When Privacy Meets Partial Information: A Refined Analysis of Differentially Private 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

We study the problem of multi-armed bandits with ϵ-global Differential Privacy (DP). First, we prove *the minimax and problem-dependent regret lower bounds* for stochastic and linear bandits that quantify the hardness of bandits with ϵ-global DP. These bounds suggest the existence of two hardness regimes depending on the privacy budget ϵ. In the high-privacy regime (small ϵ), the hardness depends on a coupled effect of privacy and partial information about the reward distributions. In the low-privacy regime (large ϵ), bandits with ϵ-global DP are not harder than the bandits without privacy. For stochastic bandits, *we further propose a generic framework to design a near-optimal ϵ-global DP extension of an index-based optimistic bandit algorithm*. The framework consists of three ingredients: the Laplace mechanism, arm-dependent adaptive episodes, and usage of only the rewards collected in the last episode for computing private statistics. Specifically, we instantiate ϵ-global DP extensions of UCB and KL-UCB algorithms, namely AdaP-UCB and AdaP-KLUCB. AdaP-KLUCB *is the first algorithm that both satisfies ϵ-global DP and yields a regret upper bound that matches the problem-dependent lower bound up to multiplicative constants.*

1 Introduction

Multi-armed bandit problems, in short *bandits*, are a model for sequential decision-making with partial information (Lattimore and Szepesvári, 2018). In bandits, a learner sequentially interacts with an environment, which is a set of unknown distributions (or arms or actions), over $T \in \mathbb{N}$ steps. T is referred to as the horizon. At each step $t \in \{1, \dots, T\}$, the learner chooses an arm A_t from $\{1, \dots, K\}$ and the environment reveals a reward r_t from the distribution ν_{A_t}. The learner's objective is to maximise its cumulative reward $\sum_{t=1}^{T} r_t$. An equivalent performance metric for a bandit algorithm π is *regret*. Regret is the difference between the expected cumulative reward collected by pulling the optimal arm a^* for T times and the expected cumulative reward obtained by using π. Regret is the price paid by the bandit algorithm due to partial information about the reward distributions. *The goal of a bandit algorithm is to minimise its regret*. Bandits, introduced in (Thompson, 1933) for medical trials, are widely studied and deployed in real-life applications, such as online advertising (Schwartz et al., 2017), recommendation systems (Silva et al., 2022) and investment portfolio design (Silva et al., 2022). Sensitive data of individuals, such as health conditions, personal preferences, financial status etc., are utilised in these applications, which raises the concern about privacy.

Example 1. *In bandit-based recommendation systems, the learning algorithm selects an item for each user. The user decides whether or not to click the recommendation. Based on the click feedback, the bandit algorithm improves its subsequent recommendations. Here, an action corresponds to an*

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

item and a reward corresponds to whether a user clicks on the item (1) or not (0). A change in a user's preference prompts changes in the algorithm's output. As a result, even if the click feedback is kept private, the user's private information, i.e. their preference over items, is revealed. Privacy protection aims to prevent the algorithm's output from revealing the preferences of any specific user.

This example demonstrates the need for privacy in bandits. In this paper, we use Differential Privacy (DP) (Dwork and Roth, 2014) as the framework for privacy. DP ensures that an algorithm's output is unaffected by changes in input at a single data point. By limiting the amount of sensitive information that an adversary can deduce from the output, DP renders an individual corresponding to a data point 'indistinguishable'. A calibrated amount of noise is injected into an algorithm to ensure DP. The noise scale is set to be proportional to the algorithm's sensitivity and inversely proportional to the privacy budget ϵ.

To address the privacy issues in bandit applications, the problem of Differentially Private Bandits is coined and studied under different settings, such as stochastic bandits (Mishra and Thakurta, 2015; Tossou and Dimitrakakis, 2016; Sajed and Sheffet, 2019; Hu et al., 2021), adversarial bandits (Tossou and Dimitrakakis, 2017), and linear contextual bandits (Shariff and Sheffet, 2018; Neel and Roth, 2018). Also, multiple adaptations of DP, namely local and global, are proposed for bandits (Basu et al., 2019). *Local DP* aims to preserve the privacy of the sequence of rewards obtained by sending noisy rewards to the algorithm (Duchi et al., 2013). *Global DP* allows the algorithm to access rewards without noise and aims to keep the sequence of rewards private while only the sequence of actions taken by the algorithm is produced to the public (Basu et al., 2019). Though local DP provides stronger privacy as the data curator has no access to the original reward stream, it injects too much noise that leads to higher regret. Also, the fundamental hardness of local DP in bandits in terms of regret lower bound and also corresponding optimal algorithms are well-understood (Zheng et al., 2020). Thus, in this paper, *we focus on the bandit problems with ϵ-global DP.* We aim to address two questions:

1. *What is the fundamental hardness of differentially private bandits with global DP expressed in terms of the regret lower bound?*
2. *How to design an algorithmic framework that converts an optimistic and near-optimal bandit algorithm into a near-optimal bandit algorithm satisfying global DP?*

Our Contributions. These questions have led to the following contributions:

1. *Hardness as Lower Bounds:* We derive both the minimax (worst-case) and problem-dependent lower bounds on the regret of bandits with ϵ-global DP for stochastic bandits (Thm. 2 and 3, Sec. 3). Both the bounds show that the hardness depends on a trade-off between the privacy budget ϵ and the distinguishability gap of a bandit environment. If the ϵ is bigger than the distinguishability gap, a bandit with global DP is not harder than a non-private bandit problem. Additionally, our problem-dependent regret lower bound (Thm. 3) provides a novel observation that the difficulty of a bandit problem with global DP depends on the TV-indistinguishability gap (t_{inf}, Thm. 3). This was not known in the regret lower bound in (Shariff and Sheffet, 2018), where the privacy dependent term in regret, i.e. $\frac{K\log(T)}{\epsilon}$ is independent of the hardness of the bandit instance. Our lower bounds explicates this missing link between the interaction of privacy and partial information. We also extend our techniques to derive the lower bounds for linear bandits with a finite number of arms (Thm. 4 and 5). These lower bounds also reflect the same transition of hardness depending on ϵ and distinguishability gaps.

2. *Algorithm Design:* Optimistic bandit algorithms used the empirical mean and variance of observed rewards to compute indexes and use them to select an action. We propose three fundamental strategies to design a private bandit algorithm from an optimistic algorithm (Sec. 4)– a) add Laplacian noise to the empirical mean of each arm calibrated by the corresponding sensitivity, b) use adaptive episodes to compute the private empirical mean less number of times, and c) use only the observed rewards of the arm's *last active episode*[1] and forget everything before to keep the sensitivity of the empirical mean low. We deploy these techniques with UCB (Auer et al., 2002) and KL-UCB (Garivier and Cappé, 2011) algorithms for non-private bandits to propose two near-optimal and ϵ-global DP bandit algorithms, AdaP-UCB and AdaP-KLUCB (Sec. 4). Both of these algorithms achieve near-optimal problem-dependent regret and also reflect a transition in hardness from a low to a high privacy regime

[1]The last active episode of an arm is the last episode in which that arm was played.

(Thm. 7 and 8). Both theoretical (Table 1) and experimental (Sec. 5) results demonstrate optimality of AdaP-KLUCB than existing algorithms.

3. *Technical Tools:* (a) To derive the lower bound, we extend the Karwa-Vadhan lemma (Lemma 6.1, (Karwa and Vadhan, 2017)) to the sequential setting (Lemma 2). (b) We present a novel sequential information processing lemma under ϵ-global DP (Thm. 10) that controls the difference between the outcome streams of a differentially private policy when interacting with two different bandit instances. (c) This leads to a generic proof structure, utilised to generate refined regret lower bounds under ϵ-global DP for different settings. (d) To derive the regret upper bound, we develop a novel and general analysis of optimistic algorithms with adaptive episodes.

Our regret lower and upper bounds close the open question posed in (Tenenbaum et al., 2021), i.e. the problem-dependent regret bounds of differentially private bandits should depend on the KL-divergence between the reward distributions.

Related Works: Lower Bound. (Shariff and Sheffet, 2018) first proposed a problem-dependent lower bound on regret, $\Omega(\max\{\sum_{a\neq a^*}\frac{\log T}{\Delta_a}, \frac{K\log(T)}{\epsilon}\})$[2], for stochastic bandits with ϵ-global DP. But this bound is restricted to Bernoulli distributions of reward. In this paper, *we provide the first problem-dependent regret lower bound that is valid for any reward distribution. This lower bound shows that the effect of privacy does not only depend on ϵ but also on the total variation distance corresponding to the environment*. Thus, *it explicates the coupled effect of privacy and partial information in a high-privacy regime*, which was not observed before. To derive this tighter lower bound, we extend the Karwa-Vadhan lemma (Lemma 6.1, (Karwa and Vadhan, 2017)) to a sequential setting and also propose a generic proof structure leading to *problem-dependent and minimax lower bounds for stochastic and linear bandits*. To our knowledge, *these are the first regret lower bounds for linear bandits with ϵ-global DP*. Basu et al. (2019) also proposed a minimax regret lower bound for stochastic bandits with ϵ-global DP. However, our minimax lower bound (Thm. 2) is tighter and does not need to assume that the reward distributions are Lipschitz continuous.

Related Works: Bandit Algorithms with ϵ-global DP. DP-UCB (Mishra and Thakurta, 2015; Tossou and Dimitrakakis, 2016) was the first global DP version of UCB. DP-UCB uses the tree-based mechanism (Dwork et al., 2010a; Chan et al., 2011) to compute the sum of rewards. The tree mechanism maintains a binary tree of depth $\log(T)$ over the T streaming observations, where each node in this tree holds an i.i.d sample from a Laplace distribution with zero mean and scale $(\log(T)/\epsilon)$. At each step t, the mechanism yields the sum of the first t observations and the $\log(T)$ nodes on the root-to-the-leaf path in the binary tree as the private empirical mean. As a result, the noise added to the UCB index per time-step is $\mathcal{O}\left(\log(T)^{2.5}/\epsilon\right)$, which is responsible for the extra multiplicative factor $\log(T)^{1.5}$ in regret compared to the lower bound. DP-SE (Sajed and Sheffet, 2019) was the first ϵ-global DP algorithm to eliminate the additional multiplicative factor $\log(T)^{1.5}$. DP-SE is an ϵ-global DP version of the Successive Elimination algorithm (Even-Dar et al., 2002). However, the drawbacks were that the algorithm was not anytime, and was optimal only asymptotically, i.e. the horizon should be big enough. On the other hand, a careful analysis of the algorithm suggests that what made the algorithm optimal was the fact that the algorithm was run in doubling episodes, where the private means computed at the end of the episode to decide which arms to eliminate, were only computed using the samples collected from that episode. *We detect these ingredients and generalise them to propose a general framework to make any optimist index-based policy optimal followed by upper bounds matching the lower bounds (Table 1).* Similar techniques have been deployed to design a global DP extension of UCB in (Hu et al., 2021) but they have an additive factor $\log(T)$ extra in regret than AdaP-UCB and do not match our proposed regret lower bound. Also, DP-TS (Hu and Hegde, 2022) aims to achieve ϵ-global DP with a Thompson sampling based approach. The "lazy" version of DP-TS achieves a similar regret as AdaP-UCB, but does not achieve the refined lower bound of Thm. 3 with the indistinguishability gaps. Even in non-private bandits, Thompson Sampling (Agrawal and Goyal, 2013) is not known to achieve the Kullback-Leibler (KL) indistinguishability gap while KL-UCB (Garivier and Cappé, 2011) is known to achieve it. *Our work leads to the first near-optimal bandit algorithm with ϵ-global DP, namely* AdaP-KLUCB.

[2]Here, a^* is the optimal arm with mean reward μ^* and $\Delta_a \triangleq \mu^* - \mu_a$ is the suboptimality gap of arm a.

[3]The number of private empirical mean reward statistics computed by the algorithm.

[4]Not using the full history of rewards to compute the private empirical mean reward statistics.

Table 1: A comparison of ϵ-global DP algorithms for bandits.

Algorithm	Regret	# Private Means [3]	Anytime	Forgetfulness [4]
DP-UCB (Mishra and Thakurta (2015); Tossou and Dimitrakakis (2016))	$\mathcal{O}\left(\frac{\log(T)^{2.5}}{\epsilon} + \sum_{a \neq a^*} \frac{\log(T)}{\Delta_a}\right)$	KT	Yes	No
DP-SE (Sajed and Sheffet, 2019)	$\mathcal{O}\left(\frac{K\log(T)}{\epsilon} + \sum_{a \neq a^*} \frac{\log(T)}{\Delta_a}\right)$	$\mathcal{O}(\log(T))$	No	Yes
AdaP-UCB	$\mathcal{O}\left(\sum_{a \neq a^*} \frac{\Delta_a \log(T)}{\min(\Delta_a^2, \epsilon\Delta_a)}\right)$	$\mathcal{O}(\log(T))$	Yes	Yes
AdaP-KLUCB	$\mathcal{O}\left(\sum_{a \neq a^*} \frac{\Delta_a \log(T)}{\min(d(\mu_a, \mu^*), \epsilon\Delta_a)}\right)$	$\mathcal{O}(\log(T))$	Yes	Yes

2 Background: Differential Privacy and Bandits

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

Definition 1 ((ϵ, δ)-DP (Dwork and Roth, 2014))**.** *A randomised algorithm $\mathcal{A}$ satisfies (ϵ, δ)-Differential Privacy (DP) if for any two neighbouring datasets[5] $\mathcal{D}$ and $\mathcal{D}'$, and for all sets of output $\mathcal{O} \subseteq \mathrm{Range}(\mathcal{A})$*

$$\Pr[\mathcal{A}(\mathcal{D}) \in \mathcal{O}] \leq e^{\epsilon} \Pr\left[\mathcal{A}\left(\mathcal{D}'\right) \in \mathcal{O}\right] + \delta, \tag{1}$$

where the probability space is over the coin flips of the mechanism $\mathcal{A}$ and for some $(\epsilon, \delta) \in \mathbb{R}^{\geq 0} \times \mathbb{R}^{\geq 0}$. If $\delta = 0$, we say that $\mathcal{A}$ satisfies ϵ-differential privacy.

The Laplace mechanism (Dwork et al., 2010b; Dwork and Roth, 2014) ensures ϵ-DP by injecting a random noise to the output of the algorithm that is sampled from a calibrated Laplace distribution (as specified in Theorem 1). We use $Lap(b)$ to denote the Laplace distribution with mean 0 and variance $2b^2$.

Theorem 1 (ϵ-DP of Laplace Mechanism (Theorem 3.6, Dwork and Roth (2014)))**.** *Let us consider an algorithm $f : \mathcal{X} \to \mathbb{R}^d$ with sensitivity $s(f) \triangleq \max_{\mathcal{D}, \mathcal{D}' \; s.t \; \|\mathcal{D} - \mathcal{D}'\|_{\text{Hamming}} = 1} \|f(\mathcal{D}) - f(\mathcal{D}')\|_1$. Here, $\|\cdot\|_1$ is the L_1 norm on $\mathbb{R}^d$. If d noise samples $\{N_i\}_{i=1}^d$ are generated independently from $Lap(\frac{s(f)}{\epsilon})$, the output injected with the noise, i.e. $f(\mathcal{D}) + [N_1, \ldots, N_d]$ satisfies ϵ-DP.*

The Laplace mechanism is originally proposed to ensure DP when the input database is static. In a sequential setting like bandits, a mechanism must update the published statistics as new data items arrive, and thus DP definitions are extended accordingly (Basu et al., 2019).

Stochastic Bandits with Global DP. Now, we describe the general canonical bandit model (Section 4.6, (Lattimore and Szepesvári, 2018)) and define global Differential Privacy (global DP) in this setting. A bandit algorithm (or policy) interacts with an environment ν consisting of K arms with reward distributions $\{\nu_a\}_{a=1}^K$ for a given horizon T and produces a history $\mathcal{H}_T \triangleq \{(A_t, R_t)\}_{t=1}^T$. At each step t, the choice of the arm depends on the previous history $\mathcal{H}_{t-1}$. The reward R_t is sampled from the reward distribution ν_{A_t} and is conditionally independent of the previous history $\mathcal{H}_{t-1}$. A bandit algorithm (or policy) π can be represented by a sequence $(\pi_t)_{t=1}^T$, where $\pi_t : \mathcal{H}_{t-1} \to [K]$ is a probability kernel. Thus, we denote the probability of occurrence of a sequence of actions $a^T \triangleq [a_1, \ldots, a_T]$ given a sequence of rewards $r^T \triangleq [r_1, \ldots, r_T]$ as $\pi(a^T \mid r^T) \triangleq \prod_{t=1}^T \pi_t(a_t \mid a_1, r_1, \ldots, a_{t-1}, r_{t-1})$. *If we consider a^T as the output of a bandit algorithm and r^T as the input, $\pi(a^T \mid r^T)$ defines the corresponding probability distribution on the output space required for defining DP* (Eq. (1)). In order to define DP in the sequential setup of bandits, the event-level privacy under continuous observations (Dwork et al., 2010b) framework is adopted. In this framework, if two data sequences differ on a single entry at a single time-step t and are identical on all other time-steps, they are called neighbouring data sequences. A sequential algorithm is differentially private if its output sequence is not distinguishable from two neighbouring input sequences. This framework leads to the *global DP* definition for bandits (Mishra and Thakurta, 2015; Basu et al., 2019).

Definition 2 (Global DP for Bandits)**.** *A bandit algorithm $\pi \in \Pi^{\epsilon}$ satisfies ϵ-global DP, if $\pi(a^T \mid r^T) \leq e^{\epsilon}\pi(a^T \mid r'^T)$ for every action sequence $a^T \in [K]^T$ and every two neighbouring reward sequences $r^T, r'^T \in \mathbb{R}^T$, which means that $\exists j \in [1, T]$ such that $r_j \neq r'_j$ and $\forall\, t \neq j \; r_t = r'_t$. The set of all bandit algorithms satisfying ϵ-global DP, i.e. Π^{ϵ}, is called the ϵ-global DP policy class.*

[5]Neighbouring datasets differ only in one entry, i.e $\|\mathcal{D} - \mathcal{D}'\|_{\text{Hamming}} = 1$.

Table 2: Regret lower bounds for bandits with ϵ-global DP

	Minimax	**Problem Dependent**
Stochastic Multi-armed bandit	$\max\left(\frac{1}{27}\sqrt{T(K-1)}, \frac{1}{131}\frac{K-1}{\epsilon}\right)$	$\sum_{a:\Delta_a>0} \frac{\Delta_a \log(T)}{\min(d_a, 6\epsilon t_a)}$
Stoachastic Linear bandit	$\max\left(\exp(-2), \exp(-6\epsilon d\sqrt{T})\right)\frac{d}{8}\sqrt{T}$	$\inf_{\alpha\in[0,\infty)^{\mathcal{A}}} \sum_{a\in\mathcal{A}} \alpha(a)\Delta_a \log(T)$ s.t. $\lVert a\rVert^2_{H_\alpha^{-1}} \leq \Delta_a \min\left(\frac{\Delta_a}{2}, 3\epsilon\rho(\mathcal{A})\right)$

Now, we formally define the regret of a bandit algorithm with ϵ-global DP, namely π^ϵ, as

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

$N_a(T) \triangleq \sum_{t=1}^{T} \mathbb{1}\{A_t = a\}$ is the number of times the arm a is played till T. $\Delta_a \triangleq \mu^*(\nu) - \mu_a$ is the suboptimality gap of the arm a. The expectation is taken with respect to the probability measure on action-reward sequences induced by the interaction of π^ϵ and ν. *The objective of the bandit algorithm π^ϵ is to satisfy ϵ-global DP while minimising the regret over horizon T. A regret lower bound of any ϵ-global DP bandit algorithm in Π^ϵ quantifies the fundamental hardness of this problem.*

3 Regret Lower Bounds for Bandits with ϵ-global DP

The first question is how much additional regret we have to endure in bandits with ϵ-global DP than the bandits without privacy. The lower bound on regret provides insight into the intrinsic hardness of the problem and serves as a target for the optimal algorithm design. In this section, *we prove minimax and problem-dependent lower bounds for stochastic and linear bandits under ϵ-global DP.* The proposed lower bounds are summarised in Table 2. We defer the proof details to the supplementary.

