Journal of Machine Learning Research 27 (2026) 1-85 Submitted 6/24; Revised 1/26; Published 4/26

# Differentially Private Best-Arm Identification

**Achraf Azize**[*†] ACHRAF.AZIZE@ENSAE.FR
*FairPlay Joint Team, CREST, ENSAE Paris*

**Marc Jourdan**[*‡] MARC.JOURDAN@EPFL.CH
*EPFL, Lausanne, Switzerland*

**Aymen Al Marjani**[§] ALMARJAN@AMAZON.LU
*Amazon*

**Debabrota Basu** DEBABROTA.BASU@INRIA.FR
*Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France*

**Editor:** Raef Bassily

## Abstract

Best Arm Identification (BAI) problems are progressively used for data-sensitive applications, such as designing adaptive clinical trials, tuning hyper-parameters, and conducting user studies. Motivated by the data privacy concerns invoked by these applications, we study the problem of BAI with fixed confidence in both the local and central models, i.e. under $\epsilon$-local and $\epsilon$-global Differential Privacy (DP). First, to quantify the cost of privacy, we derive lower bounds on the sample complexity of any $\delta$-correct BAI algorithm satisfying $\epsilon$-global DP or $\epsilon$-local DP. Our lower bounds suggest the existence of two privacy regimes. In the high-privacy regime, the hardness depends on a coupled effect of privacy and novel information-theoretic quantities involving the Total Variation distance. In the low-privacy regime, the lower bounds reduce to the non-private lower bounds. We propose $\epsilon$-local DP and $\epsilon$-global DP variants of a Top Two algorithm, namely CTB-TT and AdaP-TT$^\star$, respectively. For $\epsilon$-local DP, CTB-TT is asymptotically optimal by plugging in a private estimator of the means based on Randomised Response. For $\epsilon$-global DP, our private estimator of the mean runs in arm-dependent adaptive episodes and adds Laplace noise to ensure a good privacy-utility trade-off. By adapting the transportation costs, the expected sample complexity of AdaP-TT$^\star$ reaches the asymptotic lower bound in the asymptotic high-privacy regime, up to a small multiplicative constant.

**Keywords:** differential privacy, multi-armed bandits, best arm identification, fixed confidence, top two algorithm

∗. Equal Contribution
†. This work was done when Achraf Azize was a PhD student in the Scool team of Inria Lille.
‡. This work was done when Marc Jourdan was a PhD student in the Scool team of Inria Lille.
§. This work was done when Aymen Al Marjani was a PhD student at ENS Lyon.

## 1. Introduction

We study the stochastic multi-armed *bandit* problem (Lattimore and Szepesvári, 2020), which allows us to reflect on fundamental information-utility trade-offs involved in interactive sequential learning. Specifically, in a bandit problem, a learning *agent* is exposed to interact with $K$ unknown probability distributions $\{\nu_1, \dots, \nu_K\}$ with bounded expectations, referred to as the *reward distributions* (or *arms*). $\boldsymbol{\nu} \triangleq \{\nu_1, \dots, \nu_K\}$ is called *a bandit instance*. At every step $n > 0$, the agent chooses to interact with one of the reward distributions $\nu_{a_n}$ for an arm $a_n \in [K]$, and obtains a sample (or *reward*) $r_n$ from it. The goal of the agent can be of two types: (a) maximise the reward accumulated over time, or equivalently to minimise the regret, and (b) to find the reward distribution (or arm) with the highest expected reward. The first problem is called the regret-minimisation problem (Auer et al., 2002), while the second one is called the *Best Arm Identification (BAI)* problem (Kaufmann et al., 2016). In this paper, we focus on the BAI problem, i.e. to compute $a^\star(\boldsymbol{\nu}) \triangleq \arg\max_{a\in[K]} \mathop{\mathbb{E}}_{r\sim\nu_a}[r] \triangleq \arg\max_{a\in[K]} \mu_a$.

With its advent in 1950s (Bechhofer, 1954, 1958) and recent resurgence (Mannor and Tsitsiklis, 2004; Gabillon et al., 2012; Jamieson et al., 2014; Kaufmann et al., 2016), BAI has been extensively studied with different structural assumptions: Fixed-confidence (Jamieson and Nowak, 2014); Fixed-budget (Carpentier and Locatelli, 2016); Non-stochastic (Jamieson and Talwalkar, 2016); Best-of-both-worlds (Abbasi-Yadkori et al., 2018); Linear (Soare et al., 2014). In this paper, we specifically investigate the *Fixed Confidence BAI problem*, in brief FC-BAI, that yields a $\delta$-correct recommendation $\widehat{a} \in [K]$, i.e. the probability that the algorithm stops and returns $\widehat{a} \neq a^\star(\boldsymbol{\nu})$ is upper bounded by $\delta$. FC-BAI is increasingly deployed for different applications, such as clinical trials (Aziz et al., 2021), hyper-parameter tuning (Li et al., 2017), communication networks (Lindståhl et al., 2022), online advertisement (Chen et al., 2014), crowd-sourcing (Zhou et al., 2014), user studies (Losada et al., 2022), and pandemic mitigation (Libin et al., 2019) to name a few. All of these applications often involve the sensitive and personal data of users, which raises serious data privacy concerns (Tucker et al., 2016), as illustrated in Example 1.

**Example 1 (Adaptive dose finding trial)** *In a dose-finding trial, one physician decides $K$ possible dose levels of a medicine based on preliminary studies—$K \in \{3, \dots, 10\}$ in practice (Aziz et al., 2021). At each step $n$, a patient is chosen from a local pool of volunteers and a dose level $a_n \in [K]$ is applied to the patient. Following that, the effectiveness of the dose on the patient, i.e. $r_n \in \mathbb{R}$ is observed. The goal of the physician is to recommend after the trial, which dose level is most effective on average, i.e. the dose level $a^\star$ that maximises the expected reward. Here,* every application of a dose level and the patient's reaction to it exposes information regarding the medical conditions of the patient. *Additionally, at each step $n$ of an adaptive sequential trial, the physician can use an FC-BAI algorithm that observes the previous history of dose levels $\{a_t\}_{t<n}$ and their effectiveness $\{r_t\}_{t<n}$ to decide on the next dose level $a_n$ to test. When releasing the experimental findings of the trial to health authorities, the physician should thoroughly detail the experimental protocol. This includes the dose allocated to each patient $\{a_t\}_{t\leq n}$ and the final recommended dose level $a^\star$. Thus, even if the sequence of reactions to doses $\{r_t\}_{t\leq n}$ is kept secret,* publishing the sequence of chosen dose levels $\{a_t\}_{t\leq n}$ and the final recommended dose level $a^\star$ computed using the history can leak information regarding patients involved in the trial.

This example demonstrates the need for privacy in best-arm identification. In this paper, we investigate *privacy-utility trade-offs for a privacy-preserving algorithm in FC-BAI*. Specifically, we use the celebrated Differential Privacy (DP) (Dwork and Roth, 2014) as the framework to preserve data privacy. DP ensures that an algorithm's output is unaffected by changes in input by 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*'. Popular ways to achieve DP include Randomised Response (Warner, 1965) or injecting a calibrated amount of noise, from a Laplace (Dwork and Roth, 2014) or Gaussian distribution (Dong et al., 2022), into the algorithm. The scale of the noise is set to be proportional to the algorithm's sensitivity and inversely proportional to the privacy budget $\epsilon$. Specifically, we study $\epsilon$-*local DP*, where users do not trust the data curator, and $\epsilon$-*global DP*, where users trust the centralised decision-maker with access to the raw sensitive rewards. For example, in an adaptive dose-finding trial, the patients could trust the physician conducting the trial. In that case, at any time $n$, she has access to all the true history $\{a_t, r_t\}_{t<n}$, and it is her duty to design an algorithm such that *publishing* $\{a_t\}_{t\leq n}$ *and the recommended optimal dose* $a^\star$ *obeys* $\epsilon$*-global DP given the sensitive input*, i.e. the effectiveness of the dose levels on the patients $\{r_t\}_{t\leq n}$. Without this trust from the user, she has only access to a perturbed history $\{a_t, \tilde{r}_t\}_{t<n}$, where $\tilde{r}_t$ is a perturbed observation of the true observation $r_t$ which ensures $\epsilon$-local DP. We define the notions of $\epsilon$-*local DP* and $\epsilon$-*global DP* for BAI rigorously in Section 2.

For different settings of bandits, the costs of $\epsilon$-local DP or $\epsilon$-global DP and optimal algorithm design techniques are widely studied in the regret-minimisation problem (Mishra and Thakurta, 2015; Tossou and Dimitrakakis, 2016; Sajed and Sheffet, 2019; Shariff and Sheffet, 2018; Neel and Roth, 2018; Basu et al., 2019; Azize and Basu, 2022, 2024). Recently, a problem-dependent lower bound on regret of stochastic multi-armed bandits with $\epsilon$-global DP and an algorithm matching the regret lower bound is proposed by Azize and Basu (2022). In contrast, DP is meagerly studied in the FC-BAI problem of bandits (Sajed and Sheffet, 2019; Kalogerias et al., 2021). Though *efficient* algorithm design in FC-BAI literature is traditionally propelled by deriving tight lower bounds, we do not have any explicit sample complexity lower bound for FC-BAI satisfying $\epsilon$-local DP or $\epsilon$-global DP. By "efficient" algorithm, we refer to the FC-BAI algorithms that aim to minimise the expected number of samples required (i.e. *sample complexity*) to find a $\delta$-correct recommendation. Presently, we know neither the minimal cost in terms of sample complexity for ensuring DP in FC-BAI, nor the feasibility of efficient algorithm design to achieve the minimal cost.

### 1.1 Contributions

Motivated by this gap in the literature, this paper answers the following two questions:

*A. How many additional samples a BAI strategy must need to ensure $\epsilon$-local DP?*

*B. How many additional samples a BAI strategy must need to ensure $\epsilon$-global DP?*

Gathered in Appendix B for convenience, the notation are introduced in context.

#### 1.1.1 Lower Bounds

First, we derive a lower bound on the expected sample complexity of any $\delta$-correct FC-BAI algorithm to ensure either $\epsilon$-local DP (Theorem 9 and Corollary 10) or $\epsilon$-global DP

(Theorem 16 and Corollary 17). Due to the $\delta$-correctness and the DP constraints, each of the lower bounds corresponds to the minimum of two characteristic times. The first one is the KL characteristic time $T^\star_{\mathrm{KL}}(\boldsymbol{\nu})$ of the non-private FC-BAI (Kaufmann et al., 2016) (Lemma 4). The second one depends on the privacy $\epsilon$ and novel information-theoretic quantities depending on the Total Variation (TV) distance: the $\mathrm{TV}^2$ characteristic time $T^\star_{\mathrm{TV}^2}(\boldsymbol{\nu})$ for $\epsilon$-local DP and the TV characteristic time $T^\star_{\mathrm{TV}}(\boldsymbol{\nu})$ for $\epsilon$-global DP. As for $\epsilon$-global DP regret minimisation (Azize and Basu, 2022), the lower bound indicates that there are two regimes of hardness depending on $\epsilon$ and the aforementioned characteristic times. For lower levels of privacy (i.e. higher $\epsilon$), the expected sample complexity matches the non-private FC-BAI. However, for higher levels of privacy (i.e. lower $\epsilon$), the expected sample complexity depends both on the privacy budget $\epsilon$ and the $\mathrm{TV}^2$ or TV characteristic time. To derive the lower bound for $\epsilon$-global DP, we provide an $\epsilon$-global DP version of the "change-of-measure" lemma (Kaufmann and Kalyanakrishnan, 2013) (Lemma 15), which we prove using a sequential coupling argument.

### 1.1.2 Algorithm Design

We propose algorithms which are $\delta$-correct, and either $\epsilon$-local DP or $\epsilon$-global DP. While most existing asymptotically optimal FC-BAI algorithms can be modified to tackle DP, we consider the class of Top Two algorithms (Russo, 2016) due to their good empirical performances, low computational cost, and easy implementation. As a case study, we consider the TTUCB meta-algorithm based on the work of (Jourdan and Degenne, 2024). Table 1 summarises the different instances that we propose. We highlight that our private wrappers could be used for other FC-BAI algorithms.

*A. For $\epsilon$-local DP,* we propose the CTB-TT algorithm. CTB-TT plugs in the CTB($\epsilon$) estimator of the means, which is $\epsilon$-local DP (Ren et al., 2020, Lemma 11), into the TTUCB algorithm for $\sigma$-sub-Gaussian distributions, which is $\delta$-correct.

*B. For $\epsilon$-global DP,* we propose the DAF($\epsilon$) estimator of the means, which is $\epsilon$-global DP (Lemma 19). It relies on three ingredients: *adaptive episodes with doubling per arm*, *forgetting*, and *adding calibrated Laplacian noise*. Using the DAF($\epsilon$) estimator in the TTUCB meta-algorithm, we propose the AdaP-TT and the AdaP-TT$^\star$ algorithms. As a plug-in approach, AdaP-TT uses the non-private transportation costs both in the TC challenger and in the generalised likelihood ratio (GLR) stopping rule, which is shown to be $\delta$-correct by adding a privacy term to the stopping threshold (Lemma 20). As a lower bound based approach, AdaP-TT$^\star$ adapts the transportation costs to account for $\epsilon$-global privacy both in the TC challenger and in the GLR stopping rule, which is shown to be $\delta$-correct by modifying the privacy term in the stopping threshold (Lemma 22).

### 1.1.3 Upper Bounds

We show that the proposed algorithms exhibit upper bounds that match the lower bounds up to multiplicative constants. We highlight that our generic asymptotic analysis can be applied to any Top Two algorithms, since it builds on the one of Jourdan et al. (2022).

*A.* As CTB-TT is equivalent to running TTUCB on a modified Bernoulli instance $\boldsymbol{\nu}_\epsilon$, it recovers the asymptotic and non-asymptotic upper bounds on the expected sample complexity derived in Jourdan and Degenne (2024). The asymptotic upper bound matches the

asymptotic lower bound up to a constant multiplicative term, 2 when $\epsilon \to +\infty$ and 4 when $\epsilon \to 0$. Our experiments confirm the good performance of CTB-TT, and the existence of two hardness regimes for $\epsilon$-local DP (Section 5.1).

*B.* Using the DAF($\epsilon$) estimator yields a batched algorithm with adaptive and data-dependent changes of episodes. While our analysis is inspired by the one of Jourdan et al. (2022), studying AdaP-TT and AdaP-TT$^\star$ requires carefully quantifying the effects of doubling, forgetting, and adding noise. We derive an asymptotic upper bound on the expected sample complexity of AdaP-TT (Theorem 21) and AdaP-TT$^\star$ (Theorem 23). In the non-private regime of $\epsilon \to +\infty$, both algorithms recover the asymptotic lower bound for Gaussian distributions, up to multiplicative constants (16 and 8 respectively), with solely $\mathcal{O}\left(K \log_2(T^\star_{\mathrm{KL}}(\boldsymbol{\nu}) \log(1/\delta))\right)$ rounds of adaptivity. When $\epsilon \to 0$, AdaP-TT$^\star$ achieves the asymptotic lower bound up to a multiplicative constant 48, while AdaP-TT only recovers it for instances where the mean gaps have the same order of magnitude. Our experiments show the good performance of our algorithms compared to DP-SE (Sajed and Sheffet, 2019), which can be adapted for FC-BAI (see Section 4.4 for a detailed comparison). They confirm the existence of two hardness regimes for $\epsilon$-global DP, as well as the empirical superiority of AdaP-TT$^\star$ over AdaP-TT when $\epsilon \to 0$ (Section 5.2).

