal{A}_{t+1} \times \mathcal{A}_{t+2} \times \cdots \times \mathcal{A}_n$ *of sequence of actions ranging from step* $t+1$ *to the end of the sequence, it holds that*

$$\Pr\{\pi(S) \in E_{>t}\} \leq e^{\epsilon} \Pr\{\pi(S') \in E_{>t}\} + \delta.$$

Joint DP requires that changing the context at step t does not affect *only the future rounds* $(> t)$. In contrast, the standard notion of DP would require that the change does not have any effect on the full sequence of actions, including the one chosen at step t. [11] show that the standard notion of DP for linear contextual bandits, where both the reward and contexts are private, always leads to linear regret.

In light of the discussion on the difference between Table DP and View DP, the Joint DP as expressed in [11] is similar to the View DP definition. This is because the input considered is a sequence of context and *observed* rewards. To define a Table DP counterpart of it, we consider a joint table of contexts and rewards, i.e. $S \triangleq \{(\mathcal{A}_1, x_1), (\mathcal{A}_2, x_2), \dots, (\mathcal{A}_n, x_n)\}$ and $S' \triangleq \{(\mathcal{A}'_1, x'_1), \dots, (\mathcal{A}'_n, x'_n)\}$ as input. Here, x_t is the row of potential rewards of user u_t. Hence, the modified Table JDP definition protects the user by protecting all the potential responses rather than only the observed ones.

Definition 8 (Joint "Table" DP)**.** *We say two sequences* $S \triangleq \{(\mathcal{A}_1, x_1), (\mathcal{A}_2, x_2), \dots, (\mathcal{A}_T, x_T)\}$ *and* $S' \triangleq \{(\mathcal{A}'_1, x'_1), \dots, (\mathcal{A}'_T, x'_T)\}$ *are* t-neighbours *if for all* $s \neq t$ *and* $s \in \{1, \dots, T\}$*,* $(\mathcal{A}_s, x_s) = (\mathcal{A}'_s, x'_s)$*.*

A randomised algorithm π *for the contextual bandit problem is* (ϵ, δ)-Jointly "Table" Differentially Private *(Table JDP) if for any* $t \in \{1, \dots, T\}$ *and any pair of* t*-neighbouring sequences* S *and* S'*, and any subset* $E_{>t} \subset \mathcal{A}_{t+1} \times \mathcal{A}_{t+2} \times \cdots \times \mathcal{A}_n$ *of sequence of actions ranging from step* $t+1$ *to the end of the sequence, it holds that*

$$\Pr\{\pi(S) \in E_{>t}\} \leq e^{\epsilon} \Pr\{\pi(S') \in E_{>t}\} + \delta.$$

In Section IV-B, we only consider the rewards to be private, while the contexts are supposed to be public. Thus, we do not need to adhere to Table JDP, and the definitions of Section III can readily be applied. This assumption can make sense in applications where the context does not contain users' private information. For example, in clinical trials, one can take the context to be some set of patient's public features. In this case, the only private information to be protected is the reaction of the patient to the medicine, which is the reward.

When contexts contain sensitive users' information, AdaC-OFUL and its analysis do not hold anymore. In this case, a private bandit algorithm should verify the stronger Joint Table DP constraint. In paragraph "2. Adapting AdaC-OFUL for private context" of Section V-B2, we explain how to derive a Table JDP version of AdaC-OFUL. However, the present regret analysis does not hold anymore. It would be an interesting future work to address this open question.

APPENDIX C
FINITE-ARMED BANDITS WITH ZCDP

In this section, we first specify the setting of finite-armed bandits with ρ-Interactive zCDP. Then, we present AdaC-UCB and analyse its regret to quantify the cost of ρ-Interactive zCDP.

A. Setting

Let $\nu = (P_a : a \in [K])$ be a bandit instance with K arms and means $(\mu_a)_{a\in[K]}$. The goal is to design a ρ-Interactive zCDP policy π that maximises the cumulative reward, or minimises regret over a horizon T:

$$\mathrm{Reg}_T(\pi,\nu) \triangleq T\mu^\star - \mathbb{E}\left[\sum_{t=1}^{T} r_t\right] = \sum_{a=1}^{K} \Delta_a \mathbb{E}\left[N_a(T)\right] \quad (13)$$

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 and $N_a(T) \triangleq \sum_{t=1}^{T} \mathbb{1}\{a_t = a\}$ 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.

B. Algorithm

AdaC-UCB is an extension of the generic algorithmic wrapper proposed by [8] for bandits with ρ-Interactive zCDP. Following [8], AdaC-UCB relies on three ingredients: *arm-dependent doubling*, *forgetting*, and *adding calibrated Gaussian noise*. First, the algorithm runs in episodes. The *same arm* is played for a whole episode, and *double* the number of times it was last played. Second, at the beginning of a new episode, the index of arm a, as defined in Eq. (14), is computed only using samples from the last episode, where arm a was played, while forgetting all the other samples. In a given episode, the arm with the highest index is played for all the steps. Due to these two ingredients, namely *doubling* and *forgetting*, each empirical mean computed in the index of Eq. (14) only needs to be ρ-zCDP for the algorithm to be ρ-Interactive zCDP, avoiding the need of composition theorems.

Algorithm 4 AdaC-UCB

1: **Input:** Privacy budget ρ, an environment ν with K arms, optimism parameter $\beta > 3$
2: **Output:** Actions satisfying ρ-Interactive zCDP
3: **Initialisation:** Choose each arm once and let $t = K$
4: **for** $\ell = 1, 2, \ldots$ **do**
5: Let $t_\ell = t + 1$
6: Compute $a_\ell = \mathrm{argmax}_a \, \mathrm{I}^\rho_a(t_\ell - 1, \beta)$ (Eq. (14))
7: Choose arm a_ℓ until round t such that $N_{a_\ell}(t) = 2N_{a_\ell}(t_\ell - 1)$
8: **end for**

For AdaC-UCB, we use the private index to select the arms (Line 6 of Algorithm 4) as

$$\mathrm{I}^\rho_a(t_\ell - 1, \beta) \triangleq \hat{\mu}^\ell_a + \mathcal{N}\left(0, \sigma^2_{a,\ell}\right) + B_a(t_\ell - 1, \beta). \quad (14)$$

Here, $\hat{\mu}^\ell_a$ is the empirical mean of rewards collected in the last episode in which arm a was played, the variance of the Gaussian noise is

$$\sigma^2_{a,\ell} \triangleq \frac{1}{2\rho \times \left(\frac{1}{2}N_a(t_\ell - 1)\right)^2}$$

and the exploration bonus $B_a(t_\ell - 1, \beta)$ is defined as

$$\sqrt{\left(\frac{1}{2 \times \frac{1}{2}N_a(t_\ell - 1)} + \frac{1}{\rho \times \left(\frac{1}{2}N_a(t_\ell - 1)\right)^2}\right)\beta \log(t_\ell)}.$$

The term in blue rectifies the non-private confidence bound of UCB for the added Gaussian noise.

C. Concentration inequalities

Lemma 2. *Assume that $(X_i)_{1\le i\le n}$ are iid random variables in $[0,1]$, with $\mathbb{E}(X_i) = \mu$. Then, for any $\delta \ge 0$,*

$$\mathbb{P}\left(\hat{\mu}_n + Z_n - \sqrt{\left(\frac{1}{2n} + \frac{1}{\rho n^2}\right)\log\left(\frac{1}{\delta}\right)} \ge \mu\right) \le \delta, \quad (15)$$

and

$$\mathbb{P}\left(\hat{\mu}_n + Z_n + \sqrt{\left(\frac{1}{2n} + \frac{1}{\rho n^2}\right)\log\left(\frac{1}{\delta}\right)} \le \mu\right) \le \delta, \quad (16)$$

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

Proof. Let $Y = (\hat{\mu}_n + Z_n - \mu)$.
Using Properties 2 and 3 of Lemma 9, we get that Y is $\sqrt{\frac{1}{4n} + \frac{1}{2\rho n^2}}$-subgaussian.
We conclude using the concentration on subgaussian random variables, i.e. Lemma 8. □

D. Regret analysis

Theorem 10 (Part a: Problem-dependent regret)**.** *For rewards in $[0,1]$ and $\beta > 3$, AdaC-UCB yields a regret upper bound of*

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

Proof. By the generic regret decomposition of Theorem 11 in [8], for every sub-optimal arm a, we have that

$$\mathbb{E}[N_a(T)] \le 2^{\ell+1} + \mathbb{P}\left(G^c_{a,\ell,T}\right)T + \frac{\beta}{\beta - 3}, \quad (17)$$

where

$$G_{a,\ell,T} = \left\{\hat{\mu}_{a,2^\ell} + Z_\ell + b_{\ell,T} < \mu_1\right\}.$$

such that $b_{\ell,T} \triangleq \sqrt{\left(\frac{1}{2\times 2^\ell} + \frac{1}{\rho\times(2^\ell)^2}\right)\beta\log(T)}$ and $Z_\ell \sim \mathcal{N}\left(0, 1/\left(2\rho \times \left(2^\ell\right)^2\right)\right)$.

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

$$\begin{aligned}\mathbb{P}(G^c_{a,\ell,T}) &= \mathbb{P}\left(\hat{\mu}_{a,2^\ell} + Z_\ell + b_{\ell,T} \ge \mu_1\right)\\ &= \mathbb{P}\left(\hat{\mu}_{a,2^\ell} + Z_\ell - b_{\ell,T} \ge \mu_a + \epsilon\right)\end{aligned}$$

for $\epsilon = \Delta_a - 2b_{\ell,T}$.

The idea is to choose ℓ big enough so that $\epsilon \ge 0$.

Let us consider the contrary, i.e.

$$\epsilon < 0 \Rightarrow 2^\ell < \frac{2\beta\log(T)}{\Delta_a^2}\left(1 + \Delta_a\sqrt{\frac{1}{\rho\beta\log(T)}}\right)$$

$$\Rightarrow 2^\ell < \frac{2\beta}{\Delta_a^2}\log(T) + 2\sqrt{\frac{\beta}{\rho\Delta_a^2}}\sqrt{\log(T)} \qquad (18)$$

Thus, by choosing

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

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

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

The last inequality is due to Equation 15 of Lemma 2, with $n = 2^\ell$ and $\delta = T^{-\beta}$.

Step 2: The regret bound. Plugging the choice of ℓ and the upper bound on $\mathbb{P}(G^c_{a,\ell,T})$ in Inequality 17 gives

$$\begin{aligned}\mathbb{E}[N_a(T)] &\leq \frac{\beta}{\beta-3} + 2^{\ell+1} + T \times \frac{1}{T^\beta} \\ &\leq \frac{8\beta}{\Delta_a^2}\log(T) + 8\sqrt{\frac{\beta}{\rho\Delta_a^2}}\sqrt{\log(T)} + \frac{2\beta}{\beta-3}. \quad (19)\end{aligned}$$

Plugging this upper bound back in the definition of problem-dependent regret, we get that the regret $\mathrm{Reg}_T(\mathsf{AdaC\text{-}UCB}, \nu)$ is upped bounded by

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

□

Theorem 11 (Part b: Minimax regret)**.** *For rewards in* $[0,1]$ *and* $\beta > 3$*,* AdaC-UCB *yields a regret upper bound of*

$$\mathcal{O}\left(\sqrt{KT\log(T)}\right) + \mathcal{O}\left(K\sqrt{\frac{1}{\rho}\log(T)}\right).$$

Proof. Let Δ be a value to be tuned later.

We observe that

$$\begin{aligned}&\mathrm{Reg}_T(\mathsf{AdaP\text{-}UCB}, \nu) = \sum_a \Delta_a \mathbb{E}[N_a(T)] \\ &= \sum_{a:\Delta_a\leq\Delta} \Delta_a \mathbb{E}[N_a(T) + \sum_{a:\Delta_a>\Delta} \Delta_a \mathbb{E}[N_a(T)] \\ &\leq T\Delta + \sum_{a:\Delta_a>\Delta} \Delta_a\left(\frac{8\beta}{\Delta_a^2}\log(T) + 8\sqrt{\frac{\beta\log(T)}{\rho\Delta_a^2}} + \frac{2\beta}{\beta-3}\right) \\ &\leq T\Delta + \frac{8\beta K\log(T)}{\Delta} + 8K\sqrt{\frac{\beta\log(T)}{\rho}} + \frac{3\beta}{\beta-3}\sum_a \Delta_a \\ &\leq 4\sqrt{2\beta KT\log(T)} + 8K\sqrt{\frac{\beta\log(T)}{\rho}} + \frac{3\beta}{\beta-3}\sum_a \Delta_a\end{aligned}$$

Here, the last step is tuning $\Delta = \sqrt{\frac{8\beta K\log(T)}{T}}$. □

Theorem 12 (Privacy of AdaC-UCB)**.** *For rewards in* $[0,1]$*,* AdaC-UCB *satisfies* ρ*-Interactive zCDP.*

The privacy proof is provided in Appendix D.

