Interactive and Concentrated Differential Privacy for Bandits

Achraf Azize and Debabrota Basu
Équipe Scool, Univ. Lille, Inria,
CNRS, Centrale Lille, UMR 9189- CRIStAL
F-59000 Lille, France
{achraf.azize,debabrota.basu}@inria.fr

Abstract

Bandits play a crucial role in interactive learning schemes and modern recommender systems. However, these systems often rely on sensitive user data, making privacy a critical concern. This paper investigates privacy in bandits with a trusted centralized decision-maker through the lens of *interactive* Differential Privacy (DP). While bandits under pure ϵ-global DP have been well-studied, we contribute to the understanding of bandits under zero Concentrated DP (zCDP). We provide minimax and problem-dependent lower bounds on regret for finite-armed and linear bandits, which quantify the cost of ρ-global zCDP in these settings. These lower bounds reveal two hardness regimes based on the privacy budget ρ and suggest that ρ-global zCDP incurs less regret than pure ϵ-global DP. We propose two ρ-global zCDP bandit algorithms, AdaC-UCB and AdaC-GOPE, for finite-armed and linear bandits respectively. Both algorithms use a common recipe of Gaussian mechanism and adaptive episodes. We analyze the regret of these algorithms to show that AdaC-UCB achieves the problem-dependent regret lower bound up to multiplicative constants, while AdaC-GOPE achieves the minimax regret lower bound up to poly-logarithmic factors. Finally, we provide experimental validation of our theoretical results under different settings.

1 Introduction

Multi-armed bandit (in brief, *bandits*) (Lattimore and Szepesvári, 2020) is the archetypal setting of reinforcement learning consisting of K actions corresponding to K unknown reward distributions $\{\nu_a\}_{a\in[K]}$. We call $\{\nu_a\}_{a\in[K]} \triangleq \boldsymbol{\nu}$ 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 are increasingly used in a wide range of sequential decision-making tasks under uncertainty, such as recommender systems (Silva et al., 2022), strategic pricing (Bergemann and Välimäki, 1996), clinical trials (Thompson, 1933) 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) c_t 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

16th European Workshop on Reinforcement Learning (EWRL 2023).

Algorithm 1 Sequential interaction between a policy and users

Input: A policy $\pi = \{\pi_t\}_{t=1}^T$ and Users $\{u_t\}_{t=1}^T$ represented by the table $\mathbf{d} \triangleq \{\mathbf{x}_1, \dots, \mathbf{x}_T\} \in (\mathbb{R}^K)^T$
Output: A sequence of actions $a_1, \dots, a_T$
for $t = 1, \dots, T$ **do**
    π recommends action $a_t \sim \pi_t(. \mid a_1, r_1, \dots, a_{t-1}, r_{t-1})$
    u_t sends the **sensitive** reward $r_t \triangleq \mathbf{x}_{t,a_t}$ to π
end for

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.*

Differential privacy for bandits. Motivated by such data-sensitive scenarios, privacy issues are widely studied for bandits for different settings, such as stochastic bandits (Mishra and Thakurta, 2015; Tossou and Dimitrakakis, 2016; Sajed and Sheffet, 2019; Azize and Basu, 2022; Hu and Hegde, 2022), adversarial bandits (Tossou and Dimitrakakis, 2017), and linear contextual bandits (Shariff and Sheffet, 2018; Neel and Roth, 2018; Hanna et al., 2022). All these works adhere to Differential Privacy (DP) Dwork et al. (2014) 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 Basu et al. (2019). 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 existing works on global DP preserving bandits consider pure ϵ-DP and assume that the action sequence is published non-interactively in one-shot. In this paper, *we extend the study of privacy in bandits to the settings*, where *an adversary interacts with a policy at each step* (Vadhan and Wang, 2021), and *the algorithm aims to achieve* popular relaxations of pure DP, e.g. *zero Concentrated DP (zCDP)* (Dwork and Rothblum, 2016).

Interactive DP. A bandit algorithm induces an interactive process (Algorithm 1). At each step of this interaction, an adversary can manipulate the arm suggested by the algorithm and return it a reward from another arm. This situation is invoked in non-compliant bandits, where the user deploys an arm other than the recommended one, and in poisoning attacks, where a manipulated version of reward is sent to the policy either to leak information or to destroy its performance. This motivates us to define *Interactive DP for bandits*. A bandit policy π satisfying Interactive DP *protects all possible outcomes corresponding to a user by making the view of the adversary indistinguishable when interacting with the policy on neighbouring reward datasets*. Our effort resonates with the recent works in DP (Vadhan and Wang, 2021; Vadhan and Zhang, 2022; Lyu, 2022), where an analyst interacts with an offline dataset through an adaptive sequence of queries. The goal is to preserve privacy while responding to these adaptive queries. Our work extends the study of Interactive DP to the online setting, where a bandit algorithm generates its data by sequentially interacting with the environment (Section 2).

Relaxations of pure DP for bandits. Pure DP is widely studied for different settings of bandits. 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 (Azize and Basu, 2022). This still leaves open the question that what will be the minimal cost of preserving relaxations of pure DP in bandits as stated in (Shariff and Sheffet, 2018; Azize and Basu, 2022). Additionally, pure DP is often achieved by using Laplace noise to perturb the statistics computed on history. While in practice, Gaussian noise is widely used to perturb statistics computed on the dataset that leads to preserving relaxations of pure DP, namely (ϵ, δ)-DP (Dwork et al., 2014), Rényi DP (RDP) (Mironov, 2017), and zero Concentrated DP (zCDP) (Dwork and Rothblum, 2016), but not pure DP. *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 (Azize and Basu, 2022). Since zCDP scales well under group privacy, we adopt zCDP as the relaxation of pure DP. In this work, we investigate zCDP in two settings of bandits: *stochastic bandits with finitely many arms*, and *stochastic linear bandits with (fixed) finitely many arms*. To our knowledge, we are the first to study the complexity of zCDP for bandits with global DP.

The central questions that we aim to address are:

1. *What is the minimal cost to pay in terms of regret to achieve ρ-global zCDP for bandits?*
2. *How to design bandit algorithms that can achieve these regret lower bounds order-optimally?*

Table 1: The complexity of bandits with ρ-global zCDP. Each lower bound is the maximum of the classical non-private bound and the corresponding bound in the third column.

Setting	Type	Regret Lower Bound due to ρ zCDP [1]	Regret Upper Bound
Finite-armed	Minimax	$\rho^{-1/2}K$ (Thm 2, a)	$\mathcal{O}\left(\sqrt{KT\log(T)}\right) + \mathcal{O}\left(\rho^{-1/2}K\sqrt{\log(T)}\right)$ (Thm 5, a)
	Problem Dependent [2]	$\rho^{-1/2}\sum_{a:\Delta_a>0}\left(\Delta_a t_a^{-1}\right)\sqrt{\log(T)}$ (Thm 2, b)	$\mathcal{O}\left(\sum_a \frac{\log T}{\Delta_a}\right) + \mathcal{O}\left(\rho^{-1/2}K\sqrt{\log(T)}\right)$ (Thm 5, b)
Linear	Minimax	$\rho^{-1/2}d$ (Thm 3)	$\mathcal{O}\left(\sqrt{dT\log(KT)}\right) + \mathcal{O}\left(\rho^{-1/2}d\log^{\frac{3}{2}}(KT)\right)$ (Thm 7)

Our contributions. Answering these questions leads us to:

1. *Hardness as regret lower bounds:* First, addressing the open problem of (Shariff and Sheffet, 2018; Azize and Basu, 2022), we prove minimax and problem-dependent lower bounds for finite-armed bandits, and minimax lower bound for linear bandits with ρ-global zCDP that quantify the cost to ensure ρ-global zCDP in these settings (Section 3). 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. In the problem-dependent analysis, the additional regret due to ρ-global zCDP in finite-armed bandits appears as a lower order term, i.e. $\Omega\left(\sqrt{(\log T)/\rho}\right)$, with respect to the non-private lower bound $\Omega(\log T)$. In contrast, the regret due to ϵ-global DP, $\Omega\left((\log T)/\epsilon\right)$, is not a lower order term.

2. *Order-optimal algorithm design:* We propose two algorithms, AdaC-UCB and AdaC-GOPE, that preserves ρ-global zCDP for finite-armed and linear bandits, respectively (Section 4). Both algorithms share the same algorithmic blueprint. First, they add a calibrated *Gaussian noise* to reward statistics. Second, they run in *adaptive episodes*, with the number of episodes being logarithmic in T. We analyse the regret of both algorithms and show that they match the lower bounds up to multiplicative factors. AdaC-UCB achieves the problem-dependent lower up to multiplicative constants, while both AdaC-UCB and AdaC-GOPE match the corresponding minimax lower bounds up to poly-logarithmic factors. We summarise all the lower and upper bounds in Table 1. In Section 5, we numerically validate their performances in different settings.

3. *Technical tools:* We propose a novel technique to generate lower bounds for bandits with ρ-zCDP using coupling arguments. We adapt this technique to the sequential bandit setup and use it to derive regret lower bounds with a generic proof. We also discuss in depth the effect of partial information (bandit feedback) on the definition of DP for bandits (Appendix A). We also prove a lower bound on the cost of reward poisoning against an Interactive DP bandit algorithm (Theorem 8). This opens up a direction to bridge privacy defences and attacks for bandits.

2 Bandits with Interactive DP: The formulation

First, we formalise Interactive DP for bandits with a centralised decision-maker. We adopt the Interactive DP definition as studied in (Vadhan and Zhang, 2022; Lyu, 2022), where a mechanism $\mathcal{M}$ is viewed as a party in an interactive protocol, interacting with a possibly adversarial analyst or users.

We represent each user u_t by the vector $\mathbf{x}_t \triangleq (x_{t,1}, \dots, x_{t,K}) \in \mathbb{R}^K$, where $x_{t,a}$ represents the **potential** reward observed if action a was recommended to user u_t. Due to the bandit feedback, if at step t action $q_t \in [K]$ was queried from the environment, only the reward $r_t = x_{t,q_t}$ is observed at step t. Thus, the set of users $\{u_t\}_{t=1}^T$ is represented by **the table of potential rewards** $\mathbf{d} \triangleq \{\mathbf{x}_1, \dots, \mathbf{x}_T\} \in (\mathbb{R}^K)^T$. We view the policy π as an interactive mechanism, taking as input the table of potential rewards $\mathbf{d}$ while interacting with a possibly adversarial analyst B. The interactive protocol is described in Definition 1.

Definition 1 (The bandit-adversary interactive protocol)**.** *A policy $\pi = \{\pi_t\}_{t=1}^T$, with its input $\boldsymbol{d}$ the table of potential rewards, and the adversary $B = \{B_t\}_{t=1}^T$ follow the interactive process:*

For $t = 1, \dots, T$:
1. The bandit algorithm selects an action $o_t \sim \pi(\boldsymbol{d}; q_1, q_2, \dots, q_{t-1})$,
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] Here, we only express the private part of the lower bound

[2] For Bernoulli bandits, $t_a = \Delta_a$ and the lower bound reduces to $K\rho^{-1/2}\sqrt{\log(T)}$.

Here, we denote $\pi(\mathbf{d}; q_1, q_2, \ldots, q_{t-1}) \triangleq \pi_t(\cdot \mid q_1, x_{1,q_1}, \ldots, q_{t-1}, x_{t-1,q_{t-1}})$. At each step t, the policy π recommends an action o_t. The adversary B observes the recommended action o_t, and chooses (adversarially) a query action q_t, based on the history of recommended actions $(o_s)_{s=1}^{t-1}$. At step $t+1$, the policy recommends the next action o_{t+1} based only on its private input $\mathbf{d}$ containing all the sensitive rewards information about the users, and the adversarially chosen query actions $(q_s)_{s=1}^{t}$.

Following the Interactive DP framework, the policy π is a differentially private interactive mechanism if the view of adversary B, i.e. $\mathrm{View}(B \leftrightarrow \pi(\mathbf{d})) \triangleq (o_1, \ldots, o_T)$, is indistinguishable when the interaction is run on two neighbouring tables of rewards $\mathbf{d}$ and $\mathbf{d}'$. They represent two sets of users differing by only one individual, i.e. one row. Formally, it implies that the Hamming distance between the two tables $\mathbf{d}$ and $\mathbf{d}'$ is one, $d_{\mathrm{Ham}}(\mathbf{d}, \mathbf{d}') \triangleq \sum_{t=1}^{T} \mathbb{1}\{x_t \neq x'_t\}$. We denote them by $\mathbf{d} \sim \mathbf{d}'$.

Definition 2 (Interactive DP policy)**.** *A policy $\pi = \{\pi_t\}_{t=1}^{T}$ is said to be*

a. ***Interactive*** (ϵ, δ)***-DP policy*** *for a given $\delta \in [0,1)$ if, for every pair of neighboring* ***table of potential rewards*** *datasets $\boldsymbol{d}, \boldsymbol{d}' \in \mathcal{X}$, every adversary $B \in \mathcal{B}$, and every subset of possible views $\mathcal{S} \subseteq [K]^T$,*

$$\Pr[\mathrm{View}(B \leftrightarrow \pi(\boldsymbol{d})) \in \mathcal{S}] \leq \exp(\epsilon) \cdot \Pr[\mathrm{View}(B \leftrightarrow \pi(\boldsymbol{d}')) \in \mathcal{S}] + \delta.$$

b. ***Interactive*** (α, ϵ)***-RDP policy*** *for an $\alpha > 1$ if, for every adversary $B \in \mathcal{B}$,*

$$\sup_{\boldsymbol{d} \sim \boldsymbol{d}'} D_\alpha(\mathrm{View}(B \leftrightarrow \pi(\boldsymbol{d})) \| \mathrm{View}(B \leftrightarrow \pi(\boldsymbol{d}'))) \leq \epsilon.$$

c. ***Interactive*** (ξ, ρ)***-zCDP policy*** *if, for every $\alpha \in (1, \infty)$, and every adversary $B \in \mathcal{B}$,*

$$\sup_{\boldsymbol{d} \sim \boldsymbol{d}'} D_\alpha(\mathrm{View}(B \leftrightarrow \pi(\boldsymbol{d})) \| \mathrm{View}(B \leftrightarrow \pi(\boldsymbol{d}'))) \leq \xi + \rho\alpha. \tag{1}$$

Here, $D_\alpha(P \| Q) \triangleq \frac{1}{\alpha - 1} \log \mathbb{E}_Q \left[\left(\frac{\mathrm{d}P}{\mathrm{d}Q} \right)^{\alpha} \right]$ denotes the Rényi divergence of order α between P and Q. *We define ϵ-pure global DP as $(\epsilon, 0)$-DP and ρ-global zCDP to be $(0, \rho)$-zCDP.*

Implications of ensuring Interactive DP for bandits: We elaborate on three interesting implications of the interactive definition of privacy in bandits, compared to the non-interactive definition adopted in the literature. We recall the non-interactive definition in detail in Appendix A.