Stochastic Bandits. First, we consider the stochastic bandit problem with K-arms (as in Section 2).

Minimax Regret. We derive the minimax regret lower bound in this bandit setting with ϵ-global DP. Minimax regret is the lowest achievable regret by any bandit algorithm under the worst environment among a family of environments $\mathcal{E}^K$ with K arms under consideration.

$$\mathrm{Reg}^{\text{minimax}}_{T,\epsilon} = \inf_{\pi^\epsilon\in\Pi^\epsilon} \sup_{\nu\in\mathcal{E}^K} \mathrm{Reg}_T(\pi^\epsilon, \nu)$$

Theorem 2 (Minimax lower bound)**.** *For any $K > 1$ and $T \geq K-1$, and $\epsilon > 0$, the minimax regret of stochastic bandits with ϵ-global DP satisfies*

$$\mathrm{Reg}^{\textit{minimax}}_{T,\epsilon} \geq \max\left\{\frac{1}{27}\underbrace{\sqrt{T(K-1)}}_{\textit{without global DP}}, \ \frac{1}{131}\underbrace{\frac{K-1}{\epsilon}}_{\textit{with } \epsilon\textit{-global DP}}\right\}. \tag{3}$$

Consequences of Theorem 2. Equation (3) shows two regimes of hardness: *high-privacy*, corresponding to lower values of ϵ, and *low-privacy*, corresponding to higher values of ϵ. In the high-privacy regime, specifically for $\epsilon < \frac{131}{27}\sqrt{\frac{(K-1)}{T}}$, the hardness depends only on the number of the arms K and the privacy budget ϵ and is higher than the lower bound for bandits without privacy. In the low-privacy regime, i.e. for $\epsilon \geq \frac{131}{27}\sqrt{\frac{(K-1)}{T}}$, the lower bound coincides with that of the bandits without privacy. This indicates the phenomena that for higher values of ϵ, signifying lower privacy, bandits with global DP is not harder than the bandits without privacy. Specially, for significantly large T, the threshold between low and high privacy regime is smaller than most of the practically used privacy budget values. For example, if $T = 10^6$ and $K = 100$, the bandits with and without global DP are equivalently hard for any privacy budget $\epsilon \geq 0.05$. This shows that for stochastic bandits, we can deploy very low privacy budgets (ϵ) without losing anything in performance.

Problem-dependent Regret: Similar to the lower bounds specific to the Bernoulli reward distributions (Shariff and Sheffet, 2018), the minimax bound indicates a separation between hardness due to global DP and partial information. To understand whether this observation is an artefact of the Bernoulli distribution and the worst-case environment considered in minimax bounds, we derive a lower bound on the problem-dependent regret (Equation (2)) for general reward distributions.

Before deriving the lower bound, we define two information-theoretic terms that quantify the distinguishability of the specific bandit environment ν where the algorithm is operating from all other bandit environments with K arms and finite means. If $\mathcal{M}$ is a set of distributions with finite means, and $\mu(P)$ is the mean of the reward distribution $P \in \mathcal{M}$, we define the *KL-distinguishability gap* as $d_{\inf}(P, \mu^*, \mathcal{M}) \triangleq \inf_{P' \in \mathcal{M}} \{D_{\mathrm{KL}}(P \parallel P') : \mu(P') > \mu^*\}$ and the *TV-distinguishability gap* as $t_{\inf}(P, \mu^*, \mathcal{M}) \triangleq \inf_{P' \in \mathcal{M}} \{\mathrm{TV}(P \parallel P') : \mu(P') > \mu^*\}$. These two quantities indicate the minimum dissimilarity, in terms of KL-divergence and Total Variation (TV) distance, of the reward distribution of the optimal arm a^* with any other distribution with finite mean higher than μ^*. For bandits without DP, inverse of the KL-distinguishability gap quantifies the hardness to identify the optimal arm of the environment ν under partial information (Lai and Robbins, 1985).

Theorem 3 (Problem-dependent Regret Lower Bound)**.** *Let the environment $\mathcal{E}$ be a set of K reward distributions with finite means and a policy $\pi^\epsilon \in \Pi_{cons}(\mathcal{E}) \cap \Pi^\epsilon$ be a consistent policy*[6] *over $\mathcal{E}$ satisfying ϵ-global DP . Then, for all $\nu = (P_i)_{i=1}^K \in \mathcal{E}$, it holds that*

$$\liminf_{T\to\infty} \frac{\mathrm{Reg}_T(\pi^\epsilon, \nu)}{\log(T)} \geq \sum_{a:\Delta_a>0} \frac{\Delta_a}{\min\Big(\underbrace{d_{\inf}(P_a, \mu^*, \mathcal{M}_a)}_{\textit{without global DP}}, \underbrace{6\,\epsilon\, t_{\inf}(P_a, \mu^*, \mathcal{M}_a)}_{\textit{with } \epsilon\textit{-global DP}}\Big)}. \tag{4}$$

Consequences of Theorem 3. Now, we summarise the interesting observations led by the lower bound.

1. *Universality:* The lower bound of Theorem 3 holds for any environment with K arms and reward distributions with finite means. This is the first general lower bound for bandits with ϵ-global DP.

2. *For Bernoulli distributions of reward:* TV-distinguishability gap $t_{\inf}(P_a, \mu^*, \mathcal{M}_a) = \Delta_a$, i.e. the suboptimality gap of arm a, and the KL-distinguishability gap $d_{\inf}(P_a, \mu^*, \mathcal{M}_a) \geq 2\Delta_a^2$. Thus, our problem-dependent lower bound reduces to $\Omega(\sum_{a\neq a^*} \frac{\log T}{\min\{\Delta_a, \epsilon\}})$. For Bernoulli rewards, our lower bound is able to retrieve the lower bound of (Shariff and Sheffet, 2018) with explicit constants.

3. *High- and Low-privacy Regimes:* Like the minimax regret bound, the problem-dependent regret also indicates two clear regimes in regret due to high and low privacy (resp. small and large privacy budgets ϵ). *In the low-privacy regime, i.e. for $\epsilon \geq \frac{d_{\inf}(P_a,\mu^*,\mathcal{M}_a)}{6t_{\inf}(P_a,\mu^*,\mathcal{M}_a)}$, the regret achievable by bandits with global DP and without global DP are same.* Thus, there is no loss in performance due to privacy in this regime. *In the high-privacy regime, i.e. for $\epsilon < \frac{d_{\inf}(P_a,\mu^*,\mathcal{M}_a)}{6t_{\inf}(P_a,\mu^*,\mathcal{M}_a)}$, the regret depends on a coupled effect of privacy and partial information.* This effect is quantified by the inverse of the privacy budget times inverse of the TV-distinguishability gap. This interaction between privacy and partial information was not explicated by any of the existing lower bounds. In Section 4, we propose the AdaP-KLUCB algorithm that demonstrates the same interactive effect in its regret upper bound.

Stochastic Linear Bandits. In order to illustrate the generality of our results, and also to investigate how partial information and privacy in structured bandits, we derive minimax and problem-dependent regret lower bounds for stochastic linear bandits (Lattimore and Szepesvari, 2017). To perform the regret analysis, we consider a simple linear model with parameter $\theta \in \mathbb{R}^d$ and Gaussian noise. It implies that for an action $A_t \in \mathcal{A} \subseteq \mathbb{R}^d$ the reward is $R_t = \langle A_t, \theta\rangle + \eta_t$, where $\eta_t \sim \mathcal{N}(0,1)$ is a sequence of independent Gaussian noises. The regret of a policy is $\mathrm{Reg}_T(\mathcal{A}, \theta) \triangleq \mathbb{E}_\theta\left[\sum_{t=1}^T \Delta_{A_t}\right]$, where suboptimality gap $\Delta_a \triangleq \max_{a'\in\mathcal{A}} \langle a' - a, \theta\rangle$, and $\mathbb{E}_\theta[\cdot]$ is the expectation with respect to the measure on outcomes induced by the interaction of the policy and the linear bandit determined by θ. Given this structure, we state the minimax and problem-dependent regret lower bounds for stochastic linear bandits.

Theorem 4 (Minimax regret lower bound)**.** *Let $\mathcal{A} = [-1,1]^d$ and $\Theta = \left\{-T^{-1/2}, T^{-1/2}\right\}^d$. Then, there exists a vector $\theta \in \Theta$ leading to a minimax regret of linear bandits with ϵ-global DP such that*

$$\mathrm{Reg}_{T,\epsilon}^{\textit{minimax}}(\mathcal{A}, \Theta) \geq \max\Big\{\underbrace{\exp(-2)}_{\textit{without global DP}}, \underbrace{\exp(-6\epsilon d\sqrt{T})}_{\textit{with } \epsilon\textit{-global DP}}\Big\}\frac{d}{8}\sqrt{T}. \tag{5}$$

[6] 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 $\lim_{T\to\infty} \frac{R_T(\pi,\nu)}{T^p} = 0$. We denote the class of consistent policies over a set of environments $\mathcal{E}$ as $\Pi_{\text{cons}}(\mathcal{E})$.

Theorem 5 (Problem-dependent regret lower bound)**.** *Let $\mathcal{A} \subset \mathbb{R}^d$ be a finite set spanning $\mathbb{R}^d$ and $\theta \in \mathbb{R}^d$ be such that there is a unique optimal action. Then, for any consistent and ϵ-global DP bandit algorithm satisfies*

$$\liminf_{T\to\infty} \frac{\mathrm{Reg}_T(\mathcal{A},\theta)}{\log(T)} \geq c(\mathcal{A},\theta), \tag{6}$$

where the structural distinguishability gap *is the solution of a constraint optimisation*

$$c(\mathcal{A},\theta) \triangleq \inf_{\alpha\in[0,\infty)^{\mathcal{A}}} \sum_{a\in\mathcal{A}} \alpha(a)\Delta_a, \text{ such that } \|a\|^2_{H_\alpha^{-1}} \leq \min\left\{ \underbrace{0.5\Delta_a^2}_{\text{without global DP}} \;,\; \underbrace{3\epsilon\rho_a(\mathcal{A})\Delta_a}_{\text{with } \epsilon\text{-global DP}} \right\}$$

for all $a \in \mathcal{A}$ with $\Delta_a > 0$, $H_\alpha = \sum_{a\in\mathcal{A}} \alpha(a)aa^\top$, and a structure dependent constant $\rho_a(\mathcal{A})$.

Remarks. Theorems 4 and 5 are the first minimax and problem-dependent regret lower bounds for linear bandits. The minimax regret bound also shows a clear distinction between high- and low-privacy regimes for $\epsilon < 1/(3d\sqrt{T})$ and $\epsilon \geq 1/(3d\sqrt{T})$. Interestingly, the threshold between two regimes decreases with increase in dimension d and allows us to use quite low values of ϵ without losing any performance. The difference between high- and low-privacy regimes is more subtle. The first constraint on $\|a\|^2_{H_\alpha^{-1}}$ is reminiscent of the Graves-Lai bound for structural bandits without privacy. The second constraint, i.e. $3\epsilon\rho_a(\mathcal{A})\Delta_a$, captures the interaction between privacy and partial information under linear structure and in the high-privacy regime.

Proof Technique. In order to prove the lower bounds, we adopt the general canonical bandit model. The high level idea of proving bandit lower bounds is selecting two *hard* environments, which are hard to be statistically distinguished but are conflicting, i.e. actions that may be optimal in one is suboptimal in other. This is quantified by upper bounding the per-step KL-divergence between action distributions for such hard environments, and then plugging in this upper bound in the Bretagnolle-Huber inequality to obtain a regret lower bound. Though this proof technique works to quantify the hardness due to partial information, we need to upper bound the "confusion" created due to global DP. Existing proofs use the information-processing lemma of Karwa-Vadhan (Karwa and Vadhan, 2017). Though this works for Bernoulli bandits, to proof a general bound, we derive a sequential version of this lemma. The sequential version leads to an upper bound of the KL-divergence dependent on both total-variation distinguishability gap and privacy budget, and enables us to show the coupled effect of privacy and partial-information in the high-privacy regimes. For details of the technical results, we refer to the Supplementary Material.

4 Stochastic Bandit Algorithms with ϵ-global DP: AdaP-UCB & AdaP-KLUCB

In this section, *we present a framework (Algorithm 1) to convert index-based optimistic algorithms into algorithms satisfying ϵ-global DP*. We instantiate this framework by proposing AdaP-UCB and AdaP-KLUCB, and derive corresponding regret upper bounds.

Algorithm 1 A framework for ϵ-global DP extension of an index-based bandit algorithm

Input: Privacy budget ϵ, an environment ν with K arms, parameter $\alpha > 3$
Output: A sequence of T-actions satisfying ϵ-global DP
Initialisation: Choose each arm once and assign $t = K$
for $\ell = 1, 2, \ldots$ **do** ▷ Adaptive episodes of reward subsequences
  Let $t_\ell = t + 1$
  Compute $A_\ell = \mathrm{argmax}_a \, \mathrm{I}^\epsilon_a(t_\ell - 1, \alpha)$ ▷ Arm selection with private indexes (Eqn. (7)-(8))
  Choose arm A_ℓ until round t such that $N_{A_\ell}(t) = 2N_{A_\ell}(t_\ell - 1)$ ▷ Doubling of episodes
end for

Index-Based Algorithms and Private Empirical Mean. We focus on index-based bandit algorithms that compute the empirical mean of rewards of each arm at each step. *Using the empirical means, they compute an optimistic index for each arm that serves as a high-probability upper confidence bound on the true mean of the corresponding arm.* Examples of such algorithms with empirical mean-based indexes include UCB (Auer et al., 2002), MOSS (Audibert and Bubeck, 2010), KL-UCB (Garivier and Cappé, 2011), IMED (Honda and Takemura, 2015) etc. Here, the empirical mean is the main

statistics of reward sequences used by the algorithms. Thus, by post-processing lemma, *designing an ϵ-global DP bandit algorithm reduces to computing the empirical means privately.*

By Theorem 1, we know that adding a Laplacian noise with scale $s(\hat{\mu}_t)/\epsilon$ to each empirical mean turns it ϵ-DP, where $s(\hat{\mu}_t)$ is the sensitivity of empirical mean $\hat{\mu}_t$ at step t. For K arms and horizon T, an index-based bandit algorithm computes the empirical mean KT times. Using the naïve composition of ϵ-DP, a first baseline is to make each computed empirical mean $\frac{\epsilon}{KT}$-DP. This needs us to add at each step noise scaled linearly to T. This might lead the corresponding bandit algorithm to yield a linear regret. Thus, we need to add noise less number of times and with lower sensitivity.

Empirical Mean using Reward Sub-sequence. The improvement is invoked by the observation that a bandit algorithm does not need to calculate the empirical mean at each time-step using all the rewards observed till that step. Thus, motivated by the DP-SE (Sajed and Sheffet, 2019), we calculate the empirical mean less number of times by exploiting an episodic structure that we will explicit later. Specifically, we use only the rewards of the last episode. We formally express this trick in Lemma 1.

Lemma 1. *Let us define the private empirical mean of the rewards between steps i and j $(i < j)$ as*

$$f^{\epsilon}\{r_i, \ldots, r_j\} \triangleq \frac{1}{j-i} \sum_{t=i}^{j} r_t + Lap\left(\frac{1}{(j-i)\epsilon}\right).$$

If $1 < t_1 < \cdots < t_\ell < T$ and $r_t \in [0,1]$, the mechanism g^{ϵ} mapping the sequence of rewards $(r_1, r_2, \ldots, r_T)$ to $(\ell+1)$-private empirical means $(f^{\epsilon}\{r_1, \ldots, r_{t_1-1}\}, f^{\epsilon}\{r_{t_1}, \ldots, r_{t_2-1}\}, \ldots, f^{\epsilon}\{r_{t_\ell}, \ldots, r_T\})$ satisfies ϵ-DP.

Lemma 1 implies that, if we calculate the empirical mean of each arm $\ell+1$ $(\ll T)$ times on non-overlapping sub-sequences of the reward stream, we only need to ensure that each empirical mean is ϵ-DP with respect to the corresponding reward sub-sequence. These $(\ell+1)$ private empirical means together ensure that the sequence of $(\ell+1)$ computed indexes and the resultant action sequence of length T satisfy ϵ-global DP by the post-processing lemma (Dwork and Roth, 2014). *Thus, we divide the horizon into $(\ell+1)$-episodes. This allows us to take T actions only by computing $\ell+1$ private empirical means with the rewards observed in the last active episode of the arm.* Specifically, for each $t \in [t_i, t_{i+1}-1]$, we play the same arm that was decided at the beginning of the episode, i.e. at t_i, with the private empirical mean computed from the last active episode of that arm.

Adaptive Episodes with Doubling. In order to set the episode lengths, we focus on the specific structure of the bandit process. Specifically, we know that a sub-optimal arm a should be sampled at least $n_a = \mathcal{O}\left(\frac{\log(T)}{\Delta_a^2}\right)$ times before discarding. Since the empirical means are computed using only rewards of one episode, to yield a near-optimal regret, the length of the episode should be greater at any given time than n_a. *To satisfy this criterion, we deploy adaptive episodes with doubling.*

Example 2 (Illustration of Algortihm 1.)**.** *To clarify the schematic, we illustrate a few steps of executing Algorithm 1 in Figure 1. After playing each arm once, the first episode begins at t_1. To observe different ingredients, we focus on step $t_4 = 7$. The index of Arm 1 at t_4 uses the private empirical mean $\frac{r_4+r_5}{2} + Lap(\frac{1}{2\epsilon})$ to build a high probability upper bound of the real mean μ_1 with confidence $t_4^{-\alpha}$. The index of Arm 2 uses $r_6 + Lap(\frac{1}{\epsilon})$. If we assume that the index of Arm 1 is higher at t_4, Arm 1 is played for a full episode from t_4 until $t_5 - 1$. The last time, when Arm 1 was played, the length of the episode was 2. Thus, following t_4, the length of the episode is doubled to 4.*

Theorem 6 (ϵ-global DP)**.** *If I_a^{ϵ} is computed using only the private empirical mean of the rewards collected in the last active episode of arm a, Algorithm 1 satisfies ϵ-global DP.*

Proof. The main idea is that a change in reward will only affect the empirical mean calculated in one episode, which is made private using the Laplace Mechanism and Lemma 1. Since the actions are only calculated using the private empirical means, the algorithm is ϵ-global DP following the post-processing lemma. We refer to Appendix D for a complete proof. □

Now, to concretise an algorithm, we only need to explicit how the indexes are calculated. In other words, we need to build a high-probability upper confidence bound on the true mean of an arm only using a private empirical mean. We instantiate the design details for AdaP-UCB and AdaP-KLUCB.

AdaP-UCB **and** AdaP-KLUCB**.** Let $\hat{\mu}_a^{\ell}$ be the empirical mean reward of arm a computed using the samples collected between $t_{\psi_a(\ell)}$ and $t_{\psi_a(\ell)+1} - 1$. For an episode ℓ, $\psi_a(\ell) = \ell_a$ is the last active

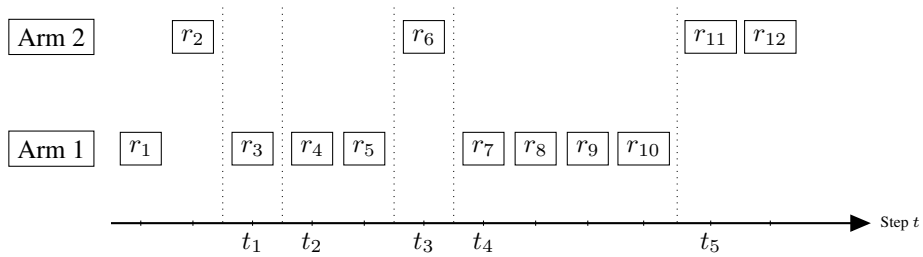


Figure 1: An illustration of adaptive episodes with per-arm doubling.

episode of arm a. In Example 2, $\psi_1(4) = 2$ and $\psi_2(4) = 3$. Thus, due to the doubling of episode length, the empirical mean corresponds to $\frac{1}{2}N_a(t_\ell - 1)$ samples of arm a. Since the rewards are in $[0, 1]$, the private empirical mean as $\tilde{\mu}^{\ell}_{a,\epsilon} = \hat{\mu}^{\ell}_a + Lap\left(\frac{2}{\epsilon N_a(t_\ell - 1)}\right)$ satisfies ϵ-DP (Theorem 1). Now, we want to ensure that $\mathrm{I}^{\epsilon}_a(t_\ell - 1, \alpha)$, computed using only $\tilde{\mu}^{\ell}_{a,\epsilon}$, is a high-probability upper bound on the true mean. Here, we introduce two specific indexes that satisfy this criterion.