E. Extensions to (ϵ,δ)*-Interactive DP and* (α,ϵ)*-Interactive RDP*

The difference comes from the different calibrations of the Gaussian Mechanism (Thm 18). Adapting the analysis from ρ-zCDP reduces to changing the $\frac{1}{2\rho}$ factor to $\frac{2}{\epsilon^2}\log(\frac{1.25}{\delta})$ for (ϵ,δ)-DP and to $\frac{\alpha}{2\epsilon}$ for (α,ϵ)-RDP, i.e. varying the constant b in Theorem 18.

APPENDIX D
PRIVACY PROOFS

In this section, we give complete proof of the privacy of AdaC-UCB, AdaC-GOPE and AdaC-OFUL. The three algorithms share the same blueprint. The intuition behind the blueprint is formalised in Lemma 1, then a generic proof of privacy and specification for each algorithm are given after.

A. The privacy lemma of non-overlapping sequences

Remark 7. *The Privacy Lemma shows that when the mechanism* $\mathcal{M}$ *is applied to non-overlapping subsets of the input dataset, there is no need to use the composition theorems. Plus, there is no additional cost in the privacy budget.*

Lemma 1 (Privacy Lemma)**.** *Let* $\mathcal{M}$ *be a mechanism that takes a* ***set*** *as input. 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

$$\mathcal{G} : \{x_1, \dots, x_T\} \to \bigotimes_{i=1}^{\ell} \mathcal{M}_{\{x_{t_i}, \dots, x_{t_{i+1}-1}\}}$$

$\mathcal{G}$ *is the mechanism we get by applying* $\mathcal{M}$ *to the partition of the input dataset* $\{x_1, \dots, x_T\}$ *according to* $t_1 < \cdots < t_\ell < t_{\ell+1}$*, i.e.*

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_T \end{pmatrix} \xrightarrow{\mathcal{G}} \begin{pmatrix} o_1 \\ \vdots \\ o_\ell \end{pmatrix}$$

where $o_i \sim \mathcal{M}_{\{x_{t_i}, \dots, x_{t_{i+1}-1}\}}$.

We have that

(a) If $\mathcal{M}$ *is* (ϵ,δ)*-DP then* $\mathcal{G}$ *is* (ϵ,δ)*-DP*
(b) If $\mathcal{M}$ *is* ρ*-zCDP then* $\mathcal{G}$ *is* ρ*-zCDP*

Proof. Let $x \triangleq \{x_1, \dots, x_T\}$ and $x' \triangleq \{x'_1, \dots, x'_T\}$ be two neighboring datasets. This implies that $\exists j \in [1,T]$ such that $x_j \neq x'_j$ and $\forall t \neq j$, $x_t = x'_t$.

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

We denote $\{x\}_{t_i}^{t_{i+1}} \triangleq \{x_{t_i}, \dots, x_{t_{i+1}-1}\}$ the records in x corresponding to the episode from t_i until $t_{i+1} - 1$.

(a) Suppose that $\mathcal{M}$ is (ϵ,δ)-DP.

For every output event $E = E_1 \times \cdots \times E_\ell$, we have that

$$\mathcal{G}_x(E) = \prod_{i=1}^{\ell} \mathcal{M}_{\{x\}_{t_i}^{t_{i+1}}}(E_i)$$

$$
\begin{aligned}
&= \mathcal{M}_{\{x\}_{t_{\ell'}}^{t_{\ell'+1}}}(E_{\ell'}) \prod_{i=1, i\neq \ell'}^{\ell} \mathcal{M}_{\{x\}_{t_i}^{t_{i+1}}}(E_i) \\
&\leq \left(e^{\epsilon} \mathcal{M}_{\{x'\}_{t_{\ell'}}^{t_{\ell'+1}}}(E_{\ell'}) + \delta \right) \prod_{i=1, i\neq \ell'}^{\ell} \mathcal{M}_{\{x\}_{t_i}^{t_{i+1}}}(E_i) \\
&= e^{\epsilon} \mathcal{G}_{x'}(E) + \delta \times \prod_{i=1, i\neq \ell'}^{\ell} \mathcal{M}_{\{x\}_{t_i}^{t_{i+1}}}(E_i) \\
&\leq e^{\epsilon} \mathcal{G}_{x'}(E) + \delta
\end{aligned}
$$

since $\prod_{i=1, i\neq \ell'}^{\ell} \mathcal{M}_{\{x\}_{t_i}^{t_{i+1}}}(E_i) \leq 1$

Which gives that $\mathcal{G}$ is (ϵ, δ)-DP.

(b) Suppose that $\mathcal{M}$ is ρ-zCDP. Let denote $o^{\ell} \triangleq (o_1, \ldots, o_{\ell})$ We have that

$$
D_{\alpha}(\mathcal{G}_x \| \mathcal{G}_{x'}) = \frac{1}{\alpha - 1} \log \left(\int_{o^{\ell}} \mathcal{G}_{x'}(o) \left(\frac{\mathcal{G}_x(o)}{\mathcal{G}_{x'}(o)} \right)^{\alpha} \right)
$$

Since

$$
\mathcal{G}_x(o) = \prod_{i=1}^{\ell} \mathcal{M}_{\{x\}_{t_i}^{t_{i+1}}}(o_i)
$$

and

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

we get

$$
\frac{\mathcal{G}_x(o)}{\mathcal{G}_{x'}(o)} = \frac{\mathcal{M}_{\{x\}_{t_{\ell'}}^{t_{\ell'+1}}}(o_i)}{\mathcal{M}_{\{x'\}_{t_{\ell'}}^{t_{\ell'+1}}}(o_i)}
$$

Thus,

$$
D_{\alpha}(\mathcal{G}_x \| \mathcal{G}_{x'}) = D_{\alpha}(\mathcal{M}_{\{x\}_{t_{\ell'}}^{t_{\ell'+1}}} \| \mathcal{M}_{\{x'\}_{t_{\ell'}}^{t_{\ell'+1}}}) \leq \alpha\rho
$$

Which gives that $\mathcal{G}$ is ρ-zCDP.

□

For each of the three algorithms proposed, the final actions can be seen as a post-processing of some private quantity of interest (empirical means for AdaC-UCB or the parameter $\hat{\theta}$ for linear and contextual bandits). However, we cannot directly conclude the privacy of the proposed algorithms using just a post-processing argument and Lemma 1. This is because the steps corresponding to the start of an episode in the algorithms $t_1 < \cdots < t_{\ell} < t_{\ell+1}$ are adaptive and depend on the dataset itself, while for Lemma 1, those have been fixed before.

To deal with the adaptive episode, we propose a generic privacy proof.

B. Generic privacy proof

In this section, we give one generic proof that works for the two proposed algorithms.

First, we give a summary of the intuition of the proof for dealing with adaptive episodes. By fixing two neighbouring tables of rewards d and d' that only differ at some user u_j, and a deterministic adversary B, we have that

- the view of the adversary B from the beginning of the interaction until step j will be the same
- the adaptive episodes generated by the policy in the first j steps will be the same, which means that step j will fall in the same episode in the view of B when interacting with $\pi(d)$ or $\pi(d')$
- for these fixed similar episodes, we use the privacy Lemma 1
- the view of B from step $j+1$ until T will be private by post-processing

Let $d = \{x_1, \ldots, x_T\}$ and $d' = \{x'_1, \ldots, x'_T\}$ two neighbouring reward tables in $(\mathbb{R}^K)^T$. Let $j \in [1, T]$ such that, for all $t \neq j$, $x_t = x'_t$.

Let B be a deterministic adversary.

We want to show that $D_{\alpha}(\text{View}(B \leftrightarrow \pi(d)) \| \text{View}(B \leftrightarrow \pi(d'))) \leq \alpha\rho$.

Step 1. Sequential decomposition of the view of the adversary B

We observe that due to the sequential nature of the interaction, the view of B can be decomposed to a part that depends on $d_{<j} \triangleq \{x_1, \ldots, x_{j-1}\}$, which is identical for both d and d' and a second conditional part on the history.

First, let us denote $\mathcal{P}_d^{B,\pi} \triangleq \text{View}(B \leftrightarrow^d \pi)$, $\mathbf{o}_{\leq j} \triangleq (o_1, \ldots, o_j)$ and $\mathbf{o}_{>j} \triangleq (o_{j+1}, \ldots, o_T)$.

We have that, for every sequence of actions $\mathbf{o} \triangleq (o_1, \ldots, o_T) \in [K]^T$

$$
\begin{aligned}
&\mathcal{P}_d^{B,\pi}(\mathbf{o}) \\
&= \prod_{t=1}^{T} \pi_t \left(o_t \mid B(o_1), x_{1,B(o_1)}, \ldots, B(\mathbf{o}_{\leq t-1}), x_{t-1,B(\mathbf{o}_{\leq t-1})} \right) \\
&\triangleq \mathcal{P}_{d_{<j}}^{B,\pi}(\mathbf{o}_{\leq j}) \mathcal{P}_d^{B,\pi}(\mathbf{o}_{>j} \mid \mathbf{o}_{\leq j})
\end{aligned}
$$

where

$$
\begin{aligned}
&\mathcal{P}_{d_{<j}}^{B,\pi}(\mathbf{o}_{\leq j}) \\
&\triangleq \prod_{t=1}^{j} \pi_t \left(o_t \mid B(o_1), x_{1,B(o_1)}, \ldots, B(\mathbf{o}_{\leq t-1}), x_{t-1,B(\mathbf{o}_{\leq t-1})} \right)
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathcal{P}_d^{B,\pi}(\mathbf{o}_{>j} \mid \mathbf{o}_{\leq j}) \\
&\triangleq \prod_{t=j+1}^{T} \pi_t \left(o_t \mid B(o_1), x_{1,B(o_1)}, \ldots, B(\mathbf{o}_{\leq t-1}), x_{t-1,B(\mathbf{o}_{\leq t-1})} \right)
\end{aligned}
$$

Similarly

$$
\mathcal{P}_{d'}^{B,\pi}(\mathbf{o}) = \mathcal{P}_{d_{<j}}^{B,\pi}(\mathbf{o}_{\leq j}) \mathcal{P}_{d'}^{B,\pi}(\mathbf{o}_{>j} \mid \mathbf{o}_{\leq j})
$$

since $d'_{<j} = d_{<j}$.

Step 2. Decomposing the Rényi divergence.

We have that

$$
\begin{aligned}
e^{(\alpha-1)D_{\alpha}(\mathcal{P}_d^{B,\pi} \| \mathcal{P}_{d'}^{B,\pi})} &= \sum_{\mathbf{o} \in [K]^T} \mathcal{P}_{d'}^{B,\pi}(\mathbf{o}) \left(\frac{\mathcal{P}_d^{B,\pi}(\mathbf{o})}{\mathcal{P}_{d'}^{B,\pi}(\mathbf{o})} \right)^{\alpha} \\
&= \sum_{\mathbf{o} \in [K]^T} \mathcal{P}_{d'}^{B,\pi}(\mathbf{o}) \left(\frac{\mathcal{P}_d^{B,\pi}(\mathbf{o}_{>j} \mid \mathbf{o}_{\leq j})}{\mathcal{P}_{d'}^{B,\pi}(\mathbf{o}_{>j} \mid \mathbf{o}_{\leq j})} \right)^{\alpha}
\end{aligned}
$$

$$= \sum_{\mathbf{o}_{\leq j} \in [K]^j} \mathcal{P}^{B,\pi}_{d'_{<j}}(\mathbf{o}_{\leq j}) \sum_{\mathbf{o}_{>j} \in [K]^{T-j}} \mathcal{P}^{B,\pi}_{d'}(\mathbf{o}_{>j} \mid \mathbf{o}_{\leq j}) \left(\frac{\mathcal{P}^{B,\pi}_{d}(\mathbf{o}_{>j} \mid \mathbf{o}_{\leq j})}{\mathcal{P}^{B,\pi}_{d'}(\mathbf{o}_{>j} \mid \mathbf{o}_{\leq j})} \right)^{\alpha}$$

$$= \sum_{\mathbf{o}_{\leq j} \in [K]^j} \mathcal{P}^{B,\pi}_{d_{<j}}(\mathbf{o}_{\leq j}) e^{(\alpha-1) D_\alpha(\mathcal{P}^{B,\pi}_{d}(.|\mathbf{o}_{\leq j}) \| \mathcal{P}^{B,\pi}_{d'}(.|\mathbf{o}_{\leq j}))}$$

$$= \mathbb{E}_{\mathbf{o}_{\leq j} \sim \mathcal{P}^{B,\pi}_{d_{<j}}} \left[e^{(\alpha-1) D_\alpha(\mathcal{P}^{B,\pi}_{d}(.|\mathbf{o}_{\leq j}) \| \mathcal{P}^{B,\pi}_{d'}(.|\mathbf{o}_{\leq j}))} \right]$$

Step 3. The adaptive episodes are the same, before step j.