1. *Interactive adversarial hypothesis testing:* Interactive DP (Def. 2) defends against an online adversary, who can manipulate the actions recommended by the policy. The adversary participates in the interaction between the algorithm and environment and can run an interactive (or sequential) hypothesis testing (Wald, 1992) to distinguish between two neighbouring datasets based on its view of the interaction. In contrast, the adversary in the non-interactive definition only observes the sequence of actions $\mathbf{a} = (a_t)_{t=1}^{T}$ that the policy π recommends without any interference. Based on the sequence $\mathbf{a}$, the adversary runs a one-shot hypothesis testing to distinguish between two neighbouring datasets (Kairouz et al., 2015). Hence, Interactive DP allows us to defend against a stronger and more realistic adversary.

2. *Protecting non-compliant users:* Interactive DP (Def. 2) protects the privacy of the users even if they are non-compliant (Kallus, 2018; Stirn and Jebara, 2018), i.e. the users decide to ignore the recommendations of the policy and choose a different arm. Specifically, the policy recommends an action o_t at step t, but the adversary could choose another query action q_t, different than o_t. The reward $r_t = \mathbf{d}_{t,q_t}$ and the decision of the policy o_{t+1} in the next step depend on the query action q_t and not o_t. In contrast, the non-interactive definition only protects the privacy of compliant users.

3. *Defending against online poisoning attacks:* In the Interactive DP definition, the policy recommends action o_t at step t and expects to receive the corresponding reward $\mathbf{d}_{t,o_t}$. However, an adversary may intentionally query a different action q_t, resulting in the observed reward $\mathbf{d}_{t,q_t}$. This can be viewed as poisoning the reward from $\mathbf{d}_{t,o_t}$ to $\mathbf{d}_{t,q_t}$ (Liu and Shroff, 2019). The interactive definition inherently provides robustness against online reward poisoning. In Appendix A.4, we show that if a policy π is consistent and ρ-global zCDP, an **online oracle attacker has to incur** $\Omega(\sqrt{(\log T)/\rho})$ **cost** to make the policy choose a non-optimal target arm linearly. Thus, for a smaller privacy budget ρ, i.e. high privacy, an attacker has to poison further more to succeed.

Remark: Handling bandit feedback. Ensuring privacy in bandit settings requires careful consideration of how the bandit feedback, or partial information, is handled. When a policy selects arms a_t, it only observes the reward $r_t = \mathbf{d}_{t,a_t}$ associated with arm a_t, while the other rewards $\mathbf{d}_{t,a}$ for $a \neq a_t$ remain unobserved. A fundamental question in defining privacy for bandits is whether the private input dataset of the policy π should be considered as the set of observed rewards $\{r_t\}_{t=1}^{T}$ (referred to as *View DP*), or the entire table of potential rewards $\mathbf{d}$ (referred to as *Table DP*). In

App. A, we compare both definitions and demonstrate that Table DP is a stronger notion, as being Table DP implies being View DP. This intuition stems from the fact that Table DP protects users u_t by safeguarding all their potential responses. We also establish that, for pure DP, the two notions are equivalent. However, for approximate DP and its variations, transitioning from View DP to Table DP causes a significant loss in the privacy budget due to group privacy considerations. For a detailed comparison and discussion, we refer to App. A. In Definition 2, we adopt the Table DP framework.

Goal: Regret minimisation. Hereafter, we adopt ρ-global zCDP (Eq. (1)) as the privacy definition. The goal is to design a ρ-global zCDP policy that minimises regret. To define regret, we adhere to the classic interaction of bandits as explained in Algorithm 1. Specifically, the adversary is no longer part of the interaction. The policy π directly interacts with the set of users $\{u_1, \ldots, u_T\}$ as in Algorithm 1, and the goal is to maximize the expected cumulative reward, or equivalently minimize the expected regret. We study two settings: finite-armed stochastic bandits and stochastic linear bandits. Now, we formally define regrets for them.

Finite-armed stochastic bandits. The environment $\nu \triangleq (\nu_a : a \in [K])$ consists of K arms (or reward distributions) with finite means $(\mu_a)_{a\in[K]}$. For any horizon T, regret is defined as

$$\mathrm{Reg}_T(\pi, \nu) \triangleq T\mu^\star - \mathbb{E}\left[\sum_{t=1}^{T} r_t\right] = \sum_{a=1}^{K} \Delta_a \mathbb{E}\left[N_a(T)\right], \tag{2}$$

where $\mu^\star \triangleq \max_{a\in[K]} \mu_a$ 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 in the interaction of Algorithm 1. The expectation is taken both on the randomness of the environment ν and the policy π, using the canonical bandit model (Chapter 4.6 of (Lattimore and Szepesvári, 2020)).

Stochastic linear bandits. 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 $(\mathcal{A}, \theta^\star)$.

Remark. There are two interaction protocols: The bandit-adversary interactive protocol of Definition 1 and the sequential interaction between a policy and users of Algorithm 1. The bandit-adversary interactive protocol is used to analyse the privacy of the policy. Specifically, we want to design a policy for which the view of an adversary is "similar" when only one user changes in the interaction of Definition 1. On the other hand, to analyse the accuracy of the policy, we adhere to the "classic" sequential interaction between a policy and users. In this interaction (Algorithm 1), the policy recommends at each time-step an action a_t and observes the reward r_t corresponding to the user u_t in the table $\mathbf{d}$, i.e. $r_t = \mathbf{x}_{t,a_t}$. There is no adversary in this interaction, and the goal of the policy is to maximize the expected cumulative reward or equivalently minimize the expected regret, *when interacting with users, without the presence of the adversary*. In brief, we want to design a policy that verifies the "adversarial" privacy constraint and minimizes the classic "expected" regret.

3 Lower bounds on regret of bandits with ρ-global zCDP

In this section, we quantify the cost of ρ-global zCDP for bandits by providing regret lower bounds for any ρ-global 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-global 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 1, while the proof details are deferred to Appendix C.

KL decomposition lemma. To proceed with the lower bounds, first, we are interested to control the Kullback-Leibler (KL) divergence between marginal distributions induced by a ρ-zCDP mechanism

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

$$M_\nu(A) \triangleq \int_{d \in \mathcal{X}^n} \mathcal{M}(A \mid d)\, d\mathcal{P}_\nu(d), \tag{4}$$

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

Theorem 1 (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 $t_i \triangleq \mathrm{TV}(p_{1,i} \parallel p_{2,i})$. If $\mathcal{M}$ is ρ-zCDP, then*

$$\mathrm{KL}(M_1 \parallel M_2) \leq \rho \left(\sum_{i=1}^n t_i \right)^2 + \rho \sum_{i=1}^n t_i(1 - t_i) \tag{5}$$

This is a centralised ρ-zCDP version of the KL-decomposition lemma under local DP (Duchi et al., 2013, Theorem 1), and a ρ-zCDP version of the Sequential Karwa-Vadhan lemma (Azize and Basu, 2022). In Appendix B, we elaborate on the new proof technique, which can be of parallel interest.

Leveraging this decomposition, now, we derive two flavours of regret lower bounds, namely *minimax* and *problem-dependent*. The minimax lower bound expresses the best regret achievable by a policy on the corresponding worst-case environment. The problem-dependent lower bound controls the regret of a 'reasonable' (consistent) policy for a specific environment that the policy interacts with.

Lower bounds on regret for finite-armed bandits

Theorem 2 (Minimax and problem-dependent lower bounds for finite-armed bandits)**.**
(a) Minimax. *Let Π^ρ be the set of ρ-zCDP policies. For any $K > 1$, $T \geq K - 1$, and $0 < \rho \leq 1$,*

$$\mathrm{Reg}_{T,\rho}^{minimax} \triangleq \inf_{\pi \in \Pi^\rho} \sup_{\nu \in \mathcal{E}^K} \mathrm{Reg}_T(\pi, \nu) \geq \max \left\{ \frac{1}{27} \underbrace{\sqrt{T(K-1)}}_{\text{without } \rho\text{-global zCDP}}, \frac{1}{44} \underbrace{\frac{K-1}{\sqrt{\rho}}}_{\text{with } \rho\text{-global zCDP}} \right\}.$$

(b) Problem-dependent. *Let $\mathcal{E} = \mathcal{M}_1 \times \cdots \times \mathcal{M}_K$ be a class of environments with K arms, where $\mathcal{M}_a$ is a set of reward distributions with finite means. Let π be a consistent policy[3] over $\mathcal{E}$ satisfying ρ-global zCDP . Then, for all $\nu = (P_a)_{a=1}^K \in \mathcal{E}$, (i.e. $P_a \in \mathcal{M}_a$), it holds that*

$$\liminf_{T \to \infty} \frac{\mathrm{Reg}_T(\pi, \nu)}{\sqrt{\log(T)}} \geq \sum_{a : \Delta_a > 0} \frac{\Delta_a}{\sqrt{\rho}\, t_{\inf}(P_a, \mu^*, \mathcal{M}_a)}.$$

where $t_{\inf}(P, \mu^*, \mathcal{M}) \triangleq \inf_{P' \in \mathcal{M}} \{\mathrm{TV}(P \parallel P') : \mu(P') > \mu^*\}$

Comments on the minimax bound. The minimax regret lower bound suggests the existence of two hardness regimes depending on ρ and T. When $\rho < (27/44)(K-1)/T$, i.e. the **high-privacy regime**, the lower bound becomes $\Omega\left(K/\sqrt{\rho}\right)$, and ρ-global zCDP bandits incur more regret than non-private ones. When $\rho > (27/44)(K-1)/T$, i.e. in the **low-privacy regime**, the lower bound retrieves the non-private lower bound, i.e. $\Omega(\sqrt{KT})$, and thus, we can achieve privacy for free.

Comments on the problem-dependent bound. The problem-dependent lower bound shows that the **price of privacy is a lower order term** $\Omega\left(\sqrt{\log(T)/\rho}\right)$. For a fixed privacy budget ρ and asymptotically in T, this is negligible compared to the non-private problem-dependent regret lower bound of $\Omega\left(\sum_a \log(T)/\Delta_a\right)$. In contrast, for pure ϵ-global DP, the price of privacy in the problem-dependent regret is $\Omega\left(\log(T)/\epsilon\right)$, which is not a second-order term (Azize and Basu, 2022). Thus, in a problem-dependent perspective, privacy is 'free' only for ρ-global zCDP, but not for ϵ-global DP.

Lower bound on regret for linear bandits

Theorem 3 (Minimax Lower Bounds for Linear Bandits)**.** *Let $\mathcal{A} = [-1, 1]^d$ and $\Theta = \mathbb{R}^d$. Then, for any ρ-global zCDP policy, we have that*

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

[3]A policy π is called *consistent* over a class of environments $\mathcal{E}$, if $\forall \nu \in \mathcal{E}$ and $p > 0$, $\lim_{T \to \infty} \frac{R_T(\pi, \nu)}{T^p} = 0$.

Algorithm 2 AdaC-UCB. *Changes due to privacy are in blue.*

1: **Input:** Privacy budget ρ, an environment ν with K arms, optimism parameter $\beta > 3$
2: **Output:** Actions satisfying ρ-global 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 $a_\ell = \text{argmax}_a \, \mathrm{I}_a^\rho(t_\ell - 1, \beta)$ (Eq. (6))
7: Choose arm a_ℓ until round t such that $N_{a_\ell}(t) = 2N_{a_\ell}(t_\ell - 1)$
8: **end for**

Two privacy regimes. Similar to the finite-arm case, the minimax regret lower bound for linear bandits suggests the existence of two hardness regimes for $\rho \geq \frac{4\exp(-0.5)}{T}$ and $\rho \leq \frac{4\exp(-0.5)}{T}$.

A generic proof technique. In order to prove the lower bounds, we deploy the KL upper bound of Theorem 1 in the classic proof scheme of regret lower bounds (Lattimore and Szepesvári, 2020). 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 is sub-optimal in other. The KL upper bound of Theorem 1 allows us to quantify the extra-hardness to statistically distinguish environments due to the additional 'blurriness' created by the ρ-zCDP constraint.

4 Algorithm design: AdaC-UCB and AdaC-GOPE

In this section, we propose AdaC-UCB and AdaC-GOPE, two algorithms that satisfy ρ-global zCDP for finite-armed and linear bandits respectively. The two algorithms share a similar blueprint: **the Gaussian mechanism** and **adaptive episodes**. For each setting, we present the algorithm, provide a privacy and a regret analysis, and compare the regret upper bounds to the regret lower bounds.

4.1 Stochastic finite-armed bandits

Now, we study the setting of finite-armed bandits under ρ-zCDP as detailed in Section 2.

Algorithm. AdaC-UCB is an extension of the generic algorithmic wrapper proposed by Azize and Basu (2022) for bandits with ρ-global zCDP. Following (Azize and Basu, 2022), 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. (6), 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. (6) only needs to be ρ-zCDP for the algorithm to be ρ-global zCDP, avoiding the need of composition theorems. We formalise this intuition in Lemma 7 of Appendix D.

For AdaC-UCB, we use the private index to select the arms (Line 6 of Algorithm 2) as

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

Here, $\hat{\mu}_a^\ell$ is the empirical mean of rewards collected in the last episode in which arm a was played, $\sigma_{a,\ell}^2 \triangleq \frac{1}{2\rho \times \left(\frac{1}{2}N_a(t_\ell - 1)\right)^2}$ is the variance of the Gaussian noise. Finally, the exploration bonus is defined as $B_a(t_\ell - 1, \beta) \triangleq \sqrt{\left(\frac{1}{2 \times \frac{1}{2}N_a(t_\ell - 1)} + \frac{1}{\rho \times \left(\frac{1}{2}N_a(t_\ell - 1)\right)^2}\right)\beta \log(t_\ell)}$. The term in blue rectifies the non-private confidence bound of UCB for the added Gaussian noise.