For AdaP-UCB:
$$\mathrm{I}^{\epsilon}_a(t_\ell - 1, \alpha) = \tilde{\mu}^{\ell}_{a,\epsilon} + \sqrt{\frac{\alpha \log(t_\ell)}{2 \times \frac{1}{2}N_a(t_\ell - 1)}} + \frac{\alpha \log(t_\ell)}{\epsilon \times \frac{1}{2}N_a(t_\ell - 1)} \tag{7}$$

For AdaP-KLUCB:
$$\mathrm{I}^{\epsilon}_a(t_\ell - 1, \alpha) = \max\left\{q \in [0, 1] : d\left(\breve{\mu}^{\ell,\alpha}_{a,\epsilon}, q\right) \leq \frac{\alpha \log(t_\ell)}{\frac{1}{2}N_a(t_\ell - 1)}\right\}, \tag{8}$$

where $\breve{\mu}^{\ell,\alpha}_{a,\epsilon} = \mathrm{Clip}_{0,1}\left(\tilde{\mu}^{\ell}_{a,\epsilon} + \frac{\alpha \log(t_\ell)}{\epsilon \frac{1}{2}N_a(t_\ell - 1)}\right) = \min\{\max\{0, \tilde{\mu}^{\ell}_{a,\epsilon} + \frac{\alpha \log(t_\ell)}{\epsilon \times \frac{1}{2}N_a(t_\ell - 1)}\}, 1\}$ is the private empirical mean clipped between zero and one.

Theorem 7. *For rewards in* $[0, 1]$, AdaP-UCB *satisfies* ϵ*-global DP, and for* $\alpha > 3$*, it yields a regret*

$$\mathrm{Reg}_T(\text{AdaP-UCB}, \nu) \leq \sum_{a:\Delta_a > 0} \left(\frac{16\alpha}{\min\{\Delta_a, \epsilon\}} \log(T) + \frac{3\alpha}{\alpha - 3}\right).$$

Theorem 8. *When the rewards are sampled from Bernoulli distributions,* AdaP-KLUCB *satisfies* ϵ*-global DP, and for* $\alpha > 3$ *and constants* $C_1(\alpha), C_2 > 0$*, it yields a regret*

$$\mathrm{Reg}_T(\text{AdaP-KLUCB}, \nu) \leq \sum_{a:\Delta_a > 0} \left(\frac{C_1(\alpha)\Delta_a}{\min\{d_{\inf}(\mu_a, \mu^*), C_2\epsilon\Delta_a\}} \log(T) + \frac{3\alpha}{\alpha - 3}\right).$$

Both the upper bounds show that for low-privacy regime, the regrets of AdaP-UCB and AdaP-KLUCB are independent of ϵ, and in high-privacy regime, they depend on the coupled effect. In Appendix E.4, we also derive problem-independent or minimax regret upper bounds for AdaP-UCB and AdaP-KLUCB, which are $\mathcal{O}(\sqrt{KT\log(T)} + \frac{K\log(T)}{\epsilon})$.

Remark 1. *α controls the width of the optimistic confidence bound. Specifically, it dictates that the real mean is smaller than the optimistic index with high probability, i.e. with probability $1 - \frac{1}{t^\alpha}$ at step t. The requirement that $\alpha > 3$ is due to our analysis of the algorithm. To be specific, the requirement that $\alpha > 3$ is needed to use a sum-integral inequality in the proof of Theorem 11. Since the dominant terms in the regret upper bounds of both* AdaP-UCB *and* AdaP-KLUCB *are multiplicative in α, making α smaller will tighten the bound. We leave it for future work. We refer to Section G.3 for a detailed discussion on choosing α.*

Summary of Algorithm Design. We propose three ingredients to design an ϵ-global DP version of a bandit algorithm (Algo. 1). Firstly, add Laplace noise to make the empirical means private (Lemma 1). Secondly, compute the index of an arm using the private empirical mean of rewards collected from the last active episode of that arm and forget all the samples obtained before that. Thirdly, use adaptive episodes with doubling. At the beginning of an episode choose the arm with the highest index and play it double number of times than the length of the last active episode of the arm.

5 Experimental Analysis

In this section, we perform empirical evaluations to test two hypotheses: (i) AdaP-KLUCB is the most optimal algorithm among the existing bandit algorithms with ϵ-global DP, and (ii) the

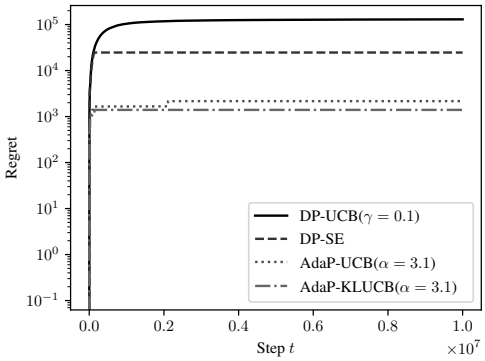


Figure 2: Evolution of regret over time for DP-UCB, DP-SE, AdaP-UCB, and AdaP-KLUCB with $\epsilon = 1$. Each algorithm is run 20 times with $T = 10^7$, and Bernoulli distributions with means $\{0.75, 0.625, 0.5, 0.375, 0.25\}$. AdaP-KLUCB achieves the lowest regret.

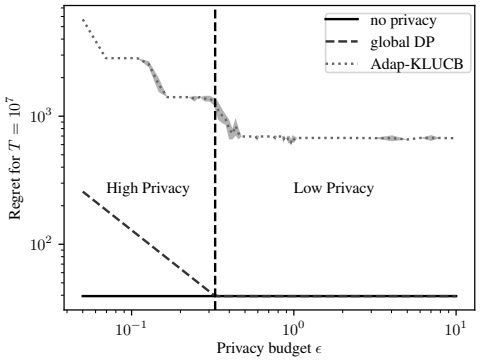


Figure 3: Dependence of lower bounds and regret of AdaP-KLUCB with respect to the privacy budget ϵ. We run AdaP-KLUCB for 20 runs with $T = 10^7$. Echoing the theoretical analysis, regret of AdaP-KLUCB transits between privacy regimes and is independent of ϵ for low-privacy.

transition between high and low-privacy regimes is reflected in the empirical performance. Additional experimental results are deferred to Appendix G.

Efficiency in Performance. First, we compare performances of AdaP-UCB and AdaP-KLUCB with those of DP-SE and DP-UCB. We set $\alpha = 3.1$ to comply with the regret upper bounds of AdaP-UCB and AdaP-KLUCB. We assign $\gamma = 0.1$ for DP-UCB and $\beta = 1/T$ for DP-SE. We implement all the algorithms in Python (version 3.8) and on an 8 core 64-bits Intel i5@1.6 GHz CPU. 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)). We run each algorithm 20 times for a horizon $T = 10^7$, and plot corresponding average and standard deviations of regrets in Figure 2. AdaP-KLUCB achieves the lowest regret followed by AdaP-UCB. Both of them achieve 10 times lower regret than the competing algorithms.

Impact of Privacy Regimes. In Figure 3, we plot regret of AdaP-KLUCB at $T = 10^7$ for a Bernoulli bandit with mean rewards $\{0.8, 0.1, 0.1, 0.1, 0.1\}$. We plot the average regret over 20 runs as a function of the privacy budget $\epsilon \in [0.05, 10]$. As indicated by the theoretical regret lower bounds and upper bounds, the experimental performance of AdaP-KLUCB demonstrates two regimes: a high-privacy regime (for $\epsilon < 0.3$), where the regret of AdaP-KLUCB depends on the privacy budget ϵ, and a low privacy regime (for $\epsilon > 0.3$), where the regret of AdaP-KLUCB does not depend on ϵ.

In brief, *our experimental results validate that* AdaP-KLUCB *is the most optimal algorithm for stochastic bandits that satisfies ϵ-global DP, and performance of* AdaP-KLUCB *transits from high- to low-privacy regimes, where its performance turns independent of the privacy budget ϵ.*

6 Conclusion and Future Works

We revisit bandits with ϵ-global DP. We prove the minimax and problem-dependent regret lower bounds for stochastic and linear bandits indicating two regimes of hardness. We propose a framework to design an ϵ-global DP version of an index-based optimistic bandit algorithm in a near-optimal fashion by utilising three ingredients: adding noise with Laplace Mechanism, using the empirical mean of rewards collected in the last active episode of an arm, and adapting episodes with doubling.

One limitation of our analysis that the lower and upper bounds defer by multiplicative constants. It would be a technical challenge to merge this gap. Another future direction is to derive regret lower bounds for other variants of DP, namely (ϵ, δ)-DP and Rényi-DP (Mironov, 2017), especially in the structured bandits compatible with Gaussian noise injection and also to extend the proposed proof techniques in these settings. For the algorithm design, it would be interesting to see how the proposed ingredients generalise to linear and contextual bandits (Shariff and Sheffet, 2018), and Markov Decision Processes (Vietri et al., 2020).

Acknowledgments and Disclosure of Funding

This work is supported by the AI_PhD@Lille grant. We also thank Philippe Preux for his support.

References

Agrawal, S. and Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR.

Audibert, J.-Y. and Bubeck, S. (2010). Regret bounds and minimax policies under partial monitoring. *The Journal of Machine Learning Research*, 11:2785–2836.

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

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*.

Chan, T.-H. H., Shi, E., and Song, D. (2011). Private and continual release of statistics. *ACM Trans. Inf. Syst. Secur.*, 14(3).

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

Dwork, C., Naor, M., Pitassi, T., and Rothblum, G. N. (2010a). Differential privacy under continual observation. In *Proceedings of the Forty-Second ACM Symposium on Theory of Computing*, STOC '10, page 715–724, New York, NY, USA. Association for Computing Machinery.

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

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

Even-Dar, E., Mannor, S., and Mansour, Y. (2002). PAC bounds for multi-armed bandit and Markov decision processes. In *Proceedings of the 15th Annual Conference on Computational Learning Theory*, COLT '02, page 255–270, Berlin, Heidelberg. Springer-Verlag.

Garivier, A. and Cappé, O. (2011). The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376. JMLR Workshop and Conference Proceedings.

Honda, J. and Takemura, A. (2015). Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards. *J. Mach. Learn. Res.*, 16:3721–3756.

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.

Hu, B., Huang, Z., and Mehta, N. A. (2021). Optimal algorithms for private online learning in a stochastic environment.

Karwa, V. and Vadhan, S. (2017). Finite sample differentially private confidence intervals.

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.

Lattimore, T. and Szepesvari, C. (2017). The end of optimism? An asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737. PMLR.

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

Mironov, I. (2017). Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE.

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.

Schwartz, E. M., Bradlow, E. T., and Fader, P. S. (2017). Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522.

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.

Tenenbaum, J., Kaplan, H., Mansour, Y., and Stemmer, U. (2021). Differentially private multi-armed bandits in the shuffle model. *Advances in Neural Information Processing Systems*, 34:24956–24967.

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*.

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

Zheng, K., Cai, T., Huang, W., Li, Z., and Wang, L. (2020). Locally differentially private (contextual) bandits learning.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] It is a purely theoretical paper
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] We put proof sketch in the main paper and the full proofs are in Supplementary Material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The full code is in the Supplementary Material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] The details are in Section 5 and in the Supplementary Materials.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] We test the algorithms in a work PC with an 8 core 64-bits Intel i5@1.6 GHz CPU and without using any GPU or cloud.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

Table of Contents

A The Canonical Model of Bandits **15**

B Distinguishing Environments with Partial Information and Global DP **16**
B.1 Sequential Karwa-Vadhan Lemma 16
B.2 KL-divergence Decomposition with ϵ-global DP 19

C Lower Bounds on Regret: Stochastic and Linear Bandits with ϵ-global DP **20**
C.1 Stochastic Bandits: Minimax Lower Bound 20
C.2 Stochastic Bandits: Problem-dependent Lower Bound 21
C.3 Stochastic Linear Bandits: Minimax Lower Bound 23
C.4 Stochastic Linear Bandits: Problem-dependent Lower Bound 24

D Privacy Analysis of Algorithm 1 **28**

E Upper Bounds on Regret: AdaP-UCB and AdaP-KLUCB **29**
E.1 Concentration Inequalities 29
E.2 Generic Regret Analysis for Algorithm 1 30
E.3 Regret Analysis for AdaP-UCB and AdaP-KLUCB 33
E.4 Problem-independent Regret Bounds 37

F Existing Technical Results and Definitions **38**

G Extended Experimental Analysis **41**
G.1 Experimental Setup 41
G.2 Experimental Results 41
G.3 Choice of α 41

A The Canonical Model of Bandits

We extend the general canonical model of bandits (Chapter 4, Lattimore and Szepesvári (2018)) with ϵ-global differential privacy. The canonical model with ϵ-global DP consists of *a privacy-preserving policy* π^ϵ and *an environment* ν. The policy interacts with the environment 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 are:

1. the probability of choosing an action $A_t = a$ at time t is dictated only by the policy $\pi^\epsilon(a|\mathcal{H}_{t-1})$,
2. the distribution of reward R_t is P_{A_t} and is conditionally independent of the previous observed history $\mathcal{H}_{t-1}$.

Let us formalise this interaction by defining an ϵ-global DP policy, the environment and the probability space produced by this interaction.

Let $T \in \mathbb{N}$ be the horizon. Let $\nu = (P_a : a \in [K])$ a bandit instance with K arms. For each $t \in [T]$, let $\Omega_t = ([K] \times \mathbb{R})^t \subset \mathbb{R}^{2t}$ and $\mathcal{F}_t = \mathfrak{B}(\Omega_t)$ with $\mathfrak{B}$ being the Borel set.

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

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

and for a sequence of actions $a^T \triangleq [a_1, \ldots, a_T]$ and a sequence of rewards $r^T \triangleq [r_1, \ldots, r_T]$:

$$\pi(a^T \mid r^T) = \prod_{t=1}^T \pi_t(a_t \mid a_1, r_1, \ldots, a_{t-1}, r_{t-1})$$

A policy π^ϵ is ϵ-global DP, if

$$\pi^\epsilon(a^T \mid r^T) \leq e^\epsilon \pi^\epsilon(a^T \mid r'^T)$$

for every sequence of actions a^T and every two neighbouring reward streams r^T, r'^T: $\exists j \in [1, T]$ such that $r_j \neq r'_j$ and $\forall\, t \neq j\ r_t = r'_t$.

Let λ be a σ-finite measure on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ for which P_a is absolutely continuous with respect to λ for all $a \in [K]$. Let $p_a = dP_a/d\lambda$ be the Radon–Nikodym derivative of P_a with respect to λ, which is a function $p_a : \mathbb{R} \to \mathbb{R}$ such that $\int_B p_a d\lambda = P_a(B)$ for all $B \in \mathfrak{B}(\mathbb{R})$. Letting ρ be the counting measure with $\rho(B) = |B|$, the density $p_{\nu\pi^\epsilon} : \Omega_T \to \mathbb{R}$ can now be defined with respect to the product measure $(\rho \times \lambda)^T$ by

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

and $\mathcal{P}_{\nu\pi^\epsilon}$ be defined by

$$\mathcal{P}_{\nu\pi^\epsilon}(B) \triangleq \int_B p_{\nu\pi^\epsilon}(\omega)(\rho \times \lambda)^T(\,\mathrm{d}\omega) \quad \text{for all } B \in \mathcal{F}_T$$

Hence $(\Omega_T, \mathcal{F}_T, \mathcal{P}_{\nu\pi^\epsilon})$ is a probability space over histories induced by the interaction between π^ϵ and ν.

We define also a marginal distribution over sequence of actions by

$$m_{\nu\pi^\epsilon}(a_1, \ldots, a_T) \triangleq \int_{r_1, \ldots, r_T} p_{\nu\pi^\epsilon}(a_1, r_1, \ldots, a_T, r_T)\,\mathrm{d}r_1 \ldots \mathrm{d}r_T,$$

and for all $C \in \mathcal{P}([K]^T)$,

$$M_{\nu\pi^\epsilon}(C) \triangleq \sum_{(a_1, \ldots, a_T) \in C} m_{\nu\pi^\epsilon}(a_1, a_2, \ldots, a_T).$$

Hence, $([K]^T, \mathcal{P}([K]^T), M_{\nu\pi^\epsilon})$ is a probability space over sequence of actions produced when π^ϵ interacts with ν for T time-steps.

B Distinguishing Environments with Partial Information and Global DP

In this section, we first revisit the Karwa-Vadhan Lemma (Lemma 6.1, (Karwa and Vadhan, 2017)) that bounds the multiplicative distance between marginal distributions induced by a differentially private mechanism, when the datasets are generated using two different distributions $\mathbb{P}$ and $\mathbb{Q}$. *We generalise this result to the setting where the inputs are not identically distributed.* We call this Sequential Karwa-Vadhan Lemma (Lemma 2) and apply it to upper bound the Kullback-Leibler (KL) divergence between the marginal distributions $M_{\nu\pi^\epsilon}$ and $M_{\nu'\pi^\epsilon}$, when π^ϵ is an ϵ-global DP policy, and ν and ν' are two different environments (Theorem 10).

Karwa-Vadhan Lemma. Let $\mathbb{P}$ and $\mathbb{Q}$ be two distributions, and $\mathrm{TV}\left(\mathbb{P} \parallel \mathbb{Q}\right)$ be the total variation distance between these two distributions. Let $\mathcal{M}$ be an (ϵ, δ)-differentially private mechanism that runs on the set of samples $\{x_1, \ldots, x_T\}$. For any event E in $\mathcal{M}$'s output space, $\mathcal{M}(E|X_1 = x_1, \ldots, X_T = x_T)$ denotes the probability that $\mathcal{M}$ outputs an element in E given the input $x_1, \ldots, x_T$, and

$$\mathbb{M}_{\mathbb{P}}(E) \triangleq \int \mathcal{M}(E|X_1, \ldots, X_T)\, \mathrm{d}\mathbb{P}(X_1, \ldots, X_T)$$

is the marginal distribution induced by the DP mechanism when the data is generated from the distribution $\mathbb{P}$.

Theorem 9 (Lemma 6.1, Karwa and Vadhan (2017))**.** *If a mechanism $\mathcal{M}$ satisfies (ϵ, δ)-DP, then for every event E in the output space of $\mathcal{M}$, the marginal distributions induced by distributions $\mathbb{P}$ and $\mathbb{Q}$ satisfy*

$$\mathbb{M}_{\mathbb{P}}(E) \leq e^{\epsilon'} \mathbb{M}_{\mathbb{Q}}(E) + \delta',$$

where $\epsilon' \triangleq (6\epsilon T)\mathrm{TV}\left(\mathbb{P} \parallel \mathbb{Q}\right)$ and $\delta' \triangleq (4e^{\epsilon'} T\delta)\mathrm{TV}\left(\mathbb{P} \parallel \mathbb{Q}\right)$.

We extend this result for the setting where the data is not identically distributed.

B.1 Sequential Karwa-Vadhan Lemma

Let $\{\mathbb{P}_1, \ldots, \mathbb{P}_T\}$ and $\{\mathbb{Q}_1, \ldots, \mathbb{Q}_T\}$ two sets of independent distributions.