Let ℓ such that $t_\ell \leq j < t_{\ell+1}$ in the view of B when interacting with d. Let us call it $\psi^\pi_d(j) \triangleq \ell$. Similarly, let ℓ' such that $t_{\ell'} \leq j < t_{\ell'+1}$ in the view of B when interacting with d. Let us call it $\psi^\pi_{d'}(j) \triangleq \ell'$.

Since $\psi^\pi_d(j)$ only depends on $d_{<j}$, which is identical for d and d', we have that $\psi^\pi_d(j) = \psi^\pi_{d'}(j)$ with probability 1.

We call ξ_j the last **time-step** of the episode $\psi^\pi_d(j)$, i.e $\xi_j \triangleq t_{\psi^\pi_d(j)+1} - 1$.

Step 4. Private sufficient statistics.

Fix $\mathbf{o}_{\leq j}$.

Let $r_s \triangleq x_{s,B(o_1,\dots,o_s)}$, for $s \in [1, j]$, be the reward corresponding to the action chosen by B in the table d. Similarly, $r'_s \triangleq x'_{s,B(o_1,\dots,o_s)}$ for d'.

Let us define $L_j \triangleq \mathcal{G}_{\{r_1,\dots,r_{\xi_j}\}}$ and $L'_j \triangleq \mathcal{G}_{\{r'_1,\dots,r'_{\xi_j}\}}$, where $\mathcal{G}$ is defined as in Eq. 7, using the same episodes for d and d'. The underlying mechanism $\mathcal{M}$, used to define $\mathcal{G}$, will be specified for each algorithm in Section D-C.

In addition, the specified mechanism $\mathcal{M}$ will verify ρ-zCDP with respect to its set input.

Using the structure of the policy π, there exists a randomised mapping $f_{x_{\xi_j+1},\dots,x_T}$ such that $\mathcal{P}^{B,\pi}_d(. \mid \mathbf{o}_{\leq j}) = f_{x_{\xi_j+1},\dots,x_T}(L_j)$ and $\mathcal{P}^{B,\pi}_{d'}(. \mid \mathbf{o}_{\leq j}) = f_{x_{\xi_j+1},\dots,x_T}(L'_j)$.

In other words, the view of the adversary B from step ξ_j+1 until T only depends on the sufficient statistics L_j and the new inputs $x_{\xi_j+1}, \dots, x_T$, which are the same for d and d'.

For example, the sufficient statistics are the private mean estimate of the active arm in each episode for AdaC-UCB and the noisy parameter estimate $\hat{\theta}$ for AdaC-GOPE.

Step 5. Concluding with Lemma 1 and post-processing.

Using Lemma 1, we have that

$$D_\alpha(L_j, L'_j) \leq \alpha\rho$$

Using the post-processing property of D_α (Lemma 4), we get that

$$D_\alpha(\mathcal{P}^{B,\pi}_d(. \mid \mathbf{o}_{\leq j}) \| \mathcal{P}^{B,\pi}_{d'}(. \mid \mathbf{o}_{\leq j})) = D_\alpha(f_{x_{\xi_j+1},\dots,x_T}(L_j) \| f_{x_{\xi_j+1},\dots,x_T}(L'_j)) \leq D_\alpha(L_j, L'_j) \leq \alpha\rho$$

Finally, we conclude by taking the expectation with respect to $\mathbf{o}_{\leq j} \sim \mathcal{P}^{B,\pi}_{d_{<j}}$

$$e^{(\alpha-1) D_\alpha(\mathcal{P}^{B,\pi}_{d} \| \mathcal{P}^{B,\pi}_{d'})} = \mathbb{E}_{\mathbf{o}_{\leq j} \sim \mathcal{P}^{B,\pi}_{d_{<j}}} \left[e^{(\alpha-1) D_\alpha(\mathcal{P}^{B,\pi}_{d}(.|\mathbf{o}_{\leq j}) \| \mathcal{P}^{B,\pi}_{d'}(.|\mathbf{o}_{\leq j}))} \right] \leq e^{(\alpha-1)\alpha\rho}$$

Thus, we conclude

$$D_\alpha(\mathcal{P}^{B,\pi}_d \| \mathcal{P}^{B,\pi}_{d'}) \leq \alpha\rho$$

Remark 8. *The same proof could be adapted to other relaxations of Pure DP.*

C. Instantiating the specifics of privacy proof for each algorithm

In this section, we instantiate Step 4 of the generic proof for each algorithm, by specifying the mechanism $\mathcal{M}$ in the proof and showing that they are ρ-zCDP.

• **For** AdaC-UCB, the mechanism $\mathcal{M}$ is the private empirical mean statistic, i.e $\mathcal{M}_{\{r_1,\dots,r_t\}} \triangleq \frac{1}{t}\sum_{s=1}^t r_s + \mathcal{N}\left(0, \frac{1}{2\rho t^2}\right)$. Since rewards are in $[0,1]$, by the Gaussian Mechanism (i.e. Theorem 18) $\mathcal{M}$ is ρ-DP.

• **For** AdaC-GOPE, the mechanism $\mathcal{M}$ is a private estimate of the linear parameter θ, i.e $\mathcal{M}_{\{r_{t_\ell},\dots,r_{t_{\ell+1}-1}\}} \triangleq V_\ell^{-1}\left(\sum_{t=t_\ell}^{t_{\ell+1}-1} a_s r_s\right) + V_\ell^{-\frac{1}{2}} N_\ell$ where $V_\ell = \sum_{a \in \mathcal{S}_\ell} T_\ell(a) a a^\top$, $N_\ell \sim \mathcal{N}\left(0, \frac{2}{\rho} g_\ell^2 I_d\right)$ and $g_\ell = \max_{b \in \mathcal{A}_\ell} \|b\|_{V_\ell^{-1}}$.

To show that $\mathcal{M}$ is ρ-zCDP, we rewrite $\hat{\theta}_\ell = V_\ell^{-1}\left(\sum_{t=t_\ell}^{t_{\ell+1}-1} a_s r_s\right) = V_\ell^{-\frac{1}{2}} \phi_\ell$ where $\phi_\ell \triangleq V_\ell^{-\frac{1}{2}}\left(\sum_{t=t_\ell}^{t_{\ell+1}-1} a_s r_s\right)$.

Let $\{r_s\}_{s=t_\ell}^{t_{\ell+1}-1}$ and $\{r'_s\}_{s=t_\ell}^{t_{\ell+1}-1}$ two neighbouring sequence of rewards that differ at only step $j \in [t_\ell, t_{\ell+1}-1]$. We have that

$$\|\phi_\ell - \phi'_\ell\|_2 = \|V_\ell^{-\frac{1}{2}}\left[a_j(r_s - r'_s)\right]\|_2 \leq 2\|V_\ell^{-\frac{1}{2}} a_j\|_2 \leq 2 g_\ell$$

since $r_j, r'_j \in [-1, 1]$.

Using the Gaussian Mechanism (i.e. Theorem 18), this means that $\phi_\ell + N_\ell$ is ρ-zCDP and $\mathcal{M}$ is too by post-processing.

• **For** AdaC-OFUL, the mechanism $\mathcal{M}$ is the private estimate of the sum $\sum_{s=t_\ell}^{t_{\ell+1}-1} a_s r_s$, i.e $\mathcal{M}_{\{r_{t_\ell},\dots,r_{t_{\ell+1}-1}\}} \triangleq \sum_{t=t_\ell}^{t_{\ell+1}-1} a_s r_s + \mathcal{N}(0, \frac{2}{\rho} I_d)$.

Since rewards are in $[-1, 1]$ and $\|a\|_2 \leq 1$, the L2 sensitivity of $\sum_{t=t_\ell}^{t_{\ell+1}-1} a_s r_s$ is 2. By Theorem 18, $\mathcal{M}$ is ρ-zCDP.

We need an extra step of cumulatively summing the outputs of $\mathcal{G}$, which is still private by post-processing, i.e

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_T \end{pmatrix} \xrightarrow{\mathcal{G}} \begin{pmatrix} o_1 \\ \vdots \\ o_\ell \end{pmatrix} \to \begin{pmatrix} o_1 \\ o_1 + o_2 \\ \vdots \\ o_1 + o_2 + \dots + o_\ell \end{pmatrix}$$

Then, we have that $\left(\sum_{t=1}^{t_j} a_s r_s + \sum_{m=1}^{j} Y_m\right)_{j \in [1,\ell]}$ is ρ-zCDP, where $Y_m \overset{\text{i.i.d}}{\sim} \mathcal{N}\left(0, \frac{2}{\rho} I_d\right)$

This shows that the price of not forgetting is, for each estimate at the end of an episode j, to have to sum all the previous independent noises i.e. $\sum_{m=1}^{j} Y_j$, compared to just Y_j when forgetting.

APPENDIX E
LINEAR BANDITS WITH ZCDP

A. Concentration inequalities

Let $a_1, \dots, a_t$ be deterministically chosen without the knowledge of $r_1, \dots, r_t$. Let 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 $\tilde{\theta}_t = \hat{\theta}_t + V_t^{-\frac{1}{2}} N_t$ where $N_t \sim \mathcal{N}\left(0, \frac{2}{\rho} g_t^2 I_d\right)$, where $g_t \triangleq \max_{b \in \mathcal{A}} \|b\|_{V_t^{-1}}$.

Theorem 13. *Let $\delta \in [0,1]$ and $\beta_t \triangleq g_t \sqrt{2 \log\left(\frac{4}{\delta}\right)} + g_t^2 \sqrt{\frac{2}{\rho} f(d,\delta)}$, where $f(d,\delta) \triangleq d + 2\sqrt{d \log\left(\frac{2}{\delta}\right)} + 2\log\left(\frac{2}{\delta}\right)$. For every $a \in \mathcal{A}$, we have that*

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

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

$$\begin{aligned}\left\langle \tilde{\theta}_t - \theta^\star, a \right\rangle &= \left\langle \hat{\theta}_t - \theta^\star, a \right\rangle + a^T V_t^{-\frac{1}{2}} N_t \\ &= \left\langle \hat{\theta}_t - \theta^\star, a \right\rangle + Z_t\end{aligned}$$

where $Z_t \triangleq a^T V_t^{-\frac{1}{2}} N_t$.

Step 1: Concentration of the least square estimate. Using Eq.(20.2) from Chapter 20 of [2], we have that

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

Step 2: Concentration of the injected Gaussian noise. On the other hand, using Cauchy-Schwartz, we have that

$$|Z_t| = \left|a^T V_t^{-\frac{1}{2}} N_t\right| \leq \|V_t^{-\frac{1}{2}} a\| . \|N_t\| \leq g_t \|N_t\|$$

using that $\|V_t^{-\frac{1}{2}} a\| = \|a\|_{V_t^{-1}} \leq g_t$.

Here, $N_t = \sqrt{\frac{2}{\rho}} g_t \mathcal{N}(0, I_d)$. Thus, using Lemma 10, we get

$$\mathbb{P}\left(|Z_t| \geq g_t^2 \sqrt{\frac{2}{\rho} f(d,\delta)}\right) \leq \frac{\delta}{2}$$

Steps 1 and 2 together conclude the proof. □

Corollary 1. *Let β be a confidence level. If each action $a \in \mathcal{A}$ is chosen for $N_a(t) \triangleq \lceil c_t \pi(a) \rceil$ where*

$$c_t \triangleq \frac{8d}{\beta^2} \log\left(\frac{4}{\delta}\right) + \frac{2d}{\beta} \sqrt{\frac{2}{\rho} f(d,\delta)}$$

and $f(d,\delta) \triangleq d + 2\sqrt{d \log\left(\frac{2}{\delta}\right)} + 2\log\left(\frac{2}{\delta}\right)$.

then, for $t = \sum_{a \in \mathrm{Supp}(\pi)} N_a(t)$, 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

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

This means that

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

where the last equality is because π is an optimal design for $\mathcal{A}$.