### 1.2 Outline

After presenting Differential Privacy and Best-Arm Identification in the Fixed-Confidence Setting in Section 2, we formulate the problem of private best-arm identification. We present lower bounds and matching upper bounds for $\epsilon$-local DP FC-BAI (Section 3) and $\epsilon$-global DP FC-BAI (Section 4). Our algorithms are studied empirically in Section 5. For clarity, standalone pseudocodes of our algorithms are given in Appendix C.

## 2. Differential Privacy and Best-Arm Identification

In this section, we provide relevant background information on Differential Privacy (DP) in Section 2.1, and Best-Arm Identification in the Fixed-Confidence Setting (FC-BAI) in Section 2.2. Then, we formulate the problem of private best-arm identification (FC-BAI with DP) in Section 2.3, under both the local and global trust models.

### 2.1 Background: Differential Privacy

Differential Privacy (DP) ensures the protection of an individual's sensitive information when her data is used for analysis. A randomised algorithm satisfies DP if the output of the algorithm stays almost the same, regardless of whether any single individual's data is included in or excluded from the input. One way of achieving DP is by adding controlled noise to the algorithm's output.

**Definition 1 ($(\epsilon, \delta)$-DP, Dwork et al., 2006)** *A randomised algorithm $\mathcal{A}$ is $(\epsilon, \delta)$-DP (Differential Privacy) if for any two neighbouring data sets $\mathcal{D}$ and $\mathcal{D}'$ that differ only in one entry, i.e. $d_{Ham}(\mathcal{D}, \mathcal{D}') = 1$, and for all sets of output $\mathcal{O} \subseteq \mathrm{Range}(\mathcal{A})$,*

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

*where the probability space is over the coin flips of the mechanism $\mathcal{A}$, and $(\epsilon, \delta) \in \mathbb{R}^{\geq 0} \times \mathbb{R}^{\geq 0}$. If $\delta = 0$, we say that $\mathcal{A}$ satisfies $\epsilon$-DP. A lower privacy budget $\epsilon$ implies higher privacy.*

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

**Theorem 2 (Laplace mechanism, Theorem 3.6, Dwork and Roth, 2014)** *Let $f : \mathcal{X} \to \mathbb{R}^d$ be an algorithm with sensitivity $s(f) \triangleq \max_{\mathcal{D},\mathcal{D}' \ s.t \ |\mathcal{D}-\mathcal{D}'|_{\text{Hamming}}=1} \|f(\mathcal{D}) - f(\mathcal{D}')\|_1$, where $\|\cdot\|_1$ is the $L_1$ norm. If samples $\{N_i\}_{i=1}^d$ are generated independently from $Lap\left(\frac{s(f)}{\epsilon}\right)$, then the output injected with the noise, i.e. $f(\mathcal{D}) + [N_1, \ldots, N_d]$, satisfies $\epsilon$-DP.*

We also study the setting of *local differential privacy*, where users do not trust the data curator, i.e. the entity collecting the data. Local DP is one of the oldest formulations of privacy, dating back to Warner (1965), who advocated it as a solution to what he called "evasive answer bias" in survey sampling.

**Definition 3 ($\epsilon$-local DP, Dinur and Nissim, 2003; Evfimievski et al., 2003)** *A randomised algorithm $\mathcal{M}$ satisfies $\epsilon$-local DP if for any pair of input values $x, x' \in \mathcal{D}$, and for all sets of output $\mathcal{O} \subseteq \text{Range}(\mathcal{M})$,*

$$\Pr[\mathcal{M}(x) \in \mathcal{O}] \leq e^{\epsilon} \Pr\left[\mathcal{M}(x') \in \mathcal{O}\right],$$

*where the probability space is over the coin flips of the mechanism $\mathcal{M}$, and for some $\epsilon \in \mathbb{R}^{\geq 0}$. The perturbation mechanism $\mathcal{M}$ is applied to each user record independently.*

For binary attributes, the Randomised Response (RR) mechanism (Warner, 1965) is a popular way to achieve $\epsilon$-local DP. The idea is to output the true value of a user's response with probability $e^{\epsilon}/(e^{\epsilon}+1)$ and output the opposite value with probability $1/(e^{\epsilon}+1)$. To make it suitable for larger discrete domains, a Generalised Randomised Response (GRR) is proposed in Kairouz et al. (2016). For continuous numerical data statistics, adding Laplace noise to each data record achieves local DP as well.

### 2.2 Background: Best Arm Identification in the Fixed-Confidence Setting

In this section, we first present the Best-arm identification (BAI) problem, a BAI strategy and $\delta$-correctness. Then, we present a lower bound on the sample complexity of any $\delta$-correct BAI strategy. Finally, we discuss algorithms in the BAI literature which match the sample complexity lower bound. We focus on the Top Two family of algorithms since they enjoy both theoretical optimality and good empirical performance.

#### 2.2.1 The Best Arm Identification Problem

Best-arm identification (BAI) is a pure exploration problem that aims to identify the optimal arm. It has been studied in two major theoretical frameworks (Audibert et al., 2010;

Gabillon et al., 2012; Jamieson and Nowak, 2014; Garivier and Kaufmann, 2016): the fixed-confidence and fixed-budget setting. In the fixed-budget setting, the objective is to minimise the probability of misidentifying a correct answer with a fixed number of samples $T$. We consider the fixed-confidence setting (FC-BAI), in which the learner aims at minimising the number of samples used to identify a correct answer with confidence $1-\delta \in (0,1)$. [1] To achieve this, the learner defines an FC-BAI strategy to interact with the bandit instance $\boldsymbol{\nu} = \{\nu_a\}_{a\in[K]} \in \mathcal{F}^K$, consisting of $K$ arms with finite means $\{\mu_a\}_{a\in[K]} \in (0,1)^K$. We assume that there is a unique best arm $a^\star(\boldsymbol{\nu})$ defined as $a^\star(\boldsymbol{\nu}) = \arg\max_{a\in[K]} \mu_a$. The set of distributions $\mathcal{F}$ will depend on the considered result, e.g. Bernoulli distributions, bounded distributions on $[0,1]$ or $\sigma$-sub-Gaussian distributions. A distribution $\kappa$ is $\sigma$-sub-Gaussian if it satisfies $\mathbb{E}_{X\sim\kappa}[e^{\lambda(X-\mathbb{E}_{X\sim\kappa}[X])}] \leq e^{\sigma^2\lambda^2/2}$ for all $\lambda \in \mathbb{R}$.

We denote the action played at step $n$ by $a_n$, and the corresponding observed reward by $r_n \sim \nu_{a_n}$. The $\sigma$-algebra $\mathcal{H}_n = \sigma(a_1, r_1, \ldots, a_n, r_n)$ is the history of actions played and rewards collected at the end of time $n$. We augment the action set by a *stopping action* $\top$, and write $a_n = \top$ to denote that the algorithm has stopped before step $n$. A FC-BAI strategy $\pi$ is composed of

*i. A pair of sampling and stopping rules* $(\mathrm{S}_n : \mathcal{H}_{n-1} \to \mathcal{P}([|1,K|]\cup\{\top\}))_{n\geq 1}$. For an action $a \in [K]$, $\mathrm{S}_n(a \mid \mathcal{H}_{n-1})$ denotes the probability of playing action $a$ given history $\mathcal{H}_{n-1}$. On the other hand, $\mathrm{S}_n(\top \mid \mathcal{H}_{n-1})$ is the probability of the algorithm halting given $\mathcal{H}_{n-1}$. For any history $\mathcal{H}_{n-1}$, a consistent sampling and stopping rule $\mathrm{S}_n$ satisfies $\mathrm{S}_n(\top \mid \mathcal{H}_{n-1}) = 1$ if $\top$ has been played before $n$.

*ii. A recommendation rule* $(\mathrm{Rec}_n : \mathcal{H}_{n-1} \to \mathcal{P}([|1,K|]))_{n>1}$. A recommendation rule dictates $\mathrm{Rec}_n(a \mid \mathcal{H}_{n-1})$, i.e. the probability of returning action $a$ as a guess for the best action given $\mathcal{H}_{n-1}$.

We denote by $\tau_\delta$ the **stopping time** (or **sample complexity**) of the algorithm, i.e. the first step $n$ demonstrating $a_n = \top$. A FC-BAI strategy $\pi$ is called $\delta$**-correct** for a class of bandit instances $\mathcal{M} \subseteq \mathcal{F}^K$, if for every instance $\boldsymbol{\nu} \in \mathcal{M}$, $\pi$ recommends $\widehat{a}$ as the optimal action $a^\star(\boldsymbol{\nu})$ with probability at least $1-\delta$, i.e. $\mathbb{P}_{\boldsymbol{\nu}}(\tau_\delta < +\infty, \widehat{a} = a^\star(\boldsymbol{\nu})) \geq 1-\delta$.

### 2.2.2 Lower Bound on the Expected Sample Complexity

Being $\delta$-correct imposes a lower bound on the expected sample complexity on any instance.

**Lemma 4 (Garivier and Kaufmann 2016)** *For all $\delta$-correct FC-BAI strategy and all instances $\boldsymbol{\nu} \in \mathcal{M}$ with unique best arm, we have that $\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta] \geq T^\star_{\mathrm{KL}}(\boldsymbol{\nu}) \log(1/(2.4\delta))$ with*

$$T^\star_{\boldsymbol{d}}(\boldsymbol{\nu})^{-1} \triangleq \sup_{\omega\in\Sigma_K} \inf_{\boldsymbol{\lambda}\in \mathrm{Alt}(\boldsymbol{\nu})} \sum_{a\in[K]} \omega_a \boldsymbol{d}(\nu_a, \lambda_a)\,, \tag{1}$$

*where the probability simplex is denoted by $\Sigma_K \triangleq \{\omega \in [0,1]^K \mid \sum_{a=1}^K \omega_a = 1\}$ and the set of alternative instances is $\mathrm{Alt}(\boldsymbol{\nu}) \triangleq \{\boldsymbol{\lambda} \in \mathcal{M} \mid a^\star(\boldsymbol{\lambda}) \neq a^\star(\boldsymbol{\nu})\}$, i.e. the bandit instances with a different optimal arm than $\boldsymbol{\nu}$. For two probability distributions $\mathbb{P}, \mathbb{Q}$ on $(\Omega, \mathcal{F})$, the KL divergence is $\mathrm{KL}(\mathbb{P} \parallel \mathbb{Q}) \triangleq \int \log\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(\omega)\right) \mathrm{d}\mathbb{P}(\omega)$, when $\mathbb{P} \ll \mathbb{Q}$, and $+\infty$ otherwise.*

1. We remind not to confuse risk level $\delta$ with the $\delta$ of $(\epsilon,\delta)$-DP. Hereafter, we consider $\epsilon$-global DP as the privacy definition, and $\delta$ always represents the risk (or probability of mistake) of the BAI strategy.

The characteristic time $T^\star_{\mathbf{d}}(\boldsymbol{\nu})$ in the lower bound is the value of a two-player zero-sum game. The MIN player plays instances $\boldsymbol{\lambda} \in \mathcal{M}$ having an answer different from $a^\star(\boldsymbol{\nu})$, while staying close to $\boldsymbol{\nu}$ in terms of allocation-reweighted sum of $\mathbf{d}(\nu_a, \lambda_a)$, in order to confuse the MAX player. The MAX player plays an arm allocation $w \in \Sigma_K$ to distinguish between $\boldsymbol{\nu}$ and $\boldsymbol{\lambda}$. The measure $\mathbf{d}$ captures the "distinguishability" between instances. In the non-private lower bounds, this is captured by the KL divergence for parametric distributions (Lemma 4) and by the Kinf (i.e., inf KL under mean constraint) for non-parametric distributions (Agrawal et al., 2020).

Early FC-BAI algorithms failed to reach the lower bound of Theorem 4, e.g. Successive Elimination (SE) based algorithms (Even-Dar et al., 2006) or confidence bounds based algorithms, e.g. LUCB (Kalyanakrishnan et al., 2012) or lil'UCB (Jamieson et al., 2014). Inspired by this lower bound, many algorithms have been designed to tackle FC-BAI. The Track-and-Stop algorithm (Garivier and Kaufmann, 2016) is the first algorithm to reach asymptotic optimality, by sequentially solving the optimisation problem $T^\star_{\mathrm{KL}}(\boldsymbol{\nu_n})$ and tracking the associated optimal weights. To reduce the computational cost of Track-and-Stop, several asymptotically optimal algorithms have been proposed recently: online optimisation-based approach, e.g. game-based algorithm (Degenne et al., 2019) or FWS (Wang et al., 2021), and Top Two algorithms (Russo, 2016). While most algorithms can be modified to tackle $\epsilon$-DP, we consider the Top Two algorithms due to their great empirical performance and easy implementation. At every step, a Top Two sampling rule selects the next arm to sample from among two candidate arms, a leader and a challenger. In recent years, numerous variants of Top Two algorithms have been analysed and shown to be asymptotically optimal (Russo, 2016; Qin et al., 2017; Shang et al., 2020; Jourdan et al., 2022; You et al., 2023; Jourdan et al., 2024). In particular, we consider one particular case study, i.e. the TTUCB algorithm (Jourdan and Degenne, 2024), but our approach can be directly adapted to any other Top Two algorithms.

**Remark 5** *The private estimators of the means and the private stopping rules presented in Sections 3 and 4 could be used with most existing existing FC-BAI algorithm. The resulting algorithms will be private, $\delta$-correct and have near-optimal asymptotic sample complexity. In this work, we consider and rigorously analyse the Top Two algorithms since they are simple algorithms enjoying both strong theoretical guarantees and empirical performance.*

### 2.2.3 The TTUCB Meta-algorithm

Since we will propose several private algorithm building on top of TTUCB, we propose a TTUCB meta-algorithm (Algorithm 1). To instantiate it, one should specify: a parameter $\beta \in (0,1)$ (e.g. $\beta = 1/2$), confidence bonuses $(b_a)_{a\in[K]}$ where $b_a : \mathbb{N}^K \to \mathbb{R}_+$, transportation costs $(W_{a,b})_{(a,b)\in[K]^2}$ where $W_{a,b} : \mathbb{R}^K \times \mathbb{N}^K \to \mathbb{R}_+$, estimator mechanisms $(\mathrm{ESTIMATOR}_a)_{a\in[K]}$ computing data-dependent estimators $(\mu_{n,a})_{(n,a)\in\mathbb{N}\times[K]}$ based on local counts $(\tilde{N}_{n,a})_{(n,a)\in\mathbb{N}\times[K]}$, and generalised likelihood ratio (GLR) stopping conditions

$$\forall (a,b) \in [K]^2, \quad \mathrm{STOP}_{a,b}(\tilde{\mu}, \omega, \delta) = W_{a,b}(\tilde{\mu}, \omega) - c_{a,b}(\omega, \delta)\,, \tag{2}$$

where $(c_{a,b})_{(a,b)\in[K]^2}$ where $c_{a,b} : \mathbb{N}^K \times (0,1) \to \mathbb{R}_+$ are stopping thresholds.