Given the samples $X_1, \ldots, X_T$ generated from the distributions $\mathbb{P}_1, \ldots, \mathbb{P}_T$, we define the corresponding marginal distribution induced by $\mathcal{M}$ as

$$\mathbb{M}_{\mathbb{P}_1,\ldots,\mathbb{P}_T}(E) \triangleq \int \mathcal{M}(E|X_1, \ldots, X_T)\, \mathrm{d}\mathbb{P}_1(X_1)\, \mathrm{d}\mathbb{P}_2(X_2) \ldots \mathrm{d}\mathbb{P}_T(X_T)$$

Lemma 2 (Sequential Karwa-Vadhan Lemma)**.** *If $\mathcal{M}$ is a mechanism satisfying (ϵ, δ)-DP, then for every event E in the output space of $\mathcal{M}$, the marginal distributions induced by the two sets of independent distributions $\{\mathbb{P}_1, \ldots, \mathbb{P}_T\}$ and $\{\mathbb{Q}_1, \ldots, \mathbb{Q}_T\}$ satisfy*

$$\mathbb{M}_{\mathbb{P}_1,\ldots,\mathbb{P}_T}(E) \leq e^{\epsilon'} \mathbb{M}_{\mathbb{Q}_1,\ldots,\mathbb{Q}_T}(E) + \delta',$$

where $\epsilon' = 6\epsilon \sum_{i=1}^{T} \mathrm{TV}\left(\mathbb{P}_i \parallel \mathbb{Q}_i\right)$ and $\delta' = 4e^{\epsilon'}\delta \sum_{i=1}^{T} \mathrm{TV}\left(\mathbb{P}_i \parallel \mathbb{Q}_i\right)$

Proof. We extend the proof proposed by (Karwa and Vadhan, 2017) to the non-identical distribution setting. The main observation is that the proof follows naturally if the data is generated from different distributions by just adapting the coupling to the case of different distributions. For completeness, we present the whole proof with all the adapted changes.

We construct a coupling between $\bigotimes_{i=1}^{T} \mathbb{P}_i$ and $\bigotimes_{i=1}^{T} \mathbb{Q}_i$ that allows us to control the hamming distance between samples generated from this distributions.

Let us denote $p_i \triangleq \mathrm{TV}\left(\mathbb{P}_i \parallel \mathbb{Q}_i\right)$, $F_i \triangleq \max(\mathbb{P}_i - \mathbb{Q}_i, 0)$, $G_i \triangleq \max(\mathbb{Q}_i - \mathbb{P}_i, 0)$, and $C_i \triangleq \min(\mathbb{P}_i, \mathbb{Q}_i)$. It is easy to see that $\mathbb{P}_i = F_i + C_i$ and $\mathbb{Q}_i = G_i + C_i$.

Given the aforementioned notations, we consider the following algorithm to generate $2T$ samples:

For $i = 1$ to T, generate H_i from Bernoulli(p_i)

(a) If $H_i = 1$, sample $X_i \propto F_i$ and $X_i' \propto G_i$
(b) If $H_i = 0$, sample $X_i \propto C_i$ and set $X_i' = X_i$.

Here $X_i \propto F_i$ means that X_i is generated from a distribution defined by normalizing F_i.
This construction satisfies the following properties:

1. $\underline{X} \triangleq (X_1, \ldots, X_T) \sim \bigotimes_{i=1}^{T} \mathbb{P}_i \triangleq \mathbb{D}_0$.
2. $\underline{X}' \triangleq (X_1', \ldots, X_T') \sim \bigotimes_{i=1}^{T} \mathbb{Q}_i \triangleq \mathbb{D}_1$.
3. $\|\underline{X} - \underline{X}'\|_{\text{Hamming}} = \sum_{i=1}^{T} H_i \triangleq H$.

Now, we introduce the following shorthand for the marginal distributions at step h

$$m_j(h) \triangleq \int_{\underline{x}} \mathcal{M}(E|\underline{X} = \underline{x})\, \mathrm{d}\mathbb{D}_j(\underline{X}|H = h)$$

for $j \in \{0, 1\}$ and $p(h) = \mathbb{P}(H = h)$. For $j \in \{0, 1\}$ and any event E, we have, by definition,

$$\mathbb{M}_j(E) = \sum_{h=0}^{T} m_j(h)p(h)$$

Fact 1: For $j \in \{0, 1\}$, $m_j(h) \leq e^{\epsilon} m_j(h-1) + \delta$ for $h = 1, \ldots, T$, and $m_1(0) = m_0(0)$.
We defer the proof of Fact 1 to the end of this proof.
By Fact 1, for $j \in \{0, 1\}$, we have

$$m_j(h) \leq e^{h\epsilon} m_j(0) + \frac{e^{h\epsilon} - 1}{e^{\epsilon} - 1}\delta$$

Now, we obtain

$$\begin{aligned}
\mathbb{M}_j(E) &= \sum_{h=0}^{T} p(h)m_j(h) \\
&= \mathbb{E}[m_j(H)] \\
&\leq \mathbb{E}[e^{H\epsilon} m_j(0) + \frac{e^{H\epsilon} - 1}{e^{\epsilon} - 1}\delta] \\
&= m_j(0) \cdot \mathbb{E}[e^{H\epsilon}] + \frac{\delta}{e^{\epsilon} - 1} \cdot \big(\mathbb{E}[e^{H\epsilon}] - 1\big) \\
&= m_j(0) \cdot \prod_{i=1}^{T}(1 - p_i + p_i \cdot e^{\epsilon}) + \frac{\delta}{e^{\epsilon} - 1} \cdot \left(\prod_{i=1}^{T}(1 - p_i + p_i \cdot e^{\epsilon}) - 1\right) \qquad (9)
\end{aligned}$$

The last equality holds due to that fact that for any $t > 0$, $\mathbb{E}[e^{tH}] = \prod_{i=1}^{T}(1 - p_i + p_i \cdot e^{t})$.
Similarly, we obtain

$$\mathbb{M}_j(E) \geq m_j(0) \prod_{i=1}^{T}(1 - p_i + p_i \cdot e^{-\epsilon}) + \frac{\delta}{e^{-\epsilon} - 1} \cdot \left(\prod_{i=1}^{T}(1 - p_i + p_i \cdot e^{-\epsilon}) - 1\right) \qquad (10)$$

Combining inequalities 9 and 10, we get

$$\mathbb{M}_0(E) \leq \left[\prod_{i=1}^{T}\left(\frac{1 - p_i + p_i \cdot e^{\epsilon}}{1 - p_i + p_i \cdot e^{-\epsilon}}\right)\right] \cdot \left(\mathbb{M}_1(E) + \frac{1 - \prod_{i=1}^{T}(1 - p_i + p_i \cdot e^{-\epsilon})}{1 - e^{-\epsilon}} \cdot \delta\right)$$

$$+ \frac{\prod_{i=1}^{T}(1-p_i+p_i\cdot e^{-\epsilon})-1}{e^{\epsilon}-1}\cdot\delta \tag{11}$$

From Lemma 6.1 of (Karwa and Vadhan, 2017), we know that

$$\log\left(\frac{1-p_i+p_i\cdot e^{\epsilon}}{1-p_i+p_i\cdot e^{-\epsilon}}\right) \leq 6\epsilon p_i,$$

Thus,

$$\prod_{i=1}^{T}\left(\frac{1-p_i+p_i\cdot e^{\epsilon}}{1-p_i+p_i\cdot e^{-\epsilon}}\right) \leq e^{6\epsilon\sum_{i=1}^{T}p_i} \triangleq e^{\epsilon'}, \tag{12}$$

and

$$e^{\epsilon'}\cdot\frac{1-\prod_{i=1}^{T}(1-p_i+p_i\cdot e^{-\epsilon})}{1-e^{-\epsilon}}\cdot\delta+\frac{\prod_{i=1}^{T}(1-p_i+p_i\cdot e^{\epsilon})-1}{e^{\epsilon}-1}\cdot\delta \tag{13}$$

$$\leq e^{\epsilon'}\cdot\frac{1-\exp(2(\sum_{i=1}^{T}p_i)\cdot(e^{-\epsilon}-1))}{1-e^{-\epsilon}}\cdot\delta+\frac{\exp(2(\sum_{i=1}^{T}p_i)\cdot(e^{\epsilon}-1))-1}{e^{\epsilon}-1}\cdot\delta \tag{14}$$

$$\leq e^{\epsilon'}\cdot 2(\sum_{i=1}^{T}p_i)\cdot\delta+2(\sum_{i=1}^{T}p_i)\cdot\delta \tag{15}$$

$$\leq e^{\epsilon'}\cdot 4\sum_{i=1}^{T}p_i\cdot\delta. \tag{16}$$

Substituting Equations (12) and (16) in Equation 11, we obtain

$$\mathbb{M}_0(E) \leq e^{\epsilon'}\mathbb{M}_1(E)+\delta',$$

where $\epsilon' = 6\epsilon(\sum_{i=1}^{T}p_i)$ and $\delta' = 4e^{\epsilon'}\delta(\sum_{i=1}^{T}p_i)$. □

Now, we prove Fact 1.

Fact 1. *For $j\in\{0,1\}$, $m_j(h)\leq e^{\epsilon}m_j(h-1)+\delta$ for $h=1,\ldots,T$, and $m_1(0)=m_0(0)$.*

Proof. We prove the claim for $j=0$, the other case is similar.

First, let us introduce some notations. Fix a $(h_1,\ldots,h_T)\in\{0,1\}^T$. Let $I'\triangleq\{i:h_i=1\}$, $J\triangleq\{i:h_i=0\}$, and r be any fixed index in I'. Let $I=I'/\{r\}$ and consider the following partition of $\underline{X}$ into three parts:

$$\underline{X}=(\underline{X}_I,X_r,\underline{X}_J),$$

where $\underline{X}_I$ is the vector $\underline{X}$ specified by the indices in I. By definition of the coupling, $\underline{X}_I\sim\bigotimes_{i\in I}F_i\triangleq F_I$, $X_r\sim F_r$, $\underline{X}_J\sim\bigotimes_{i\in J}C_i\triangleq C_J$. Now, let $X_r'\sim C_r$ and

$$\underline{X}'=(\underline{X}_I,X_r',\underline{X}_J).$$

Also, let $h_1',\ldots,h_T'$ be the binary indicators corresponding to $\underline{X}'$. By construction, we have the following properties:

1. $h_i=h_i'$ for all $i\neq r$
2. $h_r=1$ and $h_r=0$
3. $\sum_{i=1}^{T}h_i=h$ and $\sum_{i=1}^{T}h_i'=h-1$
4. $\mathbb{D}_j(\underline{X}|H_1=h_1,\ldots,H_T=h_T)=\mathbb{P}_{F_I}(\underline{X}_I)\mathbb{P}_{F_r}(X_r)\mathbb{P}_{C_J}(\underline{X}_J)$
5. $\mathbb{D}_j(\underline{X}'|H_1=h_1',\ldots,H_T=h_T')=\mathbb{P}_{F_I}(\underline{X}_I)\mathbb{P}_{C_r}(X_r')\mathbb{P}_{C_J}(\underline{X}_J)$

Thus, we obtain

$$
\begin{aligned}
&\int_{\underline{x}} \mathcal{M}(E|\underline{X}=\underline{x})\,\mathrm{d}\mathbb{D}_j(\underline{X}|H_1=h_1,\dots H_T=h_T)\\
&=\int_{\underline{x}_I}\int_{x_r}\int_{\underline{x}_J} \mathcal{M}(E|\underline{x}_I,x_r,\underline{x}_J)\,\mathrm{d}\mathbb{P}_{F_I}(\underline{X}_I)\,\mathrm{d}\mathbb{P}_{F_r}(X_r)\,\mathrm{d}\mathbb{P}_{C_J}(\underline{X}_J)\\
&\le \int_{x'_r}\int_{\underline{x}_I}\int_{x_r}\int_{\underline{x}_J} (e^{\epsilon}\mathcal{M}(E|\underline{x}_I,x'_r,\underline{x}_J)+\delta)\;\mathrm{d}\mathbb{P}_{F_I}(\underline{X}_I)\,\mathrm{d}\mathbb{P}_{F_r}(X_r)\,\mathrm{d}\mathbb{P}_{C_r}(X'_r)\,\mathrm{d}\mathbb{P}_{C_J}(\underline{X}_J)\\
&\le \int_{\underline{x}_I}\int_{x'_r}\int_{\underline{x}_J} (e^{\epsilon}\mathcal{M}(E|\underline{x}_I,x'_r,\underline{x}_J)+\delta)\;\mathrm{d}\mathbb{P}_{F_I}(\underline{X}_I)\,\mathrm{d}\mathbb{P}_{C_r}(X'_r)\,\mathrm{d}\mathbb{P}_{C_J}(\underline{X}_J)\\
&\le e^{\epsilon}\int_{\underline{x}'} \mathcal{M}(E|\underline{X}=\underline{x}')\,\mathrm{d}\mathbb{D}_j(\underline{X}'|H_1=h'_1,\dots H_T=h'_T)+\delta.
\end{aligned}
$$

Taking expectations on both sides with respect to $(H_1,\dots,H_T)$ proves the claim.

□

B.2 KL-divergence Decomposition with ϵ-global DP

The Sequential Karwa-Vadhan Lemma (Lemma 2) allows us to show the maximum KL-divergence induced in the distributions of actions by a global DP policy π^{ϵ}. The upper bound allows us to show how different the final distributions over actions induced by a global DP policy are for two different environments. Thus, in turn, it provides an information-theoretic limit on distinguishability of two environments if π^{ϵ} is played.

Theorem 10 (Upper Bound on KL-divergence for Bandits with ϵ-global DP)**.** *When an ϵ-global DP policy π^{ϵ} interacts with two bandit instances $\nu=(P_a : a\in[K])$ and $\nu'=(P'_a : a\in[K])$ we have:*

$$
D_{\mathrm{KL}}\left(M_{\nu\pi^{\epsilon}} \parallel M_{\nu'\pi^{\epsilon}}\right) \le 6\epsilon\mathbb{E}_{\nu\pi^{\epsilon}}\left[\sum_{t=1}^{T}\mathrm{TV}\left(P_{a_t}\parallel P'_{a_t}\right)\right]
$$

Proof. We define the marginal over the sequence of actions induced by π^{ϵ} for a given environment ν as

$$
m_{\nu\pi^{\epsilon}}(a_1,\dots,a_T) \triangleq \int_{r_1,\dots,r_T} \pi^{\epsilon}(a_1,\dots,a_T \mid r_1,\dots,r_T)P_{a_1}\,\mathrm{d}r_1\dots P_{a_T}\,\mathrm{d}r_T
$$

Since π^{ϵ} is ϵ-global DP, using Lemma 2, we obtain

$$
\log\left(\frac{m_{\nu\pi^{\epsilon}}(a_1,a_2,\dots,a_T)}{m_{\nu'\pi^{\epsilon}}(a_1,a_2,\dots,a_T)}\right) \le 6\epsilon\sum_{t=1}^{T}\mathrm{TV}\left(P_{a_t}\parallel P'_{a_t}\right)
$$

for every action sequence $(a_1,\dots,a_T)\in[K]^T$.

Thus,

$$
\begin{aligned}
D_{\mathrm{KL}}\left(M_{\nu\pi^{\epsilon}} \parallel M_{\nu'\pi^{\epsilon}}\right) &= \mathbb{E}_{\nu\pi^{\epsilon}}\left[\log\left(\frac{m_{\nu\pi^{\epsilon}}(A_1,A_2,...,A_T)}{m_{\nu'\pi^{\epsilon}}(A_1,A_2,...,A_T)}\right)\right]\\
&\le 6\epsilon\mathbb{E}_{\nu\pi^{\epsilon}}\left[\sum_{t=1}^{T}\mathrm{TV}\left(P_{a_t}\parallel P'_{a_t}\right)\right]
\end{aligned}
$$

□

This lemma explicates how the distinguishability of two environments ν and ν' under π^{ϵ} is dictated by a joint effect of global DP, in terms of the privacy budget ϵ, and the partial information available in bandits, in terms of the total variation distance between the rewards of the arms $\mathbb{E}_{\nu\pi^{\epsilon}}\left[\sum_{t=1}^{T}\mathrm{TV}\left(P_{a_t}\parallel P'_{a_t}\right)\right]$. We leverage this lemma further to construct the minimax and problem-dependent regret lower bounds for stochastic and linear bandits with ϵ-global DP.

C Lower Bounds on Regret: Stochastic and Linear Bandits with ϵ-global DP

In order to prove the lower bounds, we adopt the general canonical bandit model introduced in Section A. The high level idea of proving bandit lower bounds is selecting two problem instances that are similar (the policy cannot statistically distinguish between them) but conflicting (actions that may be good in one instance are not good for the other).

Under ϵ-global differential privacy, a new source of "confusion" is added to the problem, i.e. any sequence of actions induced by neighbouring reward streams must be ϵ-indistinguishable. In the canonical bandit framework, this is expressed by our Theorem 10.

In the following, we plug this upper bound on KL-divergences in the classic proofs of regret lower bounds in bandits Lattimore and Szepesvári (2018) to derive our minimax and problem-dependent regret lower bounds.

Notations. Let Π be the set of all policies, and Π^ϵ be the set of all ϵ-global DP policies.

C.1 Stochastic Bandits: Minimax Lower Bound

Theorem 2 (Minimax lower bound)**.** *For any $K > 1$ and $T \geq K-1$, and $\epsilon > 0$, the minimax regret of stochastic bandits with ϵ-global DP satisfies*

$$\mathrm{Reg}_{T,\epsilon}^{\textit{minimax}} \geq \max\left\{ \frac{1}{27}\underbrace{\sqrt{T(K-1)}}_{\textit{without global DP}}, \ \frac{1}{131}\underbrace{\frac{K-1}{\epsilon}}_{\textit{with } \epsilon\textit{-global DP}} \right\}.$$

Proof. We denote the environment corresponding to the set of K-Gaussian reward distributions with unit variance and means $\mu \in \mathbb{R}^K$ as $\mathcal{E}_{\mathcal{N}}^K(1) \triangleq \left\{ (\mathcal{N}(\mu_i, 1))_{i=1}^K : \mu = (\mu_1, \ldots, \mu_K) \in \mathbb{R}^K \right\}$.

Since $\Pi^\epsilon \subset \Pi$, we can have that

$$\mathrm{Reg}_{T,\epsilon}^{\text{minimax}} \geq \inf_{\pi \in \Pi} \sup_{\nu \in \mathcal{E}_{\mathcal{N}}^K(1)} \mathrm{Reg}_T(\pi, \nu) \geq \frac{1}{27}\sqrt{T(K-1)}$$

The second inequality is due to Theorem 15.2 in (Lattimore and Szepesvári, 2018).

Step 1: Choosing the 'Hard-to-distinguish' Environments. First, we fix a policy π^ϵ in Π^ϵ.

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 $i \triangleq \arg\min_{a>1} \mathbb{E}_{\nu,\pi^\epsilon}[N_a(T)]$ 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 $\mu'_i = 2\Delta$.

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

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

$$\mathbb{E}_{\nu\pi^\epsilon}[N_i(T)] \leq \frac{T}{K-1}$$

Step 2: From Lower Bounding Regret to Upper Bounding KL-divergence. Now by the classic regret decomposition and Markov Inequality 6, we get[7]

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

and

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

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

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

By applying the Bretagnolle–Huber inequality, we have:

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

Step 3: KL-divergence Decomposition with ϵ-global DP. Now, we apply Theorem 10 to upper-bound the KL-Divergence between the marginals.

$$\begin{aligned}D_{\mathrm{KL}}\left(M_{\nu\pi^\epsilon} \parallel M_{\nu'\pi^\epsilon}\right) &\leq 6\epsilon\mathbb{E}_{\nu\pi^\epsilon}\left[\sum_{t=1}^{T}\mathrm{TV}\left(P_{a_t} \parallel P'_{a_t}\right)\right] \\ &\leq 6\epsilon\mathbb{E}_{\nu\pi^\epsilon}\left[N_i(T)\right]\mathrm{TV}\left(p_i \parallel p'_i\right)\end{aligned}$$

since ν and ν' only differ in the arm i.

Finally, using Pinsker's Inequality 9, we obtain

$$\mathrm{TV}\left(p_i \parallel p'_i\right) \leq \sqrt{\frac{1}{2}D_{\mathrm{KL}}\left(\mathcal{N}(0,1) \parallel \mathcal{N}(2\Delta, 1)\right)} = \Delta$$

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

$$\begin{aligned}\mathrm{Reg}_T(\pi^\epsilon, \nu) + \mathrm{Reg}_T(\pi^\epsilon, \nu') &\geq \frac{T\Delta}{4}\exp\left(-6\epsilon\mathbb{E}_{\nu\pi^\epsilon}\left[N_i(T)\right]\Delta\right) \\ &\geq \frac{T\Delta}{4}\exp\left(-\frac{6\epsilon T\Delta}{K-1}\right)\end{aligned}$$

By optimising for Δ, we choose $\Delta = \frac{K-1}{6\epsilon T}$.