Recall that

$$\beta_t \triangleq g_t \sqrt{2 \log\left(\frac{4}{\delta}\right)} + g_t^2 \sqrt{\frac{2}{\rho} f(d,\delta)}$$

Thus,

$$\begin{aligned}\beta_t &\leq \sqrt{\frac{d}{c_t}} \sqrt{2 \log\left(\frac{4}{\delta}\right)} + \frac{d}{c_t} \sqrt{\frac{2}{\rho} f(d,\delta)} \\ &\leq \frac{\sqrt{2d \log\left(\frac{4}{\delta}\right)}}{\sqrt{\frac{8d}{\beta^2} \log\left(\frac{4}{\delta}\right)}} + \frac{d \sqrt{\frac{2}{\rho} f(d,\delta)}}{\frac{2d}{\beta} \sqrt{\frac{2}{\rho} f(d,\delta)}} \\ &= \frac{\beta}{2} + \frac{\beta}{2} = \beta\end{aligned}$$

The final inequality is due to $c_t \geq \frac{8d}{\beta^2} \log\left(\frac{4}{\delta}\right)$, and $c_t \geq \frac{2d}{\beta} \sqrt{\frac{2}{\rho} f(d,\delta)}$.

We conclude the proof using Theorem 13. □

B. Regret analysis

Theorem 4 (Regret analysis of AdaC-GOPE)**.** *Under Assumption 1 and for $\delta \in (0,1)$, with probability at least $1-\delta$, the regret R_T of* AdaC-GOPE *is upper-bounded by*

$$A \sqrt{dT \log\left(\frac{K \log(T)}{\delta}\right)} + \frac{Bd}{\sqrt{\rho}} \sqrt{\log\left(\frac{K \log(T)}{\delta}\right)} \log(T)$$

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

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 1, 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 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 under the good event E,

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

where the first inequality is because $\langle \theta^\star, b - a^\star \rangle \leq 0$ by definition of the optimal arm $a^\star$.
This means that $a^\star$ is never eliminated. □

- 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 under E,

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

  which means that a get eliminated at the round ℓ_a. □
- for $a \in \mathcal{A}_{\ell+1}$, we have that $\Delta_a \leq 4\beta_\ell$.

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

Step 3: Regret decomposition under E.
Fix Δ to be optimised 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 the last inequality is thanks to the third bullet point in **Step 2**, i.e. $\Delta_a \leq 4\beta_{\ell-1}$ for $a \in \mathcal{A}_\ell$.

Also $\ell(T)$ is the total number of episodes played until timestep T.

Step 4: Upper-bounding T_ℓ and $\ell(T)$ under E.
Let $\delta_{K,\ell} \triangleq \frac{\delta}{K\ell(\ell+1)}$. We recall that $f(d,\delta) \triangleq d + 2\sqrt{d\log\left(\frac{2}{\delta}\right)} + 2\log\left(\frac{2}{\delta}\right)$.
We have that

$$\begin{aligned}T_\ell &= \sum_{a \in S_\ell} T_\ell(a) \\ &= \sum_{a \in S_\ell} \left\lceil \frac{8d\pi_\ell(a)}{\beta_\ell^2} \log\left(\frac{4}{\delta_{K,\ell}}\right) + \frac{2d\pi_\ell(a)}{\beta_\ell}\sqrt{\frac{2}{\rho} f(d, \delta_{K,\ell})} \right\rceil \\ &\leq \frac{d(d+1)}{2} + \frac{8d}{\beta_\ell^2}\log\left(\frac{4}{\delta_{K,\ell}}\right) + \frac{2d}{\beta_\ell}\sqrt{\frac{2}{\rho} f(d,\delta_{K,\ell})}.\end{aligned}$$

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 $\ell(T) \leq C\log(T)$. In other words, the length of the episodes is at least doubling so their number is logarithmic.

Which means that, for $\ell \leq \ell(T)$, there exists a constant C' such that

$$\log\left(\frac{4}{\delta_{K,\ell}}\right) = \log\left(\frac{4K\ell(\ell+1)}{\delta}\right) \leq C'\log\left(\frac{K\log(T)}{\delta}\right).$$

Define $\alpha_T \triangleq \log\left(\frac{K\log(T)}{\delta}\right)$

$$T_\ell \leq \frac{d(d+1)}{2} + \frac{8d}{\beta_\ell^2} C'\alpha_T + \frac{4d}{\beta_\ell}\sqrt{\frac{1}{\rho} C'\alpha_T}$$

Step 5: Upper-bounding regret under E.
Under E

$$\begin{aligned}&\sum_{\ell=1}^{\ell_\Delta \wedge \ell(T)} 4\beta_{\ell-1} T_\ell \\ &\leq \sum_{\ell=1}^{\ell_\Delta \wedge \ell(T)} 8\beta_\ell \left(\frac{d(d+1)}{2} + \frac{8d}{\beta_\ell^2} C'\alpha_T + \frac{4d}{\beta_\ell}\sqrt{\frac{1}{\rho} C'\alpha_T} \right) \\ &\leq 4d(d+1) + 64dC'\alpha_T \left(\sum_{\ell=1}^{\ell_\Delta} 2^\ell \right) + 32d\sqrt{\frac{1}{\rho} C'\alpha_T}\,\ell(T) \\ &\leq 4d(d+1) + 16dC'\alpha_T \left(\frac{16}{\Delta}\right) + 32d\sqrt{\frac{1}{\rho} C'\alpha_T}\,\ell(T) \\ &\leq 4d(d+1) + C_1 d\alpha_T \frac{1}{\Delta} + C_2 d\sqrt{\frac{1}{\rho}\alpha_T}\log(T)\end{aligned}$$

All in all, we have that

$$R_T \leq 4d(d+1) + C_2 d\sqrt{\frac{1}{\rho}\alpha_T}\log(T) + C_1 d\alpha_T \frac{1}{\Delta} + T\Delta$$

Step 6: Optimizing for Δ. We take

$$\Delta = \sqrt{\frac{C_1 d}{T}\alpha_T}.$$

We get an upper bound on R_T of

$$A\sqrt{dT\log\left(\frac{k\log(T)}{\delta}\right)} + Bd\sqrt{\frac{1}{\rho}\log\left(\frac{k\log(T)}{\delta}\right)}\log(T)$$

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

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

□

C. Adding noise at different steps of GOPE

In order to make the GOPE algorithm differentially private, the main task is to derive a private estimate of the linear parameter θ at each phase ℓ, i.e. $\hat{\theta}_\ell$. If the estimate is private with respect to the samples used to compute it, i.e. $\hat{\theta}_\ell = V_\ell^{-1}\left(\sum_{t=t_\ell}^{t_{\ell+1}-1} a_s r_s\right)$ w.r.t $\{r_s\}_{s=t_\ell}^{t_{\ell+1}-1}$, then due to forgetting and post-processing, the algorithm turns private too.

We discuss three different ways to make the empirical estimate $\hat{\theta}_\ell$ private.

1) Adding noise in the end: A first attempt would be to analyse the L_2 sensitivity of $\hat{\theta}_\ell$ directly, and adding Gaussian noise calibrated by the L_2 sensitivity of $\hat{\theta}_\ell$.

Let $\{r_s\}_{s=t_\ell}^{t_{\ell+1}-1}$ and $\{r'_s\}_{s=t_\ell}^{t_{\ell+1}-1}$ two neighbouring sequence of rewards that differ at only step $j \in [t_\ell, t_{\ell+1}-1]$. Then, we have that

$$\begin{aligned}\|\hat{\theta}_\ell - \hat{\theta'}_\ell\|_2 &= \|V_\ell^{-1}\left[a_j(r_s - r'_s)\right]\|_2 \\ &\leq 2\|V_\ell^{-1} a_j\|_2\end{aligned}$$

since $r_j, r'_j \in [-1, 1]$.

However, it is hard to control the quantity $\|V_\ell^{-1} a_j\|_2$ without additional assumptions. The G-optimal design permits only to control another related quantity, i.e. $\|a_j\|_{V_\ell^{-1}} = \|V_\ell^{-\frac{1}{2}} a_j\|_2$. Thus, it is better to add noise at a step before if one does not want to add further assumption.

2) Adding noise in the beginning: Since $\hat{\theta}_\ell = V_\ell^{-1}\left(\sum_{t=t_\ell}^{t_{\ell+1}-1} a_s r_s\right)$, another way to make $\hat{\theta}_\ell$ private is by adding noise directly to the sum of observed rewards.

Specifically, one can rewrite the sum

$$\sum_{t=t_\ell}^{t_{\ell+1}-1} a_s r_s = \sum_{a \in S_\ell} a \sum_{a_t = a, t \in [t_\ell, t_{\ell+1}-1]} r_t \,.$$

Since rewards are in $[-1, 1]$, the L_2 sensitivity of $\sum_{a_t=a, t\in[t_\ell, t_{\ell+1}-1]} r_t$ is 2.

Thus, by Theorem 18, this means that the noisy sum of rewards $\sum_{a_t=a, t\in[t_\ell, t_{\ell+1}-1]} r_t + \mathcal{N}\left(0, \frac{2}{\rho}\right)$ is ρ-zCDP. Hence, by post-processing lemma, the corresponding noisy estimate $\hat{\theta}_\ell + V_\ell^{-1}\left(\sum_{a\in S_\ell} a\mathcal{N}\left(0, \frac{2}{\rho}\right)\right)$ is a ρ-zCDP estimate of $\hat{\theta}_\ell$.

This is exactly how both [13] and [21] derive a private version of GOPE for different privacy definitions, i.e. pure ϵ-DP for [13] and (ϵ, δ)-DP for [21], respectively. The drawback of this approach is that the variance of the noise depends on the size of the support S_ℓ of the G-optimal design.

To deal with this, both [13] and [21] solve a variant of the G-optimal design to get a solution where $|S_\ell| \leq 4d \log\log d + 16$ rather than the full $d(d+1)/2$ support of AdaC-GOPE's optimal design. And still, the dependence on d in the private part of the regret achieved by both these algorithms are d^2 in [13, Eq (18)], and $d^{\frac{3}{2}}$ in [21, Eq (56)], respectively. Thus, both of these existing algorithms do not achieve to the linear dependence on d in the regret term due to privacy, as suggested by the minimax lower bound.

3) Adding noise at an intermediate level: In contrast, AdaC-GOPE adds noise to the statistic

$$\phi_\ell = V_\ell^{-\frac{1}{2}}\left(\sum_{t=t_\ell}^{t_{\ell+1}-1} a_s r_s\right).$$

ϕ_ℓ is an intermediate quantity between the sum of rewards $\sum_{t=t_\ell}^{t_{\ell+1}-1} a_s r_s$, and the parameter $\hat{\theta}_\ell$, whose L_2 sensitivity can be controlled directly using the G-optimal Design. Due to this subtle observation, the private estimation $\tilde{\theta}_\ell$ of AdaC-GOPE is independent of the size of the support S_ℓ. Hence, the regret term of AdaC-GOPE due to privacy enjoys a linear dependence on d, as suggested by the minimax lower bound.

4) Conclusion: In brief, to achieve the same DP guarantee with the same budget, one may arrive at it by adding noise at different steps, and the resulting algorithms may have different utilities. In general, adding noise at an intermediate level of computation (not directly to the input, i.e. local and not output perturbation) generally gives the best results.

Remark 9. *We also compare the empirical performance of* AdaC-GOPE *with a variant where the noise is added to the sum statistic i.e.* $\tilde{\theta}_\ell \triangleq \hat{\theta}_\ell + V_\ell^{-1}\left(\sum_{a\in S_\ell} a\mathcal{N}\left(0, \frac{2}{\rho}\right)\right)$. *The results are presented in Appendix H validating that* AdaC-GOPE *yields the lowest regret with respect to the other noise perturbation strategy.*

APPENDIX F
LINEAR CONTEXTUAL BANDITS WITH ZCDP

A. Confidence Bound for the Private Least Square Estimator

Theorem 14. *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

$$\beta_t = \mathcal{O}\left(\sqrt{d\log(t)}\right) \text{ and } \gamma_t = \mathcal{O}\left(\sqrt{\frac{1}{\rho}} d\log(t)\right)$$

and β_t and γ_t are increasing in t.

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

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

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

Which gives that

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

Step 2: Defining the Good Event E. We call E_1, E_2 and E_3 respectively the events

$$\left\{\forall t \in [T] : \|S_t\|_{V_t^{-1}} \leq \sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(V_t)}{\lambda^d}\right)}\right\},$$
$$\left\{\forall t \in [T] : \lambda_{\min}(G_t) \geq g(t, \lambda_0, \delta, d)\right\},$$
$$\left\{\forall t \in [T] : \|N_t\| \leq \sqrt{\frac{2\ell(t)}{\rho} f\left(d, \frac{\delta}{T}\right)}\right\}$$

where $G_t \triangleq \sum_{s=1}^{t} A_s A_s^T$, $g(t, \lambda_0, \delta, d) \triangleq \frac{\lambda_0 t}{4} - 8\log\left(\frac{t+3}{\delta/d}\right) - 2\sqrt{t\log\left(\frac{t+3}{\delta/d}\right)}$ and $f(d, \delta) \triangleq d + 2\sqrt{d\log\left(\frac{1}{\delta}\right)} + 2\log\left(\frac{1}{\delta}\right)$.