Theorem 4 (Privacy of AdaC-UCB)**.** *For rewards in* $[0, 1]$, AdaC-UCB *satisfies ρ-global zCDP.*

Proof sketch. The main idea is that a change in one user *only affects* the empirical mean calculated in one episode, which is made private using the Gaussian Mechanism and Lemma 7. Since the actions are computed only using the private empirical means, AdaC-UCB is ρ-global zCDP thanks to the post-processing lemma. We refer to Appendix D for the complete proof.

Theorem 5 (Regret analysis of AdaC-UCB)**.** *For rewards in* $[0, 1]$ *and* $\beta > 3$, AdaC-UCB *yields*

(a) a problem-dependent regret upper bound $\sum_{a:\Delta_a > 0}\left(\frac{8\beta}{\Delta_a}\log(T) + 8\sqrt{\frac{\beta}{\rho}}\sqrt{\log(T)} + \frac{2\beta}{\beta - 3}\right)$*, and*

Algorithm 3 AdaC-GOPE. *Changes due to privacy are in blue.*

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

$$\max_{\substack{\pi \in \mathcal{P}(\mathcal{A}_\ell) \\ |\mathrm{Supp}(\pi)| \leq d(d+1)/2}} \log\det V(\pi). \tag{7}$$

  Step 2: $\mathcal{S}_\ell \leftarrow \mathrm{Supp}\,(\pi_\ell)$
  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 (8).
  Observe rewards $\{r_t\}_{t=t_\ell}^{t_\ell + \sum_a T_\ell(a)}$
  $T_\ell \leftarrow \sum_{a \in \mathcal{S}_\ell} T_\ell(a)$ and $t_{\ell+1} \leftarrow t_\ell + T_\ell + 1$
  Step 3: Estimate the parameter as $\hat{\theta}_\ell = V_\ell^{-1} \sum_{t=t_\ell}^{t_{\ell+1}-1} a_t r_t$ with $V_\ell = \sum_{a \in \mathcal{S}_\ell} T_\ell(a) a a^\top$
  Step 4: Make the parameter estimate private $\tilde{\theta}_\ell = \hat{\theta}_\ell + V_\ell^{-\frac{1}{2}} N_\ell$, where $N_\ell \sim \mathcal{N}\left(0, \frac{2d}{\rho c_\ell} I_d\right)$.
  Step 4: Eliminate low rewarding arms: $\mathcal{A}_{\ell+1} = \left\{ a \in \mathcal{A}_\ell : \max_{b \in \mathcal{A}_\ell} \left\langle \tilde{\theta}_\ell, b - a \right\rangle \leq 2\beta_\ell \right\}$.
end for

(b) a minimax regret upper bound $\mathcal{O}\left(\sqrt{KT\log(T)}\right) + \mathcal{O}\left(K\rho^{-1/2}\sqrt{\log(T)}\right)$.

Order-optimality of AdaC-UCB. The problem-dependent regret upper bound of AdaC-UCB matches the problem-dependent regret lower bound of Theorem 2 up to multiplicative constants for Bernoulli Bandits. On the other hand, the minimax regret upper bound of AdaC-UCB achieves the minimax regret lower bound of Theorem 2 up to an extra $\sqrt{\log T}$ term, which is usually the extra cost to pay in minimax regret for the UCB algorithm.

Discussion on related bounds. Under a distributed setting and for (α, ϵ)-RDP, Chowdhury and Zhou (2022) propose a variant of Successive Elimination (SE) with Skellam noise, which achieves a $\mathcal{O}(K\sqrt{\log(T)}/\epsilon)$ private regret. However, for non-private bandits, optimism-based strategies achieve optimality and have better performance than SE. This is shown by Azize and Basu (2022) while comparing their adaptive mechanism, AdaP-UCB, with DP-SE in the case of ϵ-pure DP. Similar reasoning follows here. Second, Skellam Noise is less practical to sample from than Gaussian Noise.

4.2 Stochastic Linear Bandits

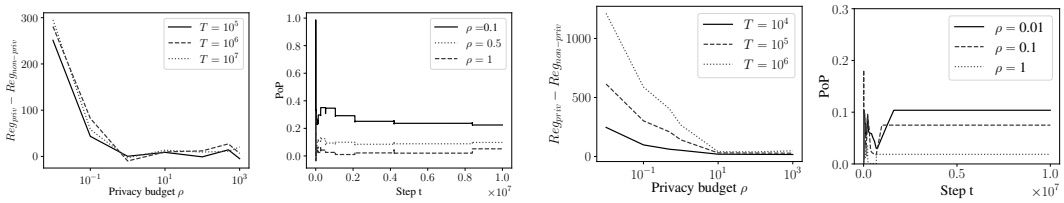
Here, we study ρ-global zCDP for stochastic linear bandits with a finite number of arms, as in Sec. 2.

Algorithm. We propose AdaC-GOPE (Algorithm 3), which is a ρ-global zCDP extension of the G-Optimal design-based Phased Elimination (GOPE) algorithm (Lattimore and Szepesvári, 2020, 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 (7)) that helps to exploit the structure of arms and to minimise the number of samples needed to eliminate a suboptimal arm.

In particular, if π_ℓ is the G-optimal solution for $\mathcal{A}_\ell$ at phase ℓ, then each action $a \in \mathcal{A}_\ell$ is played $T_\ell(a) \triangleq c_\ell \pi_\ell(a)$ times, where for $\delta' \triangleq \frac{\delta}{K\ell(\ell+1)}$,

$$c_\ell \triangleq \frac{8d}{\beta_\ell^2} \log\left(\frac{4}{\delta'}\right) + \frac{2d}{\beta_\ell}\sqrt{\frac{2}{\rho}}\left(d + 2\sqrt{d\log\left(\frac{2}{\delta'}\right)} + 2\log\left(\frac{2}{\delta'}\right)\right)^{1/2}. \tag{8}$$

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 F.2), and (b) avoiding privacy composition theorems and using, therefore, Lemma 7 to make the algorithm private. Note that AdaC-GOPE can be seen as a generalisation of DP-SE (Sajed and Sheffet, 2019) to the linear bandit setting.



(a) Finite-armed Bandits (b) Linear Bandits

Figure 1: For each bandit setting, the left figure represents the evolution of the difference between the private and non-private regret with respect to the privacy budget ρ. The right figure represents the evolution of the price of privacy (PoP) with respect to the time step.

Assumption 1 (Boundedness)**.** *We assume that: (1) actions are bounded: $\|a\|_2 \leq 1$ for all a in $\mathcal{A}$, (2) rewards are bounded: $|r_t| \leq 1$, and (3) the unknown parameter is bounded: $\|\theta^\star\|_2 \leq 1$.*

Theorem 6 (Privacy of AdaC-GOPE)**.** *Under Assumption 1,* AdaC-GOPE *satisfies ρ-global zCDP.*

Proof sketch. Similar to Theorem 4, a change in one user only affects the estimate $\hat{\theta}_\ell$ in one episode. Thanks to Lemma 7, it is enough that each $\hat{\theta}_\ell$ is ρ-zCDP with respect to the sequence of rewards collected in the corresponding episode. Since the actions only depend on the estimates $\{\hat{\theta}_\ell\}_\ell$, the algorithm is ρ-global zCDP by the post-processing lemma. We refer to Appendix D for the proof.

Theorem 7 (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 *(Algorithm 3) 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 $\mathbb{E}(R_T) \leq \mathcal{O}\left(\sqrt{dT\log(KT)}\right) + \mathcal{O}\left(\frac{d}{\sqrt{\rho}}(\log(KT))^{\frac{3}{2}}\right)$.*

Order-optimality of AdaC-GOPE. The minimax regret upper bound of AdaC-GOPE matches with the minimax regret lower bound of Theorem 3 up to an extra $(\log KT)^{\frac{3}{2}}$ factor.

Related algorithms and bounds. Hanna et al. (2022) and Li et al. (2022) study private variants of the GOPE algorithm for pure ϵ-global DP and (ϵ, δ)-DP, respectively. However, both algorithms differ in how they privatize the estimated parameter $\hat{\theta}$ compared to AdaC-GOPE. They add noise to each sum of rewards $\sum_{t=t_\ell}^{t_{\ell+1}-1} r_t$ (Line 11, Alg. 3), whereas we add noise in $\hat{\theta}_l$ (Line 12, Alg. 3). As a result, though we achieve linear dependence on the dimension d as suggested by the lower bound, others do not (d^2 for (Hanna et al., 2022) and $d^{3/2}$ for (Li et al., 2022)). In Appendix F, we analyse in detail the impact of adding noise at different steps of GOPE, both theoretically and experimentally.

5 Experimental analysis

Now, we empirically verify whether AdaC-UCB and AdaC-GOPE can achieve privacy for free.

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 the algorithms for Bernoulli bandits with 5-arms and means $\{0.75, 0.625, 0.5, 0.375, 0.25\}$ (as in (Sajed and Sheffet, 2019)). For linear bandits, 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 (Lattimore and Szepesvári, 2020). 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 both settings, we run the private and non-private algorithms 100 times for a horizon $T = 10^7$, and compare the average regret between the private and non-private algorithms in Figure 1.

Results and analysis. We reach two conclusions from the results of both 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. $\text{PoP} \triangleq \frac{Reg_{priv} - Reg_{non-priv}}{Reg_{non-priv}}$ 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.

6 Conclusion and future works

We study bandits with interactive ρ-global zCDP. First, we demonstrate the benefits of adopting the Interactive DP definition for bandits. Then, we prove the minimax and problem-dependent regret lower bounds for finite-armed and linear bandits, showing that the additional regret due to ρ-global zCDP is less compared to pure ϵ-global DP. The minimax bound additionally shows the existence of two hardness regimes and privacy can be achieved for free in the low-privacy regime. We propose AdaC-UCB and AdaC-GOPE, which satisfy ρ-global zCDP using a generic algorithmic blueprint, and match the regret lower bounds up to constants and poly-logarithmic factors respectively.

A possible 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 (Bun et al., 2014; Kamath et al., 2022). For the algorithm design, it would be also interesting to see how to close the multiplicative gaps.

Acknowledgments and Disclosure of Funding

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

Azize, A. and Basu, D. (2022). When privacy meets partial information: A refined analysis of differentially private bandits. *Advances in Neural Information Processing Systems*, 35:32199–32210.

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

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

Bun, M. and Steinke, T. (2016). Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography*, pages 635–658, Berlin, Heidelberg. Springer Berlin Heidelberg.

Bun, M., Ullman, J., and Vadhan, S. (2014). Fingerprinting codes and the price of approximate differential privacy. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 1–10.

Chowdhury, S. R. and Zhou, X. (2022). Distributed differential privacy in multi-armed bandits. *arXiv preprint arXiv:2206.05772*.

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

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

Dwork, C. and Rothblum, G. N. (2016). Concentrated differential privacy. *ArXiv*, abs/1603.01887.

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

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

Jun, K.-S., Li, L., Ma, Y., and Zhu, J. (2018). Adversarial attacks on stochastic bandits. *Advances in Neural Information Processing Systems*, 31.

Kairouz, P., Oh, S., and Viswanath, P. (2015). The composition theorem for differential privacy. In *International conference on machine learning*, pages 1376–1385. PMLR.

Kallus, N. (2018). Instrument-armed bandits. In *Algorithmic Learning Theory*, pages 529–546. PMLR.

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

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

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Li, F., Zhou, X., and Ji, B. (2022). 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)*, pages 41–48. IEEE.

Liu, F. and Shroff, N. (2019). Data poisoning attacks on stochastic bandits. In *International Conference on Machine Learning*, pages 4042–4050. PMLR.

Lyu, X. (2022). Composition theorems for interactive differential privacy. In *Advances in Neural Information Processing Systems*.

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

Mishra, N. and Thakurta, A. (2015). (Nearly) optimal differentially private stochastic multi-arm bandits. In *UAI*.

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

Sajed, T. and Sheffet, O. (2019). An optimal private stochastic-MAB algorithm based on an optimal private stopping rule.

Shariff, R. and Sheffet, O. (2018). Differentially private contextual linear bandits. In *Advances in Neural Information Processing Systems*, pages 4296–4306.

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

Stirn, A. and Jebara, T. (2018). Thompson sampling for noncompliant bandits. *arXiv preprint arXiv:1812.00856*.

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

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

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

Vadhan, S. and Wang, T. (2021). 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*, pages 582–604. Springer.

Vadhan, S. and Zhang, W. (2022). Concurrent composition theorems for all standard variants of differential privacy. *arXiv preprint arXiv:2207.08335*.

Wald, A. (1992). *Sequential tests of statistical hypotheses*. Springer.

Appendix

Table of Contents

A Privacy definitions for bandits **13**
- A.1 Non-interactive DP for bandits . . . 13
- A.2 Interactive DP for bandits . . . 17
- A.3 Consequences of the Interactive DP definition . . . 17
- A.4 Poisoning attacks against Interactive DP . . . 19

B Lower bounds via couplings for concentrated DP **20**
- B.1 From the KL to a transport problem . . . 20
- B.2 Proxy solution to the transport Problem . . . 21

C Regret lower bounds for bandits under ρ-global zCDP **23**
- C.1 Stochastic finite-armed bandits: Minimax lower bound . . . 23
- C.2 Stochastic finite-armed bandits: Problem-dependent lower bound . . . 24
- C.3 Stochastic linear bandits: Minimax lower bound . . . 25

D Privacy proofs **30**
- D.1 The privacy lemma of non-overlapping sequences . . . 30
- D.2 Generic privacy proof of AdaC-UCB and AdaC-GOPE . . . 31

E Stochastic bandits with global zCDP **35**
- E.1 Concentration inequalities . . . 35
- E.2 Regret analysis . . . 35
- E.3 Extensions to (ϵ, δ)-global DP and (α, ϵ)-global RDP . . . 37

F Linear Bandits with global zCDP **38**
- F.1 Basic definitions of optimal design . . . 38
- F.2 Concentration inequalities . . . 38
- F.3 Regret analysis . . . 39
- F.4 Extensions to (ϵ, δ)-global DP and (α, ϵ)-global RDP . . . 42
- F.5 Adding noise at different steps of AdaC-GOPE . . . 43

G Existing technical results and definitions **45**

H Extended experimental analysis **46**

A Privacy definitions for bandits

In this section, we discuss different ways to adopt Differential Privacy (DP) in bandits. The main ingredients to specify in order to have a complete definition are (1) the **mechanism** in question, (2) the **input dataset**, (3) the **neighbouring** relationship between the input datasets, and (4) the **output** of the mechanism.