We conclude the proof by lower bounding $\exp(-1)$ with $\frac{48}{131}$, and using $2\max(a,b) \geq a + b$. □

C.2 Stochastic Bandits: Problem-dependent Lower Bound

Theorem 3 (Problem-dependent Regret Lower Bound)**.** *Let the environment $\mathcal{E}$ be a set of K reward distributions with finite means and a policy $\pi^\epsilon \in \Pi_{cons}(\mathcal{E}) \cap \Pi^\epsilon$ be a consistent policy*[8] *over $\mathcal{E}$ satisfying ϵ-global DP . Then, for all $\nu = (P_i)_{i=1}^K \in \mathcal{E}$, it holds that*

$$\liminf_{T\to\infty}\frac{\mathrm{Reg}_T(\pi^\epsilon, \nu)}{\log(T)} \geq \sum_{a:\Delta_a>0}\frac{\Delta_a}{\min\left(\underbrace{d_{\inf}\left(P_a, \mu^*, \mathcal{M}_a\right)}_{\text{without global DP}}, \underbrace{6\,\epsilon\, t_{\inf}\left(P_a, \mu^*, \mathcal{M}_a\right)}_{\text{with }\epsilon\text{-global DP}}\right)}.$$

Proof. Let μ_a be the mean of the a-th arm in ν, $t_a = t_{\inf}\left(P_a, \mu^*, \mathcal{M}_a\right)$ and $\pi^\epsilon \in \Pi_{\text{cons}}(\mathcal{E}) \cap \Pi^\epsilon$.

Since π^ϵ is consistent, by (Theorem 16.2, Lattimore and Szepesvári (2018)), it holds that

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

The theorem will follow by showing, for every suboptimal arm a:

$$\liminf_{T\to\infty}\frac{\mathbb{E}_{\nu\pi^\epsilon}\left[N_a(T)\right]}{\log(T)} \geq \frac{1}{6\,\epsilon\, t_a}$$

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

[8] 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 $\lim_{T\to\infty}\frac{R_T(\pi,\nu)}{T^p} = 0$. We denote the class of consistent policies over a set of environments $\mathcal{E}$ as $\Pi_{\text{cons}}(\mathcal{E})$.

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 $\mathrm{TV}\left(P_a \parallel P'_a\right) \leq t_a + \alpha$ 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'.

Step 2: From Lower Bounding Regret to Upper Bounding KL-divergence. For simplicity of notations, we use $\mathrm{Reg}_T = \mathrm{Reg}_T(\pi^\epsilon, \nu)$, $\mathrm{Reg}'_T = \mathrm{Reg}_T(\pi^\epsilon, \nu)$, 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 6, we obtain

$$\begin{aligned}
\mathrm{Reg}_T + \mathrm{Reg}'_T &\geq \frac{T}{2}\left(M_{\nu\pi^\epsilon}(A)\Delta_a + M_{\nu'\pi^\epsilon}\left(A^c\right)\left(\mu'_a - \mu^*\right)\right) \\
&\geq \frac{T}{2}\min\left\{\Delta_a, \mu'_a - \mu^*\right\}\left(M_{\nu\pi^\epsilon}(A) + M_{\nu'\pi^\epsilon}\left(A^c\right)\right) \\
&\geq \frac{T}{4}\min\left\{\Delta_a, \mu'_a - \mu^*\right\}\exp(-D_{\mathrm{KL}}\left(M_{\nu\pi^\epsilon} \parallel M_{\nu'\pi^\epsilon}\right))
\end{aligned} \tag{17}$$

Step 3: KL-divergence Decomposition with ϵ-global DP. By Theorem 10 and the construction of the 'hard-to-distinguish' environments, we obtain

$$\begin{aligned}
D_{\mathrm{KL}}\left(M_{\nu\pi^\epsilon} \parallel M_{\nu'\pi^\epsilon}\right) &\leq 6\epsilon\mathbb{E}_{\nu\pi^\epsilon}\left[N_a(T)\right]\mathrm{TV}\left(P_a \parallel P'_a\right) \\
&\leq 6\epsilon\mathbb{E}_{\nu\pi^\epsilon}\left[N_a(T)\right]\left(t_a + \alpha\right)
\end{aligned}$$

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

$$\mathrm{Reg}_T + \mathrm{Reg}'_T \geq \frac{T}{4}\min\left\{\Delta_a, \mu'_a - \mu^*\right\}\exp\left(-6\epsilon\mathbb{E}_{\nu\pi^\epsilon}\left[N_a(T)\right]\left(t_a + \alpha\right)\right)$$

Now, taking the limit inferior on both sides leads to

$$\begin{aligned}
\liminf_{T\to\infty}\frac{\mathbb{E}_{\nu\pi^\epsilon}\left[N_a(T)\right]}{\log(T)} &\geq \frac{1}{6\epsilon\left(t_a + \alpha\right)}\liminf_{T\to\infty}\frac{\log\left(\frac{T\min\left\{\Delta_a, \mu'_a - \mu^*\right\}}{4\left(\mathrm{Reg}_T + \mathrm{Reg}'_T\right)}\right)}{\log(T)} \\
&= \frac{1}{6\epsilon\left(t_a + \alpha\right)}\left(1 - \limsup_{T\to\infty}\frac{\log\left(\mathrm{Reg}_T + \mathrm{Reg}'_T\right)}{\log(T)}\right) = \frac{1}{6\epsilon\left(t_a + \alpha\right)}.
\end{aligned}$$

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 α tends to zero.

□

Remark 2. *For Bernoulli distributions, t_a is equal to Δ_a, so the private lower bound simplifies to:*

$$O\left(\sum_{a:\Delta_a > 0}\Delta_a\frac{1}{\epsilon\Delta_a}log(T)\right) = O\left(\frac{K\log(T)}{\epsilon}\right)$$

Thus, our problem-dependent regret lower bound retrieves as a special case the lower bound found in (Shariff and Sheffet, 2018) and established for Bernoulli distributions of rewards.

C.3 Stochastic Linear Bandits: Minimax Lower Bound

Theorem 4 (Minimax Regret Lower Bound)**.** *Let $\mathcal{A} = [-1,1]^d$ and $\Theta = \left\{-T^{-1/2}, T^{-1/2}\right\}^d$. Then, there exists a vector $\theta \in \Theta$ leading to a minimax regret of linear bandits with ϵ-global DP such that*

$$\mathrm{Reg}_T^{\textit{minimax}}(\mathcal{A}, \Theta) \geq \max\Big\{ \underbrace{\exp(-2)}_{\textit{without global DP}} \ , \ \underbrace{\exp(-6\epsilon d\sqrt{T})}_{\textit{with } \epsilon\textit{-global DP}} \Big\} \frac{d}{8}\sqrt{T}.$$

Proof. Due to Theorem 24.1,(Lattimore and Szepesvári, 2018), it holds that for any policy class Π,

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

Now, we focus on proving the ϵ-global DP part of the lower bound.

For $\theta, \theta' \in \Theta$, let ν and ν' be the bandit instances corresponding resp. to θ and θ'. We denote $\mathbb{M}_\theta = \mathbb{M}_{\nu,\pi^\epsilon}$ and $\mathbb{M}_{\theta'} = \mathbb{M}_{\nu',\pi^\epsilon}$. 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

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

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 6.

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,

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

Step 2: KL-divergence Decomposition with ϵ-global DP. From Theorem 10, we obtain that

$$\begin{aligned}
D_{\mathrm{KL}}\left(\mathbb{M}_\theta \parallel \mathbb{M}_{\theta'}\right) &\leq 6\epsilon\mathbb{E}_{\nu\pi^\epsilon}\left[\sum_{t=1}^T \mathrm{TV}\left(\mathcal{N}\left(\langle A_t, \theta\rangle, 1\right) \parallel \mathcal{N}\left(\langle A_t, \theta'\rangle, 1\right)\right)\right] \\
&\leq 6\epsilon\mathbb{E}_{\nu\pi^\epsilon}\left[\sum_{t=1}^T \sqrt{\frac{1}{2}D_{\mathrm{KL}}\left(\mathcal{N}\left(\langle A_t, \theta\rangle, 1\right) \parallel \mathcal{N}\left(\langle A_t, \theta'\rangle, 1\right)\right)}\right] \\
&= 6\epsilon\mathbb{E}_{\nu\pi^\epsilon}\left[\sum_{t=1}^T \sqrt{\frac{1}{4}\left[\langle A_t, \theta - \theta'\rangle^2\right]}\right] \\
&= 3\epsilon\mathbb{E}_{\nu\pi^\epsilon}\left[\sum_{t=1}^T \left|\langle A_t, \theta - \theta'\rangle\right|\right] \qquad (18)
\end{aligned}$$

$$\leq 3\epsilon \mathbb{E}_{\nu\pi^\epsilon}\left[\sum_{t=1}^{T} \|A_t\| \left(\|\theta - \theta'\|\right)\right]$$
$$\leq 3\epsilon \mathbb{E}_{\nu\pi^\epsilon}\left[T\sqrt{d} \times 2\sqrt{d}\frac{1}{\sqrt{T}}\right] = 6\epsilon d\sqrt{T}$$

Here, the second inequality is a consequence of Pinsker's inequality (Lemma 9). The third inequality is due to Cauchy-Scwhartz. The last inequality holds true because $A_t \in [-1,1]^d$ and $\theta, \theta' \in \left\{-T^{-1/2}, T^{-1/2}\right\}^d$

Step 3: Choosing the 'Hard-to-distinguish' θ. We already have that

$$p_{\theta,i} + p_{\theta',i} \geq \frac{1}{2}\exp\left(-3\epsilon\mathbb{E}_{\nu\pi^\epsilon}\left[\sum_{t=1}^{T} |\langle A_t, \theta - \theta'\rangle|\right]\right) \geq \frac{1}{2}\exp(-6\epsilon d\sqrt{T})$$

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

$$\sum_{\theta\in\Theta}\frac{1}{|\Theta|}\sum_{i=1}^{d} p_{\theta,i} = \frac{1}{|\Theta|}\sum_{i=1}^{d}\sum_{\theta\in\Theta} p_{\theta,i} \geq \frac{d}{4}\exp(-6\epsilon d\sqrt{T}).$$

This implies that there exists a $\theta \in \Theta$ such that $\sum_{i=1}^{d} p_{\theta,i} \geq d\exp(-6\epsilon d\sqrt{T})/4$.

Step 4: Plugging Back θ in the Regret Decomposition. With this choice of θ, we conclude that

$$\mathrm{Reg}_T(\mathcal{A}, \theta) \geq \frac{\sqrt{T}}{2}\sum_{i=1}^{d} p_{\theta,i}$$
$$\geq \frac{d\sqrt{T}}{8}\exp(-6\epsilon d\sqrt{T})$$

□

C.4 Stochastic Linear Bandits: Problem-dependent Lower Bound

Theorem 5 (Problem-dependent Regret Lower Bound)**.** *Let $\mathcal{A} \subset \mathbb{R}^d$ be a finite set spanning $\mathbb{R}^d$ and $\theta \in \mathbb{R}^d$ be such that there is a unique optimal action. Then, any consistent and ϵ-global DP bandit algorithm π^ϵ satisfies*

$$\liminf_{T\to\infty}\frac{\mathrm{Reg}_T(\mathcal{A}, \theta)}{\log(T)} \geq c(\mathcal{A}, \theta),$$

where the structural distinguishability gap *is the solution of a constraint optimisation*

$$c(\mathcal{A}, \theta) \triangleq \inf_{\alpha\in[0,\infty)^{\mathcal{A}}}\sum_{a\in\mathcal{A}}\alpha(a)\Delta_a, \text{ such that } \|a\|^2_{H_\alpha^{-1}} \leq \min\left\{\underbrace{0.5\Delta_a^2}_{\text{without global DP}}, \underbrace{3\epsilon\rho_a(\mathcal{A})\Delta_a}_{\text{with } \epsilon\text{-global DP}}\right\}$$

for all $a \in \mathcal{A}$ with $\Delta_a > 0$, $H_\alpha = \sum_{a\in\mathcal{A}}\alpha(a)aa^\top$, and an arm-structure dependent constant $\rho_a(\mathcal{A})$.

Proof. Let $a^* = \mathrm{argmax}_{a\in\mathcal{A}}\langle a, \theta\rangle$ be the optimal action, which we assumed to be unique.

By Theorem 25.1, Lattimore and Szepesvári (2018),

$$\limsup_{T\to\infty}\log(T)\,\|a - a^*\|^2_{G_T^{-1}} \leq \frac{1}{2}\Delta_a^2. \tag{19}$$

Let $\mathbb{M}$ and $\mathbb{M}'$ be the measures on the sequence of outcomes $A_1, \ldots, A_T$ induced by θ and θ' respectively. Let $\mathbb{E}[\cdot]$ and $\mathbb{E}'[\cdot]$ be the expectation operators of $\mathbb{M}$ and $\mathbb{M}'$, respectively.

Step 1: Choosing the 'Hard to distinguish' θ'. Let $\theta' \in \mathbb{R}^d$ be an alternative parameter to be chosen subsequently. We follow the usual plan of choosing θ' to be close to θ, but also ensuring that the

optimal action in the bandit determined by θ' is not a^*. Let $\Delta_{\min} = \min\{\Delta_a : a \in \mathcal{A}, \Delta_a > 0\}$, $\alpha \in (0, \Delta_{\min})$ and H be a positive definite matrix (to be chosen later) such that $\|a - a^*\|_H^2 > 0$.

Given this setting, we define

$$\theta' \triangleq \theta + \frac{\Delta_a + \alpha}{\|a - a^*\|_H^2} H(a - a^*),$$

which is chosen such that $\langle a - a^*, \theta' \rangle = \langle a - a^*, \theta \rangle + \Delta_a + \alpha = \alpha$.

This means that a^* is α-suboptimal for the environment corresponding to θ'.

Step 2: From Lower Bounding Regret to Upper Bounding KL-divergence. For simiplicity, we abbreviate $\mathrm{Reg}_T = \mathrm{Reg}_T(\mathcal{A}, \theta)$ and $\mathrm{Reg}'_T = \mathrm{Reg}_T(\mathcal{A}, \theta')$.

Then, by applying the classic regret decomposition and Markov's inequality 6, we obtain

$$\mathrm{Reg}_T = \mathbb{E}\left[\sum_{a \in \mathcal{A}} N_a(T)\Delta_a\right] \geq \frac{T\Delta_{\min}}{2}\mathbb{M}\left(N_{a^*}(T) < T/2\right) \geq \frac{T\alpha}{2}\mathbb{M}\left(N_{a^*}(T) < T/2\right),$$

Since a^* is α-suboptimal in bandit θ', it implies that

$$\mathrm{Reg}'_T \geq \frac{T\alpha}{2}\mathbb{M}'\left(N_{a^*}(T) \geq T/2\right).$$

Now, Bretagnolle–Huber inequality implies that

$$\begin{aligned}
\mathrm{Reg}_T + \mathrm{Reg}'_T &\geq \frac{T\alpha}{2}\left(\mathbb{M}\left(N_{a^*}(T) < T/2\right) + \mathbb{M}'\left(N_{a^*}(T) \geq T/2\right)\right) \\
&\geq \frac{T\alpha}{4}\exp\left(-D_{\mathrm{KL}}\left(\mathbb{M} \parallel \mathbb{M}'\right)\right)
\end{aligned}$$

Step 3: KL-divergence Decomposition with ϵ-global DP. By Equation 18, we have that

$$\begin{aligned}
D_{\mathrm{KL}}\left(\mathbb{M} \parallel \mathbb{M}\right) &\leq 3\epsilon\mathbb{E}_{\nu\pi^\epsilon}\left[\sum_{t=1}^{T}\left|\langle A_t, \theta - \theta' \rangle\right|\right] \\
&= 3\epsilon\mathbb{E}_{\nu\pi^\epsilon}\left[\sum_{t=1}^{T}\left|\left\langle A_t, \frac{\Delta_a + \alpha}{\|a - a^*\|_H^2} H(a - a^*)\right\rangle\right|\right] \\
&= 3\epsilon\frac{\Delta_a + \alpha}{\|a - a^*\|_{\bar{G}_T^{-1}}^2}\rho_T(H),
\end{aligned}$$

where we define

$$\rho_T(H) \triangleq \frac{\|a - a^*\|_{\bar{G}_T^{-1}}^2}{\|a - a^*\|_H^2}\mathbb{E}_{\nu\pi^\epsilon}\left[\sum_{t=1}^{T}\left|\langle A_t, H(a - a^*)\rangle\right|\right]$$

Thus, after re-arrangement, we get

$$\frac{3\epsilon\left(\Delta_a + \alpha\right)}{\log(T)\|a - a^*\|_{\bar{G}_T^{-1}}^2}\rho_T(H) \geq 1 - \frac{\log\left(\left(4R_T + 4R'_T\right)/\alpha\right)}{\log(T)}. \tag{20}$$

Step 4: Choosing H and Taking the Limit. The definition of consistency means that Reg_T and Reg'_T are both sub-linear in T. This implies that the second term in Equation (20) tends to zero for large T. Thus, by tending T to ∞ and α to zero, we obtain

$$\liminf_{T \to \infty}\frac{\rho_T(H)}{\log(T)\|a - a^*\|_{\bar{G}_T^{-1}}^2} \geq \frac{1}{3\epsilon\Delta_a}.$$

We now choose H to be a cluster point of the sequence $\left(\bar{G}_T^{-1}/\left\|\bar{G}_T^{-1}\right\|\right)_{T \in S}$ where $\left\|\bar{G}_T^{-1}\right\|$ is the spectral norm of the matrix $\bar{G}_T^{-1}$.

Fact 2: For this choice of H,

$$\liminf_{T\to\infty} \rho_T(H) \leq \rho_a(\mathcal{A}),$$

where

$$\rho_a(\mathcal{A}) \triangleq \sum_{j=1,\|a_j\|\neq 0}^{K} \frac{\left|a_j^T(a-a^*)\right|}{\|a_j\|^2}.$$

Finally,

$$\limsup_{T\to\infty} \log(T)\, \|a-a^*\|^2_{\bar{G}_T^{-1}} \leq 3\epsilon\Delta_a\rho_a(\mathcal{A}).$$

Combined with Equation 19, we get that

$$\limsup_{T\to\infty} \log(T)\, \|a-a^*\|^2_{\bar{G}_T^{-1}} \leq \min\left(\frac{1}{2}\Delta_a^2, 3\epsilon\Delta_a\rho_a(\mathcal{A})\right).$$

Using that

$$\lim_{T\to\infty} \frac{\|a-a^*\|_{\bar{G}_T^{-1}}}{\|a\|_{\bar{G}_T^{-1}}} = 1$$

from Theorem 25.1, Lattimore and Szepesvári (2018), we get that

$$\limsup_{T\to\infty} \log(T)\, \|a\|^2_{\bar{G}_T^{-1}} \leq \min\left(\frac{1}{2}\Delta_a^2, 3\epsilon\Delta_a\rho_a(\mathcal{A})\right).$$

Step 5: Getting Back to the Regret. We conclude using the same steps as in the Corollary 2 (Lattimore and Szepesvari, 2017). □

Now, we prove Fact 2.

Fact 2. *If H is a cluster point of the sequence $\left(\bar{G}_T^{-1}/\left\|\bar{G}_T^{-1}\right\|\right)_{T\in S}$ and $\left\|\bar{G}_T^{-1}\right\|$ is the spectral norm of the matrix $\bar{G}_T^{-1}$, then the following inequality holds true:*

$$\liminf_{T\to\infty} \rho_T(H) \leq \rho_a(\mathcal{A}),$$

where

$$\rho_a(\mathcal{A}) \triangleq \sum_{j=1,\|a_j\|\neq 0}^{K} \frac{\left|a_j^T(a-a^*)\right|}{\|a_j\|^2}.$$

Proof. We let S be a subset so that $\bar{G}_T^{-1}/\left\|\bar{G}_T^{-1}\right\|$ converges to H on $T\in S$. Then,

$$\begin{aligned}
\liminf_{T\to\infty} \rho_T(H) &\leq \liminf_{T\in S} \rho_T(\bar{G}_T^{-1}/\left\|\bar{G}_T^{-1}\right\|) \\
&= \liminf_{T\in S} \mathbb{E}_\theta\left[\sum_{t=1}^{T} \left|\left\langle A_t, \bar{G}_T^{-1}(a-a^*)\right\rangle\right|\right] \\
&= \liminf_{T\in S} \sum_{j=1}^{K} \mathbb{E}_\theta(N_j(T)) \left|a_j^T \bar{G}_T^{-1}(a-a^*)\right| \\
&= \liminf_{T\in S} \sum_{j=1,\|a_j\|\neq 0}^{K} \mathbb{E}_\theta(N_j(T)) \left|a_j^T \bar{G}_T^{-1}(a-a^*)\right|
\end{aligned}$$

Let j be such that $\|a_j\| \neq 0$.