Let

$$E = E_1 \cap E_2 \cap E_3 \tag{20}$$

Step 3: Showing that E Happens with High Probability.

For event E_1:

By a direct application of Lemma 11, we get that

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

For event E_2:

By a direct application of Lemma 12, we get that

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

For event E_3:

Since $N_t \sim \mathcal{N}\left(0, \frac{2\ell(t)}{\rho} I_d\right)$, a direct application of Lemma 10 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, $V_t \geq (\lambda + \lambda_{\min}(G_t)) I_d \geq \lambda I_d$.

Which gives that, under E,

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

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 $\gamma_t = \mathcal{O}\left(\sqrt{\frac{1}{\rho}} d\log(t)\right)$

Step 6: Final Touch.

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 $\gamma_t = \mathcal{O}\left(\sqrt{\frac{1}{\rho}} d\log(t)\right)$ such that β_t and γ_t are increasing.

□

B. Regret Analysis

Theorem 5. *Under Assumptions 1 and 2, and for $\delta \in (0, 1]$, with probability at least $1 - \delta$, the regret R_T of* AdaC-OFUL *(Algorithm 3) is upper bounded by*

$$R_T \leq \mathcal{O}\left(d\log(T)\sqrt{T}\right) + \mathcal{O}\left(\sqrt{\frac{1}{\rho}} d^2 \log(T)^2\right)$$

Proof. Let E be the event defined in equation 20.

Step 1: Regret Decomposition.

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

We have that

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

Step 2: Upper-bounding Instantaneous Regret 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 3 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 13.

(d) By definition of τ_t and Line 6 of Algorithm 3 , 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

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

Step 3: Upper-bounding Regret under E.

Under E, we have that

$$\begin{aligned} R_T &= \sum_{t=1}^T r_t \\ &\leq 2\sqrt{1+C}\beta_T \sum_{t=1}^T \left(1 \wedge \|A_t\|_{V_{t-1}^{-1}}\right) \\ &\quad + 2\sqrt{1+C}\gamma_T \sum_{t=1}^T \frac{1}{\sqrt{\tau_t}} \left(1 \wedge \|A_t\|_{V_{t-1}^{-1}}\right) \\ &\leq 2\sqrt{1+C}\beta_T \sqrt{T \sum_{t=1}^T 1 \wedge \|A_t\|^2_{V_{t-1}^{-1}}} \\ &\quad + 2\sqrt{1+C}\gamma_T \sqrt{\left(\sum_{t=1}^T \frac{1}{\tau_t}\right)\left(\sum_{t=1}^T 1 \wedge \|A_t\|^2_{V_{t-1}^{-1}}\right)} \end{aligned} \tag{21}$$

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

Step 4: The Elliptical Potential Lemma.

We use that $1 \wedge x \leq \log(1+x)$ and $\det(V_t) = \det(V_{t-1})\left(1 + \|A_t\|^2_{G_{t-1}(\lambda)^{-1}}\right)$ to have that

$$\begin{aligned} \sum_{t=1}^T \left(1 \wedge \|A_t\|^2_{V_{t-1}^{-1}}\right) &\leq 2\sum_{t=1}^T \log\left(1 + \|A_t\|^2_{V_{t-1}^{-1}}\right) \\ &= 2\log\left(\frac{\det(V_T)}{\det(V_0)}\right) \\ &\leq 2d\log\left(1 + \frac{T}{\lambda d}\right) \end{aligned} \tag{22}$$

often known as the elliptical potential lemma (Lemma 19.4, [2].

Step 5: Upper-bounding the Length of Every Episode .

Episode ℓ starts at t_ℓ and ends at $t_{\ell+1} - 1$, so we have that

$$\frac{\det(V_{t_{\ell+1}-1})}{\det(V_{t_\ell})} \leq 1 + C \tag{23}$$

On the other hand,

$$\frac{\det(V_{t_{\ell+1}-1})}{\det(V_{t_\ell})} = \prod_{t=t_\ell+1}^{t_{\ell+1}-1} \left(1 + \|A_t\|^2_{V_{t-1}^{-1}}\right) \tag{24}$$

Under E, we use that

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

since $\lambda_{\max}(G_{t-1}) \leq \mathrm{trace}(G_{t-1}) \leq t-1$.

which gives that

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

Plugging in Equation 24, we get that

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

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

Finally using the upper bound of Equation 23, we get that

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

Which gives that

$$\begin{aligned} \sum_{t=1}^T \frac{1}{\tau_t} &= \sum_{\ell=1}^{\ell(T)} \sum_{t=t_\ell}^{t_{\ell+1}-1} \frac{1}{t_\ell} \\ &= \sum_{\ell=1}^{\ell(T)} \frac{t_{\ell+1} - t_\ell}{t_\ell} \leq (1+C)(1+\lambda)\ell(T) \end{aligned} \tag{25}$$

Step 6: Final Touch.

Plugging the upper bounds of Equation 22 and 25 in the regret upper bound of Equation 21, 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 finalise by using that

$$\beta_T = \mathcal{O}\left(\sqrt{d\log(T)}\right), \gamma_T = \mathcal{O}\left(\sqrt{\frac{1}{\rho}}d\log(T)\right) \quad \text{and} \quad \ell(T) = \mathcal{O}\left(d\log(T)\right)$$

We get that

$$R_T \leq \mathcal{O}\left(d\log(T)\sqrt{T}\right) + \mathcal{O}\left(\sqrt{\frac{1}{\rho}}d^2\log(T)^2\right)$$

□

C. Rectifying LinPriv Regret Analysis

[12] 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 [31], [38].

Theorem 5 in [12] 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, page 25, 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 [22] 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. (22)).

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 10. *In the proof of Theorem 5 [12], to bound the sum $\sum w_{i,t} \leq \mathcal{O}(\sqrt{\log T})\sum_{t=1}^{n}\|x_{i,t}\|_{V_{i,t}^{-1}}$, the correct bound has been used on the sum $\sum_{t=1}^{n}\|x_{i,t}\|_{V_{i,t}^{-1}}$ with the $\sqrt{T}$ appearing. However, it is misused for the private part.*

APPENDIX G
REGRET LOWER BOUNDS FOR ZCDP

In this section, we provide missing proofs from Section VI.

A. KL decomposition

We recall the definition of an f-divergence.

Definition 9 (f-divergence)**.** *Let $f:(0,\infty)\to\mathbb{R}$ be a convex function with $f(1)=0$. Let P and Q be two probability distributions on a measurable space $(\mathcal{X},\mathcal{F})$. If $P \ll Q$, i.e. P is absolutely continuous with respect to Q then the f-divergence is defined as*

$$D_f(P\|Q) \triangleq \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right]$$

where $\frac{dP}{dQ}$ is a Radon-Nikodym derivative and $f(0) \triangleq f(0+)$.

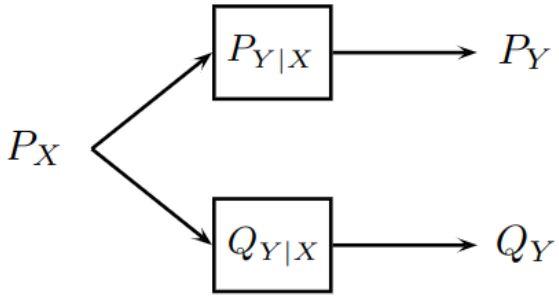
Let $\mathcal{P}_1$ and $\mathcal{P}_2$ two distributions over $\mathcal{X}^n$. Define $\mathcal{C}$ as a coupling of $(\mathcal{P}_1,\mathcal{P}_2)$, i.e. the marginals of $\mathcal{C}$ are $\mathcal{P}_1$ and $\mathcal{P}_2$. We denote by $\Pi(\mathcal{P}_1,\mathcal{P}_2)$ the set of all the couplings between $\mathcal{P}_1$ and $\mathcal{P}_2$. Let M_1 and M_2 be defined as in (Eq. 8). We recall the definition of an f-divergence.

Theorem 15. *We have that*

$$D_f(M_1\|M_2) \leq \inf_{\mathcal{C}\in\Pi(\mathcal{P}_1,\mathcal{P}_2)}\mathbb{E}_{(d,d')\sim\mathcal{C}}[D_f(\mathcal{M}_d\|\mathcal{M}_{d'})]. \quad (26)$$

Proof. Let $\mathcal{C}$ be a coupling of $\mathcal{P}_1$ and $\mathcal{P}_2$. We provide a visual proof of the theorem.

First, we recall Theorem 19.



If $P_X \overset{P_{Y|X}}{\longrightarrow} P_Y$ and $P_X \overset{Q_{Y|X}}{\longrightarrow} Q_Y$, then

$$D_f\left(P_Y\|Q_Y\right) \leq \mathbb{E}_{X\sim P_X}\left[D_f\left(P_{Y|X}\|Q_{Y|X}\right)\right].$$

The idea is to use Theorem 19, where the input is a pair of datasets (d,d') sampled from the coupling $\mathcal{C}$, the first channel applies the private mechanism to the first dataset, the second channel applies the mechanism to the second dataset. In other words,

- $X=(d,d')$ a pair of datasets in $\mathcal{X}^n$
- the input distribution is $P_X=\mathcal{C}$ the coupling distribution.
- the first channel is the mechanism applied to the first dataset $P_{Y|X}=\mathcal{M}(Y\mid d)$.
- the second channel is the mechanism applied to the second dataset $Q_{Y|X}=\mathcal{M}(Y\mid d')$.
- Y is the output of the mechanism

Using this notation, we have that

- $P_Y = M_1$
- $Q_Y = M_2$
- $D_f\left(P_{Y|X} \| Q_{Y|X}\right) = D_f(\mathcal{M}_d \| \mathcal{M}_{d'})$.

Using Theorem 19, we have that

$$D_f(M_1 \| M_2) \leq \mathbb{E}_{(d,d') \sim \mathcal{C}}[D_f(\mathcal{M}_d \| \mathcal{M}_{d'})].$$

which is true for every coupling $\mathcal{C}$. Taking the infimum over the couplings concludes the proof. □

We will use the group privacy property of ρ-zCDP to upper bound the RHS of Equation 26.

Theorem 16 (Group Privacy for ρ-zCDP, Proposition 27, [23])**.** *If $\mathcal{M}$ is ρ-CDP, then*

$$\forall d, d' \in \mathcal{X}^n,\ \forall \alpha \geq 1,\ D_\alpha(\mathcal{M}_d \| \mathcal{M}_{d'}) \leq \rho d_{Ham}(d, d')^2 \alpha.$$

Combining the last two theorems gives the proof of Theorem 6 as a corollary.

Theorem 6 (KL upper bound as a transport problem)**.** *If $\mathcal{M}$ is ρ-CDP, then*

$$\mathrm{KL}\left(M_1 \parallel M_2\right) \leq \rho \inf_{\mathcal{C} \in \Pi(\mathcal{P}_1, \mathcal{P}_2)} \mathbb{E}_{(d,d') \sim \mathcal{C}}[d_{Ham}(d, d')^2].$$

Proof. Let $\mathcal{M}$ be ρ-CDP. Applying Theorem 15, with $f(x) = x \log(x)$ gives that

$$\mathrm{KL}\left(M_1 \parallel M_2\right) \leq \rho \inf_{\mathcal{C} \in \Pi(\mathcal{P}_1, \mathcal{P}_2)} \mathbb{E}_{(d,d') \sim \mathcal{C}}[\mathrm{KL}\left(\mathcal{M}_d \parallel \mathcal{M}_{d'}\right)].$$

Applying Theorem 16 with $\alpha = 1$ gives that

$$\mathrm{KL}\left(\mathcal{M}_d \parallel \mathcal{M}_{d'}\right) \leq \rho d_{\mathrm{Ham}}(d, d')^2$$

Combining both inequalities gives the final bound. □

Using maximal coupling for data-generating distributions that are product distributions yields the proof of Theorem 7.