**Algorithm 1** TTUCB Meta-algorithm

**Input:** setting parameter $\delta \in (0,1)$, algorithmic hyperparameter $\beta \in (0,1)$, e.g., $\beta = 1/2$, confidence bonuses $(b_a)_{a\in[K]}$, transportation costs $(W_{a,b})_{(a,b)\in[K]^2}$, estimator mechanisms $(\text{ESTIMATOR}_a)_{a\in[K]}$ and stopping conditions $(\text{STOP}_{a,b})_{(a,b)\in[K]^2}$.
**Initialisation:** Observe $r_a \sim \nu_a$ for all arms $a \in [K]$; Initialise the estimators $\tilde{\mu}_{n,a} = r_a$ and $N_{n,a} = \tilde{N}_{n,a} = 1$ where $n = K+1$;
**for** $n > K$ **do**
  Set arm $\hat{a}_n = \arg\max_{a\in[K]} \tilde{\mu}_{n,a}$; ▷ Recommendation rule
  **if** $\text{STOP}_{\hat{a}_n,a}(\tilde{\mu}_n, \tilde{N}_n, \delta) \geq 0$ for all $a \neq \hat{a}_n$ **then** ▷ Stopping rule
    Set $a_n = \top$ and **return** $\hat{a}_n$;
  **end if**
  Set arm $B_n = \arg\max_{a\in[K]} \left\{\tilde{\mu}_{n,a} + b_a(\tilde{N}_n)\right\}$; ▷ UCB leader
  Set arm $C_n = \arg\min_{a\neq B_n} W_{B_n,a}(\tilde{\mu}_n, N_n)$; ▷ TC challenger
  Set arm $a_n = B_n$ if $N_{n,B_n}^{B_n} \leq \beta L_{n+1,B_n}$, and $a_n = C_n$ otherwise; ▷ $\beta$-tracking
  Pull $a_n$ and observe $r_n \sim \nu_{a_n}$; ▷ Sampling rule
  Set $N_{n+1,a_n} \leftarrow N_{n,a_n} + 1$, $L_{n+1,B_n} \leftarrow L_{n,B_n} + 1$, $N_{n+1,B_n}^{B_n} \leftarrow N_{n,B_n}^{B_n} + \mathbb{1}(B_n = a_n)$;
  Get $(\tilde{\mu}_{n+1,a_n}, \tilde{N}_{n+1,a_n}) = \text{ESTIMATOR}_{a_n}(\mathcal{H}_n)$ and $n \leftarrow n+1$; ▷ Update rule
**end for**

**Algorithm 2** Maximum Likelihood Estimator (MLE)

**Input:** History $\mathcal{H}_n$, arm $a \in [K]$.
**Return** $(\hat{\mu}_{n,a}, N_{n,a})$ where $\hat{\mu}_{n,a} = N_{n,a}^{-1} \sum_{t\in[n-1]} r_t \mathbb{1}\{a_t = a\}$;

For $\sigma$-sub-Gaussian distributions, TTUCB in Jourdan and Degenne (2024) is an instance of Algorithm 1 using the MLE (Algorithm 2) and

$$W_{a,b}^G(\tilde{\mu}, \omega) = \frac{(\tilde{\mu}_a - \tilde{\mu}_b)_+^2}{2\sigma^2(1/\omega_a + 1/\omega_b)} \text{ and } b_a^G(\omega) = \sqrt{\frac{2\sigma^2\alpha(1+s)\log\|\omega\|_1}{\omega_a}} \text{ with } s, \alpha > 1\,. \quad (3)$$

In practice, they take $s = \alpha = 1.2$. The standalone pseudocode of TTUCB is detailed in Algorithm 1 (Appendix C). The GLR stopping rule has to ensure $\delta$-correctness. This is done by choosing the stopping threshold as

$$c_{a,b}^G(\omega, \delta) = 2\mathcal{C}_G(\log((K-1)/\delta)/2) + 2\log(4 + \log\omega_a) + 2\log(4 + \log\omega_b)\,, \quad (4)$$

where the function $\mathcal{C}_G$ is defined in Eq. (20). It satisfies $\mathcal{C}_G(x) \approx x + \log(x)$. For bounded distributions on $[0,1]$ such as Bernoulli, we take $\sigma = 1/2$.

At each step, a Top Two algorithm selects two arms called leader and challenger, and samples one arm among them. TTUCB uses a UCB-based leader and a Transportation Cost (TC) challenger. The theoretical motivation behind the TC challenger comes from the theoretical lower bound in FC-BAI (Lemma 4), which involves the KL-characteristic time $T_{\text{KL}}^\star(\boldsymbol{\nu}) = \min_{\beta\in(0,1)} T_{\text{KL},\beta}^\star(\boldsymbol{\nu})$. For Gaussian distributions $\nu_a = \mathcal{N}(\mu_a, \sigma^2)$, it writes as

$$T_{\text{KL},\beta}^\star(\boldsymbol{\nu})^{-1} = \max_{\omega\in\Sigma_K, \omega_{a^\star}=\beta} \frac{(\mu_{a^\star} - \mu_a)^2}{2\sigma^2(1/\beta + 1/\omega_a)} \quad \text{and} \quad T_{\text{KL},1/2}^\star(\boldsymbol{\nu}) \leq 2T_{\text{KL}}^\star(\boldsymbol{\nu})\,. \quad (5)$$

Note that $H(\boldsymbol{\nu}) \leq T^\star_{\mathrm{KL}}(\boldsymbol{\nu}) \leq 2H(\boldsymbol{\nu})$ where $H(\boldsymbol{\nu}) = 2\sigma^2 \sum_{a\in[K]} \Delta_a^{-2}$ with $\Delta_a = \mu_{a^\star} - \mu_a$ for all $a \neq a^\star$ and $\Delta_{a^\star} = \Delta_{\min} = \min_{a\neq a^\star}(\mu_{a^\star} - \mu_a)$ for all $a \neq a^\star$. The maximiser of (5) is denoted by $\omega^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu})$, and is further referred to as the $\beta$-optimal allocation as it is unique. Let $N^a_{n,b}$ denote the number of times arm $b$ was pulled when $a$ was the leader, and $L_{n,a}$ denotes the number of times arm $a$ was the leader. In order to select the next arm to sample $a_n$, TTUCB relies on $K$ tracking procedures, i.e. set $a_n = B_n$ if $N^{B_n}_{n,B_n} \leq \beta L_{n+1,B_n}$, else $a_n = C_n$. This ensures that $\max_{a\in[K],n>K} |N^a_{n,a} - \beta L_{n,a}| \leq 1$ (Degenne et al., 2020).

### 2.3 Problem Statement: FC-BAI with DP

Now, we formally extend DP to BAI. We consider two trust models: (1) $\epsilon$-local DP BAI, where each user sends her reward to the BAI strategy, using an $\epsilon$-local DP perturbation mechanism, and (2) $\epsilon$-global DP BAI, where the BAI strategy, a.k.a. the centralised decision maker, is trusted with all the intermediate rewards. We summarise the BAI strategy-Users interaction in Algorithm 3, under global DP and local DP.

#### 2.3.1 Local DP FC-BAI

We represent each user $u_t$ by the vector $\mathbf{x}_t \triangleq (x_{t,1}, \dots, x_{t,K}) \in \mathbb{R}^K$, where $x_{t,a}$ represents the **potential** reward observed, if action $a$ was recommended to user $u_t$. Due to the bandit feedback, only $r_t = x_{t,a_t} \sim \nu_{a_t}$ is observed at step $t$. The user observes the real reward $r_t = \mathbf{x}_{t,a_t}$ but only sends a noisy version $z_t$ to the BAI strategy, by sampling $z_t$ from the perturbation mechanism, i.e. $z_t \sim \mathcal{M}(r_t)$. The BAI strategy only has access to the noisy rewards $(z_t)$ to make its decisions.

**Definition 6 ($\epsilon$-local DP for BAI)** *A pair $(\mathcal{M}, \pi)$ of perturbation mechanism and BAI strategy satisfies $\epsilon$-**local DP**, if they satisfy*
*(a) The perturbation mechanism $\mathcal{M}$ is $\epsilon$-local DP with respect to each reward record, i.e. for all $T$, all rewards $r_t, r'_t$ and all noisy outputs $z_t$, $\Pr[\mathcal{M}(r_t) = z_t] \leq e^\epsilon \Pr[\mathcal{M}(r'_t) = z_t]$.*
*(b) The BAI strategy only has access to the noisy rewards $z_t \sim \mathcal{M}(r_t)$ to make its decisions.*

For a pair $(\mathcal{M}, \pi)$ to be $\delta$-correct with respect to an environment $\nu$, under a local DP interaction protocol, the pair should verify: (a) the perturbation mechanism $\mathcal{M}$ should not change the identity of the optimal arm, i.e. $a^\star(\nu) = a^\star(\nu^{\mathcal{M}})$ and (b) the BAI strategy $\pi$ should be $\delta$-correct for the noisy environment $\nu^{\mathcal{M}}$. The goal in $\epsilon$-local DP FC-BAI is to design a $\delta$-correct $\epsilon$-local DP pair $(\mathcal{M}, \pi)$ of perturbation mechanism and BAI strategy, with $\mathbb{E}[\tau_\delta]$ as small as possible.

#### 2.3.2 Global DP BAI

Again, we represent each user $u_t$ by the vector $\mathbf{x}_t \triangleq (x_{t,1}, \dots, x_{t,K}) \in \mathbb{R}^K$, where $x_{t,a}$ represents the **potential** reward observed, if action $a$ was recommended to user $u_t$. Due to the bandit feedback, only $r_t = x_{t,a_t} \sim \nu_{a_t}$ is observed at step $t$. We use an underline to denote any sequence. Thus, we denote the sequence of sampled actions until $T$ as $\underline{a}^T = (a_1, \dots, a_T)$. We further represent a set of users $\{u_t\}_{t=1}^T$ until $T$ by **the table of potential rewards** $\underline{\mathbf{d}}^{\mathbf{T}} \triangleq \{\mathbf{x}_1, \dots, \mathbf{x}_T\} \in (\mathbb{R}^K)^T$. First, we observe that $\underline{\mathbf{d}}^T$ is the sensitive input data set to be made private, and $(\underline{a}^T, \widehat{a}, T)$ is the output of the BAI strategy. Hence, we

**Algorithm 3** Sequential Interaction Between a BAI Strategy and Users

**Input:** A BAI strategy $\pi$, Users $\{u_t\}_{n\geq 1}$ represented by the table $\underline{\mathbf{d}}$ and a perturbation mechanism $\mathcal{M}$
**Output:** A stopping time $\tau$, a sequence of samples actions $\underline{a}^\tau = (a_1, \ldots, a_\tau)$ and a recommendation $\hat{a}$ satisfying $\epsilon$-DP
**for** $t = 1, \ldots$ **do**
  $\pi$ recommends action $a_t \sim S_t(. \mid a_1, z_1, \ldots, a_{t-1}, z_{t-1})$
  **if** $a_t = \top$ **then**
    Halt. Return $\tau = t$ and $\hat{a} \sim \mathrm{Rec}_t(. \mid a_1, z_1, \ldots, a_{t-1}, z_{t-1})$
  **else**
    **if** Global DP **then**
      $u_t$ observes the **sensitive** reward $r_t \triangleq \underline{\mathbf{d}}_{t,a_t}$
      $u_t$ sends the **sensitive** reward $z_t \triangleq r_t$ to $\pi$
    **else** Local DP
      $u_t$ observes the **sensitive** reward $r_t \triangleq \underline{\mathbf{d}}_{t,a_t}$
      $u_t$ sends the **noisy** reward $z_t \sim \mathcal{M}(r_t)$ to $\pi$
    **end if**
  **end if**
**end for**

define the probability that the BAI strategy $\pi$ samples the action sequence $\underline{a}^T$, recommends the action $\widehat{a}$, and halts at time $T$, as

$$\pi(\underline{a}^T, \widehat{a}, T \mid \underline{\mathbf{d}}^T) \triangleq \mathrm{Rec}_{T+1}\left(\widehat{a} \mid \mathcal{H}_T\right) \mathrm{S}_{T+1}\left(\top \mid \mathcal{H}_T\right) \prod_{t\in[T]} \mathrm{S}_t\left(a_t \mid \mathcal{H}_{t-1}\right) , \tag{6}$$

where $T$ users under interaction are represented by the table of potential rewards $\underline{\mathbf{d}}^T$. A BAI strategy satisfies $\epsilon$-global DP if the probability in Eq. (6) is similar when the BAI strategy interacts with two neighbouring tables of rewards differing by one user (i.e. a row in $\underline{\mathbf{d}}^T$). Definition 7 can be seen as a BAI counterpart of the $\epsilon$-global DP definition proposed in Azize and Basu (2022) for regret minimisation.

**Definition 7 ($\epsilon$-global DP for BAI)** *A BAI strategy satisfies $\epsilon$-**global DP**, if for all $T \geq 1$, all neighbouring table of rewards $\underline{\boldsymbol{d}}^T$ and $\underline{\boldsymbol{d}'}^T$, i.e. $d_{Ham}(\underline{\boldsymbol{d}}^T, \underline{\boldsymbol{d}'}^T) = 1$, all sequences of sampled actions $\underline{a}^T \in [K]^T$ and recommended actions $\widehat{a} \in [K]$ we have that*

$$\pi(\underline{a}^T, \widehat{a}, T \mid \underline{\boldsymbol{d}}^T) \leq e^\epsilon \pi(\underline{a}^T, \widehat{a}, T \mid \underline{\boldsymbol{d}'}^T) .$$

The goal in $\epsilon$-global DP FC-BAI is to design a $\delta$-correct $\epsilon$-global DP BAI strategy $\pi$, with $\mathbb{E}[\tau_\delta]$ as small as possible.

**Remark 8** *It is possible to consider that the output of a BAI strategy is only the final recommended action $\hat{a}$, i.e. not publishing the intermediate actions $\underline{a}^T$. This gives a weaker definition of privacy compared to Definition 7, since the latter defends against adversaries that may look inside the execution of the BAI strategy, i.e. pan-privacy (Dwork et al., 2010b). Also, Definition 7 is needed in practice. For example, in the case of dose-finding (Example 1), the experimental protocol, i.e. the intermediate actions, needs to be published too.*

## 3. Local Differentially Private Best-Arm Identification

In this section, we answer the following question: *How many additional samples a BAI strategy must select to ensure $\epsilon$-local DP?* We provide a lower bound on the expected sample complexity of any $\delta$-correct $\epsilon$-local DP pair of perturbation mechanism and BAI strategy. We complement the sample complexity lower bound with a matching upper bound.

### 3.1 Lower Bound on the Expected Sample Complexity

We derive a lower bound on the expected sample complexity in $\epsilon$-local DP FC-BAI, which features problem-dependent characteristic times as in the FC-BAI setting.