Now, we aim to upper bound the term $\left|a_j^T \bar{G}_T^{-1}(a - a^*)\right|$

First, we decompose $a - a^*$ into two orthogonal components, which are aligned and orthogonal to a_j respectively.

$$a - a^* = \alpha_j a_j + b_j,$$

where $a_j^\top b_j = 0$ and $\alpha_j = \frac{a_j^T(a-a^*)}{\|a_j\|^2}$.

On the other hand, we have that

$$\bar{G}_T = \mathbb{E}_\theta\left[\sum_{t=1}^{T} A_t A_t^\top\right] = \sum_{j=1}^{K} \mathbb{E}_\theta(N_j(T)) a_j a_j^\top \succeq \mathbb{E}_\theta(N_j(T)) a_j a_j^\top$$

Since

$$\left(\mathbb{E}_\theta(N_j(T)) a_j a_j^\top\right)^\dagger = \frac{1}{\mathbb{E}_\theta(N_j(T))(a_j^\top a_j)^2} a_j a_j^\top,$$

and

$$\left(\mathbb{E}_\theta(N_j(T)) a_j a_j^\top\right)^\dagger b_j = 0,$$

only the component of $a - a^*$ in the direction of a_j matters in the dot product $a_j^T \bar{G}_T^{-1}(a - a^*)$. Thus,

$$\begin{aligned}\left|a_j^T \bar{G}_T^{-1}(a - a^*)\right| &\leq \frac{|\alpha_j|}{\mathbb{E}_\theta(N_j(T))(a_j^\top a_j)^2} a_j^T a_j a_j^T a_j \\ &= \frac{|\alpha_j|}{\mathbb{E}_\theta(N_j(T))}\end{aligned}$$

Consequently,

$$\liminf_{T\to\infty} \rho_T(H) \leq \sum_{j=1, \|a_j\|\neq 0}^{K} \frac{\left|a_j^T(a - a^*)\right|}{\|a_j\|^2} \triangleq \rho_a(\mathcal{A})$$

□

Example 3 ($\rho_a(\mathcal{A})$ for an orthogonal set of arms)**.** *If the action space is the orthogonal basis, then* $\rho_a(\mathcal{A}) = 2$*, because:*

$$\bar{G}_T = \begin{bmatrix} \mathbb{E}(N_1(T)) & & \\ & \ddots & \\ & & \mathbb{E}(N_d(T)) \end{bmatrix}$$

and:

$$\left|\left\langle A_t, \bar{G}_T^{-1}(a - a^*)\right\rangle\right| = \frac{1}{\mathbb{E}(N_a(T))}\mathbb{I}_{A_t = a} + \frac{1}{\mathbb{E}(N_{a^\star}(T))}\mathbb{I}_{A_t = a^\star}$$

so:

$$\mathbb{E}\left[\sum_{t=1}^{T}\left|\left\langle A_t, \bar{G}_T^{-1}(a - a^*)\right\rangle\right|\right] = 2$$

D Privacy Analysis of Algorithm 1

In this section, we prove that any bandit algorithm designed using the framework of Algorithm 1 satisfies ϵ-global DP. We establish the claim by proving ϵ-global DP for the set of private indices computed in Algorithm 1 and the final result is a consequence of the post-processing property of DP (Lemma 5).

Lemma 1 (Privacy of the $(l+1)$-means Computed in Algorithm 1)**.** *Let us define the private empirical mean of the rewards between steps i and j $(i < j)$ as*

$$f^{\epsilon}\{r_i, \ldots, r_j\} \triangleq \frac{1}{j-i}\sum_{t=i}^{j} r_t + Lap\left(\frac{1}{(j-i)\epsilon}\right). \tag{21}$$

If $1 < t_1 < \cdots < t_\ell < T$ and $r_t \in [0,1]$, the mechanism g^{ϵ} mapping the sequence of rewards $(r_1, r_2, \ldots, r_{T-1}, r_T)$ to $(\ell+1)$-private empirical means $(f^{\epsilon}\{r_1, \ldots, r_{t_1-1}\}, f^{\epsilon}\{r_{t_1}, \ldots, r_{t_2-1}\}, \ldots, f^{\epsilon}\{r_{t_{\ell-1}}, \ldots, r_{t_\ell-1}\}, f^{\epsilon}\{r_{t_\ell}, \ldots, r_T\})$ satisfies ϵ-DP.

Proof. Let $r^T \triangleq (r_1, \ldots, r_T)$ and $r'^T \triangleq (r'_1, \ldots, r'_T)$ be two neighbouring reward sequences in [0,1]. This implies that $\exists j \in [1, T]$ such that $r_j \neq r'_j$ and $\forall t \neq j$, $r_t = r'_t$.

Let ℓ' be such that $t_{\ell'} \leq j \leq t_{\ell'+1} - 1$, and follows the convention that $t_0 = 1$ and $t_{\ell+1} = T + 1$.

Let $\mu \triangleq (\mu_0, \ldots, \mu_\ell)$ a fixed sequence of outcomes obtained using Equation (21). Then,

$$\frac{\mathbb{P}(g^{\epsilon}(r^T) = \mu)}{\mathbb{P}(g^{\epsilon}(r'^T) = \mu)} = \frac{\mathbb{P}\left(f^{\epsilon}\{r_{t_{\ell'}}, \ldots, r_{t_{\ell'+1}-1}\} = \mu_{\ell'}\right)}{\mathbb{P}\left(f^{\epsilon}\{r_{t_{\ell'}}, \ldots, r_{t_{\ell'+1}-1}\} = \mu_{\ell'}\right)} \leq e^{\epsilon},$$

where the last inequality holds true because f^{ϵ} satisfies ϵ-DP following Theorem 1. □

Theorem 6 (ϵ-global DP for Algorithm 1)**.** *For any index I_a^{ϵ} computed using the private empirical mean of the rewards collected in the last active episode of arm a, Algorithm 1 satisfies ϵ-global DP.*

Proof. Fix two neighboring reward streams $r^T = \{r_1, \ldots, r_T\}$ and $r'^T = \{r'_1, \ldots, r'_T\}$.
This implies that $\exists j \in [1, T]$ such that $r_j \neq r'_j$ and $\forall t \neq j$, $r_t = r'_t$.
We also fix a sequence of actions $a^T = \{a_1, \ldots, a_T\}$.
We want to show that: $Pr(\pi(r^T) = a^T) \leq e^{\epsilon} Pr(\pi(r'^T) = a^T)$.

The main idea is that the change of reward in the j-th reward only affects the empirical mean computed in one episode, which is made private using the Laplace Mechanism and Lemma 1.

- Since $r^{j-1} = r'^{j-1}$, $Pr(\pi(r^{j-1}) = a^{j-1}) = Pr(\pi(r'^{j-1}) = a^{j-1})$.
- Let $t_\ell \leq j < t_{\ell+1}$ and $t_{\ell'} \leq j < t_{\ell'+1}$ be the episodes corresponding to the jth reward in r^T and r'^T respectively. Since $r^{j-1} = r'^{j-1}$, we get that $\ell = \ell'$. Thus, $Pr(\pi(r^{t_{\ell+1}}) = a^{t_\ell+1}) = Pr(\pi(r'^{t_\ell+1}) = a^{t_\ell+1})$.
- Let $\tilde{\mu}^{\ell}_{a_j,\epsilon}$ and $\tilde{\mu}'^{\ell}_{a,\epsilon}$ be the private means of arm a_j computed in the episode $[t_\ell, t_{\ell+1}]$, by the Laplace mechanism, for every interval $I \in \mathcal{R}$, $Pr(\tilde{\mu}^{\ell}_{a,\epsilon} \in I) \leq e^{\epsilon} Pr(\tilde{\mu}'^{\ell}_{a,\epsilon} \in I)$.
- Finally, since $\{r_{j+1}, \ldots, r_T\} = \{r'_{j+1}, \ldots, r'_T\}$, $Pr(\pi(r^T) = a^T | \tilde{\mu}^{\ell}_{a,\epsilon} \in I) = Pr(\pi(r'^T) = a^T | \tilde{\mu}'^{\ell}_{a,\epsilon} \in I)$

Now, we conclude the argument by using a chain rule.

□

Since Theorem 6 holds for any index-based bandit algorithm that uses only private empirical means of rewards (Equation (21)) of the last active episode to compute the indices, it also implies that AdaP-UCB and AdaP-KLUCB satisfy ϵ-global DP.

E Upper Bounds on Regret: AdaP-UCB and AdaP-KLUCB

E.1 Concentration Inequalities

Lemma 3. *Assume that $(X_i)_{1\le i\le n}$ are iid random variables in $[0,1]$, with $\mathbb{E}(X_i)=\mu$. Then, for any $\delta\ge 0$,*

$$\mathbb{P}\left(\hat{\mu}_n + Lap\left(\frac{1}{n\epsilon}\right) - \frac{\log\left(\frac{1}{\delta}\right)}{n\epsilon} - \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}} \ge \mu\right) \le \frac{3}{2}\delta, \tag{22}$$

and

$$\mathbb{P}\left(\hat{\mu}_n + Lap\left(\frac{1}{n\epsilon}\right) + \frac{\log\left(\frac{1}{\delta}\right)}{n\epsilon} + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}} \le \mu\right) \le \frac{3}{2}\delta, \tag{23}$$

where $\hat{\mu}_n = \frac{1}{n}\sum_{t=1}^{n} X_t$

Proof. We have that

$$\begin{aligned}
p_1 &\triangleq \mathbb{P}\left(\hat{\mu}_n + Lap\left(\frac{1}{n\epsilon}\right) - \frac{\log\left(\frac{1}{\delta}\right)}{n\epsilon} - \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}} \ge \mu\right)\\
&\le \mathbb{P}\left(\hat{\mu}_n - \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}} \ge \mu\right) + \mathbb{P}\left(Lap\left(\frac{1}{n\epsilon}\right) - \frac{\log\left(\frac{1}{\delta}\right)}{n\epsilon} \ge 0\right)\\
&\le \delta + \frac{\delta}{2} = \frac{3}{2}\delta,
\end{aligned}$$

where the last inequality is due to Lemma 11 and Lemma 10.

Similarly,

$$\begin{aligned}
p_2 &\triangleq \mathbb{P}\left(\hat{\mu}_n + Lap\left(\frac{1}{n\epsilon}\right) + \frac{\log\left(\frac{1}{\delta}\right)}{n\epsilon} + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}} \le \mu\right)\\
&\le \mathbb{P}\left(\hat{\mu}_n + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}} \le \mu\right) + \mathbb{P}\left(Lap\left(\frac{1}{n\epsilon}\right) + \frac{\log\left(\frac{1}{\delta}\right)}{n\epsilon} \le 0\right)\\
&\le \delta + \frac{\delta}{2} = \frac{3}{2}\delta,
\end{aligned}$$

where the last inequality is due to Lemma 11 and Lemma 10. □

Lemma 4. *Let $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables sampled from a Bernoulli distribution with mean μ, and let $\hat{\mu}_n = \frac{1}{n}\sum_{t=1}^{n} X_t$ be the sample mean. Let*

$$\breve{\mu}_n(\delta) \triangleq \mathrm{Clip}_{0,1}\left(\hat{\mu}_n + Lap\left(\frac{1}{n\epsilon}\right) + \frac{\log(\frac{1}{\delta})}{n\epsilon}\right) \tag{24}$$

for $\delta > 0$ be the clipped and private empirical mean.

Claim 1. *For any $\delta > 0$ and $\alpha \in [0,\mu]$, the following inequality holds:*

$$\mathbb{P}(\mu \ge \breve{\mu}_n(\delta) + \alpha) \le \exp(-nd(\mu-\alpha,\mu)) + \frac{1}{2}\delta \tag{25}$$

Claim 2. *Furthermore for $\delta \ge 0$, we define*

$$U_n(\delta) \triangleq \max\left\{q \in [0,1] : d\left(\breve{\mu}_n\left(\delta\right), q\right) \le \frac{\log\left(\frac{1}{\delta}\right)}{n}\right\} \tag{26}$$

Then,

$$\mathbb{P}(\mu \ge U_n(\delta)) \le \frac{3}{2}\delta \tag{27}$$

Proof. Here, we prove Claim 1 followed by Claim 2.

Claim 1. Since $\breve{\mu}_n(\delta) = \min\left\{\max\left\{0, \hat{\mu}_n + Lap\left(\frac{1}{n\epsilon}\right) + \frac{\log(\frac{1}{\delta})}{n\epsilon}\right\}, 1\right\}$, we have that

$$\begin{aligned}
\mu - \alpha \geq \breve{\mu}_n(\delta) &\Rightarrow \mu - \alpha \geq 1 \quad \text{or} \quad \mu - \alpha \geq \max\left\{0, \left(\hat{\mu}_n + Lap\left(\frac{1}{n\epsilon}\right) + \frac{\log(\frac{1}{\delta})}{n\epsilon}\right)\right\} \\
&\Rightarrow \mu - \alpha \geq \hat{\mu}_n + Lap\left(\frac{1}{n\epsilon}\right) + \frac{\log(\frac{1}{\delta})}{n\epsilon} \quad (\text{since } \mu \leq 1) \\
&\Rightarrow \mu - \alpha \geq \hat{\mu}_n \quad \text{or} \quad Lap\left(\frac{1}{n\epsilon}\right) + \frac{\log(\frac{1}{\delta})}{n\epsilon} \leq 0.
\end{aligned}$$

It implies that

$$\begin{aligned}
\mathbb{P}(\mu \geq \breve{\mu}_n(\delta) + \alpha) &\leq \mathbb{P}\Big(\mu \geq \hat{\mu}_n + \alpha\Big) + \mathbb{P}\Big(Lap\left(\frac{1}{n\epsilon}\right) + \frac{\log(\frac{1}{\delta})}{n\epsilon} \leq 0\Big) \\
&\leq \exp(-nd(\mu - \alpha, \mu)) + \frac{1}{2}\delta.
\end{aligned}$$

The last inequality is due to Equation 38 of Lemma 13 and Lemma 10.

Claim 2.

We have that the sets

$$\begin{aligned}
\{\mu \geq U_n(\delta)\} &\underset{(a)}{=} \{\mu \geq U_n(\delta) \geq \breve{\mu}_n(\delta)\} \\
&\underset{(b)}{=} \{d(\breve{\mu}_n(\delta), \mu) \geq d(\breve{\mu}_n(\delta), U_n(\delta)), \mu \geq \breve{\mu}_n(\delta)\} \\
&\underset{(c)}{=} \{d(\breve{\mu}_n(\delta), \mu) \geq \frac{\log(\frac{1}{\delta})}{n}, \mu \geq \breve{\mu}_n(\delta)\} \\
&\underset{(d)}{=} \{\breve{\mu}_n(\delta) \leq \mu - \alpha\}
\end{aligned}$$

Here, we chose an $\alpha > 0$ such that $d(\mu - \alpha, \mu) = \frac{\log(\frac{1}{\delta})}{n}$.

Step (a) holds because $U_n(\delta) \geq \breve{\mu}_n(\delta)$ by the definition of $U_n(\delta)$. Step (b) also holds true since $d(\breve{\mu}_n(\delta), \cdot)$ is increasing on $[\breve{\mu}_n(\delta), 1]$. Since $d(\breve{\mu}_n(\delta), U_n(\delta)) = \frac{\log(\frac{1}{\delta})}{n}$ by the definition of $U_n(\delta)$, we obtain the equality in Step (c). Finally, Step (d) is obtained by inverting the relative entropy.

We conclude the proof by

$$\begin{aligned}
\mathbb{P}\{\mu \geq U_n(\delta)\} &= \mathbb{P}\left\{\breve{\mu}_n(\delta) \leq \mu - \alpha\right\} \\
&\leq \exp(-nd(\mu - \alpha, \mu)) + \frac{1}{2}\delta \quad (\text{by } \textbf{Claim 1}) \\
&= \delta + \frac{\delta}{2} = \frac{3}{2}\delta \quad (\text{by substituting } \alpha)
\end{aligned}$$

□

E.2 Generic Regret Analysis for Algorithm 1

Algorithm 1 is a generic framework to construct an extension of any optimistic index-based bandit algorithm, which would satisfy ϵ-global DP. The algorithm is based on the index I_a^ϵ of each arm. I_a^ϵ is computed using the private empirical mean of the last active episode of arm a and is a high probability upper bound of the real mean μ_a.

To explicate the two conditions on arm indexes, we introduce the notation $I_a^\epsilon(t-1, \alpha, s)$, which is the index of arm a, at time-step t and computed using s reward samples from arm a.

Thus, we can express the index computed using just the last active episode as

$$I_a^\epsilon(t-1, \alpha) = I_a^\epsilon(t-1, \alpha, \frac{1}{2}N_a(t-1)). \tag{28}$$

Because $I_a^\epsilon(t-1,\alpha)$ only uses samples collected from the last active episode, and due to the doubling, the last active episode's size is exactly half the number of times arm a was pulled since the beginning.

The optimism of the index is ensured by the fact that

$$\mathbb{P}\left(I_a^\epsilon(t-1,\alpha,s) \leq \mu_a\right) \leq \frac{3}{2}\frac{1}{t^\alpha} \tag{29}$$

for every arm a, every sample size s and every time-step t, where α is the confidence level.

Theorem 11. *Let a be a suboptimal arm and $\ell \in \mathbb{N}$ such that $2^\ell < T$. Then, Algorithm 1 using an index I_a^ϵ satisfying Equations 28 and 29, also satisfies that for any $\alpha > 3$,*

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

where $G_{a,\ell,T} = \{\mathrm{I}_a^\epsilon(T-1,\alpha,2^\ell) < \mu^*\}$ and $G_{a,\ell,T}^c$ is the complement of $G_{a,\ell,T}$

Proof. Without loss of generality, we assume the first arm is the optimal one ($\mu^* = \mu_1$) and denote a suboptimal arm by a ($1 < a \leq K$).

We leverage the standard idea of UCB-type proofs: if arm a is chosen at the beginning of an episode ℓ, then either its index at t_ℓ is larger than the true mean of the first arm, or the true mean of the first arm is larger than the first arm's index at t_ℓ.

Since decisions, i.e. playing the arm with the highest index, are only taken at the beginning of an episode, we introduce ϕ which takes as input a time step and outputs the time step corresponding to the beginning of an episode. Formally, for each $t \in [K+1, T]$, let $\phi(t) = t_\ell$ such that $t_\ell \leq t \leq t_{\ell+1} - 1$. In Example 2, $\phi(5) = 4$ and $\phi(9) = 7$.

Formally, $\phi(t)$ is a random variable such that

$$\forall t : \phi(t) \leq t \leq 2\phi(t) \tag{30}$$

Step 1: Decomposition of $N_a(T)$. We observe that

$$\begin{aligned}
N_a(T) &= 1 + \sum_{t=K+1}^{T} \mathbb{I}\{A_t = a\} \\
&= 1 + \sum_{t=K+1}^{T} \mathbb{I}\{A_t = a \text{ and } \mathrm{I}_1^\epsilon(\phi(t)-1,\alpha) > \mu_1\} + \mathbb{I}\{A_t = a \text{ and } \mathrm{I}_1^\epsilon(\phi(t)-1,\alpha) \leq \mu_1\} \\
&\leq 1 + \underbrace{N_a'(T)}_{Term1} + \underbrace{\sum_{t=K+1}^{T} \mathbb{I}\{\mathrm{I}_1^\epsilon(\phi(t)-1,\alpha) \leq \mu_1\}}_{Term2}
\end{aligned}$$

We define $N_a'(T) \triangleq \sum_{t=K+1}^{T} \mathbb{I}\{A_t = a \text{ and } \mathrm{I}_1^\epsilon(t_{\ell'}-1,\alpha) > \mu_1\}$

Step 2: Decomposition of Term 1: $N_a'(T)$. Let $G_{a,\ell,T}$ be the 'good' event defined by

$$G_{a,\ell,T} = \{\mathrm{I}_a^\epsilon(T-1,\alpha,2^\ell) < \mu_1\}.$$

The main part of the proof is decomposing $N_a'(T)$ among the 'good' and the 'bad' events, i.e.

$$\mathbb{E}[N_a'(T)] = \mathbb{E}[\mathbb{I}\{G_{a,\ell,T}\}N_a'(T)] + \mathbb{E}[\mathbb{I}\{G_{a,\ell,T}^c\}N_a'(T)] \leq 2^{\ell+1} + \mathbb{P}(G_{a,\ell,T}^c)T.$$

$G_{a,\ell,T}^c$ denotes the complement of $G_{a,\ell,T}$.