Theorem 7 (KL decomposition for ρ-zCDP)**.** *Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be two product distributions over $\mathcal{X}^n$, i.e. $\mathcal{P}_1 = \bigotimes_{i=1}^n p_{1,i}$ and $\mathcal{P}_2 = \bigotimes_{i=1}^n p_{2,i}$, where $p_{\nu,i}$ for $\nu \in \{1, 2\}, i \in [1, n]$ are distributions over $\mathcal{X}$. Let $t_i \triangleq \mathrm{TV}\left(p_{1,i} \parallel p_{2,i}\right)$. If $\mathcal{M}$ is ρ-zCDP, then*

$$\mathrm{KL}\left(M_1 \parallel M_2\right) \leq \rho \left(\sum_{i=1}^n t_i\right)^2 + \rho \sum_{i=1}^n t_i(1 - t_i)$$

Proof. Let c_∞^i be a maximal coupling between $p_{1,i}$ and $p_{2,i}$ for all $i \in [1, n]$. We define the coupling $\mathcal{C}_\infty \triangleq \bigotimes_{i=1}^n c_\infty^i$. Then $\mathcal{C}_\infty$ is a coupling of $\mathcal{P}_1$ and $\mathcal{P}_2$.

Since $d_{\mathrm{Ham}}(d, d') = \sum_{i=1}^n \mathbb{1}\{d_i \neq d_i'\}$ we get that, for $(d, d') \sim \mathcal{C}_\infty$,

$$d_{\mathrm{Ham}}(d, d') \sim \sum_{i=1}^n \mathrm{Bernoulli}(t_i),$$

where $t_i \triangleq \mathrm{TV}\left(p_{1,i} \parallel p_{2,i}\right)$.

This further yields

$$\mathbb{E}_{(d,d') \sim \mathcal{C}_\infty}[d_{\mathrm{Ham}}(d, d')^2] = \left(\sum_{i=1}^n t_i\right)^2 + \sum_{i=1}^n t_i(1 - t_i).$$

Corollary 6 concludes the proof. □

B. Lower bounds on regret for bandits

Theorem 8 (KL decompostion for ρ-Interactive zCDP)**.** *If π is ρ-Interactive zCDP, then*

$$\mathrm{KL}\left(m_{\nu\pi} \parallel m_{\nu'\pi}\right) \leq \rho \left[\mathbb{E}_{\nu\pi}\left(\sum_{t=1}^T t_{a_t}\right)\right]^2 + \rho\, \mathbb{E}_{\nu\pi}\left(\sum_{t=1}^T t_{a_t}(1 - t_{a_t})\right) + \rho\, \mathbb{V}_{\nu\pi}\left(\sum_{t=1}^T t_{a_t}\right)$$

where $t_{a_t} \triangleq \mathrm{TV}\left(P_{a_t} \parallel P'_{a_t}\right)$ and $\mathbb{E}_{\nu\pi}$ and $\mathbb{V}_{\nu\pi}$ are the expectation and variance under $m_{\nu\pi}$ respectively.

Proof. We adapt the proofs of the first section to the bandit case, by creating a coupled bandit instance.

Let $\nu = \{P_a : a \in [K]\}$ and $\nu' = \{P'_a : a \in [K]\}$ be two bandit instances. Define c_a as the maximal coupling between P_a and P'_a. Let $\pi = \{\pi_t\}_{t=1}^T$ be a ρ-Interactive zCDP policy.

Here, we build a coupled environment γ of ν and ν'. The policy π interacts with the coupled environment γ up to a given time horizon T to produce a history $\{(A_t, R_t, R'_t)\}_{t=1}^T$. The iterative steps of this interaction process are:

1. the probability of choosing an action $A_t = a$ at time t is dictated only by the policy π_t and $A_1, R_1, A_2, R_2, \ldots, A_{t-1}, R_{t-1}$, i.e. ignores $\{R'_s\}_{s=1}^{t-1}$.
2. the distribution of rewards (R_t, R'_t) is c_{A_t} and is conditionally independent of the previous observed history $\{(A_s, R_s, R'_s)\}_{t=1}^{t-1}$.

This interaction is similar to the interaction process of policy π with the first bandit instance ν, with the addition of sampling an extra R'_t from the coupling of P_{a_t} and P'_{a_t}.

The distribution of the history induced by the interaction of π and the coupled environment can be defined as

$$p_{\gamma\pi}(a_1, r_1, r'_1 \ldots, a_T, r_T, r'_T) \triangleq \prod_{t=1}^T \pi_t(a_t \mid a_1, r_1, \ldots, a_{t-1}, r_{t-1}) c_{a_t}(r_t, r'_t)$$

To simplify the notation, let $\mathbf{a} \triangleq (a_1, \ldots, a_T)$, $\mathbf{r} \triangleq (r_1, \ldots, r_T)$ and $\mathbf{r'} \triangleq (r'_1, \ldots, r'_T)$. Also, let $c_{\mathbf{a}}(\mathbf{r}, \mathbf{r'}) \triangleq \prod_{t=1}^T c_{a_t}(r_t, r'_t)$ and $\pi(\mathbf{a} \mid \mathbf{r}) \triangleq \prod_{t=1}^T \pi_t(a_t \mid a_1, r_1, \ldots, a_{t-1}, r_{t-1})$. We put $\mathbf{h} \triangleq (\mathbf{a}, \mathbf{r}, \mathbf{r'})$. With the new notation

$$p_{\gamma\pi}(\mathbf{a}, \mathbf{r}, \mathbf{r'}) \triangleq \pi(\mathbf{a} \mid \mathbf{r}) c_{\mathbf{a}}(\mathbf{r}, \mathbf{r'})$$

Similarly, we define

$$q_{\gamma\pi}(\mathbf{a}, \mathbf{r}, \mathbf{r'}) \triangleq \pi(\mathbf{a} \mid \mathbf{r'}) c_{\mathbf{a}}(\mathbf{r}, \mathbf{r'})$$

It follows that $m_{\nu,\pi}$ is the marginal of $p_{\gamma\pi}$ when integrated over $(\mathbf{r}, \mathbf{r'})$, and $m_{\nu',\pi}$ is the marginal of $q_{\gamma\pi}$ when integrated over $(\mathbf{r}, \mathbf{r'})$, i.e.

$$m_{\nu,\pi}(\mathbf{a}) = \int_{\mathbf{r},\mathbf{r'}} p_{\gamma\pi}(\mathbf{a}, \mathbf{r}, \mathbf{r'}) \, \mathrm{d}\mathbf{r} \, \mathrm{d}\mathbf{r'}$$

and

$$m_{\nu',\pi}(\mathbf{a}) = \int_{\mathbf{r},\mathbf{r'}} q_{\gamma\pi}(\mathbf{a}, \mathbf{r}, \mathbf{r'}) \, \mathrm{d}\mathbf{r} \, \mathrm{d}\mathbf{r'}.$$

By the data-processing inequality, we get that

$$\mathrm{KL}\left(m_{\nu,\pi} \parallel m_{\nu',\pi}\right) \leq \mathrm{KL}\left(p_{\gamma\pi} \parallel q_{\gamma\pi}\right) \tag{27}$$

In the following, upper case variables refer to random variables. We have that

$$\begin{aligned}
&\mathrm{KL}\left(p_{\gamma\pi} \parallel q_{\gamma\pi}\right) \\
&\overset{(a)}{=} \mathbb{E}_{\mathbf{H}\triangleq(\mathbf{A},\mathbf{R},\mathbf{R'})\sim p_{\gamma\pi}}\left[\log\left(\frac{\pi(\mathbf{A} \mid \mathbf{R})c_{\mathbf{A}}(\mathbf{R},\mathbf{R'})}{\pi(\mathbf{A} \mid \mathbf{R'})c_{\mathbf{A}}(\mathbf{R},\mathbf{R'})}\right)\right] \\
&\overset{(b)}{=} \sum_{t=1}^{T} \mathbb{E}_{\mathbf{H}\sim p_{\gamma\pi}}\left[\log\left(\frac{\pi_t(A_t \mid \mathcal{H}_{t-1})}{\pi_t(A_t \mid \mathcal{H}'_{t-1})}\right)\right] \\
&\overset{(c)}{=} \sum_{t=1}^{T} \mathbb{E}_{\mathbf{H}\sim p_{\gamma\pi}}\left[\mathbb{E}_{\mathbf{H}\sim p_{\gamma\pi}}\left[\log\left(\frac{\pi_t(A_t \mid \mathcal{H}_{t-1})}{\pi_t(A_t \mid \mathcal{H}'_{t-1})}\right) \mid \mathcal{H}_{t-1}\right]\right] \\
&\overset{(d)}{=} \sum_{t=1}^{T} \mathbb{E}_{\mathbf{H}\sim p_{\gamma\pi}}\left[\mathbb{E}_{A_t\sim\pi_t(.|\mathcal{H}_{t-1})}\left[\log\left(\frac{\pi_t(A_t \mid \mathcal{H}_{t-1})}{\pi_t(A_t \mid \mathcal{H}'_{t-1})}\right)\right]\right] \\
&\overset{(e)}{=} \sum_{t=1}^{T} \mathbb{E}_{\mathbf{H}\sim p_{\gamma\pi}}\left[\mathrm{KL}\left(\pi_t(. \mid \mathcal{H}_{t-1}) \parallel \pi_t(. \mid \mathcal{H}'_{t-1})\right)\right]
\end{aligned}$$

where $\mathcal{H}_t \triangleq (A_1, R_1, \ldots, A_t, R_t)$ and $\mathcal{H}'_t \triangleq (A_1, R'_1, \ldots, A_t, R'_t)$, where we obtain

(a): by definition of $p_{\gamma\pi}$, $q_{\gamma\pi}$ and the KL divergence

(b): by definition of $\pi(\mathbf{A} \mid \mathbf{R})$ and $\pi(\mathbf{A} \mid \mathbf{R'})$

(c): using the towering property of the expectation

(d): using that, conditioned on the history $\mathcal{H}_{t-1}$, the distribution of A_t is $\pi_t(. \mid \mathcal{H}_{t-1})$.

(e): by definition of the KL divergence

On the other hand, Theorem 1, we have that

$$\sum_{t=1}^{T} \mathrm{KL}\left(\pi_t(. \mid \mathcal{H}_{t-1}) \parallel \pi_t(. \mid \mathcal{H}'_{t-1})\right) \leq \rho d^2_{\mathrm{Ham}}(\mathbf{R}, \mathbf{R}')$$

which means that

$$\begin{aligned}
&\mathrm{KL}\left(p_{\gamma\pi} \parallel q_{\gamma\pi}\right) \\
&\leq \mathbb{E}_{\mathbf{H}\sim p_{\gamma\pi}}\left[\rho d^2_{\mathrm{Ham}}(\mathbf{R}, \mathbf{R}')\right] \\
&\overset{(a)}{=} \mathbb{E}_{\mathbf{H}\sim p_{\gamma\pi}}\left[\mathbb{E}_{\mathbf{H}\sim p_{\gamma\pi}}\left[\rho d^2_{\mathrm{Ham}}(\mathbf{R}, \mathbf{R}') \mid \mathbf{A}\right]\right] \\
&\overset{(b)}{=} \rho\, \mathbb{E}_{\mathbf{H}\sim p_{\gamma\pi}}\left[\mathbb{E}_{\mathbf{H}\sim p_{\gamma\pi}}\left[d_{\mathrm{Ham}}(\mathbf{R}, \mathbf{R}') \mid \mathbf{A}\right]^2\right. \\
&\qquad \left. + \rho\, \mathbb{V}\left[d_{\mathrm{Ham}}(\mathbf{R}, \mathbf{R}') \mid \mathbf{A}\right]\right] \\
&\overset{(c)}{=} \rho\, \mathbb{E}_{\nu,\pi}\left[\left(\sum_{t=1}^{T} t_{a_t}\right)^2\right] + \rho\, \mathbb{E}_{\nu,\pi}\left(\sum_{t=1}^{T} t_{a_t}(1 - t_{a_t})\right) \\
&\overset{(d)}{=} \rho\left[\mathbb{E}_{\nu,\pi}\left(\sum_{t=1}^{T} t_{a_t}\right)\right]^2 \\
&\qquad + \rho\, \mathbb{E}_{\nu,\pi}\left(\sum_{t=1}^{T} t_{a_t}(1 - t_{a_t})\right) + \rho\, \mathbb{V}_{\nu,\pi}\left[\sum_{t=1}^{T} t_{a_t}\right],
\end{aligned}$$

where we obtain

(a): using the towering property of the expectation

(b) and (d): by definition of the variance

(c): using that $d_{\mathrm{Ham}}(\mathbf{R}, \mathbf{R}') = \sum_{t=1}^{T} \mathbb{1}\{R_t \neq R'_t\}$ where $\mathbb{1}\{R_t \neq R'_t\} \mid A_t \sim \text{Bernoulli}(t_{a_t})$ by the definition of the maximal coupling and the sum is i.i.d given $\mathbf{A}$.