**Theorem 9** *Let $\delta \in (0,1)$ and $\epsilon > 0$. For any $\delta$-correct $\epsilon$-local DP pair $(\mathcal{M}, \pi)$ of perturbation mechanism and BAI strategy, for all instance $\boldsymbol{\nu}$ with unique best arm, we have $\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta] \geq T^\star_\ell(\boldsymbol{\nu};\epsilon)\log(1/(2.4\delta))$ with*

$$T^\star_\ell(\boldsymbol{\nu};\epsilon)^{-1} \triangleq \sup_{\omega\in\Sigma_K} \inf_{\boldsymbol{\lambda}\in \mathrm{Alt}(\boldsymbol{\nu})} \sum_{a\in[K]} \omega_a \min\left\{ \mathrm{KL}\left(\nu_a \parallel \lambda_a\right), c(\epsilon)\left(\mathrm{TV}\left(\nu_a \parallel \lambda_a\right)\right)^2 \right\},$$

*where $c(\epsilon) \triangleq \min\{4, e^{2\epsilon}\}(e^\epsilon - 1)^2$ is a privacy term. For two probability distributions $\mathbb{P}, \mathbb{Q}$ on the measurable space $(\Omega, \mathcal{F})$, the TV distance is $\mathrm{TV}(\mathbb{P} \parallel \mathbb{Q}) \triangleq \sup_{A\in\mathcal{F}}\{\mathbb{P}(A) - \mathbb{Q}(A)\}$.*

*Proof sketch.* To prove this theorem, we first define the "noisy" environment $\boldsymbol{\nu}^{\mathcal{M}} \triangleq \{\nu_a^{\mathcal{M}} : a \in [K]\}$ *induced* by the perturbation mechanism $\mathcal{M}$, where

$$\nu_a^{\mathcal{M}}(Z) = \int_{r\in\mathbb{R}} \mathcal{M}(Z \mid r)\, \mathrm{d}\nu_a(r) dr$$

is the marginal over the noisy rewards of arm $a$. Then, we use the KL-decomposition from Lemma 1 of Garivier and Kaufmann (2016) applied to the "noisy" environment to get

$$\sum_{a=1}^K \mathbb{E}\left[N_{\tau_\delta,a}\right] \mathrm{KL}\left(\nu_a^{\mathcal{M}} \parallel \lambda_a^{\mathcal{M}}\right) \geq \mathrm{kl}(1-\delta,\delta),$$

where $\mathrm{kl}(x,y) \triangleq x\log\frac{x}{y} + (1-x)\log\frac{1-x}{1-y}$ for $x, y \in (0,1)$. Then, Theorem 1 of Duchi et al. (2013) is applied to relate the KL of rewards in the "noisy" bandit environment to the original environment to get

$$\mathrm{KL}\left(\nu_a^{\mathcal{M}} \parallel \lambda_a^{\mathcal{M}}\right) \leq c(\epsilon)(\mathrm{TV}\left(\nu_a \parallel \lambda_a\right))^2.$$

This bound shows that the perturbation mechanism $\mathcal{M}$ acts as a contraction on the space of probability measures. The rest of the proof is recovered by observing that $\mathbb{E}[\tau_\delta] = \sum_{a=1}^K \mathbb{E}\left[N_{\tau_\delta,a}\right]$ and taking the infinimum over all alterative environments. In Appendix D.2, we formally define the bandit canonical model under local DP, and provide a complete proof of the theorem. ■

Similar to the lower bound for the non-private BAI (Lemma 4, see Garivier and Kaufmann 2016), the lower bound of Theorem 9 is the value of a two-player zero-sum game between a MIN player and MAX player. For $\epsilon$-local DP, the measure of "distinguishability"

**Algorithm 4** Convert-To-Bernoulli($\epsilon$) Estimator (CTB) (Ren et al., 2020)

**Input:** History $\mathcal{H}_n$ with past perturbations $(\tilde{r}_t)_{t\in[n-2]}$, arm $a \in [K]$.
Observe $\tilde{r}_{n-1} \sim \mathrm{Ber}\left(\frac{r_{n-1}(e^\epsilon-1)+1}{e^\epsilon+1}\right)$; ▷ Randomised Response
**Return** $(\tilde{\mu}_{n,a}, N_{n,a})$ with $\tilde{\mu}_{n,a} = \frac{1}{N_{n,a}} \sum_{t=1}^{n-1} \tilde{r}_t \mathbb{1}\{a_t = a\}$;

between instances is captured by $\min\{\mathrm{KL}, c(\epsilon)\mathrm{TV}^2\}$. It interpolates between the KL stemming from the $\delta$-correctness constraint (i.e., "distinguishability" measure in the non-private lower bound) and the squared TV coming from the $\epsilon$-local DP constraint, scaled by $c(\epsilon)$.

Corollary 10 gives a lower bound on $T^\star_\ell$ as the maximum between the non-private characteristic time $T^\star_{\mathrm{KL}}$ and a privacy-rescaled characteristic time $T^\star_{\mathrm{TV}^2}$.

**Corollary 10 (Relaxing the local DP lower bound)** *Let $T^\star_\ell(\boldsymbol{\nu};\epsilon)$ as in Theorem 9 and $T^\star_{\boldsymbol{d}}(\boldsymbol{\nu})$ as in Eq. (1). Then, we have*

$$T^\star_\ell(\boldsymbol{\nu};\epsilon) \geq \max\left\{T^\star_{\mathrm{KL}}(\boldsymbol{\nu}), c(\epsilon)^{-1} T^\star_{\mathrm{TV}^2}(\boldsymbol{\nu})\right\} \quad \text{with} \quad c(\epsilon) \triangleq \min\{4, e^{2\epsilon}\}(e^\epsilon - 1)^2 \,. \tag{7}$$

*Let $\boldsymbol{\nu}_{\mathrm{G}}$ be the Gaussian instance with unit variances and the same means as the Bernoulli instance $\boldsymbol{\nu}$. Then, we have $T^\star_{\mathrm{TV}^2}(\boldsymbol{\nu}) \geq 2T^\star_{\mathrm{KL}}(\boldsymbol{\nu})$ and $T^\star_{\mathrm{TV}^2}(\boldsymbol{\nu}) = T^\star_{\mathrm{KL}}(\boldsymbol{\nu}_{\mathrm{G}})/2$.*

**Proof** The first part is true since $T^\star_\ell(\boldsymbol{\nu};\epsilon) \geq T^\star_{\mathrm{KL}}(\boldsymbol{\nu})$ and $T^\star_\ell(\boldsymbol{\nu};\epsilon) \geq \frac{T^\star_{\mathrm{TV}^2}(\boldsymbol{\nu})}{\min\{4,e^{2\epsilon}\}(e^\epsilon-1)^2}$. The second part uses that $\mathrm{KL}(\mathcal{N}(p,1) \,\|\, \mathcal{N}(q,1)) = \frac{1}{2}|p-q|^2 = \frac{1}{2}\mathrm{TV}(\mathrm{Ber}(p) \,\|\, \mathrm{Ber}(q))^2$. ■

Corollary 10 relates the $\mathrm{TV}^2$ characteristic time for Bernoulli to the KL characteristic times for Bernoulli and Gaussian. The sample complexity of FC-BAI with local DP on Bernoulli instances is reduced to the characteristic time of the non-private FC-BAI on Gaussian instances, up to a multiplicative factor which only depends on $\epsilon$.

*Two privacy regimes.* The sample complexity lower bound in Eq. (7) suggests the existence of two hardness regimes depending on $\epsilon$, $T^\star_{\mathrm{KL}}(\boldsymbol{\nu})$ and $T^\star_{\mathrm{TV}^2}(\boldsymbol{\nu})$. In the high privacy regime, as $\epsilon \to 0$, the lower bound reduces to $\epsilon^{-2} T^\star_{\mathrm{TV}^2}(\boldsymbol{\nu})$. In the low privacy regime, as $\epsilon \to \infty$, the lower bound reduces to the non-private complexity $T^\star_{\mathrm{KL}}(\boldsymbol{\nu})$. The switch between the low and the high privacy regimes happens at the $\epsilon$ verifying $\min\{4, e^{2\epsilon}\}(e^\epsilon-1)^2 = \frac{T^\star_{\mathrm{TV}^2}(\boldsymbol{\nu})}{T^\star_{\mathrm{KL}}(\boldsymbol{\nu})}$. For example, for environments where the Pinsker inequality is tight, i.e. $T^\star_{\mathrm{TV}^2}(\boldsymbol{\nu}) \approx 2T^\star_{\mathrm{KL}}(\boldsymbol{\nu})$, then the switch happens at $\epsilon \approx 0.582$.

### 3.2 A Plug-In Approach: the CTB-TT Algorithm

Ren et al. (2020) proposed the Convert-To-Bernoulli (CTB, Algorithm 4) estimator of the means, which relies on the Randomised Response mechanism to ensure $\epsilon$-local DP on $[0,1]$.

**Lemma 11 (Lemma 5 in Ren et al. 2020)** *CTB($\epsilon$) ensures $\epsilon$-local DP on $[0,1]$, and the returned value follows the Bernoulli distribution with mean $\mu_{\epsilon,a} \triangleq (2\mu_a - 1)\frac{e^\epsilon-1}{2(e^\epsilon+1)} + 1/2$.*

When the reward $r$ is generated using a Bernoulli of parameter $\mu_a$, and $r'$ is the result of the Randomised Response mechanism applied to $r$, *i.e.* $r' \sim \mathrm{Ber}\left(\frac{r(e^\epsilon-1)+1}{e^\epsilon+1}\right)$, then the marginal distribution of $r'$ is Bernoulli of parameter $\mu_{\epsilon,a}$ defined in Lemma 11.

CTB-TT *algorithm.* To solve $\epsilon$-local DP FC-BAI, we propose the CTB-TT algorithm, whose standalone pseudocode is detailed in Algorithm 7 (Appendix C). CTB-TT is an instance of Algorithm 1 using the CTB($\epsilon$) estimator (Algorithm 4), $(W^G_{a,b}, b^G_a)$ as in Eq. (3) with $\sigma = 1/2$ and $c^G_{a,b}$ as in Eq. (4).

Using Lemma 11, the CTB-TT algorithm is $\epsilon$-local DP and is equivalent to running the non-private TTUCB algorithm on a modified bandit instance $\boldsymbol{\nu}_{\boldsymbol{\epsilon}}$, where $\nu_{\epsilon,a} \triangleq \mathrm{Ber}(\mu_{\epsilon,a})$ with $\mu_{\epsilon,a} \triangleq (2\mu_a - 1)\frac{e^\epsilon - 1}{2(e^\epsilon+1)} + 1/2$ for all $a \in [K]$. While the analysis in Jourdan and Degenne (2024) is written for Gaussian distributions with unit variance, their Section 3.2 shows that the same results can be obtained for $\sigma$-sub-Gaussian distributions. As such, the theoretical guarantees obtained in Jourdan and Degenne (2024) apply to our algorithm. In particular, CTB-TT is $\delta$-correct and satisfies that, for all $\boldsymbol{\nu} \in \mathcal{M}$ such that $\min_{a \neq b} |\mu_a - \mu_b| > 0$,

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta]}{\log(1/\delta)} \leq T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}_\epsilon) = \left(1 + \frac{2}{e^\epsilon - 1}\right)^2 T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}) \,,$$

where $T^\star_{\mathrm{KL},\beta}$ as in Eq. (5) for $\sigma = 1/2$. For $\beta = 1/2$, combining Lemma 10 and (5) yields $\limsup_{\delta \to 0} \mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta]/\log(1/\delta) \leq (1 + 2/(e^\epsilon - 1))^2 \, T^\star_{\mathrm{TV}^2}(\boldsymbol{\nu})$. On top of its asymptotic guarantees, CTB-TT enjoys guarantees on its expected sample complexity at any confidence level (non-asymptotic regime). Taking $s = \alpha = 1.2$ and $\beta = 1/2$ as algorithmic parameters yields, for all $\delta \in (0,1)$ and all $\boldsymbol{\nu} \in \mathcal{M}$ such that $|a^\star(\boldsymbol{\nu})| = 1$,

$$\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta] = \mathcal{O}\left((H(\boldsymbol{\nu}_\epsilon) \log H(\boldsymbol{\nu}_\epsilon))^{1.2}\right) \quad \text{with} \quad H(\boldsymbol{\nu}_\epsilon) = (1 + 2/(e^\epsilon - 1))^2 \, H(\boldsymbol{\nu}) \,.$$

The notation $\mathcal{O}$ gives the dominating term when $H(\boldsymbol{\nu}) \to +\infty$.

In the non-private regime where $\epsilon \to +\infty$, our upper bound recovers the result of Jourdan and Degenne (2024). It matches the non-private lower bound for Gaussian distributions $T^\star_{\mathrm{KL}}(\boldsymbol{\nu})$ up to a multiplicative factor 2. Our upper bound matches the lower bound of Theorem 9 up to a multiplicative factor of $(e^\epsilon + 1)^2 \min\{4, e^{2\epsilon}\}$, whose limit is 4 when $\epsilon \to 0$. Instead of a fixed design $\beta$, we could use the optimal design IDS (You et al., 2023) which sets $\beta_n$ adaptively, i.e. $\beta_n = \frac{N_{n,C_n}}{N_{n,C_n} + N_{n,B_n}}$ for Gaussian distributions. Since this modification yields $T^\star_{\mathrm{KL}}(\boldsymbol{\nu}_\epsilon)$ as an asymptotic upper bound, it shaves a multiplicative factor 2. In the limit of $\epsilon \to 0$, it leaves a multiplicative gap of 2 between the lower and the upper bound. Closing this gap is an interesting direction for future research.

**Remark 12 (Extension to $(\epsilon, \gamma)$-Local DP)** *For approximate*[2] *$(\epsilon, \gamma)$-local DP, Zheng et al. (2020, Algorithm 1) uses the Gaussian mechanism (Dwork and Roth, 2014) to send noisy versions of the reward to the policy. Specifically, if the rewards are in $[0,1]$, at each step $t$, the noisy reward is $\tilde{r}_t = r_t + \mathcal{N}(0, \sigma^2_{\epsilon,\gamma})$, where $\sigma_{\epsilon,\gamma} \triangleq \frac{\sqrt{2\log(1.25/\gamma)}}{\epsilon}$. Then, the noisy rewards are fed to some non-private algorithm $\mathcal{A}$. Similar to our analysis of CTB-TT, Zheng et al. (2020, Theorem 12) shows that the sample complexity of the local DP algorithm (Algorithm 1) is equivalent to the sample complexity of the non-private algorithm $\mathcal{A}$ run on an instance of variance $\sigma^2_{\epsilon,\gamma} + \frac{1}{4}$. When combined with TTUCB, their Algorithm 1 inherits from the known guarantees of TTUCB on these modified Gaussian instances.*

2. We use $(\epsilon, \gamma)$-DP notation for approximate DP, since $\delta$ is used throughout the paper for correctness.

## 4. Global Differentially Private Best-Arm Identification

The central question that we address in this section is: *How many additional samples a BAI strategy must select for ensuring $\epsilon$-global DP?* In response, we prove a lower bound on the expected sample complexity of any $\delta$-correct $\epsilon$-global DP BAI strategy (Section 4.1). To design an $\epsilon$-global DP BAI algorithm, we first propose a private mean estimator (Section 4.2) based on arm-dependent doubling and forgetting. Then, in Sections 4.3 and 4.4, we plug this estimator in TTUCB to get AdaP-TT and AdaP-TT$^\star$. These two algorithms differ in how they account for the noise addition used in the mean estimation part. Specifically, AdaP-TT accounts for the noise addition by adapting the stopping threshold and UCB index of the leader. In addition to this, AdaP-TT$^\star$ also changes the transport used in TTUCB, and bases it on the lower bound of Section 4.1. This provides a tighter stopping rule and different challenger choosing rule that depends on the privacy regime.

### 4.1 Lower Bound on the Expected Sample Complexity

To prove BAI lower bounds with privacy, it is important to translate the privacy constraint to an upper bound on the KL between the marginals over the outputs, when the inputs are stochastically generated. In the following, we use coupling techniques to generate these upper bounds on the KL between marginals. We first explore the batch setting, where the data-generating distributions are product distributions. Then, we adapt the same techniques to the sequential setting of BAI.