To prove the last inequality, we only need to prove that when $G_{a,\ell,T}$ happens, $N_a'(T) \leq 2^{\ell+1}$. We prove it by contradiction.

Hence, let us assume that $G_{a,\ell,T}$ holds but $N_a'(T) > 2^{\ell+1}$.

This assumption implies that the arm a is played more than $2^{\ell+1}$ times. Thus, there must exist a round $t_{\ell'}$, where $N_a(t_{\ell'}-1) = 2^{\ell+1}$, $A_{t_{\ell'}} = i$ and $\mathrm{I}_1^\epsilon(t_{\ell'}-1,\alpha) \geq \mu_1$. Since indices are computed only

using the samples from the last active episode, $\mathrm{I}_a^\epsilon(t_{\ell'}-1,\alpha)$ is computed using exactly 2^ℓ reward samples from arm a.

Thus, we obtain

$$\begin{aligned}
\mathrm{I}_a^\epsilon(t_{\ell'}-1,\alpha) &= \mathrm{I}_a^\epsilon(t_{\ell'}-1,\alpha,2^\ell) \\
&\le \mathrm{I}_a^\epsilon(T-1,\alpha,2^\ell) \quad \text{(because } t_{\ell'} \le T \text{ and } \mathrm{I}_a^\epsilon(\cdot,\alpha,2^\ell) \text{ is increasing)} \\
&< \mu_1 \quad \text{(definition of } G_{a,\ell,T}\text{)} \\
&\le \mathrm{I}_1^\epsilon(t_{\ell'}-1,\alpha)
\end{aligned}$$

The last inequality contradicts the fact that $A_{t_{\ell'}} = i$ and thus, establishes the claim that $N'_a(T) \le 2^{\ell+1}$ under the 'good' event.

Step 3: Upper-bounding Term 2. To conclude,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{I}\{\mathrm{I}_1^\epsilon(\phi(t)-1,\alpha) \le \mu_1\}\right] &= \sum_{t=K+1}^{T} \mathbb{P}\{\mathrm{I}_1^\epsilon(\phi(t)-1,\alpha) \le \mu_1\} \\
&\le \sum_{t=K+1}^{T} \sum_{\phi=t/2}^{t} \mathbb{P}\{\mathrm{I}_1^\epsilon(\phi-1,\alpha) \le \mu_1\} \\
&\le \sum_{t=K+1}^{T} \sum_{\phi=t/2}^{t} \sum_{s=1}^{\phi} \mathbb{P}\{\mathrm{I}_1^\epsilon(\phi-1,\alpha,s) \le \mu_1\} \\
&\le \sum_{t=K+1}^{T} \sum_{\phi=t/2}^{t} \sum_{s=1}^{\phi} \frac{3}{2}\frac{1}{\phi^\alpha} \qquad \text{(Equation 29)} \\
&= \frac{3}{2} \sum_{t=K+1}^{T} \sum_{\phi=t/2}^{t} \frac{1}{\phi^{\alpha-1}} \\
&\le \frac{3}{2} \sum_{t=K+1}^{T} \frac{2^{\alpha-2}}{t^{\alpha-2}} \quad \text{(because } \phi \ge \frac{t}{2}\text{)} \\
&\le \frac{3}{2} 2^{\alpha-2} \int_K^T \frac{1}{x^{\alpha-2}} dx \quad \text{(sum-integral inequality)} \\
&\le \frac{3}{2} 2^{\alpha-2} \frac{1}{\alpha-3} \frac{1}{K^{\alpha-3}} = \frac{3}{2}\frac{2}{\alpha-3}\left(\frac{2}{K}\right)^{\alpha-3} \\
&\le \frac{3}{\alpha-3}
\end{aligned}$$

for $\alpha > 3$ and $K \ge 2$.

Here, the first inequality is due to an union bound on $\phi(t) \in [t/2, t]$ (Equation 30), and the second inequality is due to a union bound on $N_1(\phi-1)$.

Step 4: Combining the Bounds on Terms 1 and 2.

$$\begin{aligned}
\mathbb{E}[N_a(T)] &\le 1 + 2^{\ell+1} + \mathbb{P}\left(G_{a,\ell,T}^c\right) T + \frac{3}{\alpha-3} \\
&= 2^{\ell+1} + \mathbb{P}\left(G_{a,\ell,T}^c\right) T + \frac{\alpha}{\alpha-3}
\end{aligned}$$

□

Now we design indexes that satisfy the conditions of Theorem 11, namely AdaP-UCB and AdaP-KLUCB.

To obtain the final regret bounds, we only have to choose ℓ big enough such that $\mathbb{P}\left(\mathrm{I}_a(T,2^\ell) \ge \mu_1\right) T$ is negligible. This corresponds to the leading term in the regret upper-bounds, and this is where the regrets of AdaP-UCB and AdaP-KLUCB differ.

We explicate the issues of designing the indexes and choosing corresponding ℓ in the following section, which leads to the regret upper bounds of AdaP-UCB and AdaP-KLUCB.

E.3 Regret Analysis for AdaP-UCB and AdaP-KLUCB

Theorem 7. *For rewards in* $[0,1]$, AdaP-UCB *satisfies* ϵ*-global DP, and for* $\alpha > 3$*, it yields a regret*

$$\mathrm{Reg}_T(\mathsf{AdaP\text{-}UCB}, \nu) \leq \sum_{a:\Delta_a>0} \left(\frac{16\alpha}{\min\{\Delta_a, \epsilon\}} \log(T) + \frac{3\alpha}{\alpha - 3} \right).$$

Proof. The proof is constituted of three steps.

Step 1: Designing an Index satisfying Equation (28)**, Equation** (29)**, and** ϵ**-global DP.** For AdaP-UCB, the index is defined as

$$\mathrm{I}_a^{\epsilon}(t_\ell - 1, \alpha) = \tilde{\mu}_{a,\epsilon}^{\ell} + \sqrt{\frac{\alpha \log(t_\ell)}{2 \times \frac{1}{2} N_a(t_\ell - 1)}} + \frac{\alpha \log(t_\ell)}{\epsilon \times \frac{1}{2} N_a(t_\ell - 1)},$$

where

$$\tilde{\mu}_{a,\epsilon}^{\ell} = \hat{\mu}_{a, \frac{1}{2} N_a(t_\ell - 1)} + Lap\left(\frac{1}{\epsilon \times \frac{1}{2} N_a(t_\ell - 1)} \right) \tag{31}$$

is the private empirical mean of arm a computed using only samples from the last active episode, and $\hat{\mu}_{a,s}$ is the empirical mean of arm a calculated using s samples of reward from arm a.

This index verifies the first condition (Equation 28) of Theorem 11.

The second condition (Equation 29) of Theorem 11 follows directly from Equation 23 of Lemma 3

By Theorem 6,AdaP-UCB is ϵ-global DP.

By Theorem 11, for every suboptimal arm a, we have that

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

where

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

Step 2: Choosing an ℓ**.** Now, we observe that

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

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

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

Let us consider the contrary, i.e.

$$\begin{aligned}
\gamma < 0 \Rightarrow \sqrt{2^\ell} &< \sqrt{\frac{\alpha \log(T)}{2\Delta_a^2}} \left(1 + \sqrt{1 + \frac{4\Delta_a}{\epsilon}} \right) \\
\Rightarrow 2^\ell &< \frac{\alpha \log(T)}{2\Delta_a^2} \left(4 + \frac{8\Delta_a}{\epsilon} \right) \\
\Rightarrow 2^\ell &< \frac{4\alpha \log(T)}{\Delta_a \min\{\epsilon, 2\Delta_a\}}.
\end{aligned} \tag{32}$$

Thus, by choosing

$$\ell = \left\lceil \frac{1}{\log(2)} \log\left(\frac{4\alpha\log(T)}{\Delta_a \min\{\epsilon, 2\Delta_a\}}\right)\right\rceil$$

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

$$\mathbb{P}(G^c_{a,\ell,T}) \leq \mathbb{P}\left(\hat{\mu}_{a,2^\ell} + Lap\left(\frac{1}{2^\ell\epsilon}\right) - \sqrt{\frac{\alpha\log(T)}{2\times 2^\ell}} - \frac{\alpha\log(T)}{\epsilon 2^\ell} \geq \mu_a\right) \leq \frac{3}{2T^\alpha}$$

The last inequality is due to Equation 22 of Lemma 3.

Step 3: The Regret Bound. Combining Steps 1 and 2, we get that

$$\begin{aligned}\mathbb{E}[N_a(T)] &\leq \frac{\alpha}{\alpha-3} + 2^{\ell+1} + T\times\frac{3}{2T^\alpha}\\ &\leq \frac{16\alpha\log(T)}{\Delta_a\min\{\epsilon,2\Delta_a\}} + \frac{3\alpha}{\alpha-3}.\end{aligned} \tag{33}$$

Plugging this upper bound back in the definition of problem-dependent regret concludes the proof. □

Remark 3. *The leading term of the regret is* $\frac{16\alpha\log(T)}{\Delta_a\min\{\epsilon,2\Delta_a\}}$*, which is* 4 *times more than what we got from Equation 32. A multiplicative factor of* 2 *is introduced due to the doubling and another multiplicative factor of* 2 *is due to the forgetting. Thus, the combined price of doubling and forgetting is a multiplicative constant* 4 *in the leading term of regret.*

Theorem 8. *When the rewards are sampled from Bernoulli distributions,* AdaP-KLUCB *satisfies* ϵ*-global DP, and for* $\alpha > 3$ *and constants* $C_1(\alpha), C_2 > 0$*, it yields a regret*

$$\mathrm{Reg}_T(\mathsf{AdaP\text{-}KLUCB},\nu) \leq \sum_{a:\Delta_a>0}\left(\frac{C_1(\alpha)\Delta_a}{\min\{d(\mu_a,\mu^*), C_2\epsilon\Delta_a\}}\log(T) + \frac{\alpha}{\alpha-3}\right).$$

Proof. The proof is constituted of three steps.

Step 1: Designing an Index satisfying Equation (28)**, Equation** (29)**, and** ϵ**-global DP.** For AdaP-KLUCB, the index is defined as

$$\mathrm{I}^\epsilon_a(t_\ell-1,\alpha) = \max\left\{q\in[0,1] : d\left(\breve{\mu}^{\ell,\alpha}_{a,\epsilon}, q\right) \leq \frac{\alpha\log(t_\ell)}{\frac{1}{2}N_a(t_\ell-1)}\right\} \triangleq U_{a,\frac{1}{2}N_a(t_\ell-1)}\left(\frac{1}{t^\alpha_\ell}\right),$$

where $\breve{\mu}^{\ell,\alpha}_{a,\epsilon} = \mathrm{Clip}_{0,1}\left(\tilde{\mu}^\ell_{a,\epsilon} + \frac{\alpha\log(t_\ell)}{\epsilon\frac{1}{2}N_a(t_\ell-1)}\right) = \breve{\mu}_{a,\frac{1}{2}N_a(t_\ell-1)}\left(\frac{1}{t^\alpha_\ell}\right)$ as defined in Equation 24,

$\tilde{\mu}^\ell_{a,\epsilon}$ is the private empirical computed only using the samples from the last active episode (as defined for AdaP-UCB, and $U_{a,s}(\delta) = \max\left\{q\in[0,1] : d\left(\breve{\mu}_{a,s}(\delta), q\right) \leq \frac{\log\left(\frac{1}{\delta}\right)}{s}\right\}$ as defined in Equation 26

This index verifies the first condition (Equation 28) of Theorem 11.

The second condition (Equation 29) of Theorem 11 follows directly from Equation 23 of Lemma 3

By Theorem 6, AdaP-KLUCB also satisfies ϵ-global DP.

By Theorem 11, for every suboptimal arm a, we have that

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

where

$$G_{a,\ell,T} = \left\{U_{a,2^\ell}\left(\frac{1}{T^\alpha}\right) < \mu_1\right\}.$$

Step 2: Choosing an ℓ. We observe that

$$\begin{aligned}\mathbb{P}(G^c_{a,\ell,T}) &= \mathbb{P}\left(U_{a,2^\ell}\left(\frac{1}{T^\alpha}\right) \geq \mu_1\right)\\ &\leq \mathbb{P}\left(d^+\left(\breve{\mu}_{a,2^\ell}\left(\frac{1}{T^\alpha}\right), \mu_1\right) \leq \frac{\alpha\log(T)}{2^\ell}\right) \quad \text{(by definition of } U_{a,2^\ell})\end{aligned}$$

where $d^+(p,q) \triangleq d(p,q)\mathbb{I}_{p<q}$and $d(p,q)$ is the relative entropy between Bernoulli distributions as stated in Definition 5.

Let $\beta > 0$, and $c(\beta) \in [0,1]$ such that: $d(\mu_a + c(\beta)\Delta_a, \mu_1) = \frac{d(\mu_a,\mu_1)}{1+\beta}$.

Since $d(\cdot, \mu_1)$ is a bijective function from $[\mu_a, \mu_1]$ to $[0, d(\mu_a, \mu_1)]$, we get that $c(\beta)$ always exists and is unique.

In addition, $c(\beta)$ verifies: $\lim_{\beta\to 0} c(\beta) = 0$, $\lim_{\beta\to+\infty} c(\beta) = 1$ and $c(\beta)$ is an increasing function of β.

First, we choose ℓ such that

$$2^\ell \geq \frac{(1+\beta)\alpha\log(T)}{d(\mu_a,\mu_1)}. \tag{34}$$

This leads to

$$\begin{aligned}\mathbb{P}(G^c_{a,\ell,T}) &\leq \mathbb{P}\left(d^+\left(\breve{\mu}_{a,2^\ell}\left(\frac{1}{T^\alpha}\right), \mu_1\right) \leq \frac{d(\mu_a,\mu_1)}{1+\beta}\right)\\ &= \mathbb{P}\left(d^+\left(\breve{\mu}_{a,2^\ell}\left(\frac{1}{T^\alpha}\right), \mu_1\right) \leq d(\mu_a + c(\beta)\Delta_a, \mu_1)\right) \quad \text{(definition of } c(\beta))\\ &\leq \mathbb{P}\left(\breve{\mu}_{a,2^\ell}\left(\frac{1}{T^\alpha}\right) \geq \mu_a + c(\beta)\Delta_a\right) \quad (d(\cdot,\mu_1) \text{ is decreasing on } [0,\mu_1])\\ &\leq \mathbb{P}\left(\hat{\mu}_{a,2^\ell} + Lap\left(\frac{1}{2^\ell\epsilon}\right) + \frac{\alpha\log(T)}{\epsilon 2^\ell} \geq \mu_a + c(\beta)\Delta_a\right) \quad \text{(definition of } \breve{\mu})\end{aligned}$$

Let us consider $\gamma_{\ell,T}$ such that $d(\mu_a + \gamma_{\ell,T}\Delta_a, \mu_a) = \frac{\log(T)}{2^\ell}$. We prove its existence and upper bound it later in Fact 3. Thus, we obtain

$$\begin{aligned}\mathbb{P}(G^c_{a,\ell,T}) &\leq \mathbb{P}\left(\hat{\mu}_{a,2^\ell} - \gamma_{\ell,T}\Delta_a + Lap\left(\frac{1}{2^\ell\epsilon}\right) - \frac{\log(T)}{\epsilon 2^\ell} \geq \mu_a + (c(\beta) - \gamma_{\ell,T})\Delta_a - \frac{(1+\alpha)\log(T)}{\epsilon 2^\ell}\right)\\ &= \mathbb{P}\left(\hat{\mu}_{a,2^\ell} - \gamma_{\ell,T}\Delta_a + Lap\left(\frac{1}{2^\ell\epsilon}\right) - \frac{\log(T)}{\epsilon 2^\ell} \geq \mu_a + \theta\right)\end{aligned}$$

Here, $\theta \triangleq (c(\beta) - \gamma_{\ell,T})\Delta_a - \frac{(1+\alpha)\log(T)}{\epsilon 2^\ell}$.

By choosing

$$2^\ell \geq \frac{(1+\alpha)\log(T)}{(c(\beta) - \gamma_{\ell,T})\epsilon\Delta_a}, \tag{35}$$

we ensure that $\theta \geq 0$. Thus, we get

$$\begin{aligned}\mathbb{P}(G^c_{a,\ell,T}) &\leq \mathbb{P}\left(\hat{\mu}_{a,2^\ell} - \gamma_{\ell,T}\Delta_a + Lap\left(\frac{1}{2^\ell\epsilon}\right) - \frac{\log(T)}{\epsilon 2^\ell} \geq \mu_a\right)\\ &\leq \mathbb{P}\left(\hat{\mu}_{a,2^\ell} - \gamma_{\ell,T}\Delta_a \geq \mu_a\right) + \mathbb{P}\left(Lap\left(\frac{1}{2^\ell\epsilon}\right) - \frac{\log(T)}{\epsilon 2^\ell} \geq 0\right)\\ &\leq \exp\left(-2^\ell d(\mu_a + \gamma_{\ell,T}\Delta_a, \mu_a)\right) + \frac{1}{2T}\\ &= \frac{3}{2T}.\end{aligned}$$

The last inequality is due to Equation 37 of Lemma 13 and Lemma 10.

Fact 3. $\mathbf{B} \triangleq \{\beta > 0 : c(\beta) > \gamma_{\ell,T}\} \neq \varnothing$.

Combining both conditions 34 and 34, we choose ℓ to be the smallest integer such that

$$2^\ell \geq \inf_{\beta \in \mathbf{B}} \max\left\{\frac{(1+\beta)\alpha}{d(\mu_a,\mu_1)}, \frac{(1+\alpha)}{(c(\beta)-\gamma_{\ell,T})\epsilon\Delta_a}\right\}\log(T) \triangleq \frac{\frac{1}{4}C_1(\alpha)}{\min\{d(\mu_a,\mu_1), C_2\epsilon\Delta_a\}}\log(T)$$

Step 3: The Regret Bound. Combining Steps 1 and 2, we get that

$$\begin{aligned}\mathbb{E}[N_a(T)] &\leq 2^{\ell+1} + T \times \frac{3}{2T} + \frac{\alpha}{\alpha-3} \\ &\leq \frac{C_1(\alpha)}{\min\{d(\mu_a,\mu_1), C_2\epsilon\Delta_a\}}\log(T) + \frac{3\alpha}{\alpha-3}\end{aligned}$$

Plugging this upper bound back in the definition of problem-dependent regret concludes the proof. □

To conclude, we prove Fact 3.

Fact 3. $\mathbf{B} \triangleq \{\beta > 0 : c(\beta) > \gamma_{\ell,T}\} \neq \varnothing$.

Proof. **Step 1: Going from $d(\cdot,\mu_a)$ to $d(\cdot,\mu_1)$.** The difficulty of the proof lies in the fact that $\gamma_{\ell,T}$ is defined by inverting $d(\cdot,\mu_a)$ while $c(\beta)$ is defined by inverting $d(\cdot,\mu_1)$.

To handle that, we investigate the function $g(x) \triangleq d(x,\mu_a) - d(x,\mu_1)$.

g satisfies the following properties:

- g is continuous and increasing in the interval $[\mu_a,\mu_1]$,
- $g(\mu_a) = -d(\mu_a,\mu_1) < 0$, and
- $g(\mu_1) = d(\mu_1,\mu_a) > 0$.

This implies that there exists a unique root of $g(x)$, where it changes sign. Specifically, there exists a unique $z \in [\mu_a,\mu_1]$ such that:

- $g(z) = 0$
- $\forall x \in [\mu_a, z[: g(x) < 0$
- $\forall x \in]z,\mu_1] : g(x) > 0$

and consequently z verifies $d(z,\mu_a) = d(z,\mu_1)$

Step 2: Choosing β. We choose β such that $\frac{d(\mu_a,\mu_1)}{1+\beta} = d(z,\mu_a) = d(z,\mu_1)$.