For all the adaptations of DP for bandits studied in this section, the **output** of the mechanism is the same, i.e. a sequence of actions in $[K]^T$. The **mechanism** in question is always induced by the policy. For completeness, we recall the definition of the policy in Definition 3.

The main differences in the adaptations of DP originate from two sources:

1. Considering the **policy** as an *interactive* or *non-interactive* mechanism.
2. Considering the **input** of the mechanism to be the sequence of **observed rewards**, i.e. $r = \{r_1, \dots, r_T\} \in \mathbb{R}^T$, that we call **View DP**. Alternatively, considering the **input** of the mechanism to be the full table of **potential rewards**, i.e. $d = \{d_1, \dots, d_T\} \in (\mathbb{R}^K)^T$, that we call **Table DP**.

Before starting, we recall the definition of the policy. Let $T \in \mathbb{N}$ be the horizon. For each $t \in [T]$, let $\Omega_t = ([K] \times \mathbb{R})^t$ and $\mathcal{F}_t = \mathfrak{B}(\Omega_t)$ with $\mathfrak{B}$ being the Borel set.

Definition 3. *A policy π is a sequence of rules $(\pi_t)_{t=1}^T$, where each π_t is a probability kernel from the histories $(\Omega_{t-1}, \mathcal{F}_{t-1})$ to arms $([K], 2^{[K]})$. Since $[K]$ is discrete, we adopt the convention that for $i \in [K]$,*

$$\pi_t(i \mid a_1, r_1, \dots, a_{t-1}, r_{t-1}) = \pi_t(\{i\} \mid a_1, r_1, \dots, a_{t-1}, r_{t-1})$$

In the following, we first consider the non-interactive definition of privacy in bandits, which is the definition usually adopted in the private bandit literature. There, we mainly discuss the relationship between View and Table DP for different variants of DP. Then, we formalize the interactive definition of privacy in bandits. We state several consequences of the interactive definition. Also, we discuss its relation to the non-interactive definition, as well as to poisoning attacks.

A.1 Non-interactive DP for bandits

In this section, we define View and Table DP. Then, we discuss their relations. We prove that Table DP always implies View DP, with the same privacy parameters. However, the converse may not be always true. The equivalence can only be shown for pure ϵ-DP. For other variants of DP, there could be a huge increase in the privacy budget.

To commence, we recall the defintions of variants of DP for a non-interactive mechansim $\mathcal{M}$.

Definition 4 (Variants of Approximate DP (ADP) for non-interactive mechanisms)**.** *A non-interactive mechanism $\mathcal{M}$, that assigns to each dataset d a probability distribution $\mathcal{M}_d$ on some measurable space $(\mathbb{X}, \mathcal{F})$, satisfies*

1. (ϵ, δ)-DP (Dwork et al. (2014)) for a given $\delta \in [0, 1)$ if

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

2. (α, ϵ)-Rényi DP (RDP) (Mironov (2017)) for an $\alpha > 1$ if

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

3. (ξ, ρ)-zero Concentrated DP (zCDP) (Bun and Steinke (2016)) if, for all $\alpha \in (1, \infty)$,

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

Here, two datasets d and d' are said to be neighbouring (denoted by $d \sim d'$) if their Hamming distance is one. $D_\alpha(P \| Q) \triangleq \frac{1}{\alpha - 1} \log \mathbb{E}_Q \left[\left(\frac{dP}{dQ} \right)^\alpha \right]$ denotes the Rényi divergence of order α between P and Q. We define ϵ-pure DP as $(\epsilon, 0)$-DP and ρ-zCDP to be $(0, \rho)$-zCDP.

A.1.1 View DP

This is the definition usually adopted in the literature of private bandits (Mishra and Thakurta, 2015; Tossou and Dimitrakakis, 2016; Sajed and Sheffet, 2019; Azize and Basu, 2022). We formalise it by stating its main ingredients, and coin the term "View DP".

Input. The input considered is only the sequence of observed rewards $r = (r_1, \dots, r_T)$ and the neighbouring is a change in one reward in this sequence.

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

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

The mechanism $\mathcal{V}^\pi$, when applied to a sequence of observed rewards r, outputs (in one shot) a sequence of actions $a^T \triangleq (a_1, \dots, a_T) \in [K]^T$, with probability $\mathcal{V}^\pi_r(a^T) = \prod_{t=1}^T \pi_t(a_{t-1}|a_1, r_1, \dots a_t, r_{t-1})$.

Neighbouring input. 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.

$$d_{\text{Ham}}(r, r') \triangleq \sum_{t=1}^T \mathbb{1}\{r_t \neq r'_t\}$$

Privacy definition. A policy π is

- (ϵ, δ)-view DP if $\mathcal{V}^\pi$ is (ϵ, δ)-DP
- (α, ϵ)-view RDP if $\mathcal{V}^\pi$ is (α, ϵ)-RDP
- (ξ, ρ)-view zCDP if $\mathcal{V}^\pi$ is (ξ, ρ)-zCDP

A.1.2 Table DP

To formalise the intuition of Figure 2, Table DP protects the patients by considering the input of the mechanism as the table which represents all the possible outcomes of the bandit interaction. Again in the following, we explain the mechanism to be made DP, its inputs, outputs and the neighbouring relation between its input.

Interaction protocol. Let $\nu \triangleq \{P_a : a \in [K]\}$ a bandit instance with K arms. The policy π interacts with the environment ν up to a given time horizon T to produce a history $\mathcal{H}_T \triangleq \{(a_t, r_t)\}_{t=1}^T$. The iterative steps of this interaction process yielding $\mathcal{H}_T$ verify **two conditions**:

1. the conditional probability of choosing an action $a_t = a$ at time t is dictated only by the policy $\pi_t(a|\mathcal{H}_{t-1})$,
2. the conditional distribution of reward r_t given $(\mathcal{H}_{t-1}, a_t)$ is P_{a_t}.

Thus, the policy can be seen as a set of *adaptively chosen* queries, applied to an *adaptively-gathered* data set of rewards. Conceived this way, it is hard to decouple inputs from outputs to extend DP correctly and protect the privacy of users.

Input. To overcome this problem, we will adhere to the **random table model of bandits** (Lattimore and Szepesvári, 2020). Each user u_t is represented by the row vector $\mathbf{x}_t \triangleq (x_{t,1}, \dots, x_{t,K}) \in \mathbb{R}^K$, where $x_{t,a}$ represents the **potential** reward observed, if action a was recommended to user u_t. Due to the bandit feedback, only $r_t = x_{t,a_t}$ is observed at step t. One can verify that defined this way, the induced history $\mathcal{H}_T \triangleq \{(a_t, r_t)\}_{t=1}^T$ from the interaction between π and ν still verifies the two conditions 1. and 2. as defined above. The distribution of $\mathcal{H}_T$ depends both on the stochasticity of the environment ν and the randomness of the policy π and is denoted by $\mathbb{P}_{\nu,\pi}$.

Using the random table model, the input corresponding to the set of users $\{u_1, \dots, u_T\}$ is the fixed dataset $d = \{x_1, \dots, x_T\} \in (\mathbb{R}^K)^T$. This way, **the bandit interaction can be seen as applying a set of adaptively chosen queries on a *fixed* dataset**.

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

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

The mechanism $\mathcal{M}^\pi$ when applied to a dataset d outputs (in one shot) a sequence of actions $a^T \triangleq (a_1, \dots, a_T) \in [K]^T$ with probability $\mathcal{M}^\pi_d(a^T) = \prod_{t=1}^T \pi_t(a_t | a_1, x_{1,a_1}, \dots a_{t-1}, x_{t-1,a_{t-1}})$.

Neighbouring Input. A change in one user reflects as a change in one row in the table d, so we define $d_{\text{Ham}}(d, d') \triangleq \sum_{t=1}^T \mathbb{1}\{x_t \neq x'_t\} = \sum_{t=1}^T \mathbb{1}\left\{\exists i \in [K],\ x_{t,i} \neq x'_{t,i}\right\}$.

Privacy definition. A policy π is

- (ϵ, δ)-Table DP if $\mathcal{M}^\pi$ is (ϵ, δ)-DP
- (α, ϵ)-Table RDP if $\mathcal{M}^\pi$ is (α, ϵ)-RDP
- (ξ, ρ)-Table zCDP if $\mathcal{M}^\pi$ is (ξ, ρ)-zCDP

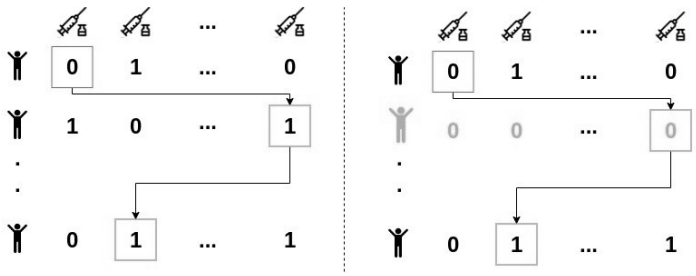


Figure 2: An example of the interaction of a policy with two sets of patients, that differ in one user only. Each row represents the potential reactions of the patient to each medicine, but only one reaction is observed by the policy, i.e. the framed one. A change in one patient reflects as a change in one row in this table of *potential* rewards.

A.1.3 Relation Between Table ADP and View ADP Definitions

Table DP is a stronger notion of privacy than View DP.

Lemma 1. *For a fixed policy π, we have that*

$$\mathcal{M}^\pi \text{ is ADP} \Rightarrow \mathcal{V}^\pi \text{ is ADP.}$$

Proof. Suppose $\mathcal{M}^\pi$ is ADP.

Let $r, r' \in \mathbb{R}^T$ be two lists of rewards such that $d_{\text{Ham}}(r, r') = 1$.

Define d such that $d_{t,i} = r_t$ for all $i \in [K]$ and all $t \in [T]$, i.e. d is the table of rewards where r is concatenated colon-wise K times. We define d' similarly with respect to r'.

For d, d' defined this way, we have that $d_{\text{Ham}}(d, d') = 1$, $\mathcal{V}^\pi_r = \mathcal{M}^\pi_d$ and $\mathcal{V}^\pi_{r'} = \mathcal{M}^\pi_{d'}$.

This means that

- If $\mathcal{M}^\pi$ is (ϵ, δ)-DP, then
$$\forall A \in P([K]^T) \quad \mathcal{V}^\pi_r(A) - e^\epsilon \mathcal{V}^\pi_{r'}(A) = \mathcal{M}^\pi_d(A) - e^\epsilon \mathcal{M}^\pi_{d'}(A) \leq \delta$$
and $\mathcal{V}^\pi$ is (ϵ, δ)-DP.

- If $\mathcal{M}^\pi$ is (α, ϵ)-RDP, then
$$D_\alpha(\mathcal{V}_r^\pi \| \mathcal{V}_{r'}^\pi) = D_\alpha(\mathcal{M}_d^\pi \| \mathcal{M}_{d'}^\pi) \leq \epsilon$$
and $\mathcal{V}^\pi$ is (α, ϵ)-RDP
- If $\mathcal{M}^\pi$ is (ξ, ρ)-zCDP, then
$$D_\alpha(\mathcal{V}_r^\pi \| \mathcal{V}_{r'}^\pi) = D_\alpha(\mathcal{M}_d^\pi \| \mathcal{M}_{d'}^\pi) \leq \xi + \rho\alpha$$
and $\mathcal{V}^\pi$ is (ξ, ρ)-zCDP.

□

For pure ϵ-DP, View DP and Table DP are equivalent.

Lemma 2. *For a fixed policy π, we have that*

$$\mathcal{M}^\pi \textit{ is } \epsilon\textit{-DP} \Leftrightarrow \mathcal{V}^\pi \ \epsilon\textit{-DP.}$$

Proof.
(Proving $\Rightarrow$) $\mathcal{M}^\pi$ is ϵ-DP $\rightarrow$ $\mathcal{V}^\pi$ ϵ-DP is true by Lemma 1, because an ϵ-DP mechanism is also (ϵ, δ)-DP for $\delta = 0$.

(Proving $\Leftarrow$) Suppose $\mathcal{V}^\pi$ is ϵ-DP. We want to show that $\mathcal{M}^\pi$ is ϵ-DP too.

Let $d, d' \in (\mathbb{R}^K)^T$ such that $d_{\text{Ham}}(d, d') = 1$.

For $a^T \in [K]^T$, let $d_{a^T} \triangleq (d_{1,a_1}, d_{2,a_2}, \ldots, d_{T,a_T}) \in \mathbb{R}^T$ be the trajectory of reward induced by a^T in d.

Since $d_{\text{Ham}}(d, d') = 1$, we have that $\forall a^T \in [K]^T d_{\text{Ham}}(d_{a^T}, d'_{a^T}) \leq 1$.

Let $a^T \in [K]^T$. We have that
$$\mathcal{M}_d^\pi(a^T) = \mathcal{V}_{d_{a^T}}^\pi(a^T) \leq e^\epsilon \mathcal{V}_{d'_{a^T}}^\pi(a^T) = e^\epsilon \mathcal{M}_{d'}^\pi(a^T)$$
where the inequality is because $\mathcal{V}^\pi$ is ϵ-DP and $d_{a^T} \sim d'_{a^T}$. Thus, $\mathcal{M}^\pi$ is ϵ-DP.

□

Remark 1. *The crux of the reciprocal proof comes from the fact that to prove ϵ-DP, you only need to check the atomic events a^T. In that case, we can link $\mathcal{M}^\pi$ and $\mathcal{V}^\pi$ easily.*

This is not the case for approximate DP. For example, for (ϵ, δ), there is a huge loss in parameters.