*Batch Setting with Product Distributions.* Let $\mathcal{M}$ be a mechanism that takes as input data set $D \in \mathcal{X}^n$, and outputs $o \in \mathcal{O}$. Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be two data-generating distributions over $\mathcal{X}^n$. We define the marginals $M_1$ and $M_2$ over the output of the mechanism $\mathcal{M}$ as

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

when the inputs are generated from $\mathcal{P}_\nu$ for $\nu \in \{1, 2\}$ and $A$ an event in the output space. The goal in this section is to provide an upper bound on the quantity $\mathrm{KL}(M_1 \parallel M_2)$ when the mechanism $\mathcal{M}$ satisfies $\epsilon$-DP. Theorem 13 uses coupling and optimal transport to provide an upper bound on the quantity $\mathrm{KL}(M_1 \parallel M_2)$.

**Theorem 13 (KL Upper Bound as a Transport Problem)** *If $\mathcal{M}$ is $\epsilon$-pure DP, then*

$$\mathrm{KL}(M_1 \parallel M_2) \leq \epsilon \inf_{\mathcal{C} \in \Pi(\mathcal{P}_1, \mathcal{P}_2)} \mathbb{E}_{(D, D') \sim \mathcal{C}}[d_{Ham}(D, D')],$$

*where $\Pi(\mathcal{P}_1, \mathcal{P}_2)$ is the set of all couplings between $\mathcal{P}_1$ and $\mathcal{P}_2$.*

*Proof sketch.* To prove this, the main idea is to think about $M_1$ as the marginal over outputs when a pair of data sets $(D, D')$ is generated through the coupling $\mathcal{C}$ and the channel is $\mathcal{M}(|\, D)$, i.e. applying the mechanism only to the first data set. On the other hand, $M_2$ is the marginal over outputs when $(D, D') \sim \mathcal{C}$ but the channel is $\mathcal{M}(|\, D')$. Combining the fact that the KL between marginals is smaller than the expected KL between the channels, and that the KL of between the channels is controlled by group privacy, the proof is concluded. The complete proof is presented in Appendix D.3.1

Deriving the sharpest upper bound for the KL requires solving the transport problem

$$\inf_{\mathcal{C}\in\Pi(\mathcal{P}_1,\mathcal{P}_2)} \mathbb{E}_{(D,D')\sim\mathcal{C}}[d_{\mathrm{Ham}}(D,D')] \,. \tag{8}$$

As a proxy, we use maximal couplings.

*Product Distributions.* Now, suppose that $\mathcal{P}_1$ and $\mathcal{P}_2$ are two product distributions over $\mathcal{X}^n$, *i.e.* $\mathcal{P}_1 = \bigotimes_{i=1}^n p_{1,i}$ and $\mathcal{P}_2 = \bigotimes_{i=1}^n p_{2,i}$, where $p_{\nu,i}$ for $\nu \in \{1,2\}$ and $i \in [1,n]$ are distributions over $\mathcal{X}$. Let $c_\infty^i$ be a maximal coupling between $p_{1,i}$ and $p_{2,i}$ for all $i \in [1,n]$. We define the coupling $\mathcal{C}_\infty \triangleq \bigotimes_{i=1}^n c_\infty^i$. Then $\mathcal{C}_\infty$ is a coupling of $\mathcal{P}_1$ and $\mathcal{P}_2$. Using the $\mathcal{C}_\infty$ coupling between the product distributions $\mathcal{P}_1$ and $\mathcal{P}_2$ as a proxy to solve the transport problem of Equation (8), we show Corollary 14.

**Corollary 14 (KL Decomposition for Product Distributions)** *If $\mathcal{M}$ is $\epsilon$-pure DP, $\mathcal{P}_1 = \bigotimes_{i=1}^n p_{1,i}$ and $\mathcal{P}_2 = \bigotimes_{i=1}^n p_{2,i}$ are product distributions, then*

$$\mathrm{KL}\left(M_1 \parallel M_2\right) \le \epsilon \sum_{i=1}^n t_i \,,$$

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

**Proof** Since $d_{\mathrm{Ham}}(D,D') = \sum_{i=1}^n \mathbb{1}\{d_i \neq d_i'\}$, we have $d_{\mathrm{Ham}}(D,D') \sim \sum_{i=1}^n \mathrm{Bernoulli}(t_i)$ for $(D,D') \sim \mathcal{C}_\infty \triangleq \bigotimes_{i=1}^n c_\infty^i$, where $t_i \triangleq \mathrm{TV}\left(p_{1,i} \parallel p_{2,i}\right)$, and the terms in the sum are mutually independent. This further yields that $\mathbb{E}_{(D,D')\sim\mathcal{C}_\infty}[d_{\mathrm{Ham}}(D,D')] = \sum_{i=1}^n t_i$. ■

Corollary 14 can be seen as a stochastic generalisation of the group privacy property of DP. Specifically, the results from Theorem 14 suggest that two random data sets $D$ and $D'$ sampled from $\mathcal{P}_1 = \bigotimes_{i=1}^n p_{1,i}$ and $\mathcal{P}_2 = \bigotimes_{i=1}^n p_{2,i}$ respectively could be thought of as $(\sum_{i=1}^n t_i)$-neighboring data sets "in expectation", where $t_i = \mathrm{TV}\left(p_{1,i} \parallel p_{2,i}\right)$.

*Relation to similar results in the literature.* Lemma 6.1 in Karwa and Vadhan (2018) shows that, for any event E, $M_1(E) \le e^{6\epsilon n \mathrm{TV}(p_1 \parallel p_2)} M_2(E)$, when the mechanism is $\epsilon$-pure DP, and the data-generating distributions are i.i.d from $p_1$ or $p_2$, *i.e.* $\mathcal{P}_\nu = \bigotimes_{i=1}^n p_\nu$ for $\nu \in \{1,2\}$. The Karwa Vadhan is a stronger result than Theorem 14 since it controls the multiplicative difference between the marginals at each event. This gives the following direct KL upper bound $\mathrm{KL}\left(M_1 \parallel M_2\right) \le 6\epsilon n \mathrm{TV}\left(p_1 \parallel p_2\right)$ for i.i.d distributions. Also, the Karwa Vadhan lemma builds explicitly the maximal coupling in their proof. Our result generalises this upper bound to product distributions and improves the dependence of factor 6 there. Also, similar coupling ideas have been developed in Lalanne et al. (2023) to derive DP and zCDP variants of LeCam and Fano inequalities, and in Azize and Basu (2024) to derive zCDP regret lower bounds for bandits.

*BAI setting.* Adapting similar coupling ideas from the batch setting, we derive an $\epsilon$-global DP version of the "change-of-measure" lemma.

**Lemma 15 (Change-of-measure lemma under $\epsilon$-global DP)** *Let $\delta \in (0,1)$ and $\epsilon > 0$. Let $\boldsymbol{\nu}$ be a bandit instance and $\lambda \in \mathrm{Alt}(\boldsymbol{\nu})$. For any $\delta$-correct $\epsilon$-global DP BAI strategy,*

$$\epsilon \sum_{a\in[K]} \mathbb{E}_{\boldsymbol{\nu}}[N_{\tau_\delta,a}] \mathrm{TV}\left(\nu_a \parallel \lambda_a\right) \ge \mathrm{kl}(1-\delta,\delta) \,.$$

*Proof Sketch.* The main technical challenge in this proof is to extend Corollary 14 to bandit distributions using the idea of "coupled environments". Then, using a classic data-processing inequality concludes the proof. The complete proof is presented in Appendix D.3.3.

Combining the non-private and $\epsilon$-global DP change of measure lemmas gives the lower bound on the sample complexity of any $\delta$-correct $\epsilon$-global DP BAI algorithm.

**Theorem 16** *Let $(\epsilon, \delta) \in \mathbb{R}_+^\star \times (0,1)$. For any $\delta$-correct and $\epsilon$-global DP FC-BAI algorithm, for all instances $\boldsymbol{\nu} \in \mathcal{M}$ with unique best arm, we have $\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta] \geq T_g^\star(\boldsymbol{\nu}; \epsilon) \log(1/(2.4\delta))$ with*

$$T_g^\star(\boldsymbol{\nu}; \epsilon)^{-1} \triangleq \sup_{\omega \in \Sigma_K} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\nu})} \min \left\{ \sum_{a \in [K]} \omega_a \mathrm{KL}\left(\nu_a \parallel \lambda_a\right), \epsilon \sum_{a \in [K]} \omega_a \mathrm{TV}\left(\nu_a \parallel \lambda_a\right) \right\}.$$

As for the non-private BAI (Lemma 4, see Garivier and Kaufmann 2016), Theorem 16 is the value of a two-player zero-sum game between a MIN player and MAX player. On top of the KL divergence present in the non-private lower bound, our bound features the TV distance that appears naturally when incorporating the $\epsilon$-global DP constraint. For $\epsilon$-global DP, our proposed measure of "distinguishability" between instances is captured by a minimum between the instance-wise non-private measure $\sum_{a \in [K]} \omega_a \mathrm{KL}\left(\nu_a \parallel \lambda_a\right)$ and an instance-wise private measure $\sum_{a \in [K]} \omega_a \mathrm{TV}\left(\nu_a \parallel \lambda_a\right)$, scaled by $\epsilon$. While it trades off between the $\delta$-correctness constraint and the $\epsilon$-global DP constraint, this interpolation occurs at the level of the instance instead of being at the level of individual arms, as it does in Theorem 9 for $\epsilon$-local DP. Deriving a measure of "distinguishability" at the arm's level is an interesting open-problem that we leave for furture work.

Corollary 17 gives a lower bound on $T_g^\star$ as the maximum between the non-private characteristic time $T_{\mathrm{KL}}^\star$ and a privacy-rescaled characteristic time $T_{\mathrm{TV}}^\star$. Compared to Corollary 10, the "distinguishability" measure for global DP is $\mathrm{TV}^2$ instead of TV for local DP. The TV characteristic time $T_{\mathrm{TV}}^\star(\boldsymbol{\nu})$ serves as the BAI counterpart to the TV-distinguishability gap ($t_{\mathrm{inf}}$) in the problem-dependent regret lower bound for bandits with $\epsilon$-global DP as in Azize and Basu (2022, Theorem 3).

**Corollary 17 (Relaxing the global DP lower bound)** *Let $T_g^\star(\boldsymbol{\nu}; \epsilon)$ be the $\epsilon$-global DP characteristic time as in Theorem 16 and $T_{\boldsymbol{d}}^\star(\boldsymbol{\nu})$ as in Eq.* (1) *be the characteristic time for the "distinguishability" measure $\boldsymbol{d}$ between probability distributions, e.g.,* KL *or* TV*. Then,*

$$T_g^\star(\boldsymbol{\nu}; \epsilon) \geq \max\{T_{\mathrm{KL}}^\star(\boldsymbol{\nu}), T_{\mathrm{TV}}^\star(\boldsymbol{\nu})/\epsilon\}. \tag{9}$$

***Bernoulli instances.*** *Let $\boldsymbol{\nu}$ be a Bernoulli instance with unique best arm $a^\star$, mean gaps $\Delta_a = \mu_{a^\star} - \mu_a$ for all $a \neq a^\star$ and $\Delta_{a^\star} = \Delta_{\min} = \min_{a \neq a^\star}(\mu_{a^\star} - \mu_a)$ for all $a \neq a^\star$. Let $\boldsymbol{\nu}_{G,\epsilon}$ be a Gaussian instance with variance $\sigma = 1/2$ and means $\mu_{\epsilon,a} = \mu_a$ for all $a \in \{a^\star\} \cup \{a \neq a^\star \mid \Delta_a \leq \epsilon/2\}$, and $\mu_{\epsilon,a} = \mu_{a^\star} - \sqrt{\epsilon \Delta_a/2}$ otherwise. Then, we have*

$$T_{\mathrm{TV}}^\star(\boldsymbol{\nu}) = \sum_{a \in [K]} \Delta_a^{-1} \quad \text{and} \quad T_g^\star(\boldsymbol{\nu}; \epsilon) \leq T_{\mathrm{KL}}^\star(\boldsymbol{\nu}_{G,\epsilon}) \leq 2H(\boldsymbol{\nu}_{G,\epsilon}), \quad \text{where we recall that}$$

$$T_{\mathrm{KL}}^\star(\boldsymbol{\nu}_{G,\epsilon})^{-1} \triangleq \max_{\omega \in \Sigma_K} \min_{a \neq a^\star} \frac{2\Delta_a \min\{\epsilon/2, \Delta_a\}}{1/\omega_{a^\star} + 1/\omega_a}, \quad 2H(\boldsymbol{\nu}_{G,\epsilon}) = \sum_{a \in [K]} (\Delta_a \min\{\Delta_a, \epsilon/2\})^{-1}. \tag{10}$$

**Remark 18** *In Section 4.1, only the results of Paragraph "**Bernoulli instances**" are specific to Bernoulli distributions. All the other results of Section 4.1 are true for any class of distributions, e.g. Gaussians, sub-Gaussians, Exponentials, etc.*

**Proof** The first part is a direct consequence of the definition of $T_g^\star(\boldsymbol{\nu};\epsilon)$. The second part uses $\mathrm{TV}\left(\mathrm{Ber}(p) \parallel \mathrm{Ber}(q)\right) = |p-q|$ to solve the optimisation problem and is detailed in Appendix D.3.5. The last part is obtained by using Pinsker's inequality. ■

Corollary 17 also upper bounds $T_g^\star$ by the non-private characteristic time $T_{\mathrm{KL}}^\star$ on a Gaussian instance $\boldsymbol{\nu}_{G,\epsilon}$ with privacy-aware means. In this modified instance, the non-private mean gaps are clipped when the privacy budget is small enough.

*Two privacy regimes.* The sample complexity lower bound in Eq. (9) suggests the existence of *two hardness regimes depending on* $\epsilon$, $T_{\mathrm{KL}}^\star(\boldsymbol{\nu})$ *and* $T_{\mathrm{TV}}^\star(\boldsymbol{\nu})$. (1) *Low-privacy regime*: When $\epsilon > T_{\mathrm{TV}}^\star(\boldsymbol{\nu})/(T_{\mathrm{KL}}^\star(\boldsymbol{\nu}))$, the lower bound retrieves the non-private lower bound, i.e. $T_{\mathrm{KL}}^\star(\boldsymbol{\nu})$, and thus, **privacy can be achieved for free**. (2) *High-privacy regime:* When $\epsilon < T_{\mathrm{TV}}^\star(\boldsymbol{\nu})/(T_{\mathrm{KL}}^\star(\boldsymbol{\nu}))$, the lower bound becomes $T_{\mathrm{TV}}^\star(\boldsymbol{\nu})/\epsilon$ and $\epsilon$-global DP $\delta$-BAI requires more samples than non-private ones. Using Pinsker's inequality, one can connect the TV and KL characteristic times by $T_{\mathrm{TV}}^\star(\boldsymbol{\nu}) \geq \sqrt{2T_{\mathrm{KL}}^\star(\boldsymbol{\nu})}$.

The global trade-off between low and high privacy regimes at the instance level given by (9) does not give any information at the level of a specific arm. For each sub-optimal arm, the transition from low to high privacy is better understood by considering (10), even though it only upper bounds $T_g^\star(\boldsymbol{\nu};\epsilon)$. For any arm $a \neq a^\star(\boldsymbol{\nu})$, the high-privacy regime corresponds to a mean gap such that $\epsilon < 2\Delta_a$, and the low-privacy regime to $\epsilon > 2\Delta_a$.