Finally, plugging the upper bound in Inequality (27) concludes the proof. □

Remark 11 (Sharper Bound than pure DP)**.** *Theorem 8 is ρ-zCDP version of the KL decomposition of Theorem 10 in [8]. The KL upper bound of Theorem 8 has a dependence on the total variation squared, emerging from the d^2_{Ham} of the group privacy. In contrast, the KL upper bound of Theorem 10 in [8] has a linear dependence in the total variation.*

1) Finite-armed bandits:

Theorem 17 (Minimax lower bounds for finite-armed bandits)**.** *Let Π^ρ be the set of ρ-zCDP policies. For any $K > 1$, $T \geq K - 1$, and $0 < \rho \leq 1$,*

$$\begin{aligned}
\mathrm{Reg}^{minimax}_{T,\rho} &\triangleq \inf_{\pi\in\Pi^\rho} \sup_{\nu\in\mathcal{E}^K} \mathrm{Reg}_T(\pi, \nu) \\
&\geq \max\left\{\frac{1}{27}\underbrace{\sqrt{T(K-1)}}_{\textit{without } \rho \textit{ zCDP}}, \quad \frac{1}{124}\underbrace{\sqrt{\frac{K-1}{\rho}}}_{\textit{with } \rho \textit{ zCDP}}\right\}.
\end{aligned}$$

Proof. The non-private part of the lower bound is due to Theorem 15.2 in [2]. To prove the private part of the lower bound, we plug our KL decomposition theorem into the proofs of regret lower bounds for bandits.

Step 1: Choosing the 'hard-to-distinguish' environments. First, we fix a ρ-zCDP policy π . Let Δ be a constant (to be specified later), and ν be a Gaussian bandit instance with unit variance and mean vector $\mu = (\Delta, 0, 0, ..., 0)$.

To choose the second bandit instance, let $a \triangleq \arg\min_{i\in[2,K]} \mathbb{E}_{\nu,\pi}[N_i(T)]$ be the least played arm in expectation other than the optimal arm 1. The second environment ν' is then chosen to be a Gaussian bandit instance with unit variance and mean vector $\mu' = (\Delta, 0, 0, \ldots 0, 2\Delta, 0 \ldots, 0)$, where $\mu'_j = \mu_j$ for every j except for $\mu'_a = 2\Delta$.

The first arm is optimal in ν and the arm i is optimal in ν'.

Since $T = \mathbb{E}_{\nu\pi}[N_1(T)] + \sum_{i>1} \mathbb{E}_{\nu\pi}[N_i(T)] \geq (K-1)\mathbb{E}_{\nu\pi}[N_a(T)]$, we observe that

$$n_a \triangleq \mathbb{E}_{\nu\pi}[N_a(T)] \leq \frac{T}{K-1}$$

Step 2: From lower bounding regret to upper bounding KL-divergence. Now by the classic regret decomposition and Markov inequality (Lemma 5), we get[4]

$$\begin{aligned}\mathrm{Reg}_T(\pi,\nu) &= \left(T - \mathbb{E}_{\nu\pi}\left[N_1(T)\right]\right)\Delta \\ &\geq \mathbb{M}_{\nu\pi}\left(N_1(T) \leq T/2\right)\frac{T\Delta}{2},\end{aligned}$$

and

$$\begin{aligned}\mathrm{Reg}_T(\pi,\nu') &= \Delta\mathbb{E}_{\nu'\pi}\left[N_1(T)\right] + \sum_{a\notin\{1,i\}} 2\Delta\mathbb{E}_{\nu'\pi}\left[N_a(T)\right] \\ &\geq \mathbb{M}_{\nu'\pi}\left(N_1(T) > T/2\right)\frac{T\Delta}{2}.\end{aligned}$$

Let us define the event $A \triangleq \{N_1(T) \leq T/2\} = \{(a_1, a_2, \ldots, a_T) : \mathrm{card}(\{j : a_j = 1\}) \leq T/2\}$.

By applying the Bretagnolle–Huber inequality, we have:

$$\begin{aligned}\mathrm{Reg}_T(\pi,\nu) + \mathrm{Reg}_T(\pi,\nu') &\geq \frac{T\Delta}{2}(M_{\nu\pi}(A) + M_{\nu'\pi}(A^c)) \\ &\geq \frac{T\Delta}{4}\exp(-\mathrm{KL}\left(M_{\nu\pi} \parallel M_{\nu'\pi}\right))\end{aligned}$$

Step 3: KL-divergence decomposition with ρ-Interactive zCDP. Since ν and ν' only differ in arm a, we get that $\sum t_{a_t} = t_a \sum \mathbb{1}\{a_t = a\}$, where $t_a \triangleq \mathrm{TV}\left(\nu_a \parallel \nu'_a\right)$.

Now, applying Theorem 8 gives

$$\begin{aligned}&\mathrm{KL}\left(M_{\nu\pi} \parallel M_{\nu'\pi}\right) \\ &\leq \rho(n_a^2 t_a^2 + n_a t_a(1-t_a) + t_a^2\,\mathbb{V}_{\nu\pi}(N_a(T))) \\ &\leq \rho(n_a^2 t_a^2 + n_a t_a + t_a^2\,\mathbb{V}_{\nu\pi}(N_a(T)))\,.\end{aligned}$$

where the last inequality is due to the fact that $1 - t_a \leq 1$.

On the other hand, we have the following upper bounds,

$$n_a \leq \frac{T}{K-1}$$

and

$$\mathbb{V}_{\nu\pi}(N_a(T)) \leq \mathbb{E}_{\nu\pi}\left[N_a(T)\right]\left(T - \mathbb{E}_{\nu\pi}\left[N_a(T)\right]\right) \leq \frac{T^2}{K-1}$$

and finally, using Pinsker's Inequality (Lemma 7)

$$t_a = \mathrm{TV}\left(\nu_a \parallel \nu'_a\right) \leq \sqrt{\frac{1}{2}\mathrm{KL}\left(\mathcal{N}(0,1) \parallel \mathcal{N}(2\Delta,1)\right)} = \Delta$$

Step 4: Choosing the worst Δ. Plugging back in the regret expression, we find

$$\begin{aligned}&\mathrm{Reg}_T(\pi,\nu) + \mathrm{Reg}_T(\pi,\nu') \\ &\geq \frac{T\Delta}{4}\exp\left(-\rho\left[\frac{T^2}{K-1}\left(1 + \frac{1}{K-1}\right)\Delta^2 + \frac{T}{K-1}\Delta\right]\right)\end{aligned}$$

Let $\alpha \triangleq \frac{T}{4}$, $\beta \triangleq \frac{\rho T^2}{K-1}(1 + \frac{1}{K-1})$ and $\gamma \triangleq \frac{\rho T}{K-1}$.

We have then

$$\begin{aligned}&\mathrm{Reg}_T(\pi,\nu) + \mathrm{Reg}_T(\pi,\nu') \\ &\geq \alpha\Delta\exp\left(-\beta\Delta^2 - \gamma\Delta\right)\end{aligned}$$

[4]In all regret lower bound proofs, we are under the probability space over the sequence of actions, produced when π interacts with ν for T time-steps. We do this to use the KL-divergence decomposition of $\mathbb{M}_{\nu\pi}$

$$\geq \alpha\Delta\exp\left(-\beta\left(\Delta + \frac{\gamma}{2\beta}\right)^2\right)$$

By optimising for Δ, we choose $\Delta = \frac{1}{\sqrt{\beta}} - \frac{\gamma}{2\beta}$. Putting back in Δ we have

$$\begin{aligned}\Delta &= \frac{1}{\sqrt{\beta}} - \frac{\gamma}{2\beta} \\ &= \sqrt{\frac{K-1}{\rho T^2\left(1 + \frac{1}{K-1}\right)}} - \frac{1}{2T\left(1 + \frac{1}{K-1}\right)} \\ &\geq \sqrt{\frac{K-1}{2\rho T^2}} - \frac{1}{2T} \\ &= \frac{\sqrt{K-1}}{T}\left(\frac{1}{\sqrt{2\rho}} - \frac{1}{2\sqrt{K-1}}\right) \\ &\geq \frac{\sqrt{K-1}}{T}\left(\frac{1}{\sqrt{2\rho}} - \frac{1}{2}\right) \\ &\geq \frac{\sqrt{K-1}}{T}\left(\frac{1}{4\sqrt{2\rho}}\right)\end{aligned}$$

where all the inequalities use that $K \geq 2$ and $\rho \leq 1$.

This gives that

$$\mathrm{Reg}_T(\pi,\nu) + \mathrm{Reg}_T(\pi,\nu') \geq \frac{\sqrt{K-1}}{4}\left(\frac{1}{4\sqrt{2\rho}}\right)\exp(-1)$$

We conclude the proof by using $\frac{1}{16\sqrt{2}}\exp(-1) \geq \frac{1}{62}$, and using $2\max(a,b) \geq a + b$. □

2) Linear bandits:

Theorem 9 (Minimax lower bounds for linear bandits)**.** *Let $\mathcal{A} = [-1,1]^d$ and $\Theta = \mathbb{R}^d$. Then, for any ρ-Interactive zCDP policy, we have that*

$$\mathrm{Reg}^{\textit{minimax}}_{T,\rho}(\mathcal{A},\Theta) \geq \max\left\{\underbrace{\frac{e^{-2}}{8}d\sqrt{T}}_{\textit{without } \rho\textit{-zCDP}},\ \underbrace{\frac{e^{-2.25}}{4}\frac{d}{\sqrt{\rho}}}_{\textit{with } \rho\textit{-zCDP}}\right\}$$

Proof. For the non-private lower bound, Theorem 24.1 of [2] gives that,

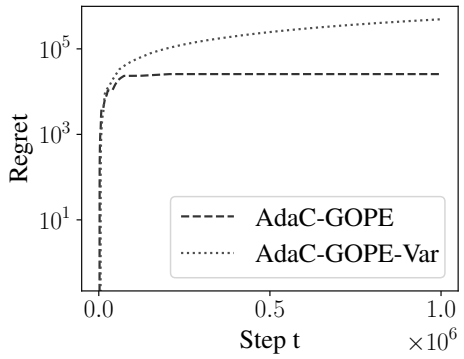
$$\mathrm{Reg}^{\text{minimax}}_T(\mathcal{A},\Theta) \geq \exp(-2)\frac{d}{8}\sqrt{T}.$$

Now, we focus on proving the ρ-zCDP part of the lower bound.

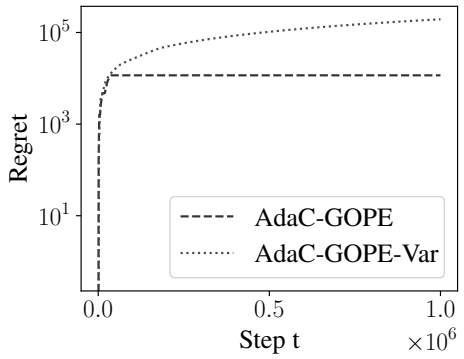
Let $\Theta = \left\{-\frac{1}{T\sqrt{\rho}}, \frac{1}{T\sqrt{\rho}}\right\}^d$. For $\theta, \theta' \in \Theta$, let ν and ν' be the bandit instances corresponding resp. to θ and θ'. We denote $\mathbb{M}_\theta = \mathbb{M}_{\nu,\pi}$ and $\mathbb{M}_{\theta'} = \mathbb{M}_{\nu',\pi}$. Let $\mathbb{E}_\theta$ and $\mathbb{E}_{\theta'}$ the expectations under $\mathbb{M}_\theta$ and $\mathbb{M}_{\theta'}$ respectively.