Lemma 3. *For a fixed policy π, we have that*

$$\mathcal{V}^\pi \textit{ is } (\epsilon, \delta)\textit{-DP} \Rightarrow \mathcal{M}^\pi \textit{ is } (\epsilon, K^T\delta)\textit{-DP.}$$

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

Let $d, d' \in (\mathbb{R}^K)^T$ such that $d_{\text{Ham}}(d, d') = 1$.

We have that, for every $a \in [K]^T$, $d_{\text{Ham}}(d_{a^T}, d'_{a^T}) \leq 1$.

Let $E \subset [K]^T$ be an event, i.e a set of sequences. We have that
$$\begin{aligned}\mathcal{M}_d^\pi(E) = \sum_{a^T \in E} \mathcal{M}_d^\pi(a^T) &= \sum_{a^T \in E} \mathcal{V}_{d_{a^T}}^\pi(a^T) \\ &\underset{(a)}{\leq} \sum_{a^T \in E} (e^\epsilon \mathcal{V}_{d'_{a^T}}^\pi(a^T) + \delta) \\ &\underset{(b)}{\leq} e^\epsilon \mathcal{M}_{d'}^\pi(E) + K^T\delta\,,\end{aligned}$$
where (a) holds true because $\mathcal{V}^\pi$ is (ϵ, δ)-DP, and (b) is true because $card(E) \leq K^T$.

This means that $\mathcal{M}^\pi$ is $(\epsilon, K^T\delta)$-DP □

All in all, Table DP is the notion of privacy that we adhere to in this paper, since it protects all the potential responses of an individual rather than just the observed one.

A.2 Interactive DP for bandits

The classic non-interactive definition of DP (Definition 4) considers only mechanisms $\mathcal{M}$ that release answers in one shot. However, data analysts often interact with a database in an adaptive fashion. This motivates the study of *interactive mechanisms* to capture full-featured privacy-preserving data analytics. Here, we adopt the Interactive DP definition as expressed in (Vadhan and Zhang, 2022). The mechanism $\mathcal{M}$ is viewed as a party in an interactive protocol, interacting with a possibly adversarial analyst. We recall the complete definition here.

Definition 5 (Interactive protocol)**.** *An **interactive protocol** (A, B) is any pair of functions on tuples of binary strings. The* interaction *between A with input x_A and B with input x_B is the following random process (denoted $(A(x_A), B(x_B))$):*

1. *Uniformly choose random coins r_A and r_B for A and B, respectively.*
2. *Repeat the following for $i = 0, 1, \ldots$.*
   (a) If i is even, let $m_i = A(x_A, m_1, m_3, \ldots, m_{i-1}; r_A)$.
   (b) If i is odd, let $m_i = B(x_B, m_0, m_2, \ldots, m_{i-1}; r_B)$.
   (c) If $m_i =$ `halt`, then exit loop.

The view of a party in an interactive protocol captures everything the party "observes" during the execution. If (A, B) is an interactive protocol, A's view of the interaction is the tuple $\mathrm{View}_A(A(x_A) \leftarrow B(x_B)) = (r_A, x_A, m_1, m_3, \ldots)$ consisting of all the messages received by A in the execution of the protocol together with the private input x_A and random coins r_A. B's view of $(A(x_A; r_A), B(x_B; r_B))$ is defined symmetrically.

In the setting of DP, Party A is the mechanism, where the input x_A is the dataset. Party B is the adversary that does not have an input x_B. Since we only care about the view of the adversary, we will drop the subscript and denote the view of the adversary as $\mathrm{View}(B \leftrightarrow \mathcal{M}(x))$. With this notation, interactive differential privacy is defined by asking for the views of an adversary on any pair of neighbouring datasets $\mathrm{View}(B \leftrightarrow \mathcal{M}(x))$ and $\mathrm{View}(B \leftrightarrow \mathcal{M}(x'))$ satisfying the same closeness notion as in non-interactive differential privacy.

Definition 6 (Variants of Approximate DP (ADP) for Interactive mechanisms)**.** *A mechanism $\mathcal{M}$ is said to be an*

1. (ϵ, δ)-DP interactive mechanism for a given $\delta \in [0, 1)$ if, for every pair of neighboring datasets $d, d' \in \mathcal{X}$, every adversary $B \in \mathcal{B}$, and every subset of possible views $\mathcal{S} \subseteq$ Range(View), we have

$$\Pr[\mathrm{View}(B \leftrightarrow \mathcal{M}(x)) \in \mathcal{S}] \leq \exp(\epsilon) \cdot \Pr[\mathrm{View}(B \leftrightarrow \mathcal{M}(x')) \in \mathcal{S}] + \delta.$$

2. (α, ϵ)-RDP interactive mechanism for an $\alpha > 1$ if, for every adversary $B \in \mathcal{B}$

$$\sup_{d \sim d'} D_\alpha(\mathrm{View}(B \leftrightarrow \mathcal{M}(d)) \| \mathrm{View}(B \leftrightarrow \mathcal{M}(d'))) \leq \epsilon.$$

3. (ξ, ρ)-zCDP interactive mechanism if, for every $\alpha \in (1, \infty)$, and every adversary $B \in \mathcal{B}$

$$\sup_{d \sim d'} D_\alpha(\mathrm{View}(B \leftrightarrow \mathcal{M}(d)) \| \mathrm{View}(B \leftrightarrow \mathcal{M}(d'))) \leq \xi + \rho\alpha.$$

The interactive protocol Definition 5 is adapted to bandits in Definition 1. Similarly, the interactive definitions of Definition 6 are formalised for bandits in Definition 2.

A.3 Consequences of the Interactive DP definition

Here, we state different corollaries and lemmas, obtained as consequences of Interactive DP.

First, we recall that to check the interactive DP condition, it is enough to only consider deterministic adversaries (Lemma 2.2 in Vadhan and Wang (2021)).

Second, it is easy to see that interactive DP implies non-interactive DP.

Lemma 4. *If π is Interactive b-ADP, π is b-global ADP.*

Proof. This is direct by taking the identity-adversary B^{id} defined by $B^{\mathrm{id}}_t(o_1, \ldots, o_t) = o_t$. □

We also provide the following lemma, that relates the interactive and non-interactive definitions, using an interactive post-processing.

Lemma 5 (Relation between interactive and non-interactive DP for bandits)**.** *π is Interactive b-ADP if and only if, for every deterministic adversary B, π^B is b-Table ADP, where $\pi^B = \{\pi_t^B\}_{t=1}^T$ and*

$$\pi_t^B(a \mid a_1, r_1 \ldots, a_{t-1}, r_{t-1}) \triangleq \pi_t(a \mid B(a_1), r_1, B(a_1, a_2), r_2 \ldots, B(a_1, \ldots, a_{t-1}), r_{t-1}) \quad (9)$$

Remark 2. *We use b-ADP as a shorthand for properties that are true for the three variants of DP. Here, b is the budget, namely $b = (\epsilon, \delta), (\alpha, \epsilon), (\xi, \rho)$.*

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

This means that any interactive policy could be simulated by interactive post-processing of a mechanism verifying non-interactive DP. If a policy is "closed" under interactive post-processing, both interactive and non-interactive DP definitions are equivalent.

Finally, we provide a "group privacy" property, verified by any Interactive DP policy.

Corollary 1. *If π is a ρ-global 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*

$$\sum_{t=1}^{T} \mathrm{KL}\left(\pi_t(. \mid a_1, r_1, \ldots, a_{t-1}, r_{t-1}) \,\middle\|\, \pi_t(. \mid a_1, r'_1, \ldots, a_{t-1}, r'_{t-1})\right) \leq \rho d_{Ham}(\boldsymbol{r}, \boldsymbol{r'})^2$$

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

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

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

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 $\pi_{\mathbf{a}} \triangleq \pi^{B_{\mathbf{a}}}$ as defined in Eq. (9).

Since π is ρ-global zCDP, using Lemma 5, then $\mathcal{M}^{\pi_{\mathbf{a}}}$ is ρ-zCDP. And Lemma 1 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 10 with $\alpha = 1$, we get that

$$\mathrm{KL}\left(\mathcal{V}_{\mathbf{r}}^{\pi_{\mathbf{a}}} \parallel \mathcal{V}_{\mathbf{r'}}^{\pi_{\mathbf{a}}}\right)) \leq \rho \, d_{\mathrm{Ham}}(\mathbf{r}, \mathbf{r'})^2. \quad (10)$$

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

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

In other words $\mathcal{V}_{\mathbf{r}}^{\pi_{\mathbf{a}}} = \bigotimes_{t=1}^{T} \pi_t(. \mid a_1, r_1, \ldots, a_{t-1}, r_{t-1})$.

Similarly, $\mathcal{V}_{\mathbf{r'}}^{\pi_{\mathbf{a}}} = \bigotimes_{t=1}^{T} \pi_t(. \mid a_1, r'_1, \ldots, a_{t-1}, r'_{t-1})$.

Hence, we get

$$\mathrm{KL}\left(\mathcal{V}_{\mathbf{r}}^{\pi_{\mathbf{a}}} \parallel \mathcal{V}_{\mathbf{r'}}^{\pi_{\mathbf{a}}}\right)) = \sum_{t=1}^{T} \mathrm{KL}\left(\pi_t(. \mid a_1, r_1, \ldots, a_{t-1}, r_{t-1}) \,\middle\|\, \pi_t(. \mid a_1, r'_1, \ldots, a_{t-1}, r'_{t-1})\right) \quad (11)$$

Plugging Equaion (11) in Inequality (10) concludes the proof. □

A.4 Poisoning attacks against Interactive DP

We recall the setting of poisoning attacks for bandits (Jun et al., 2018; Liu and Shroff, 2019).

A poisoning attacker B sits between a policy π and the real environment ν. When the policy pulls the action a_t, the environment generates the real reward $r_t^0 \sim \nu_{a_t}$ and the attacker decides on an attack α_t. The reward observed by the policy π is then $r_t = r_t^0 - \alpha_t$. The goal of the attacker B is to manipulate π to choose a sub-optimal target arm (call it K without loss of generality) while spending a minimum cumulative attack cost $\sum_{t=1}^{T} |\alpha_t|$ in expectation. The attack is successful if the number of pulls of the target arms $N_K(T) = T - o(T)$.

The Oracle attack (Jun et al., 2018) is a trivial attack when the attacker **knows the real means** of the environment ν. The attack proceeds by attacking any round t where a non-target arm $a_t \neq K$ is pulled by π. The Oracle attacker pulls down the reward of the corresponding arm by $\alpha_t = \Delta_{a_t}^{\gamma} = \max\{\mu_{a_t} - \mu_K + \gamma, 0\}$ for a small parameter $\gamma > 0$. The Oracle attack transforms the original bandit problem into one where all non-target arms have an expected reward of less than μ_k.

Theorem 8 (Defense against Oracle attack)**.** *If π is a consistent ρ-global zCDP policy, the Oracle attacker needs $\Omega\left(\sqrt{\frac{\log(T)}{\rho}}\right)$ expected cumulative cost to succeed.*

Proof. The oracle attack targets the arm K and makes it appear optimal for the policy π. Since π is a consistent policy, π will linearly pull the 'optimal arm' in the transformed bandit, which is arm K. Thus, the Oracle attacker can succeed.

On the other hand, the Oracle attack, defined by

$$\alpha_t = \Delta_{a_t}^{\gamma} \mathbb{1}\{a_t \neq K\} = \max\{\mu_{a_t} - \mu_K + \gamma, 0\} \mathbb{1}\{a_t \neq K\}$$

has an expected cumulative cost of

$$\mathbb{E}\left[\sum_{t=1}^{T} |\alpha_t|\right] = \sum_{a=1}^{K-1} \mathbb{E}[N_a(T)] \Delta_{a_t}^{\gamma}$$

Since π is a consistent ρ-global zCDP policy, the problem-dependent regret lower bound (Theorem 2) gives that for $a \neq K$, $\mathbb{E}[N_a(T)] = \Omega\left(\sqrt{\frac{\log(T)}{\rho}}\right)$ which concludes the proof. □

B Lower bounds via couplings for concentrated DP

In this section, we are interested in controlling the distance (the Kullback-Leibler, i.e. KL) between marginal distributions induced by a differentially private mechanism, 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. Our main theorem in this section relates the effect of Concentrated DP on this information-theoretic quantity.

In particular, if $\mathcal{P}_1$ and $\mathcal{P}_2$ are two data-generating distributions over $\mathcal{X}^n$, we are interested in the marginals over the output of the mechanism $\mathcal{M}$ when the inputs are generated from $\mathcal{P}_1$ and $\mathcal{P}_2$, i.e. for $\nu \in \{1, 2\}$ and $A \in \mathcal{F}$

$$M_\nu(A) \triangleq \int_{d \in \mathcal{X}^n} \mathcal{M}\left(A \mid d\right) d\mathcal{P}_\nu\left(d\right) \tag{12}$$

In the following, we will provide general results to bound the KL divergence between the distributions M_1 and M_2 defined in (12), when the mechanism $\mathcal{M}$ is ρ-zCDP. The upper bound depends on the privacy budget ρ and the per-step total variation distance between the data-generating distributions $\mathcal{P}_1$ and $\mathcal{P}_2$.

We recall the definition of an f-divergence.

Definition 7 (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 $P \ll Q$ 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 $f(0) \triangleq f(0+)$.

B.1 From the KL to a transport problem

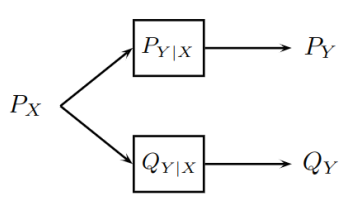
Let $\mathcal{P}_1$ and $\mathcal{P}_2$ two distributions over $\mathcal{X}^n$. Define $\mathcal{C}$ as a coupling of $(\mathcal{P}_1, \mathcal{P}_2)$, i.e. the marginals of $\mathcal{C}$ are $\mathcal{P}_1$ and $\mathcal{P}_2$. We denote by $\Pi(\mathcal{P}_1, \mathcal{P}_2)$ the set of all the couplings between $\mathcal{P}_1$ and $\mathcal{P}_2$. Let M_1 and M_2 be defined as in (12). We recall the definition of an f-divergence.