### 4.2 Private Mean Estimator

To define a sequence of mean estimators, we propose the DAF($\epsilon$) update (Algorithm 5) which relies on three ingredients: *adaptive episodes with doubling*, *forgetting*, and *adding calibrated Laplacian noise*. (1) DAF maintains $K$ episodes, i.e. one per arm. The private empirical estimate of the mean of an arm is only updated at the end of an episode, that means when the number of times that a particular arm was played doubles. (2) For each arm $a$, DAF forgets rewards from previous phases of arm $a$, i.e. the private empirical estimate of arm $a$ is only computed using the rewards collected in the last phase of arm $a$. This assures that the means of each arm are estimated using a non-overlapping sequence of rewards. (3) Thanks to this *doubling* and *forgetting*, DAF is $\epsilon$-global DP as soon as each empirical mean is made $\epsilon$-DP, and thus, avoiding any use of privacy composition. This is achieved by adding Laplace noise. We formalise this intuition in Lemma 30 of Appendix E.

**Lemma 19** *Any algorithm relying solely on the DAF($\epsilon$) update is $\epsilon$-global DP on* $[0,1]$.

**Proof** A change in one user *only affects* the empirical mean calculated at one episode of an arm, which is made private using the Laplace Mechanism and Lemma 30. Since the sampled actions, recommended action, and stopping time are computed only using the private empirical means, the algorithm satisfies $\epsilon$-global DP thanks to the post-processing lemma. ■

**Algorithm 5** Doubling-And-Forgetting($\epsilon$) Estimator (DAF)

**Input:** History $\mathcal{H}_n$, arm $a \in [K]$.
**Initialisation:** $T_1(a) = K + 1$ and $k_{K+1,a} = 1$;
**if** $N_{n,a} \geq 2N_{T_{k_{n,a}}(a),a}$ **then** ▷ Per-arm doubling update grid
  Change phase $k_{n,a} \leftarrow k_{n,a} + 1$ for arm $a$;
  Set $T_{k_{n,a}}(a) = n$ and $\tilde{N}_{k_{n,a},a} = N_{T_{k_{n,a}}(a),a} - N_{T_{k_{n,a}-1}(a),a}$;
  Set $\hat{\mu}_{k_{n,a},a} = \tilde{N}^{-1}_{k_{n,a},a} \sum_{t=T_{k_{n,a}-1}(a)}^{T_{k_{n,a}}(a)-1} r_t \mathbb{1}\{a_t = a\}$; ▷ Forgetting past observations
  Set $\tilde{\mu}_{k_{n,a},a} = \hat{\mu}_{k_{n,a},a} + Y_{k_{n,a},a}$ where $Y_{k_{n,a},a} \sim \mathrm{Lap}((\epsilon\tilde{N}_{k_{n,a},a})^{-1})$; ▷ Private estimator
**end if**
**Return** $(\tilde{\mu}_{n,a}, \tilde{N}_{n,a})$;

*Batching and forgetting for BAI.* For stochastic bandits, it is crucial to effectively use a combination of batching and forgetting (Sajed and Sheffet, 2019; Azize and Basu, 2022; Chowdhury and Zhou, 2023). These techniques help to use the parallel composition property of DP, and thus avoid using sequential composition, where more noise is needed to achieve DP. However, these batching and forgetting techniques should be adapted, depending on the setting and the "accuracy" guarantee. For example, for BAI, it is important that the batching is arm-dependent. For example, having a global doubling scheme would not work, i.e. update the means when the global count $n$ double. Also, it is important that the actions sampled by the Top Two algorithm during a fixed arm-phase ensure exploration of all the arms. This is different from the arm-dependent batching of bandits under regret (Azize and Basu, 2022), where the *same arm* is always chosen during a phase. In contrast, for BAI, only the mean estimators are updated when an arm counts doubles, but the Top Two algorithm still samples arms between the fixed leader and the potentially varying challenger. Intuitively, the challenger will evolve to ensure the equality of the transportation costs for the global counts, which is key to obtaining asymptotic optimality. In contrast to the regret minimisation algorithms, BAI algorithms require the empirical counts to be "synchronised" with respect to the same global time, as the empirical proportions should converge towards an allocation $\omega$ with dense support.

### 4.3 A Plug-In Approach: the AdaP-TT Algorithm

A natural approach is to simply plug in the private mean estimator in the non-private algorithm. The Plug-In approach is successful for $\epsilon$-local DP FC-BAI (Section 3) and for $\epsilon$-global DP regret minimisation (Azize and Basu, 2022).

AdaP-TT *algorithm.* To solve $\epsilon$-global DP FC-BAI, we propose the AdaP-TT algorithm, whose standalone pseudocode is detailed in Algorithm 9 (Appendix C). AdaP-TT is an instance of Algorithm 1 using the DAF($\epsilon$) estimator (Algorithm 5), $W^G_{a,b}$ as in Eq. (3) for $\sigma = 1/2$,

$$b^{G,\epsilon}_a(\omega) = \sqrt{\frac{k(\omega_a)}{\omega_a}} + \frac{k(\omega_a)}{\epsilon\omega_a} \quad \text{with} \quad k(x) = \log_2(x) + 2\,, \tag{11}$$

and $c^{G,\epsilon}_{a,b}$ as in Eq. (12) which yields $\delta$-correctness for any sampling rule (Lemma 20).

**Lemma 20** *Let $\delta \in (0,1)$, $\epsilon > 0$. Let $c^G_{a,b}$ as in Eq. (4) and $k(x) = \log_2 x + 2$. Given any sampling rule, combining the DAF($\epsilon$) estimator with the GLR stopping rule as in Eq. (2) with $W^G_{a,b}$ as in Eq. (3) and the stopping threshold*

$$c^{G,\epsilon}_{a,b}(\omega,\delta) = 2c^G_{a,b}(\omega, \delta(k(\omega_a)^2 k(\omega_b)^2 \pi^4/18)^{-1}) + \frac{1}{\epsilon^2\sigma^2} \sum_{c\in\{a,b\}} \frac{1}{\omega_c} \left(\log \frac{\pi^2 K k(\omega_c)^2}{3\delta}\right)^2 , \quad (12)$$

*yields a $\delta$-correct algorithm for any $\sigma$-sub-Gaussian distributions with unique best arm.*

Asymptotically, our threshold is $c^{G,\epsilon}_{a,b}(\omega,\delta) \approx_{\delta\to 0} 2\log(1/\delta) + (1/\omega_a + 1/\omega_b)\log(1/\delta)^2/(\epsilon^2\sigma^2)$.

**Proof** The proof of $\delta$-correctness of a GLR stopping rule is standard, by leveraging concentration results. Compared to the non-private proof, we also need to control the Laplace "noise", which results in additive terms in the stopping threshold. Our proof technique can be used to study any Gaussian GLR stopping rules facing additive known pertubations. Specifically, we start by decomposing the failure probability $\mathbb{P}_\mu(\tau_\delta < +\infty, \hat{a} \neq a^\star)$ into a non-private and a private part using the basic property of $\mathbb{P}(X + Y \geq a + b) \leq \mathbb{P}(X \geq a) + \mathbb{P}(Y \geq b)$. The two-factor in front of $c^G_{a,b}$ originates from the looseness of this decomposition, and we improve on it in Section 4.4. We conclude using concentration results from $\sigma$-sub-Gaussian and Laplace random variables. The proof is detailed in Appendix F.2. ■

*Algorithmic intuition.* Thanks to the implicit exploration of the UCB leader, $B_n$ and $\tilde{a}_n$ converge towards $a^\star$. Moreover, the TC challenger selects the alternative arm to balance the information between the other arms, i.e., an arm $a \neq a^\star$ stops being selected as challenger when its empirical allocation $N_{n,a}/(n-1)$ overshoots an allocation for which the empirical transportation costs $W^G_{a^\star,a}$ are all equals (Lemma 48). The per-arm geometric update grid ensures that the local empirical proportions $\tilde{N}_n/(n-1)$ behave similarly to $N_n/(n-1)$, up to a factor at most 4 due to doubling and forgetting. This synchronization of the local counts with the global time is key to studying the GLR stopping rule. While $\epsilon$-global DP regret minimization algorithms pull arms in batches, this approach fails in BAI as it would prevent convergence of $\tilde{N}_n/(n-1)$ towards a fixed allocation.

*Asymptotic upper bound.* AdaP-TT achieves an asymptotic upper bound on its expected sample complexity as $T^\star_{\mathrm{KL},\beta}$ rescaled by a privacy-aware cost (Theorem 21).

**Theorem 21** *Let $(\delta,\beta) \in (0,1)^2$ and $\epsilon > 0$. The* AdaP-TT *algorithm is $\epsilon$-global DP, $\delta$-correct and satisfies that, for any Bernoulli $\boldsymbol{\nu}$ instance with distinct means $\mu \in (0,1)^K$,*

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta]}{\log(1/\delta)} \leq 4T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}) \left(1 + \sqrt{1 + \frac{\Delta^2_{\max}}{2\sigma^4\epsilon^2}}\right) ,$$

*where $T^\star_{\mathrm{KL},\beta}$ as in Eq. (5) for $\sigma = 1/2$.*

The asymptotic analysis of Top Two algorithms is standard, by building upon the unified analysis from Jourdan et al. (2022). Compared to the non-private proof, we adapt their argument to cope for the effect of the DAF($\epsilon$) estimator on the expected sample complexity. This required two main technical novelties. First, the proof of sufficient exploration requires to reason about phases per arm, and cope for doubling and forgetting (Appendix G.2).

Second, once the convergence of the global count is established, we should argue that the phase switches of the arms happen in a "round-robin" fashion, i.e., an arm switches phase for a second time after all other arms first switch their own phases (Lemma 50). Our proof technique can be used to study algorithms relying on phases to limit the number of rounds of adaptivity (Remark 24). We sketch other high-level ideas of the proof below, and refer to Appendix G for more details.

**Proof** (1) The *non-private TTUCB algorithm* (Jourdan and Degenne, 2024) achieves a sample complexity of $T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu})$ for sub-Gaussian random variables. The proof relies on showing that the empirical pulling counts are converging towards the $\beta$-optimal allocation $\omega^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu})$. (2) The *effect of doubling and forgetting* is a multiplicative four-factor, i.e. $4T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu})$. The first two-factor is due to forgetting since we throw away half of the samples. The second two-factor is due to doubling since we have to wait for the end of an episode to evaluate the stopping condition. (3) The *Laplace noise* only affects the empirical estimate of the mean. Since the Laplace noise has no bias and a sub-exponential tail, the private means will still converge towards their true values. Therefore, the empirical counts will also converge to $\omega^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu})$ asymptotically. (4) While the *Laplace noise has little effect on the sampling rule* itself, it *changes the dependency in* $\log(1/\delta)$ *of the threshold* used in the GLR stopping rule. The private threshold $c^{G,\epsilon}_{a,b}$ has an extra factor $\mathcal{O}(\log^2(1/\delta))$ compared to the non-private one $c^G_{a,b}$. Using the convergence towards $\omega^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu})$, the stopping condition is met as soon as $\frac{n}{T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu})} \lesssim 2\log(1/\delta) + \frac{\Delta^2_{\max}}{2\sigma^4\epsilon^2}\frac{T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu})}{n}\log^2(1/\delta)$. Solving the inequality for $n$ concludes the proof while adding a multiplicative four-factor. ■

*Discussion.* In the non-private regime where $\epsilon \to +\infty$, our upper bound recovers the non-private lower bound for Gaussian distributions $T^\star_{\mathrm{KL}}(\boldsymbol{\nu})$ up to a multiplicative factor 16. For Bernoulli distributions (or bounded distributions in $[0,1]$), there is still a mismatch between the upper and lower bounds due to the mismatch between the KL divergence of Bernoulli distributions and that of Gaussian (e.g. large ratio when the means are close to 0 or 1). This is in essence, similar to the mismatch between UCB and KL-UCB in the regret-minimisation literature (e.g. Chapter 10 in Lattimore and Szepesvári 2020). To overcome this mismatch, it is necessary to adapt the transportation costs to the family of distributions considered. While the Top Two algorithms for Bernoulli distributions (or bounded distributions in $[0,1]$) have been studied in Jourdan et al. (2022), the analysis is more involved. Therefore, it would obfuscate where and how privacy is impacting the expected sample complexity.

In the asymptotic highly privacy regime where $\epsilon \to 0$, our upper bound becomes $\mathcal{O}(T^\star_{\mathrm{KL}}(\boldsymbol{\nu})\Delta_{\max}/\epsilon)$ while the lower bound is $\Omega(T^\star_{\mathrm{TV}}(\boldsymbol{\nu})/\epsilon)$. Therefore, our upper bound is only asymptotically tight for instances such that $T^\star_{\mathrm{KL}}(\boldsymbol{\nu}) = \mathcal{O}(T^\star_{\mathrm{TV}}(\boldsymbol{\nu})/\Delta_{\max})$, e.g. instances where the mean gaps have the same order of magnitude. In all the other cases, the plug-in approach is sub-optimal due to a problem-dependent gap. The major limitation of AdaP-TT lies in the fact that the transportation costs are independent of the privacy budget. Without accounting for the knowledge that there is an additional Laplace "noise", there is little hope to match the asymptotic lower bound in the high-privacy regime. We remedy this issue with AdaP-TT$^\star$ in Section 4.4.

*Non-asymptotic confidence regime for* AdaP-TT. Studying AdaP-TT for any confidence level requires to adapt the proof of Jourdan and Degenne (2024). The $\delta$-dependent term in their upper bound is worse than for an asymptotic analysis (i.e., Theorem 21). Therefore, we focus on their dominating $\delta$-independent term that stems from Jourdan and Degenne (2024, Lemma 3.2), i.e., $B_n = a^\star$ except for a sublinear number of times. While the intution behind their proof still holds, using the DAF($\epsilon$) estimator adds steps to cope for the Laplace "noise", doubling and forgetting. If $B_n = a$ with $a \neq a^\star$ and empirical means do not deviate too much from the true means, then we have $N_{n,a} = \mathcal{O}(\log(n)\Delta_{\epsilon,a}^{-2})$ where $\Delta_{\epsilon,a} \triangleq \epsilon(\sqrt{1 + 2\Delta_a/\epsilon} - 1)$ by solving for $N_{n,a}/\log n$ the quadratic inequality

$$\mu_{a^\star} \lesssim \tilde{\mu}_{n,a^\star} + b_{a^\star}^{G,\epsilon}(\tilde{N}_n) \leq \tilde{\mu}_{n,a} + b_a^{G,\epsilon}(\tilde{N}_n) \lesssim \mu_a + 2b_a^{G,\epsilon}(\tilde{N}_n) \leq \mu_a + 2\left(\sqrt{\frac{4\log n}{N_{n,a}}} + \frac{4\log n}{\epsilon N_{n,a}}\right),$$

where $\tilde{N}_{n,a} \geq N_{n,a}/4$ due to doubling and forgetting, and $\lesssim$ stems from the "heuristic" choice of $(b_a^{G,\epsilon})_{a\in[K]}$, which do not yield valid upper confidence bounds. Since $N_{n,a}$ is bounded and incremented by one half of the time (tracking for $\beta = 1/2$), the event $B_n \neq a^\star$ occurs less than $\mathcal{O}(H(\boldsymbol{\nu}, \epsilon)\log n)$ times, where $H(\boldsymbol{\nu}, \epsilon) = \sum_{a\in[K]} \Delta_{\epsilon,a}^{-2}$ and $\Delta_{\epsilon,a^\star} = \min_{a\neq a^\star} \Delta_{\epsilon,a}$. The rest of the proof can be adapted similarly, with extra terms due to the Laplace "noise", doubling and forgetting. Intuitively, it holds thanks to the use of the Gaussian transportation costs $W_{a,b}^G$ as in Eq. (3) providing time-uniform separability between the means and the allocations, i.e., ratio of the squared mean gap and the sum of the inverse allocation. In summary, we conjecture that AdaP-TT satisfies a non-asymptotic upper bound whose dominating $\delta$-independent term scale as $\mathcal{O}\left((H(\boldsymbol{\nu}, \epsilon)\log H(\boldsymbol{\nu}, \epsilon))^\alpha\right)$ for some $\alpha > 1$, where $\alpha$ would be an algorithmic hyperparameter involved in a valid choice for $(b_a^{G,\epsilon})_{a\in[K]}$. As $\Delta_{\epsilon,a}^2/\epsilon \to_{\epsilon\to 0} \Delta_a$ and $\Delta_{\epsilon,a}^2 \to_{\epsilon\to+\infty} \Delta_a^2$, the bound suffers from a suboptimal dependency due to the power $\alpha$.