Step 1: From lower bounding regret to upper bounding KL-divergence. We begin with

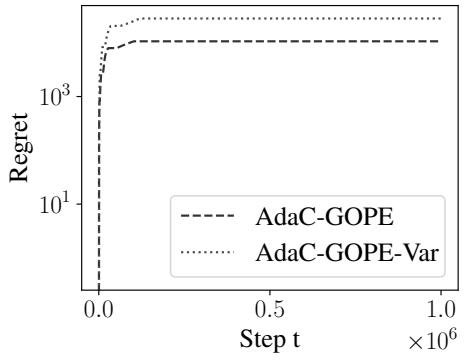
$$\begin{aligned}&\mathrm{Reg}_T(\mathcal{A},\theta) \\ &= \mathbb{E}_\theta\left[\sum_{t=1}^{T}\sum_{i=1}^{d}\left(\mathrm{sign}\left(\theta_i\right) - A_{ti}\right)\theta_i\right]\end{aligned}$$



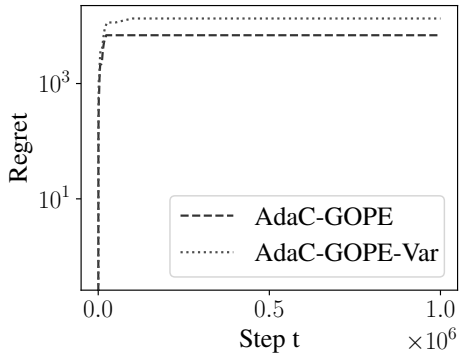
(a) $\rho = 0.001$



(b) $\rho = 0.01$



(c) $\rho = 0.1$



(d) $\rho = 1$

Fig. 5: Evolution of the regret over time for AdaC-GOPE and Adar-GOPE-Var for different values of the privacy budget ρ

$$\geq \frac{1}{T\sqrt{\rho}} \sum_{i=1}^{d} \mathbb{E}_\theta \left[\sum_{t=1}^{T} \mathbb{I}\left\{ \operatorname{sign}\left(A_{ti}\right) \neq \operatorname{sign}\left(\theta_i\right) \right\} \right]$$
$$\geq \frac{1}{\sqrt{\rho}} \sum_{i=1}^{d} \mathbb{M}_\theta \left(\sum_{t=1}^{T} \mathbb{I}\left\{ \operatorname{sign}\left(A_{ti}\right) \neq \operatorname{sign}\left(\theta_i\right) \right\} \geq T/2 \right)$$

In this derivation, the first equality holds because the optimal action satisfies $a_i^* = \operatorname{sign}(\theta_i)$ for $i \in [d]$. The first inequality follows from an observation that $(\operatorname{sign}(\theta_i) - A_{ti})\,\theta_i \geq |\theta_i|\,\mathbb{I}\{\operatorname{sign}(A_{ti}) \neq \operatorname{sign}(\theta_i)\}$. The last inequality is a direct application of Markov's inequality 5.

For $i \in [d]$ and $\theta \in \Theta$, we define

$$p_{\theta,i} \triangleq \mathbb{M}_\theta \left(\sum_{t=1}^{T} \mathbb{I}\left\{ \operatorname{sign}\left(A_{ti}\right) \neq \operatorname{sign}\left(\theta_i\right) \right\} \geq T/2 \right).$$

Now, let $i \in [d]$ and $\theta \in \Theta$ be fixed. Also, let $\theta'_j = \theta_j$ for $j \neq i$ and $\theta'_i = -\theta_i$. Then, by the Bretagnolle-Huber inequality,

$$p_{\theta,i} + p_{\theta',i} \geq \frac{1}{2} \exp\left(-\mathrm{KL}\left(\mathbb{M}_\theta \parallel \mathbb{M}_{\theta'}\right)\right).$$

Step 2: KL-divergence decomposition with ρ-Interactive zCDP.

Define $p_t \triangleq \mathrm{TV}\left(\mathcal{N}\left(\langle A_t, \theta\rangle, 1\right) \parallel \mathcal{N}\left(\langle A_t, \theta'\rangle, 1\right)\right)$.

From Lemma 8, we obtain that

$$\mathrm{KL}\left(\mathbb{M}_\theta \parallel \mathbb{M}_{\theta'}\right) \leq \rho \left(\mathbb{E}_{\nu\pi} \left[\sum_{t=1}^{T} p_t \right] \right)^2 + \rho \left(\mathbb{E}_{\nu\pi} \left[\sum_{t=1}^{T} p_t \right] \right) + \rho\, \mathbb{V}_{\nu\pi} \left[\sum_{t=1}^{T} p_t \right]$$

On the other hand, using Pinsker's inequality (Lemma 7), we have that

$$\begin{aligned}
\sum_{t=1}^{T} p_t &\leq \sum_{t=1}^{T} \sqrt{\frac{1}{2}\mathrm{KL}\left(\mathcal{N}\left(\langle A_t, \theta\rangle, 1\right) \parallel \mathcal{N}\left(\langle A_t, \theta'\rangle, 1\right)\right)} \\
&\leq \sum_{t=1}^{T} \sqrt{\frac{1}{4}\left[\langle A_t, \theta - \theta'\rangle^2\right]} \\
&\leq \frac{1}{2}\left[\sum_{t=1}^{T} \left|\langle A_t, \theta - \theta'\rangle\right|\right] \\
&\leq \frac{1}{2}\left[\sum_{t=1}^{T} \left|A_{t,i}\right| \left(2\left|\theta_i\right|\right)\right] \\
&\leq \frac{1}{2}\left[T \times 2\frac{1}{T\sqrt{\rho}}\right] = \frac{1}{\sqrt{\rho}}.
\end{aligned}$$

The last inequality holds true because $A_t \in [-1,1]^d$ and $\theta, \theta' \in \left\{-\frac{1}{T\sqrt{\rho}}, \frac{1}{T\sqrt{\rho}}\right\}^d$.

This gives that

$$\mathbb{E}_{\nu\pi}\left[\sum_{t=1}^{T} p_t\right] \leq \frac{1}{\sqrt{\rho}} \quad \text{and} \quad \mathbb{V}_{\nu\pi}\left[\sum_{t=1}^{T} p_t\right] \leq \frac{1}{4\rho}$$

Plugging back in the KL decomposition, we get that,

$$\begin{aligned}\mathrm{KL}\left(\mathbb{M}_\theta \parallel \mathbb{M}_{\theta'}\right) &\leq \rho\left(\frac{1}{\sqrt{\rho}}\right)^2 + \rho\left(\frac{1}{\sqrt{\rho}}\right) + \rho\left(\frac{1}{4\rho}\right) \\ &= 1 + \sqrt{\rho} + \frac{1}{4} \leq \frac{9}{4}\end{aligned}$$

where the last inequality is due to $\rho \leq 1$.

Step 3: Choosing the 'hard-to-distinguish' θ. Now, we have that

$$p_{\theta,i} + p_{\theta',i} \geq \frac{1}{2}\exp\left(-9/4\right)$$

Now, we apply an 'averaging hammer' over all $\theta \in \Theta$, such that $|\Theta| = 2^d$, to obtain

$$\sum_{\theta\in\Theta}\frac{1}{|\Theta|}\sum_{i=1}^{d} p_{\theta,i} = \frac{1}{|\Theta|}\sum_{i=1}^{d}\sum_{\theta\in\Theta} p_{\theta,i} \geq \frac{d}{4}\exp(-\frac{9}{4}).$$

This implies that there exists a $\theta \in \Theta$ such that $\sum_{i=1}^{d} p_{\theta,i} \geq d\exp(-\frac{9}{4})/4$.

Step 4: Plugging back θ in the regret decomposition. With this choice of θ, we conclude that

$$\begin{aligned}\mathrm{Reg}_T(\mathcal{A},\theta) &\geq \frac{1}{\sqrt{\rho}}\sum_{i=1}^{d} p_{\theta,i} \\ &\geq \frac{\exp(-\frac{9}{4})}{4}\frac{d}{\sqrt{\rho}}\end{aligned}$$

□

Remark 12 (Smaller High-Privacy Regime for ρ-zCDP)**.** *The minimax lower bound of Theorem 9 suggests that, for bandits with ρ-Interactive zCDP, as soon as the privacy budget $\rho = \Omega(T^{-1})$, it is possible to achieve privacy for free. In contrast, for ϵ-Pure DP, the minimax lower bounds of Theorem 5 in [8] show that it is possible to achieve privacy for free when $\epsilon = \Omega(T^{-1/2})$. We note that the ϵ-DP to $(\frac{1}{2}\epsilon^2)$-zCDP conversion provides a good intuition to justify this phenomenon.*

APPENDIX H
EXTENDED EXPERIMENTAL ANALYSIS

In this section, we add an experimental comparison between AdaC-GOPE and a variant of AdaC-GOPE where the way of making the estimate $\hat{\theta}_\ell$ private is different (Section E-C). In AdaR-GOPE-Var, Step 4 changes to

$$\tilde{\theta}_\ell^{\text{AdaR-GOPE-Var}} = \hat{\theta}_\ell + V_\ell^{-1}\left(\sum_{a\in S_\ell} a\mathcal{N}\left(0,\frac{2}{\rho}\right)\right).$$

We compare AdaC-GOPE and AdaR-GOPE-Var in the same experimental setup and instances as in Section VII, for different privacy budgets ρ and report the results in Figure 5.

As suggested by the regret analysis, AdaC-GOPE *achieves less regret, especially in the high privacy regime where the private part of the regret has more impact.*

APPENDIX I
EXISTING TECHNICAL RESULTS AND DEFINITIONS

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

Lemma 3 (Post-processing Lemma (Proposition 2.1, [15]))**.** *If $\mathcal{M}$ is a mechanism and f is an arbitrary randomised mapping defined on $\mathcal{M}$'s output, then*

- *If $\mathcal{M}$ is (ϵ,δ)-DP, then $f\circ\mathcal{M}$ is (ϵ,δ)-DP.*
- *If $\mathcal{M}$ is ρ-zCDP, then $f\circ\mathcal{M}$ is ρ-zCDP.*

Theorem 18 (The Gaussian Mechanism ([15], [43], [23]))**.** *Let $f:\mathcal{X}\to\mathbb{R}^d$ be a mechanism with L_2 sensitivity $s(f) \triangleq \max_{d\sim d'}\|f(d)-f(d')\|_2$. Let $g \triangleq f + Z$, such that $Z \sim \mathcal{N}(0, b\times s(f)^2 I_d)$. Here, $\mathcal{N}(\mu,\Sigma)$ denotes the Gaussian distribution with mean μ and co-variance matrix Σ, and $\|\cdot\|_2$ denotes the L_2 norm on $\mathbb{R}^d$. Then, for $b = \frac{2}{\epsilon^2}\log(\frac{1.25}{\delta})$, $\frac{\alpha}{2\epsilon}$, $\frac{1}{2\rho}$, g satisfies (ϵ,δ)-DP, (α,ϵ)-RDP and ρ-zCDP respectively.*

Lemma 4 (Post-processing property of Renyi Divergence, Lemma 2.2 [23])**.** *Let P and Q be distributions on Ω and let $f:\Omega\to\Theta$ be a function. Let $f(P)$ and $f(Q)$ denote the distributions on Θ induced by applying f to P and Q respectively. Then $\mathrm{D}_\alpha(f(P)\|f(Q)) \leq \mathrm{D}_\alpha(P\|Q)$.*

Lemma 5 (Markov's Inequality)**.** *For any random variable X and $\varepsilon > 0$,*

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

Definition 10 (Consistent Policies)**.** *A policy π is called consistent over a class of bandits $\mathcal{E}$ if for all $\nu\in\mathcal{E}$ and $p>0$, it holds that*

$$\lim_{T\to\infty}\frac{\mathrm{Reg}_T(\pi,\nu)}{T^p} = 0.$$

The class of consistent policies over $\mathcal{E}$ is denoted by $\Pi_{cons}(\mathcal{E})$.

Lemma 6 (Bretagnolle-Huber inequality)**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on the same measurable space $(\Omega,\mathcal{F})$, and let $A\in\mathcal{F}$ be an arbitrary event. Then,*

$$\mathbb{P}(A) + \mathbb{Q}(A^c) \geq \frac{1}{2}\exp(-\mathrm{D}(\mathbb{P},\mathbb{Q})),$$

where $A^c = \Omega\backslash A$ is the complement of A.

Lemma 7 (Pinsker's Inequality)**.** *For two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on the same probability space $(\Omega,\mathcal{F})$, we have*

$$\mathrm{KL}\left(\mathbb{P}\parallel\mathbb{Q}\right) \geq 2(\mathrm{TV}\left(\mathbb{P}\parallel\mathbb{Q}\right))^2.$$

Definition 11 (Sub-Gaussianity)**.** *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 8 (Concentration of Sub-Gaussian random variables)**.** *If X is σ-sub-Gaussian, then for any $\epsilon \geq 0$,*

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

Lemma 9 (Properties of Sub-Gaussian Random Variables)**.** *Suppose that X_1 and X_2 are independent and σ_1 and σ_2-sub-Gaussian, respectively, then*

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

Lemma 10 (Concentration of the χ^2-Distribution, Claim 17 of [11])**.** *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 11 (Theorem 20.4 of [2])**.** *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 12 (Lemma 2, Equation (6) of [34])**.** *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

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

Lemma 13 (Lemma 12 in [22])**.** *Let A, B and C be positive semi-definite matrices such that $A = B + C$. Then, we have that*

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

Theorem 19 (Conditioning Increases f-divergence)**.** *Let $P_X \xrightarrow{P_{Y|X}} P_Y$ and $P_X \xrightarrow{Q_{Y|X}} Q_Y$. Then,*

$$D_f\left(P_Y \| Q_Y\right) \leq \mathbb{E}_{X \sim P_X}\left[D_f\left(P_{Y|X} \| Q_{Y|X}\right)\right].$$