Theorem 9. *We have that*

$$D_f(M_1\|M_2) \leq \inf_{\mathcal{C} \in \Pi(\mathcal{P}_1, \mathcal{P}_2)} \mathbb{E}_{(d,d') \sim \mathcal{C}}[D_f(\mathcal{M}_d\|\mathcal{M}_{d'})]. \tag{13}$$

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 15.



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

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

The idea is to use Theorem 15, 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 15, 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 to upper bound the RHS of Equation 13.

Theorem 10 (Group Privacy for ρ-zCDP, Proposition 27, Bun and Steinke (2016))**.** *If $\mathcal{M}$ is ρ-CDP, then*

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

Combining the last two theorems gives the following corollary.

Corollary 2. *If $\mathcal{M}$ is ρ-CDP, then*

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

Proof. Let $\mathcal{M}$ be ρ-CDP. Applying Theorem 9, 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 10 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. □

B.2 Proxy solution to the transport Problem

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

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

As a proxy, we will use maximal couplings.

Proposition 1. *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*

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

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

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

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

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

This further yields

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

Corollary 2 concludes the proof. □

Comments on the bound of Theorem 1. This is a centralised ρ-zCDP version of the KL-decomposition lemma under local DP (Duchi et al., 2013, Theorem 1), and a ρ-zCDP version of the Sequential Karwa-Vadhan lemma (Azize and Basu, 2022). We also refer to (Lalanne et al., 2022) that uses similar coupling ideas to derive ρ-zCDP variants of LeCam and Fano inequalities.

C Regret lower bounds for bandits under ρ-global zCDP

In this section, we will use the result of Theorem 1 in classic regret lower bounds for bandits to generate multiple lower bounds, namely minimax and problem dependent for stochastic and minimax for linear bandits.

C.1 Stochastic finite-armed bandits: Minimax lower bound

Theorem 2 (Part a: Minimax lower bound for finite-armed bandits)**.** *Let Π^ρ be the set of ρ-zCDP policies. For any $K > 1$, $T \geq K-1$, and $0 < \rho \leq 1$,*

$$\mathrm{Reg}_{T,\rho}^{minimax} \triangleq \inf_{\pi \in \Pi^\rho} \sup_{\nu \in \mathcal{E}^K} \mathrm{Reg}_T(\pi, \nu) \geq \max\left\{ \frac{1}{27} \underbrace{\sqrt{T(K-1)}}_{\text{without } \rho\text{-global zCDP}} \ , \ \frac{1}{44} \underbrace{\frac{K-1}{\sqrt{\rho}}}_{\text{with } \rho\text{-global zCDP}} \right\}.$$

Proof. The non-private part of the lower bound is due to Theorem 15.2 in Lattimore and Szepesvári (2020). 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 11), we get[4]

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

and

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

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:

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

Step 3: KL-divergence decomposition with ρ-global zCDP. Now, we apply Theorem 1 along with an oracle argument similar to (Shariff and Sheffet, 2018). Since ν and ν' only differ in the distribution of arm a, the oracle coupling induces a maximal coupling only on the samples coming from arm a. Specifically, we build the following oracle coupling $\mathcal{O}$. When π samples an action $i \neq a$, the oracle $\mathcal{O}$ provides the same sample twice, i.e. $r_i \sim \nu_i$ and $r'_i = r_i$. Otherwise, for the samples coming from

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

arm a, the oracle provides, in expectation, n_a fresh iid samples from the maximal coupling between ν_a and ν'_a.

Using Theorem 1 with the oracle coupling $\mathcal{O}$, $n = n_a$ and $t_i = t_a \triangleq \mathrm{TV}\left(\nu_a \parallel \nu'_a\right)$, we get that

$$\begin{aligned}\mathrm{KL}\left(M_{\nu\pi} \parallel M_{\nu'\pi}\right) &\leq \rho(n_a^2 t_a^2 + n_a t_a(1-t_a)) \\ &\leq \rho(n_a^2 t_a^2 + n_a t_a)\,.\end{aligned}$$

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

Finally, using Pinsker's Inequality (Lemma 13), we obtain

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

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

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

By optimising for Δ, we choose $\Delta = \frac{K-1}{T}\left(\frac{1}{\sqrt{\rho}} - \frac{1}{2}\right) > 0$, since $\rho \leq 1$.

This gives that

$$\begin{aligned}\mathrm{Reg}_T(\pi,\nu) + \mathrm{Reg}_T(\pi,\nu') &\geq \frac{K-1}{4}\left(\frac{1}{\sqrt{\rho}} - \frac{1}{2}\right)\exp\left(-1\right) \\ &\geq \frac{K-1}{8\sqrt{\rho}}\exp\left(-1\right)\end{aligned}$$

We conclude the proof by lower bounding $\frac{1}{8}\exp(-1) \geq \frac{1}{22}$, and using $2\max(a,b) \geq a + b$. □

C.2 Stochastic finite-armed bandits: Problem-dependent lower bound

Theorem 2 (Part b: Problem-dependent lower bounds for finite-armed bandits)**.** *Let $\mathcal{E} = \mathcal{M}_1 \times \cdots \times \mathcal{M}_K$ be a class of environments with K arms, where $\mathcal{M}_a$ is a set of reward distributions with finite means. Let π be a consistent policy*[5] *over $\mathcal{E}$ satisfying ρ-global zCDP . Then, for all $\nu = (P_a)_{a=1}^K \in \mathcal{E}$, (i.e. $P_a \in \mathcal{M}_a$), it holds that*

$$\liminf_{T\to\infty}\frac{\mathrm{Reg}_T(\pi,\nu)}{\sqrt{\log(T)}} \geq \sum_{a:\Delta_a>0}\frac{\Delta_a}{\sqrt{\rho}\,t_{\inf}\left(P_a,\mu^*,\mathcal{M}_a\right)}.$$

where $t_{\inf}\left(P,\mu^*,\mathcal{M}\right) \triangleq \inf_{P'\in\mathcal{M}}\left\{\mathrm{TV}\left(P \parallel P'\right) : \mu\left(P'\right) > \mu^*\right\}$

Proof. Let π be a consistent policy satisfying ρ-global zCDP. Let μ_a be the mean of the a-th arm in ν, $t_a = t_{\inf}\left(P_a,\mu^*,\mathcal{M}_a\right)$.

Fix a suboptimal arm a, and let $\beta > 0$ be an arbitrary constant.

Step 1: Choosing the 'hard-to-distinguish' environment. Let $\nu' \triangleq \left(P'_j\right)_{j=1}^K \in \mathcal{E}$ be a bandit with $P'_j = P_j$ for $j \neq a$ and $P'_a \in \mathcal{M}_a$ be such that $\mathrm{TV}\left(P_a \parallel P'_a\right) \leq t_a + \beta$ and $\mu\left(P'_a\right) > \mu^*$, which exists by the definition of t_a. Let $\mu' \in \mathbb{R}^K$ be the vector of means of distributions of ν'.

[5] A policy π is called *consistent* over a class of environments $\mathcal{E}$, if $\forall\nu \in \mathcal{E}$ and $p > 0$, $\lim_{T\to\infty}\frac{R_T(\pi,\nu)}{T^p} = 0$.

Step 2: From lower bounding regret to upper bounding KL-divergence. For simplicity of notations, we use $\mathrm{Reg}_T = \mathrm{Reg}_T(\pi, \nu)$, $\mathrm{Reg}'_T = \mathrm{Reg}_T(\pi, \nu)$, and $A = \{(a_1, a_2, \ldots, a_T) : \mathrm{card}(\{j : a_j = 1\}) \leq T/2\}$.

Then, by regret decomposition and Markov Inequality 11, we obtain

$$\begin{aligned}\mathrm{Reg}_T + \mathrm{Reg}'_T &\geq \frac{T}{2}\left(M_{\nu\pi}(A)\Delta_a + M_{\nu'\pi}\left(A^c\right)\left(\mu'_a - \mu^*\right)\right) \\ &\geq \frac{T}{2}\min\left\{\Delta_a, \mu'_a - \mu^*\right\}\left(M_{\nu\pi}(A) + M_{\nu'\pi}\left(A^c\right)\right) \\ &\geq \frac{T}{4}\min\left\{\Delta_a, \mu'_a - \mu^*\right\}\exp(-\mathrm{KL}\left(M_{\nu\pi} \parallel M_{\nu'\pi}\right))\end{aligned} \tag{14}$$

Step 3: KL-divergence decomposition with ρ-global zCDP. Similar to Step 3 in the previous minimax proof, we build the oracle coupling $\mathcal{O}$ that provides a maximal coupling only on the samples coming from arm a.

Using Theorem 1 with the oracle coupling $\mathcal{O}$, $n = n_a$ and $t_i = \mathrm{TV}\left(\nu_a \parallel \nu'_a\right) = t_a + \beta$, we get that

$$\mathrm{KL}\left(M_{\nu\pi} \parallel M_{\nu'\pi}\right) \leq \rho[n_a^2(t_a + \beta))^2 + n_a(t_a + \beta))]$$

where $n_a \triangleq \mathbb{E}_{\nu\pi}\left[N_a(T)\right]$.

Step 4: Rearranging and taking the limit inferior. Thus, we get

$$\mathrm{Reg}_T + \mathrm{Reg}'_T \geq \frac{T}{4}\min\left\{\Delta_a, \mu'_a - \mu^*\right\}\exp\left(-\rho\left[n_a^2\left(t_a + \beta\right)^2 + n_a\left(t_a + \beta\right)\right]\right)$$

Solving for n_a gives that

$$n_a \geq \frac{\sqrt{4c(T) + 1} - 1}{2(t_a + \beta)}$$

where $c(T) \triangleq \frac{1}{\rho}\log\left(\frac{T\min\left\{\Delta_a, \mu'_a - \mu^*\right\}}{4\left(\mathrm{Reg}_T + \mathrm{Reg}'_T\right)}\right)$.

Now, taking the limit on both sides leads to

$$\begin{aligned}\liminf_{T\to\infty}\frac{\mathbb{E}_{\nu\pi}\left[N_a(T)\right]}{\sqrt{\log(T)}} &\geq \frac{1}{(t_a + \beta)}\liminf_{T\to\infty}\sqrt{\frac{c(T)}{\log(T)}} \\ &= \frac{1}{(t_a + \beta)}\sqrt{\frac{1}{\rho}\left(1 - \limsup_{T\to\infty}\frac{\log\left(\mathrm{Reg}_T + \mathrm{Reg}'_T\right)}{\log(T)}\right)} \\ &= \frac{1}{(t_a + \beta)}\sqrt{\frac{1}{\rho}}\end{aligned}$$

The last equality follows from the definition of consistency, which says that for any $p > 0$, there exists a constant C_p such that for sufficiently large T, $\mathrm{Reg}_T + \mathrm{Reg}'_T \leq C_p T^p$. This property implies that

$$\limsup_{T\to\infty}\frac{\log\left(\mathrm{Reg}_T + \mathrm{Reg}'_T\right)}{\log(T)} \leq \limsup_{T\to\infty}\frac{p\log(T) + \log\left(C_p\right)}{\log(T)} = p,$$

which gives the result since $p > 0$ was an arbitrary constant.

We arrive at the claimed result by taking the limit as β tends to zero.

□

C.3 Stochastic linear bandits: Minimax lower bound

First, we give a specific coupling lemma for the linear case and plug it in the minimax lower bound proofs.

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.

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

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

Lemma 6. *If π is ρ-global zCDP, then*

$$\mathrm{KL}\left(m_{\nu,\pi} \parallel m_{\nu',\pi}\right) \leq \rho \left[\mathbb{E}_{\nu,\pi}\left(\sum_{t=1}^{T} t_{a_t}\right)\right]^2 + \rho\, \mathbb{E}_{\nu,\pi}\left(\sum_{t=1}^{T} t_{a_t}(1 - t_{a_t})\right) + \rho\, \mathbb{V}_{\nu,\pi}\left(\sum_{t=1}^{T} t_{a_t}\right) \tag{15}$$

where $t_{a_t} \triangleq \mathrm{TV}\left(P_{a_t} \,\middle\|\, P'_{a_t}\right)$ 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 Appendix B 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-global 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.

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

By the data-processing inequality, we get that

$$\mathrm{KL}\left(m_{\nu,\pi} \parallel m_{\nu',\pi}\right) \leq \mathrm{KL}\left(p_{\gamma\pi} \parallel q_{\gamma\pi}\right) \tag{16}$$

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

$$
\begin{aligned}
&\mathrm{KL}\left(p_{\gamma\pi} \parallel q_{\gamma\pi}\right) \\
&\overset{(a)}{=} \mathbb{E}_{\mathbf{H} \triangleq (\mathbf{A},\mathbf{R},\mathbf{R'}) \sim p_{\gamma\pi}} \left[\log\left(\frac{\pi(\mathbf{A} \mid \mathbf{R}) c_{\mathbf{A}}(\mathbf{R},\mathbf{R'})}{\pi(\mathbf{A} \mid \mathbf{R'}) c_{\mathbf{A}}(\mathbf{R},\mathbf{R'})}\right)\right] \\
&\overset{(b)}{=} \sum_{t=1}^{T} \mathbb{E}_{\mathbf{H} \sim p_{\gamma\pi}} \left[\log\left(\frac{\pi_t(A_t \mid A_1, R_1, \dots A_{t-1}, R_{t-1})}{\pi_t(A_t \mid A_1, R'_1, \dots A_{t-1}, R'_{t-1})}\right)\right] \\
&\overset{(c)}{=} \sum_{t=1}^{T} \mathbb{E}_{\mathbf{H} \sim p_{\gamma\pi}} \left[\mathbb{E}_{\mathbf{H} \sim p_{\gamma\pi}} \left[\log\left(\frac{\pi_t(A_t \mid A_1, R_1, \dots A_{t-1}, R_{t-1})}{\pi_t(A_t \mid A_1, R'_1, \dots A_{t-1}, R'_{t-1})}\right) \mid A_1, R_1, \dots A_{t-1}, R_{t-1}\right]\right] \\
&\overset{(d)}{=} \sum_{t=1}^{T} \mathbb{E}_{\mathbf{H} \sim p_{\gamma\pi}} \left[\mathbb{E}_{A_t \sim \pi_t(.\mid A_1, R_1, \dots A_{t-1}, R_{t-1})} \left[\log\left(\frac{\pi_t(A_t \mid A_1, R_1, \dots A_{t-1}, R_{t-1})}{\pi_t(A_t \mid A_1, R'_1, \dots A_{t-1}, R'_{t-1})}\right)\right]\right] \\
&\overset{(e)}{=} \sum_{t=1}^{T} \mathbb{E}_{\mathbf{H} \sim p_{\gamma\pi}} \left[\mathrm{KL}\left(\pi_t(. \mid A_1, R_1, \dots A_{t-1}, R_{t-1}) \,\middle\|\, \pi_t(. \mid A_1, R'_1, \dots A_{t-1}, R'_{t-1})\right)\right] \ ,
\end{aligned}
$$

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 $(A_1, R_1, \dots A_{t-1}, R_{t-1})$, the distribution of A_t is $\pi_t(. \mid A_1, R_1, \dots A_{t-1}, R_{t-1})$.