### 4.4 A Modified Transportation Cost Approach: the AdaP-TT$^\star$ Algorithm

To overcome the limitation of AdaP-TT, one should adapt the transportation costs to reflect the lower bound (Theorem 16) instead of "ignoring" the privacy constraint by using the transportation costs $W_{a,b}^G$ as in Eq. (3) which are tailored for non-private FC-BAI.

AdaP-TT$^\star$ *algorithm.* We propose the AdaP-TT$^\star$ algorithm, whose standalone pseudocode is detailed in Algorithm 10 (Appendix C). It differs from AdaP-TT solely by (i) the use of a modified transportation costs in the GLR stopping rule and the TC challenger, and (ii) an adequate stopping threshold to ensure $\delta$-correctness. AdaP-TT$^\star$ is an instance of Algorithm 1 using the DAF($\epsilon$) estimator (Algorithm 5),

$$W_{a,b}^{G,\epsilon}(\tilde{\mu}, \omega) = \frac{(\tilde{\mu}_a - \tilde{\mu}_b)_+ \min\{\epsilon/2, (\tilde{\mu}_a - \tilde{\mu}_b)_+\}}{2\sigma^2(1/\omega_a + 1/\omega_b)} \quad \text{with} \quad \sigma = 1/2\,, \tag{13}$$

$b_{a,b}^{G,\epsilon}$ as in Eq. (11), and $\tilde{c}_{a,b}^{G,\epsilon}(\tilde{\mu}, \omega, \delta)$ as in Eq. (14), which yields $\delta$-correctness for any sampling rule (Lemma 22).

**Lemma 22** *Let $\delta \in (0,1)$, $\epsilon > 0$. Let $\overline{W}_{-1}(x) = -W_{-1}(-e^{-x})$ for all $x \geq 1$, where $W_{-1}$ is the negative branch of the Lambert W function. It satisfies $\overline{W}_{-1}(x) \approx x + \log x$. Let $c_{a,b}^{G,\epsilon}$*

*as in Eq.* (12), $k(x) = \log_2(x) + 2$ *and*

$$h(x, \delta) = \overline{W}_{-1}\left(2\log(\pi^2 K k(x)^2/(2\delta)) + 4\log(4 + \log x) + 1/2\right)/2\,.$$

*Given any sampling rule, combining the DAF($\epsilon$) estimator with the GLR stopping rule as in Eq.* (2) *with* $W_{a,b}^{G,\epsilon}$ *as in Eq.* (13) *and the stopping threshold* $\tilde{c}_{a,b}^{G,\epsilon}(\tilde{\mu}, \omega, \delta)$ *which is equal to*

$$\begin{cases} \frac{1}{2}c_{a,b}^{G,\epsilon}(\omega, 2\delta/3) + \frac{\sqrt{2}}{\epsilon\sigma}\sum_{c\in\{a,b\}}\sqrt{\frac{h(\omega_c,\delta)}{\omega_c}}\log\left(\frac{\pi^2 K k(\omega_c)^2}{2\delta}\right) & \text{if } (\tilde{\mu}_a - \tilde{\mu}_b)_+ < \epsilon/2 \\ \frac{1}{2\sigma^2}\log\left(\pi^2 K \max_{c\in\{a,b\}} k(w_c)/(2\delta)\right) + \frac{\epsilon}{2\sqrt{2}\sigma}\sum_{c\in\{a,b\}}\sqrt{\omega_c h(\omega_c,\delta)} & \end{cases}, \tag{14}$$

*yields a $\delta$-correct algorithm for any $\sigma$-sub-Gaussian distributions with unique best arm.*

Our threshold is $\frac{1}{2\sigma^2}\log(1/\delta) + \frac{\epsilon}{2\sqrt{2}\sigma}(\sqrt{\omega_b} + \sqrt{\omega_a})\sqrt{\log(1/\delta)}$ when $\tilde{\mu}_a - \tilde{\mu}_b \geq \epsilon/2$, and

$$\log(1/\delta) + \frac{1}{2\epsilon^2\sigma^2}(1/\omega_a + 1/\omega_b)\log(1/\delta)^2 + \frac{\sqrt{2}}{\epsilon\sigma}(\sqrt{1/\omega_a} + \sqrt{1/\omega_b})\log(1/\delta)^{3/2} \quad \text{otherwise.}$$

While the proof bares ressemblence to the one of Lemma 20 (see Appendix F.3), the main novelty in $\tilde{c}_{a,b}^{G,\epsilon}$ is that it depends on whether $(\tilde{\mu}_{n,a} - \tilde{\mu}_{n,b})_+$ lies above $\epsilon/2$ or below it. This is instrumental to switch between the high-privacy regime, where the Laplace "noise" dominates, and the low-privacy regime, where the Laplace "noise" is "negligible" In the latter case, we recover $c_{a,b}^{G,\epsilon}$, with an improved factor $1/2$ due to tighter concentration inequalities. In the former case, the dominating term in the stopping threshold stems from the control of the Laplace "noise", while the Gaussian "noise" contributes to second-order terms. To the best of our knownledge, Lemma 22 constitutes the first mean-aware stopping thresholds. This idea might be of independent interest, especially when several sources of randomness coexist in the GLR stopping rule. Echoing the discussion in Section 4.1, once a measure of "distinguishability" at the arm's level is obtained, we conjecture that controlling its empirical version in a mean-agnostic fashion is possible.

The transportation cost $W_{a,b}^{G,\epsilon}$ is inspired by $T^\star_{\mathrm{KL}}(\boldsymbol{\nu}_{G,\epsilon})$ as in Eq. (10). Recall that the associated $\beta$-characteristic time $T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}_{G,\epsilon})$ in Eq. (5) is defined as

$$T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}_{G,\epsilon})^{-1} \triangleq \max_{\omega\in\Sigma_K, \omega_{a^\star}=\beta}\min_{a\neq a^\star}\frac{(\mu_{a^\star} - \mu_a)\min\{\epsilon/2, \mu_{a^\star} - \mu_a\}}{2\sigma^2(1/\beta + 1/\omega_a)} \quad \text{with} \quad \sigma = 1/2\,. \tag{15}$$

The algorithmic intuition behind AdaP-TT$^\star$ is the same as for AdaP-TT. The main difference lies in the behavior of the TC challenger that now aims at reaching the equality for all the empirical modified transporation costs $W_{a^\star,a}^{G,\epsilon}$ compared to $W_{a^\star,a}^{G}$ for AdaP-TT. Given that AdaP-TT$^\star$ is designed to reach $T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}_{G,\epsilon})$ asymptotically, we will not achieve our lower bound, as $T^\star_{\mathrm{KL}}(\boldsymbol{\nu}_{G,\epsilon})$ is only an upper bound on $T^\star_g(\boldsymbol{\nu};\epsilon)$ (Corollary 17).

*Asymptotic upper bound.* AdaP-TT$^\star$ achieves an asymptotic upper bound on its expected sample complexity as $T^\star_{\mathrm{KL},\beta}$ evaluated at $\boldsymbol{\nu}$ in the low privacy regime and at $\boldsymbol{\nu}_{G,\epsilon}$ in the high-privacy regime, up to multiplicative privacy-aware cost (Theorem 23).

**Theorem 23** *Let $(\delta, \beta) \in (0,1)^2$ and $\epsilon > 0$. The* AdaP-TT$^\star$ *algorithm is $\epsilon$-global DP, $\delta$-correct and satisfies that, for any Bernoulli $\boldsymbol{\nu}$ instance with distinct means $\mu \in (0,1)^K$,*

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta]}{\log(1/\delta)} \leq \begin{cases} 4T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}) g_1\left(\Delta_{\max}/(\sigma^2\epsilon)\right) & \text{if } \Delta_{\max} < \epsilon/2 \\ 2T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}_{G,\epsilon}) g_2(\epsilon^2 T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}_{G,\epsilon}) \max\{\beta, 1-\beta\}/4)/\sigma^2 & \text{otherwise} \end{cases},$$

*where $T^\star_{\mathrm{KL},\beta}$ as in Eq. (5) for $\sigma = 1/2$ and $T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}_{G,\epsilon})$ as in Eq. (15). The function $g_1(y) = \sup\left\{x \mid x^2 < x + y\sqrt{2x} + \frac{y^2}{4}\right\}$ is increasing on $[0, 12]$ and satisfies that $g_1(0) = 1$ and $g_1(12) \leq 10$. The function $g_2(y) = 1 + 2(\sqrt{1+1/y} - 1)^{-1}$ is increasing on $\mathbb{R}_+^\star$ and satisfies that $\lim_{y \to 0} g_2(y) = 1$.*

**Proof** The proof is almost identical to the one of Theorem 21, hence we defer to Appendix G for more details. While $W^{G,\epsilon}_{a,b}$ is used instead of $W^G_{a,b}$, both arguments rely on similar properties holding for Gaussian-like transportation costs. The main technical novelty occurs during the "inversion" of the GLR stopping rule since the stopping threshold is mean-aware and includes additional second order terms (Appendix G.5). ■

*Discussion.* When $\Delta_{\max} < \epsilon/2$, our upper bound recovers the non-private lower bound for Gaussian distributions $T^\star_{\mathrm{KL}}(\boldsymbol{\nu})$ up to a multiplicative factor $8g_1(4\Delta_{\max}/\epsilon) \in [8, 80]$, whose limit is 8 in non-private regime where $\epsilon \to +\infty$. When $\Delta_{\min} \geq \epsilon/2$, we have $2T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}_{G,\epsilon}) \leq 4T^\star_{\mathrm{TV}}(\boldsymbol{\nu})/\epsilon$. In the asymptotic highly privacy regime where $\epsilon \to 0$, our upper bound matches the lower bound up to a multiplicative factor 16. Therefore, we close the gap left open by the algorithm in Section 4.3. While the regime $\Delta_{\max} \geq \epsilon/2 > \Delta_{\min}$ is relevant for practical application, it is harder to understand how the different quantities interact in the upper/lower bounds in transitional phases. Thus, we do not claim optimality in those phases. Having matching upper and lower bounds *only* for high privacy regimes is an interesting phenomenon that appears in different settings of differential privacy literature, such as regret minimisation (Azize and Basu, 2022), parameter estimation (Cai et al., 2021) and hidden probabilistic graphical models (Nikolakakis et al., 2019). For regret minimisation and BAI, we conjecture that deriving matching bounds at any privacy level fundamentally requires a measure of "distinguishability" at the arm's level. This is a promising research direction that would require finer technical tools to optimally merge the privacy and the correctness constraints.

*Comparison to DP-SE.* DP-SE (Sajed and Sheffet, 2019) is an $\epsilon$-global DP version of the Successive Elimination algorithm introduced for the regret minimisation setting. The algorithm samples active arms uniformly during phases of geometrically increasing length. Based on the private confidence bounds, DP-SE eliminates provably sub-optimal arms at the end of each phase. Due to its phased-elimination structure, DP-SE can be easily converted into an $\epsilon$-global DP FC-BAI algorithm, where we stop once there is only one active arm left. In particular, the proof of Theorem 4.3 of Sajed and Sheffet (2019) shows that with high probability any sub-optimal arm $a \neq a^\star$ is sampled no more than $\mathcal{O}(\Delta_a^2 + (\epsilon\Delta_a)^{-1})$. From this result, it is straightforward to extract a sample complexity upper bound for DP-SE, i.e. $\mathcal{O}(\sum_{a \neq a^\star} \Delta_a^{-2} + \sum_{a \neq a^\star} (\epsilon\Delta_a)^{-1})$. This shows that DP-SE too achieves (ignoring constants) the high-privacy lower bound $T^\star_{\mathrm{TV}}(\boldsymbol{\nu})/\epsilon$ for Bernoulli instances. However, due to its uniform sampling within the phases, DP-SE is less adaptive than TTUCB. Inside

a phase, DP-SE continues to sample arms that might already be known to be bad, while TTUCB adapts its sampling rule based on the transportation costs that reflect the amount of evidence collected in favour of the hypothesis that the leader is the best arm. Finally, TTUCB has the advantage of being anytime, i.e. its sampling strategy does not depend on the risk $\delta$.

Another adaptation of DP-SE, namely DP-SEQ, is proposed in Kalogerias et al. (2021) for the problem of privately finding the arm with the highest quantile at a fixed level, hence it is different from BAI. For multiple agents, Rio et al. (2023) studies privacy for BAI under fixed confidence. They propose and analyse the sample complexity of DP-MASE, a multi-agent version of DP-SE. They show that multi-agent collaboration leads to better sample complexity than independent agents, even under privacy constraints. While the multi-agent setting with federated learning allows tackling large-scale clinical trials taking place at several locations simultaneously, we study the single-agent setting, which is relevant for many small-scale clinical trials (see Example 1).

*Non-asymptotic confidence regime for* AdaP-TT$^\star$. AdaP-TT$^\star$ *has the same property on the leader as* AdaP-TT*, i.e.,* $B_n = a^\star$ *except for a sublinear number of times. However, the modified transportation costs* $W_{a,b}^{G,\epsilon}$ *as in Eq.* (13) *do not provide time-uniform separability between the means and the allocations. As it depends on* $\min\{\epsilon/2, (\tilde{\mu}_a - \tilde{\mu}_b)_+\}$ *in the numerator, we should justifty that* $\tilde{\mu}_{n,a^\star} - \tilde{\mu}_{n,C_n}$ *and* $\mu_{a^\star} - \mu_{C_n}$ *lies on the same side as* $\epsilon/2$*, in order to use the equality at equilibrium of the modified transportation costs to obtain* $T^\star_{\epsilon,\beta}(\boldsymbol{\nu})^{-1}$*. It is challenging to prove this property non-asymptotically without incurring a larger* $\delta$*-independent dependency. Therefore, we conjecture that* AdaP-TT$^\star$ *suffers from a worse non-asymptotic upper bound based on the proof of Jourdan and Degenne (2024). While our asymptotic analysis deals with this subtelty, the extra cost vanishes as* $\delta \to 0$*.*

**Remark 24 (On the number of rounds of adaptivity)** *Used on any existing FC-BAI algorithm, the DAF update yields a batched algorithm, which satisfies* $\epsilon$*-global DP. At the end of the episode of arm a (after updating its mean), it is possible to compute the sequence of all the arms to be pulled before the end of the next episode (for another arm), without taking the collected observations into account. In contrast to the classical batched setting where the batch size is fixed, the size of the resulting batches is adaptive and data-dependent. In the non-private setting (*$\epsilon = +\infty$*), we recover Batched Best-Arm Identification (BBAI) in the fixed-confidence setting.* AdaP-TT *and* AdaP-TT$^\star$ *are asymptotically optimal up to a multiplicative factor 4 with solely* $\mathcal{O}\left(K \log_2(T^\star_{\mathrm{KL}}(\boldsymbol{\nu}) \log(1/\delta))\right)$ *rounds of adaptivity. We refer the reader to Appendix H for more details, including comparison to existing works.*

## 5. Experimental Analysis

We perform experiments for both $\epsilon$-local DP and $\epsilon$-global DP. The code is available at `https://github.com/achraf-azize/DP-BAI`.