Step 3: Consequence of the choice of β on $c(\beta)$. Thus,

$$d(\mu_a + c(\beta)\Delta_a, \mu_1) = d(z,\mu_1),$$

which yields

$$z = \mu_a + c(\beta)\Delta_a$$

by uniqueness of z.

Step 4: Consequence of the choice of β on $\gamma_{\ell,T}$. On the other hand,

$$\begin{aligned}d(\mu_a + \gamma_{\ell,T}\Delta_a, \mu_a) &= \frac{\log(T)}{2^\ell} && \text{(by definition of } \gamma_{\ell,T}) \\ &\leq \frac{d(\mu_a,\mu_1)}{\alpha(\beta+1)} && \text{(by Equation 34)} \\ &< d(z,\mu_a) && \text{(since } \alpha > 3) \\ &= d(\mu_a + c(\beta)\Delta_a, \mu_a) && (36)\end{aligned}$$

As a consequence, we conclude that $\gamma_{\ell,T}$ exists and $\gamma_{\ell,T} < c(\beta)$ as $d(\cdot,\mu_a)$ is an increasing function in the interval $[\mu_a, 1]$ □

E.4 Problem-independent Regret Bounds

In this section, we provide problem-independent (or minimax) regret upper bounds for AdaP-UCB.

Theorem 12. *For rewards in* $[0,1]$*,* AdaP-UCB *yields a regret*

$$\mathrm{Reg}_T(\mathsf{AdaP\text{-}UCB},\nu) \leq \frac{3\alpha}{\alpha-3}\sum_a \Delta_a + 8\sqrt{\alpha K T\log(T)} + \frac{16\alpha K\log(T)}{\epsilon}$$

which achieves the minimax lower bound of Thm 2 up to $\log(T)$ *factors.*

Proof. Let Δ be a value to be tuned later.
We have that

$$\begin{aligned}
\mathrm{Reg}_T(\mathsf{AdaP\text{-}UCB},\nu) &= \sum_a \Delta_a \mathbb{E}[N_a(T)] = \sum_{a:\Delta_a\leq\Delta}\Delta_a\mathbb{E}[N_a(T) + \sum_{a:\Delta_a>\Delta}\Delta_a\mathbb{E}[N_a(T)] \\
&\leq T\Delta + \sum_{a:\Delta_a>\Delta}\Delta_a\left(\frac{16\alpha\log(T)}{\Delta_a\min\{\epsilon,\Delta_a\}} + \frac{3\alpha}{\alpha-3}\right) \quad \text{(eq. 33)} \\
&\leq T\Delta + \frac{16\alpha K\log(T)}{\Delta} + \frac{16\alpha K\log(T)}{\epsilon} + \frac{3\alpha}{\alpha-3}\sum_a\Delta_a \\
&\leq 8\sqrt{\alpha K T\log(T)} + \frac{16\alpha K\log(T)}{\epsilon} + \frac{3\alpha}{\alpha-3}\sum_a\Delta_a
\end{aligned}$$

where the last step is by taking $\Delta = 4\sqrt{\frac{\alpha K\log(T)}{T}}$.

□

Remark 4. *The same bound is achieved by* AdaP-KLUCB *(up to multiplicative constants) by using that* $d(\mu_a,\mu^*) \geq 2\Delta_a^2$ *and using the same steps in Thm 12.*

F Existing Technical Results and Definitions

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

Lemma 5 (Post-processing Lemma (Proposition 2.1, (Dwork and Roth, 2014))). *If a randomised algorithm $\mathcal{A}$ satisfies (ϵ, δ)-Differential Privacy and f is an arbitrary randomised mapping defined on $\mathcal{A}$'s output, then $f \circ \mathcal{A}$ satisfies (ϵ, δ)-DP.*

Lemma 6 (Markov's Inequality). *For any random variable X and $\varepsilon > 0$,*

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

Definition 4 (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 7 (Divergence decomposition). *Let $\nu = (P_1, \ldots, P_K)$ and $\nu' = (P'_1, \ldots, P'_K)$ be two bandit instances. Fix some policy π and let $\mathbb{P}_{\nu\pi}$ and $\mathbb{P}_{\nu'\pi}$ be the probability measures on the canonical bandit model. Then,*

$$D_{\mathrm{KL}}\left(\mathbb{P}_{\nu\pi} \parallel \mathbb{P}_{\nu'\pi}\right) = \sum_{a=1}^{K} \mathbb{E}_{\nu}\left[N_a(T)\right] \mathrm{D}\left(P_a, P'_a\right).$$

Lemma 8 (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,*

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

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

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

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

Lemma 10 (Tail Bounds for Laplacian Random Variables). *For any $a, b > 0$, we have*

$$\mathbb{P}(Lap(b) > a) = \frac{1}{2}\exp\left(-\frac{a}{b}\right) \quad \textit{and} \quad \mathbb{P}(Lap(b) < -a) = \frac{1}{2}\exp\left(-\frac{a}{b}\right).$$

Lemma 11 (Hoeffding's Bound). *Assume that $(X_i)_{1 \leq i \leq n}$ are iid random variables in $[0, 1]$, with $\mathbb{E}(X_i) = \mu$. For any $\delta, \beta \geq 0$ and, we have:*

$$\mathbb{P}\left(\hat{\mu}_n \geq \mu + \beta\right) \leq \exp\left(-2n\beta^2\right) \quad \textit{and} \quad \mathbb{P}\left(\hat{\mu}_n \leq \mu - \beta\right) \leq \exp\left(-2n\beta^2\right),$$

where $\hat{\mu}_n = \frac{1}{n}\sum_{t=1}^{n} X_t$.

Definition 5 (Relative entropy between Bernoulli distributions). *The relative entropy between Bernoulli distributions with parameters $p, q \in [0, 1]$ is*

$$d(p, q) = p\log(p/q) + (1-p)\log((1-p)/(1-q)),$$

where singularities are defined by taking limits: $d(0, q) = \log(1/(1-q))$ and $d(1, q) = \log(1/q)$ for $q \in [0, 1]$ and $d(p, 0) = 0$ if $p = 0$ and ∞ otherwise and $d(p, 1) = 0$ if $p = 1$ and ∞ otherwise.

Lemma 12 (Properties of the relative entropy between Bernoulli distributions (Lemma 10.2, (Lattimore and Szepesvári, 2018))). *Let $p, q, \varepsilon \in [0, 1]$.*

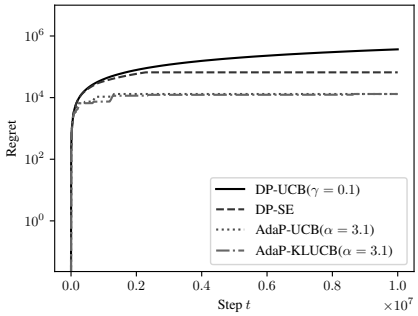
1. *The functions $d(\cdot, q)$ and $d(p, \cdot)$ are convex and have unique minimisers at q and p, respectively.*
2. *$d(p, \cdot)$ and $d(\cdot, p)$ are increasing in the interval $[p, 1]$ and decreasing in the interval $[0, p]$.*

Lemma 13 (Chernoff's Bound). *Let $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables that are Bernoulli distributed with mean μ, and let $\hat{\mu}_n = \frac{1}{n}\sum_{t=1}^{n} X_t$ be the sample mean. Then, for $\beta \in [0, 1-\mu]$, it holds that:*

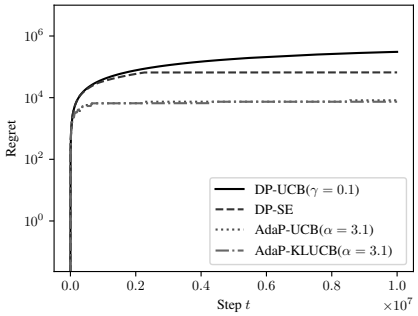
$$\mathbb{P}(\hat{\mu}_n \geq \mu + \beta) \leq \exp(-nd(\mu + \beta, \mu)), \tag{37}$$

and for $\beta \in [0, \mu]$,

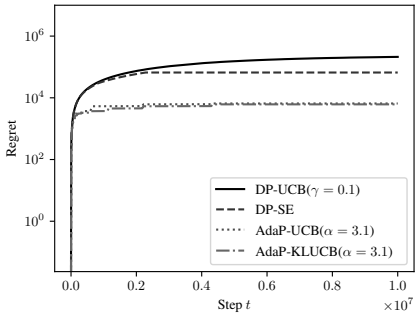
$$\mathbb{P}(\hat{\mu}_n \leq \mu - \beta) \leq \exp(-nd(\mu - \beta, \mu)). \tag{38}$$



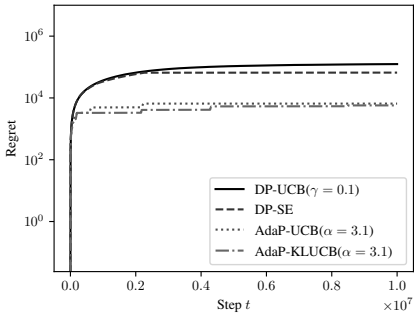
(a) $\epsilon = 0.1$



(b) $\epsilon = 0.25$

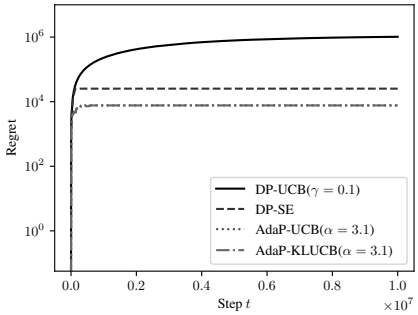


(c) $\epsilon = 0.5$

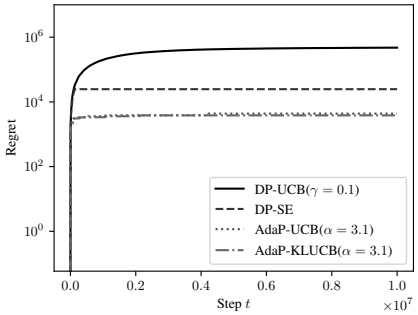


(d) $\epsilon = 1$

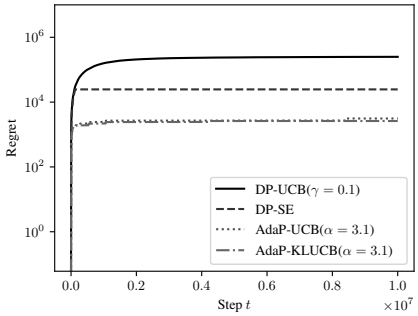
Figure 4: Evolution of regret over time for DP-UCB, DP-SE, AdaP-UCB, and AdaP-KLUCB under C_1 for different values of the privacy budget ϵ. AdaP-KLUCB achieves the lowest regret.



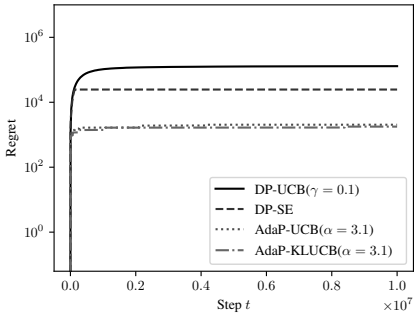
(a) $\epsilon = 0.1$



(b) $\epsilon = 0.25$

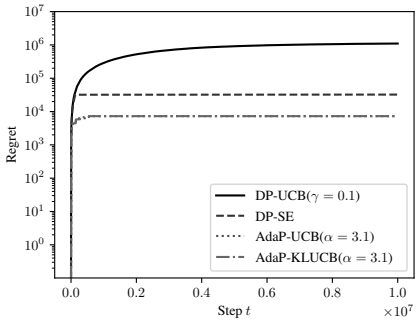


(c) $\epsilon = 0.5$

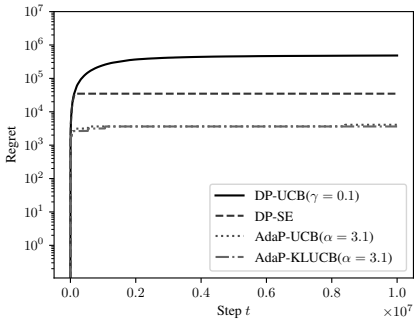


(d) $\epsilon = 1$

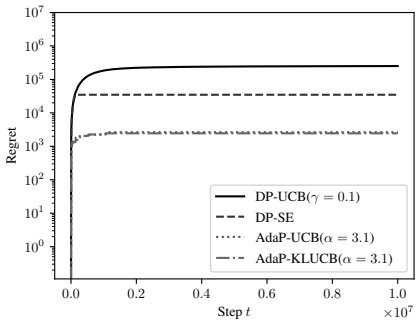
Figure 5: Evolution of regret over time for DP-UCB, DP-SE, AdaP-UCB, and AdaP-KLUCB under C_2 for different values of the privacy budget ϵ. AdaP-KLUCB achieves the lowest regret.



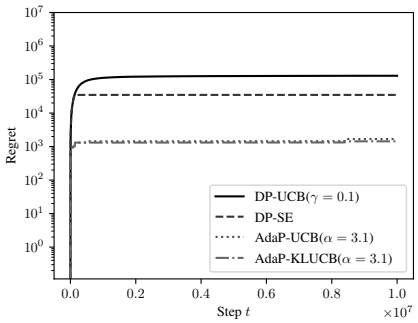
(a) $\epsilon = 0.1$



(b) $\epsilon = 0.25$

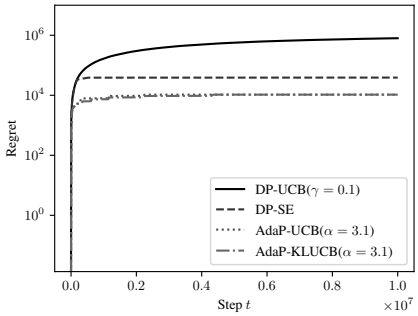


(c) $\epsilon = 0.5$

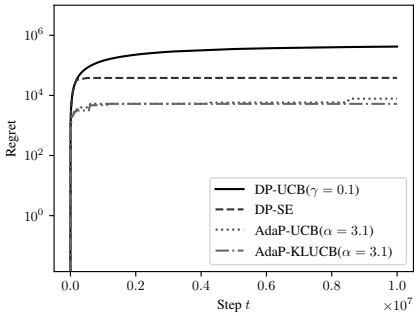


(d) $\epsilon = 1$

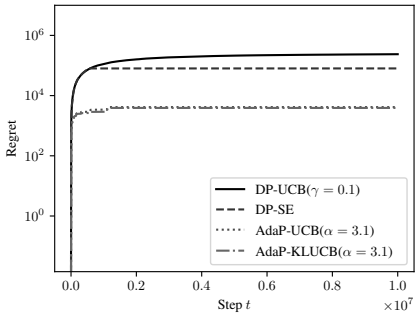
Figure 6: Evolution of regret over time for DP-UCB, DP-SE, AdaP-UCB, and AdaP-KLUCB under C_3 for different values of the privacy budget ϵ. AdaP-KLUCB achieves the lowest regret.



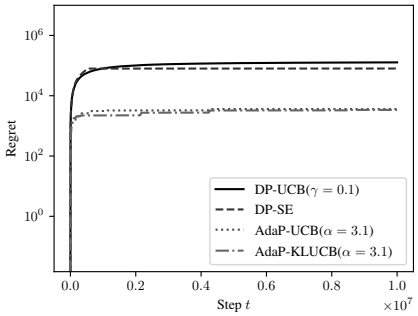
(a) $\epsilon = 0.1$



(b) $\epsilon = 0.25$



(c) $\epsilon = 0.5$



(d) $\epsilon = 1$

Figure 7: Evolution of regret over time for DP-UCB, DP-SE, AdaP-UCB, and AdaP-KLUCB under C_4 for different values of the privacy budget ϵ. AdaP-KLUCB achieves the lowest regret.

G Extended Experimental Analysis

G.1 Experimental Setup

In this section, we perform additional experiments to compare AdaP-UCB and AdaP-KLUCB with respect to the existing bandit algorithms satisfying global DP, i.e. DP-SE (Sajed and Sheffet, 2019) and DP-UCB (Mishra and Thakurta, 2015). We test the four algorithms in the four bandit environments with Bernoulli distributions, as defined by Sajed and Sheffet (2019), namely

$$C_1 = \{0.75, 0.70, 0.70, 0.70, 0.70\}, \quad C_2 = \{0.75, 0.625, 0.5, 0.375, 0.25\},$$
$$C_3 = \{0.75, 0.53125, 0.375, 0.28125, 0.25\}, \quad C_4 = \{0.75, 0.71875, 0.625, 0.46875, 0.25\}.$$

For each bandit environment, we implement the algorithms with $\epsilon \in \{0.1, 0.25, 0.5, 1\}$. We set $\alpha = 3.1$ to comply with the regret upper bounds of AdaP-UCB and AdaP-KLUCB. We set $\gamma = 0.1$ for DP-UCB and $\beta = 1/T$. All the algorithms are implemented in Python (version 3.8) and are tested with an 8 core 64-bits Intel i5@1.6 GHz CPU. We run each algorithm 20 times, and plot their average regrets over the runs in Figures 4, 5, 6, and 7.In Section 5, we include Figure 2 to illustrate the evolution of the regret for the four algorithms with environment C_2 and $\epsilon = 1$.

G.2 Experimental Results

Here, we summarise the observations obtained from the experimental results.

Comparative Performance. All the experiments validate that AdaP-KLUCB *is the most optimal algorithm satisfying ϵ-global DP for stochastic bandits.* Both AdaP-UCB and AdaP-KLUCB achieve similar regret, but AdaP-KLUCB is slightly better in all the cases studied. This observation matches the proven upper bounds, and also reflects similar improvement that KL-UCB brings over UCB in non-private bandits.

Dependence of Regret on ϵ. As predicted by the theoretical analysis, AdaP-UCB and AdaP-KLUCB have different regret depending on ϵ: the regret is smaller for low-privacy regimes. This is also the case for DP-UCB. However, DP-SE have the same performance for different choices of ϵ and echoes the experimental results presented in the original paper (Sajed and Sheffet, 2019).

The Shapes of the Regrets. DP-UCB has a regret shaped like the regret of the classic UCB algorithm. The algorithm chooses a different action at each time-step allowing it to still choose exploratory actions. On the other hand, due to the successive elimination, DP-SE "commits" at a certain step to one action (the optimal action with high probability). Thus, the shape of regret for DP-SE is piece-wise linear. On the other hand, AdaP-UCB and AdaP-KLUCB are a trade-off of both strategies: *due to the doubling, both algorithms "commit" for long episodes to near-optimal actions*, while *still explore the sub-optimal actions for short episodes*.

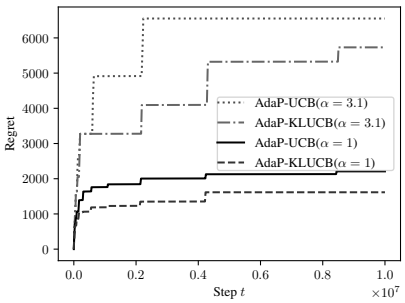


Figure 8: Evolution of regret over time for AdaP-UCB and AdaP-KLUCB for different values of α with C_1 and $\epsilon = 1$. $\alpha = 1$ performs better.

G.3 Choice of α

α controls the width of the optimistic confidence bound. Specifically, it dictates that the real mean is smaller than the optimistic index with high probability, i.e. with probability $1 - \frac{1}{t^\alpha}$ at step t. The requirement that $\alpha > 3$ is due to our analysis of the algorithm. To be specific, the requirement that $\alpha > 3$ is needed to use a sum-integral inequality to bound Term 2 of Step 3 in the proof of Theorem 11. We leave it for future work to relax this requirement.

The experiments are done with $\alpha = 3.1$ to comply with the theoretical analysis. As shown in Figure 8, choosing $\alpha = 1$ works better experimentally. This observation complies with the theoretical results, since the dominant terms in the regret upper bounds of both AdaP-UCB and AdaP-KLUCB are multiplicative in α. A tighter analysis might give us a bound for $\alpha = 1$ and close the multiplicative gap between the regret's lower and upper bound. Reflecting this phenomenon in the analysis will be an interesting future work to pursue.