(e): by definition of the KL divergence

On the other hand, Corollary 1, we have that

$$
\sum_{t=1}^{T} \mathrm{KL}\left(\pi_t(. \mid A_1, R_1, \dots A_{t-1}, R_{t-1}) \,\middle\|\, \pi_t(. \mid A_1, R'_1, \dots A_{t-1}, R'_{t-1})\right) \leq \rho d^2_{\text{Ham}}(\mathbf{R}, \mathbf{R}')
$$

which means that

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

where we obtain

(a): using the towering property of the expectation

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

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

Finally, plugging the upper bound in Inequality (16) concludes the proof. □

Theorem 3 (Minimax lower bounds for linear bandits)**.** *Let $\mathcal{A} = [-1, 1]^d$ and $\Theta = \mathbb{R}^d$. Then, for any ρ-global zCDP policy, we have that*

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

Proof. For the non-private lower bound, Theorem 24.1 of (Lattimore and Szepesvári, 2020) gives that,

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

Now, we focus on proving the ρ-global 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

$$\begin{aligned}
\mathrm{Reg}_T(\mathcal{A}, \theta) &= \mathbb{E}_\theta \left[\sum_{t=1}^{T} \sum_{i=1}^{d} \left(\mathrm{sign}\left(\theta_i\right) - A_{ti} \right) \theta_i \right] \\
&\geq \frac{1}{T\sqrt{\rho}} \sum_{i=1}^{d} \mathbb{E}_\theta \left[\sum_{t=1}^{T} \mathbb{I}\left\{ \mathrm{sign}\left(A_{ti}\right) \neq \mathrm{sign}\left(\theta_i\right) \right\} \right] \\
&\geq \frac{1}{\sqrt{\rho}} \sum_{i=1}^{d} \mathbb{M}_\theta \left(\sum_{t=1}^{T} \mathbb{I}\left\{ \mathrm{sign}\left(A_{ti}\right) \neq \mathrm{sign}\left(\theta_i\right) \right\} \geq T/2 \right)
\end{aligned}$$

In this derivation, the first equality holds because the optimal action satisfies $a_i^* = \mathrm{sign}\left(\theta_i\right)$ for $i \in [d]$. The first inequality follows from an observation that $\left(\mathrm{sign}\left(\theta_i\right) - A_{ti}\right) \theta_i \geq |\theta_i| \, \mathbb{I}\left\{ \mathrm{sign}\left(A_{ti}\right) \neq \mathrm{sign}\left(\theta_i\right) \right\}$. The last inequality is a direct application of Markov's inequality 11.

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

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

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-global 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 6, 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 13), we have that

$$\sum_{t=1}^{T} p_t \leq \sum_{t=1}^{T} \sqrt{\frac{1}{2} \mathrm{KL}\left(\mathcal{N}\left(\langle A_t, \theta \rangle, 1 \right) \parallel \mathcal{N}\left(\langle A_t, \theta' \rangle, 1 \right) \right)}$$

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

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

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

□

D Privacy proofs

In this section, we give a complete proof of the privacy of both AdaC-UCB and AdaC-GOPE. Both algorithms share the same blueprint. We first formalise the intuition behind the blueprint in Lemma 7, then give a generic proof of privacy and specify the minor differences to complete the proofs in the last section.

D.1 The privacy lemma of non-overlapping sequences

Remark 3. *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 7 (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

$$\mathcal{G} : \{x_1, \dots, x_T\} \to \bigotimes_{i=1}^{\ell} \mathcal{M}_{\{x_{t_i}, \dots, x_{t_{i+1}-1}\}} \tag{17}$$

In other words, $\mathcal{G}$ 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 (α, ϵ)-RDP then $\mathcal{G}$ is (α, ϵ)-RDP

(c) If $\mathcal{M}$ is (ξ, ρ)-zCDP then $\mathcal{G}$ is (ξ, 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 $t_{\ell'} \leq j \leq t_{\ell'+1} - 1$.

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

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

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

$$\leq e^{\epsilon}\mathcal{G}_{x'}(E) + \delta$$

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

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

(b) $\mathcal{M}$ is (α, ϵ)-RDP.

We have that

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

Since

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

and

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

we get

$$\frac{\mathcal{G}_x(o)}{\mathcal{G}_{x'}(o)} = \frac{\mathcal{M}_{\{x_{t_{\ell'}},\dots,x_{t_j},\dots,x_{t_{\ell'+1}-1}\}}(o_i)}{\mathcal{M}_{\{x_{t_{\ell'}},\dots,x'_{t_j},\dots,x_{t_{\ell'+1}-1}\}}(o_i)}$$

Thus,

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

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

(c) $\mathcal{M}$ is (ξ, ρ)-zCDP.

Similarly, we have that

$$D_\alpha(\mathcal{G}_x \| \mathcal{G}_{x'}) = D_\alpha(\mathcal{M}_{\{x_{t_{\ell'}},\dots,x_{t_j},\dots,x_{t_{\ell'+1}-1}\}} \| \mathcal{M}_{\{x_{t_{\ell'}},\dots,x'_{t_j},\dots,x_{t_{\ell'+1}-1}\}}) \leq \xi + \rho\alpha\,.$$

Thus, $\mathcal{G}$ is (ξ, 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 7. 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 7, those have been fixed before.

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

D.2 Generic privacy proof of AdaC-UCB and AdaC-GOPE

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 7

- the view of B from step $j+1$ until T will be private by post-processing

Let $d = \{d_1, \dots, d_T\}$ and $d' = \{d'_1, \dots, d'_T\}$ two neighbouring reward tables in $(\mathbb{R}^K)^T$. Let $j \in [1, T]$ such that, for all $t \neq j$, $d_t = d'_t$.

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 $d_{<j} \triangleq \{d_1, \dots, d_{j-1}\}$, which is identical for both d and d' and a second conditional part on the history.

First, let us denote

$$\text{View}(B \leftrightarrow \pi(d)) \triangleq \mathcal{P}_d^{B,\pi}.$$

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

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

where

- $\mathbf{o}_{\leq j} \triangleq (o_1, \dots, o_j)$ and $\mathbf{o}_{>j} \triangleq (o_{j+1}, \dots, o_T)$
- $\mathcal{P}_{d_{<j}}^{B,\pi}(\mathbf{o}_{\leq j}) \triangleq \prod_{t=1}^{j} \pi_t \left(o_t \mid B(o_1), d_{1,B(o_1)}, \dots, B(o_1, \dots, o_{t-1}), d_{t-1,B(o_1,\dots,o_{t-1})}\right)$
- $\mathcal{P}_d^{B,\pi}(\mathbf{o}_{>j} \mid \mathbf{o}_{\leq j}) \triangleq \prod_{t=j+1}^{T} \pi_t \left(o_t \mid B(o_1), d_{1,B(o_1)}, \dots, B(o_1, \dots, o_{t-1}), d_{t-1,B(o_1,\dots,o_{t-1})}\right)$

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

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

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 $\psi_d^\pi(j) \triangleq \ell$. 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 $\psi_{d'}^\pi(j) \triangleq \ell'$.

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

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

Step 4. Private sufficient statistics.

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

Let $r_s \triangleq d_{s,B(o_1,\dots,o_s)}$, for $s \in [1, j]$, be the reward corresponding to the action chosen by B in the table d. Similarly, $r'_s \triangleq d'_{s,B(o_1,\dots,o_s)}$ 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. 17, 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.2.1.

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 $f_{d_{\xi_j+1},\dots,d_T}$ such that $\mathcal{P}_d^{B,\pi}(. \mid \mathbf{o}_{\leq j}) = f_{d_{\xi_j+1},\dots,d_T}(L_j)$ and $\mathcal{P}_{d'}^{B,\pi}(. \mid \mathbf{o}_{\leq j}) = f_{d_{\xi_j+1},\dots,d_T}(L'_j)$.

In other words, the view of the adversary B from step $\xi_j + 1$ until T only depends on the sufficient statistics L_j and the new inputs $d_{\xi_j+1}, \dots, d_T$, 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 7 and post-processing.

Using Lemma 7, we have that

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

Using the post-processing property of D_α (Lemma 10), we get that

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

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

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

Thus, we conclude

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

Remark 4. *The same proof could be adapted to (α, ϵ)-RDP, by just showing that the Rényi divergence is smaller than ϵ rather than $\alpha\rho$. For (ϵ, δ)-DP, the same proof follows by changing D_α to the Hockey-Stick Divergence i.e $D_{f_\epsilon}(\mathbb{P}, \mathbb{Q}) = \mathbb{E}[f_\epsilon(\frac{dP}{dQ})]$ where $f_\epsilon = \max(t - e^\epsilon, 0)$. Otherwise, just rewriting the proof using the probability of events is straightforward too.*

D.2.1 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 $\mathcal{G}$ and $\mathcal{M}$ 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 14) $\mathcal{M}$ 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

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

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

Using the Gaussian Mechanism (i.e. Theorem 14), this means that $\phi_\ell + N_\ell$ is ρ-zCDP and $\mathcal{M}$ is too by post-processing.

E Stochastic bandits with global zCDP

E.1 Concentration inequalities

Lemma 8. *Assume that $(X_i)_{1\leq i\leq n}$ are iid random variables in $[0,1]$, with $\mathbb{E}(X_i)=\mu$. Then, for any $\delta\geq 0$,*

$$\mathbb{P}\left(\hat{\mu}_n + Z_n - \sqrt{\left(\frac{1}{2n}+\frac{1}{\rho n^2}\right)\log\left(\frac{1}{\delta}\right)} \geq \mu\right) \leq \delta, \tag{18}$$

and

$$\mathbb{P}\left(\hat{\mu}_n + Z_n + \sqrt{\left(\frac{1}{2n}+\frac{1}{\rho n^2}\right)\log\left(\frac{1}{\delta}\right)} \leq \mu\right) \leq \delta, \tag{19}$$

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 15, 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 14. □

E.2 Regret analysis

Theorem 5 (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 Azize and Basu (2022), for every suboptimal arm a, we have that

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

where

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

such that $Z_\ell \sim \mathcal{N}\left(0,\ \frac{1}{2\rho\times(2^\ell)^2}\right)$

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

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

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

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

Let us consider the contrary, i.e.

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

$$\Rightarrow 2^\ell < \frac{2\beta}{\Delta_a^2}\log(T) + 2\sqrt{\frac{\beta}{\rho\Delta_a^2}}\sqrt{\log(T)} \tag{20}$$

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

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

The last inequality is due to Equation 18 of Lemma 8.

Step 2: The regret bound. Combining Steps 1 and 2, we get that

$$\begin{aligned}\mathbb{E}[N_a(T)] &\leq \frac{\beta}{\beta-3} + 2^{\ell+1} + T\times\frac{1}{T^\beta}\\ &\leq \frac{8\beta}{\Delta_a^2}\log(T) + 8\sqrt{\frac{\beta}{\rho\Delta_a^2}}\sqrt{\log(T)} + \frac{2\beta}{\beta-3}.\end{aligned} \tag{21}$$

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

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

□

Theorem 5 (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

$$\begin{aligned}\mathrm{Reg}_T(\mathsf{AdaP\text{-}UCB},\nu) &= \sum_a \Delta_a\mathbb{E}[N_a(T)]\\ &= \sum_{a:\Delta_a\leq\Delta}\Delta_a\mathbb{E}[N_a(T) + \sum_{a:\Delta_a>\Delta}\Delta_a\mathbb{E}[N_a(T)]\\ &\leq T\Delta + \sum_{a:\Delta_a>\Delta}\Delta_a\left(\frac{8\beta}{\Delta_a^2}\log(T) + 8\sqrt{\frac{\beta}{\rho\Delta_a^2}}\sqrt{\log(T)} + \frac{2\beta}{\beta-3}\right) \quad \text{(Eq. 21)}\\ &\leq T\Delta + \frac{8\beta K\log(T)}{\Delta} + 8K\sqrt{\frac{\beta\log(T)}{\rho}} + \frac{3\beta}{\beta-3}\sum_a\Delta_a\\ &\leq 4\sqrt{2\beta KT\log(T)} + 8K\sqrt{\frac{\beta\log(T)}{\rho}} + \frac{3\beta}{\beta-3}\sum_a\Delta_a\,.\end{aligned}$$

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

E.3 Extensions to (ϵ, δ)-global DP and (α, ϵ)-global RDP

In this section, we specify the modifications required to make AdaC-UCB (ϵ, δ)-global DP and (α, ϵ)-global RDP. Also, we give the corresponding regret upper bounds.

The difference comes from the different calibrations of the Gaussian Mechanism (Thm 14). 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 14.

(ϵ, δ)-global DP. The private index to select the arms (Line 6 of Algorithm 2) becomes

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

where $\sigma_{a,\ell}^2 \triangleq \frac{2\log(\frac{1.25}{\delta})}{\epsilon^2 \times \left(\frac{1}{2}N_a(t_\ell - 1)\right)^2}$, and the exploration bonus is

$$B_a(t_\ell - 1, \beta) \triangleq \sqrt{\left(\frac{1}{2 \times \frac{1}{2}N_a(t_\ell - 1)} + \frac{4\log(\frac{1.25}{\delta})}{\epsilon^2 \times \left(\frac{1}{2}N_a(t_\ell - 1)\right)^2}\right)\beta\log(t_\ell)}\,.$$

Thus, the regret upper bounds become:

$$\text{Problem-dependent:} \quad \sum_{a:\Delta_a > 0}\left(\frac{8\beta}{\Delta_a}\log(T) + 8\sqrt{\frac{4}{\beta\epsilon^2}\log\left(\frac{1.25}{\delta}\right)}\sqrt{\log(T)} + \frac{2\beta}{\beta - 3}\right)\,.$$

$$\text{Problem-independent:} \quad \mathcal{O}\left(\sqrt{KT\log(T)}\right) + \mathcal{O}\left(\frac{K\sqrt{\log\left(\frac{1}{\delta}\right)}}{\epsilon}\sqrt{\log(T)}\right)\,.$$

(α, ϵ)-global RDP. The private index to select the arms (Line 6 of Algorithm 2) becomes

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

where $\sigma_{a,\ell}^2 \triangleq \frac{\alpha}{2\epsilon \times \left(\frac{1}{2}N_a(t_\ell - 1)\right)^2}$, and the exploration bonus is

$$B_a(t_\ell - 1, \beta) \triangleq \sqrt{\left(\frac{1}{2 \times \frac{1}{2}N_a(t_\ell - 1)} + \frac{\alpha}{\epsilon \times \left(\frac{1}{2}N_a(t_\ell - 1)\right)^2}\right)\beta\log(t_\ell)}\,.$$

The regret upper bounds become:

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

$$\text{Problem-independent:} \quad \mathcal{O}\left(\sqrt{KT\log(T)}\right) + \mathcal{O}\left(K\sqrt{\frac{\alpha}{\epsilon}\log(T)}\right)\,.$$

F Linear Bandits with global zCDP

F.1 Basic definitions of optimal design

Definition 8 (Optimal design)**.** *Let $\mathcal{A} \subset \mathbb{R}^d$ and $\pi : \mathcal{A} \to [0,1]$ be a distribution on $\mathcal{A}$ so that $\sum_{a\in\mathcal{A}} \pi(a) = 1$. Let $V(\pi) \in \mathbb{R}^{d\times d}$ and $f(\pi), g(\pi) \in \mathbb{R}$ be given by*

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

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

Theorem 11 (Kiefer–Wolfowitz theorem)**.** *Assume that $\mathcal{A}$ is compact and $span(\mathcal{A}) = \mathbb{R}^d$. The following are equivalent:*

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

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

F.2 Concentration inequalities

Let $a_1, \ldots, a_t$ be deterministically chosen without the knowledge of $r_1, \ldots, r_t$. Let π be an optimal design for $\mathcal{A}$.

Let $V_t \triangleq \sum_{s=1}^t a_s a_s^T = \sum_{a\in\mathcal{A}} N_a(t) aa^T$ 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 12. *Let $\delta \in [0,1]$ and $\beta_t \triangleq g_t\sqrt{2\log\left(\frac{4}{\delta}\right)} + g_t^2\sqrt{\frac{2}{\rho}\left(d + 2\sqrt{d\log\left(\frac{2}{\delta}\right)} + 2\log\left(\frac{2}{\delta}\right)\right)}$. 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 Lattimore and Szepesvári (2020), 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 16, we get

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

Steps 1 and 2 together conclude the proof. □

Corollary 3. *Let β be a confidence level. If each action $a \in \mathcal{A}$ is chosen for $N_a(t) \triangleq c_t \pi(a)$ where*

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

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

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

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

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

Thus,

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

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

We conclude the proof using Theorem 12. □

F.3 Regret analysis

Theorem 13. *Under Assumption 1 and for $\delta \in (0,1)$, with probability at least $1-\delta$, the regret R_T of* AdaC-GOPE *(Algorithm 3) 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 3, 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 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

$$R_T = \sum_{a\in\mathcal{A}} \Delta_a N_a(T)$$

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

where the last inequality is thanks to the third bullet point in **Step 2**, i.e. $\Delta_a \le 4\beta_{\ell-1}$ 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. We have that

$$
\begin{aligned}
T_\ell &= \sum_{a\in S_\ell} T_\ell(a) \\
&= \sum_{a\in S_\ell} \left[\frac{8d\pi_\ell(a)}{\beta_\ell^2} \log\left(\frac{4k\ell(\ell+1)}{\delta}\right) + \frac{2d\pi_\ell(a)}{\beta_\ell} \sqrt{\frac{2}{\rho}\left(d + 2\sqrt{d\log\left(\frac{2k\ell(\ell+1)}{\delta}\right)} + 2\log\left(\frac{2k\ell(\ell+1)}{\delta}\right)\right)} \right] \\
&\le \frac{d(d+1)}{2} + \frac{8d}{\beta_\ell^2}\log\left(\frac{4k\ell(\ell+1)}{\delta}\right) + \frac{2d}{\beta_\ell}\sqrt{\frac{2}{\rho}\left(d + 2\sqrt{d\log\left(\frac{2k\ell(\ell+1)}{\delta}\right)} + 2\log\left(\frac{2k\ell(\ell+1)}{\delta}\right)\right)}.
\end{aligned}
$$

since $\beta_{\ell+1} = \frac{1}{2}\beta_\ell$ and $\sum_{\ell=1}^{\ell(T)} T_\ell = T$, there exists a constant C such that $\ell(T) \le C\log(T)$. In other words, the length of the episodes is at least doubling so their number is logarithmic.

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

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

hence

$$
T_\ell \le \frac{d(d+1)}{2} + \frac{8d}{\beta_\ell^2} C' \log\left(\frac{k\log(T)}{\delta}\right) + \frac{4d}{\beta_\ell}\sqrt{\frac{1}{\rho} C' \log\left(\frac{k\log(T)}{\delta}\right)}
$$

Step 5: Upper-bounding regret under E.

Under E

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

All in all, we have that

$$
R_T \le 4d(d+1) + C_2 d\sqrt{\frac{1}{\rho}\log\left(\frac{k\log(T)}{\delta}\right)}\log(T) + C_1 d\log\left(\frac{k\log(T)}{\delta}\right)\frac{1}{\Delta} + T\Delta
$$

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

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

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

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

□

F.4 Extensions to (ϵ, δ)-global DP and (α, ϵ)-global RDP

In this section, we specify the modifications required to make AdaC-GOPE (ϵ, δ)-global DP and (α, ϵ)-global RDP, and provide the corresponding regret upper bounds.

The difference comes from the different calibrations of the Gaussian Mechanism (Thm 14). Adapting the analysis from ρ-**zCDP** reduces to changing the $\mathbf{\frac{1}{2\rho}}$ factor to $\mathbf{\frac{2}{\epsilon^2}}\log(\mathbf{\frac{1.25}{\delta}})$ **for** (ϵ, δ)**-DP** and to $\mathbf{\frac{\alpha}{2\epsilon}}$ **for** (α, ϵ)**-RDP**, i.e. varying the constant b in Theorem 14.

(ϵ, δ)**-global DP.** The number of times each action a is played at episode ℓ for AdaC-GOPE is $T_\ell(a) \triangleq c_\ell \pi_\ell(a)$ times, where for $\delta' \triangleq \frac{\delta}{K\ell(\ell+1)}$,

$$c_\ell \triangleq \frac{8d}{\beta_\ell^2}\log\left(\frac{4}{\delta'}\right) + \frac{2d}{\beta_\ell}\sqrt{\frac{8}{\epsilon^2}\log(\frac{1.25}{\delta})}\left(d + 2\sqrt{d\log\left(\frac{2}{\delta'}\right)} + 2\log\left(\frac{2}{\delta'}\right)\right)^{1/2}$$

The added Gaussian noise in Step 4 of AdaC-GOPE becomes $N_\ell \sim \mathcal{N}\left(0, \frac{8d}{\epsilon^2 c_\ell}\log(\frac{1.25}{\delta})I_d\right)$.

Thus, the regret upper-bound becomes

$$\mathcal{O}\left(\sqrt{dT\log(KT)}\right) + \mathcal{O}\left(\sqrt{\frac{1}{\epsilon^2}\log\left(\frac{1.25}{\delta}\right)}d(\log(KT))^{\frac{3}{2}}\right)$$

(α, ϵ)**-global RDP.** The number of times each action a is played at episode ℓ for AdaC-GOPE is $T_\ell(a) \triangleq c_\ell \pi_\ell(a)$ times, where for $\delta' \triangleq \frac{\delta}{K\ell(\ell+1)}$,

$$c_\ell \triangleq \frac{8d}{\beta_\ell^2}\log\left(\frac{4}{\delta'}\right) + \frac{2d}{\beta_\ell}\sqrt{\frac{2\alpha}{\epsilon}}\left(d + 2\sqrt{d\log\left(\frac{2}{\delta'}\right)} + 2\log\left(\frac{2}{\delta'}\right)\right)^{1/2}$$

The added Gaussian noise in Step 4 of AdaC-GOPE becomes $N_\ell \sim \mathcal{N}\left(0, \frac{2d\alpha}{\epsilon c_\ell}I_d\right)$.

Thus, the regret upper-bound becomes

$$\mathcal{O}\left(\sqrt{dT\log(KT)}\right) + \mathcal{O}\left(\sqrt{\frac{\alpha}{\epsilon}}d(\log(KT))^{\frac{3}{2}}\right).$$

F.5 Adding noise at different steps of AdaC-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 14, 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 Hanna et al. (2022) and Li et al. (2022) derive a private version of GOPE for different privacy definitions, i.e. pure ϵ-DP for Hanna et al. (2022) and (ϵ, δ)-DP for Li et al. (2022), 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 Hanna et al. (2022) and Li et al. (2022) 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 (Hanna et al., 2022, Eq (18)), and $d^{\frac{3}{2}}$ in (Li et al., 2022, 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.

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 5. *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.*

G Existing technical results and definitions

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

Lemma 9 (Post-processing Lemma (Proposition 2.1, (Dwork et al., 2014))). *If a randomised algorithm $\mathcal{A}$ satisfies (ϵ, δ)-Differential Privacy and f is an arbitrary randomised mapping defined on $\mathcal{A}$'s output, then $f \circ \mathcal{A}$ satisfies (ϵ, δ)-DP.*

Theorem 14 (The Gaussian Mechanism (Dwork et al. (2014), Mironov (2017), Bun and Steinke (2016))). *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 10 (Post-processing property of Renyi Divergence, Lemma 2.2 Bun and Steinke (2016)). *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 11 (Markov's Inequality). *For any random variable X and $\varepsilon > 0$,*

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

Definition 9 (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 12 (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 13 (Pinsker's Inequality). *For two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on the same probability space $(\Omega, \mathcal{F})$, we have*

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

Definition 10 (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 14 (Concentration of Sub-Gaussian random variables). *If X is σ-sub-Gaussian, then for any $\epsilon \geq 0$,*

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

Lemma 15 (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 16 (Concentration of the χ^2-Distribution, Claim 17 of Shariff and Sheffet (2018)). *If $X \sim \mathcal{N}(0, I_d)$ and $\delta \in (0, 1)$, then*

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

Theorem 15 (Conditioning Increases f-divergence). *Let $P_X \xrightarrow{P_{Y|X}} P_Y$ and $P_X \xrightarrow{Q_{Y|X}} Q_Y$. Then,*

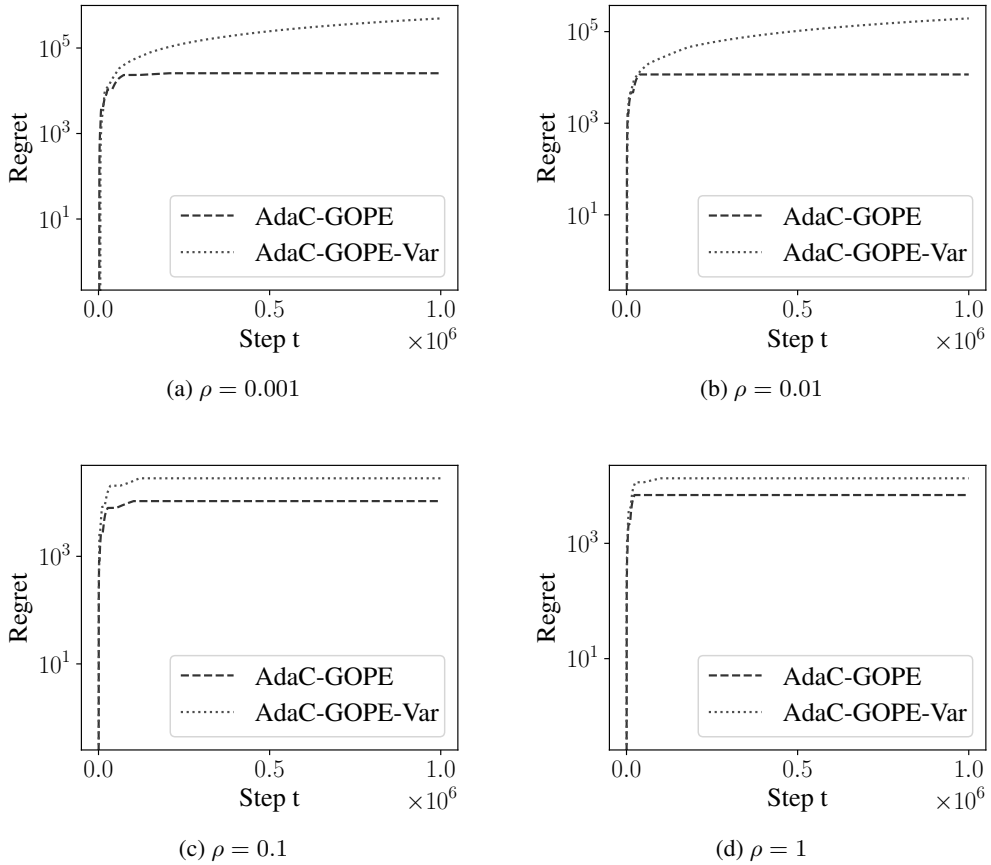
$$D_f(P_Y\|Q_Y) \leq \mathbb{E}_{X \sim P_X}\left[D_f\left(P_{Y|X}\|Q_{Y|X}\right)\right].$$

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 F.5). 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 5, for different privacy budgets ρ and report the results in Figure 3.



(a) $\rho = 0.001$ (b) $\rho = 0.01$

(c) $\rho = 0.1$ (d) $\rho = 1$

Figure 3: Evolution of the regret over time for AdaC-GOPE and Adar-GOPE-Var for different values of the privacy budget ρ

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.*