### 5.1 Local DP

We run CTB-TT in different Bernoulli instances as in Sajed and Sheffet (2019). As a benchmark, we also compare to the non-private TTUCB. As for $\epsilon$-global DP, we set the risk $\delta = 10^{-2}$, implement all the algorithms in Python (version 3.8) and run each algorithm

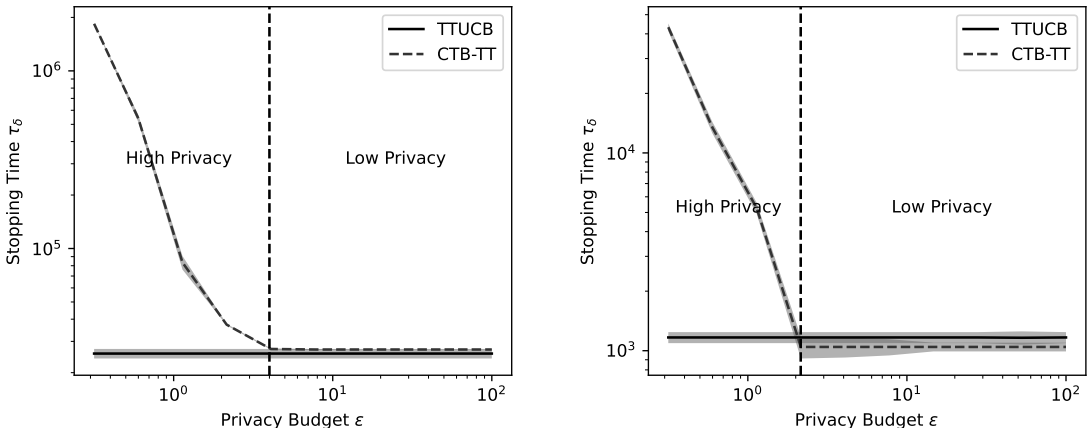



Figure 1: Empirical stopping time $\tau_\delta$ (mean $\pm$ std. over 1000 runs, $\delta = 10^{-2}$) with respect to the privacy budget $\epsilon$ for $\epsilon$-local DP on Bernoulli instance $\mu_1$ (left) and $\mu_2$ (right). The shaded vertical line separates the two privacy regimes.

1000 times. We plot the corresponding average and standard deviations of the empirical stopping times in Figure 1. We also test the algorithms on other Bernoulli instances and report the results in Appendix I.

Figure 1 shows that CTB-TT performance has two regimes. In the low privacy regime ($\epsilon > 4$ for $\mu_1$ and $\epsilon > 2$ for $\mu_2$), the CTB($\epsilon$) estimator reduces to the MLE, and CTB-TT matches exactly the performance of the non-private TTUCB. In the high privacy regime ($\epsilon < 4$ for $\mu_1$ and $\epsilon < 2$ for $\mu_2$), the price of privacy on the stopping time is a multiplicative $\epsilon^{-2}$. Therefore, the sample complexity is prohibitively large to be computed numerically for $\epsilon < 0.1$. The switching value of $\epsilon$ between the low and high privacy regimes is an order of magnitude higher for $\epsilon$-local DP compared to the one for $\epsilon$-global DP. This is predictable since local DP provides a "stronger" privacy guarantee at the cost of worse performance.

### 5.2 Global DP

We compare the performances of AdaP-TT, AdaP-TT$^\star$ and DP-SE for FC-BAI in different Bernoulli instances as in Sajed and Sheffet (2019). The first instance has means $\mu_1 = (0.95, 0.9, 0.9, 0.9, 0.5)$ and the second instance has means $\mu_2 = (0.75, 0.7, 0.7, 0.7, 0.7)$. As a benchmark, we also compare to the non-private TTUCB. We set the risk $\delta = 10^{-2}$ and implement all the algorithms in Python (version 3.8). We run each algorithm 1000 times, and plot corresponding average and standard deviations of the empirical stopping times in Figure 2. We also test the algorithms on other Bernoulli instances and report the results in Appendix I.

Figure 2 shows that: (a) AdaP-TT and AdaP-TT$^\star$ require fewer samples than DP-SE to provide a $\delta$-correct answer, for different values of $\epsilon$ and in all the instances tested. AdaP-TT and AdaP-TT$^\star$ have the same performance in the low privacy regimes, while AdaP-TT$^\star$ improves the sample complexity in the high privacy regime, as predicted theoretically. (b) The experimental performance of AdaP-TT and AdaP-TT$^\star$ demonstrate two regimes. A

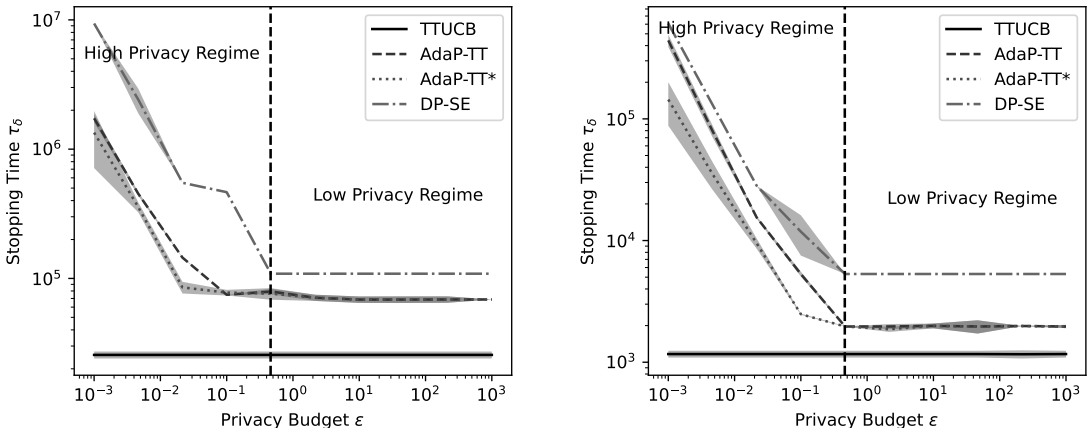



Figure 2: Empirical stopping time $\tau_\delta$ (mean $\pm$ std. over 1000 runs) with respect to the privacy budget $\epsilon$ for $\epsilon$-global DP on Bernoulli instance $\mu_1$ (left) and $\mu_2$ (right). The shaded vertical line separates the two privacy regimes.

| Algorithm | Trust model | Mean estimator | Transportation cost | Stopping threshold |
|---|---|---|---|---|
| TTUCB | None | MLE (Alg. 2) | $W^G$ as in Eq. (3) | $c^G$ as in Eq. (4) |
| CTB-TT | $\epsilon$-local DP | CTB (Alg. 4) | $W^G$ as in Eq. (3) | $c^G$ as in Eq. (4) |
| AdaP-TT | $\epsilon$-global DP | DAF (Alg. 5) | $W^G$ as in Eq. (3) | $c^{G,\epsilon}$ as in Eq. (12) |
| AdaP-TT$^\star$ | $\epsilon$-global DP | DAF (Alg. 5) | $W^{G,\epsilon}$ as in Eq. (13) | $\tilde{c}^{G,\epsilon}$ as in Eq. (14) |

Table 1: Instances of the TTUCB meta-algorithm defined in Algorithm 1.

high-privacy regime (for $\epsilon < 0.1$ for $\mu_1$ and $\epsilon < 0.4$ for $\mu_2$), where the stopping time depends on the privacy budget $\epsilon$, and a low privacy regime (for $\epsilon > 0.1$ for $\mu_1$ and $\epsilon > 0.4$ for $\mu_2$), where the performance of AdaP-TT and AdaP-TT$^\star$ does not depend on $\epsilon$, and is four times the samples required by TTUCB in the worst case, as shown theoretically.

## 6. Conclusion and Perspectives

We study FC-BAI with $\epsilon$-local DP and $\epsilon$-global DP. In both settings, we derive a lower bound on the expected sample complexity which quantifies the additional samples needed by a $\delta$-correct BAI strategy to ensure DP. The lower bounds further suggest the existence of two privacy regimes. In the *low-privacy regime*, no additional samples are needed, and *privacy can be achieved for free*. For the *high-privacy regime*, the lower bound reduces to $\Omega(\epsilon^{-2}T^\star_{\mathrm{TV}^2}(\nu))$ for $\epsilon$-local DP, and to $\Omega(\epsilon^{-1}T^\star_{\mathrm{TV}}(\nu))$ for $\epsilon$-global DP. To match those lower bounds up to multiplicative constants, we propose $\epsilon$-local DP and $\epsilon$-global DP variants of a Top Two algorithm. For $\epsilon$-local DP, the CTB-TT algorithm reaches asymptotic optimality by plugging in a private estimator of the means based on Randomised Response. For $\epsilon$-global DP, our private estimator of the mean runs in *arm-dependent adaptive episodes* and adds *Laplace noise* to ensure a good privacy-utility trade-off. By solely plugging in this estimator,

the AdaP-TT algorithm fails to recover the asymptotic lower bound for instances with highly different mean gaps. The AdaP-TT$^\star$ algorithm overcomes this limitation by adapting the transportation costs, hence reaching the asymptotic lower bound in the asymptotic high-privacy regime (up to a small multiplicative constant).

The upper bound matches the lower bound by a multiplicative constant in the high privacy regime, and is also loose in some instances in the low privacy regime, due to the mismatch between the KL divergence of Bernoulli distributions and that of Gaussian. One possible direction to solve this issue is to use transportation costs tailored to Bernoulli for both the Top Two Sampling and the stopping. Since our bounds only give a clear picture in the high and low privacy regimes, it would be interesting to provide better insights for the regime in-between where both the $\delta$-correctness and the DP constraints are of the same order. An interesting direction would be to extend the proposed technique to other variants of pure DP, namely $(\epsilon, \delta)$-DP and Rényi-DP (Mironov, 2017), or other trust models, e.g. shuffle DP (Cheu, 2021; Girgis et al., 2021).

## Acknowledgments

This work has been partially supported by the THIA ANR program "AI_PhD@Lille". A. Al-Marjani acknowledges the support of the Chaire SeqALO (ANR-20-CHIA-0020). D. Basu acknowledges the Inria-Kyoto University Associate Team "RELIANT" for supporting the project, the ANR JCJC for the REPUBLIC project (ANR-22-CE23-0003-01), and the PEPR project FOUNDRY (ANR23-PEIA-0003). We thank Emilie Kaufmann and Aurélien Garivier for the interesting conversations. We also thank Philippe Preux for his support.

## Appendix A. Outline

The appendices are organised as follows:

- Notation are summarized in Appendix B.
- For clarity, we explicit the algorithms as standalone pseudocode in Appendix C.
- In Appendix D, we detail the proofs of our lower bounds for $\epsilon$-local DP (Theorem 9 and Corollary 10) and $\epsilon$-global DP (Theorem 16 and Corollary 17).
- In Appendix E, we show that AdaP-TT and AdaP-TT$^\star$ are $\epsilon$-global DP since they use the DAF($\epsilon$) estimator of the means.
- In Appendix F, we prove that the $\epsilon$-global DP GLR stopping rules yields $\delta$-correctness regardless of the sampling rule, both when using non-private transportation costs (Lemma 20) and adapted transportation costs (Lemma 22).
- In Appendix G, we detail the proofs of the asymptotic upper bound on the expected sample complexity of AdaP-TT (Theorem 21) and AdaP-TT$^\star$ (Theorem 23).
- In Appendix H, we discuss in more details the number of rounds of adaptivity.
- Extended experiments are presented in Appendix I.

## Appendix B. Notation

We recall some commonly used notation: the set of integers $[n] \triangleq \{1, \cdots, n\}$, the $\ell_1$-norm $\|\cdot\|_1$, the complement $X^{\complement}$ and interior $\mathring{X}$ of a set $X$, the indicator function $\mathbb{1}(X)$ of an event, the probability $\mathbb{P}_{\boldsymbol{\nu}}$ and the expectation $\mathbb{E}_{\boldsymbol{\nu}}$ taken over the randomness of the observations from $\boldsymbol{\nu}$ and the randomness of the algorithm, Landau's notation $o$, $\mathcal{O}$, $\Omega$ and $\Theta$, the $(K-1)$-dimensional probability simplex $\Sigma_K \triangleq \left\{\omega \in \mathbb{R}_+^K \mid \omega \geq 0,\ \sum_{a\in[K]} \omega_a = 1\right\}$. Lap($b$) be the Laplace distribution with mean/variance $(0, 2b^2)$, and Ber($p$) Bernoulli of parameter $p$. The functions $(x)_+ = \max\{0, x\}$, $k(x) \triangleq \log_2(x) + 2$, $\overline{W}_{-1}$ in Lemma 39, $\mathcal{C}_G$ as in Eq. (20), $\zeta$ is the Riemann $\zeta$ function, $h$ in Lemma 22, $g_1$ and $g_2$ in Theorem 23. The concentration parameters $s > 1$ and $\alpha > 1$, e.g., $s = \alpha = 1.2$. Moreover, we recall the definitions: $T^\star_{\mathrm{KL}}(\boldsymbol{\nu})$, $T^\star_{\mathrm{TV}}(\boldsymbol{\nu})$, $T^\star_{\mathrm{TV}^2}(\boldsymbol{\nu})$ as in Eq. (1) for the KL divergence, the TV distance and the squared TV distance, $H(\boldsymbol{\nu}) = 2\sigma^2 \sum_{a\in[K]} \Delta_a^{-2}$, $H(\boldsymbol{\nu}, \epsilon) = 2\sigma^2 \sum_{a\in[K]} \Delta_{\epsilon,a}^{-2}$ where $\Delta_{\epsilon,a} \triangleq \epsilon(\sqrt{1 + 2\Delta_a/\epsilon} - 1)$, $\nu_{\epsilon,a} \triangleq \mathrm{Ber}(\mu_{\epsilon,a})$ with $\mu_{\epsilon,a} \triangleq (2\mu_a - 1)\frac{e^\epsilon - 1}{2(e^\epsilon + 1)} + 1/2$ for all $a \in [K]$. The estimator mechanisms $(\mathrm{ESTIMATOR}_a)_{a\in[K]}$, e.g., MLE (Alg. 2), CTB (Alg. 4), DAF (Alg. 5), see Table 1. The stopping conditions $(\mathrm{STOP}_{a,b})_{(a,b)\in[K]^2}$ as in Eq. (2). A BAI strategy is denoted by $\pi$, and is composed of a sequence of sampling and stopping rules $S_n$, and recommendation rule $\mathrm{Rec}_n$. The user interacting with the BAI strategy at step $t$ is $u_t$, has potential reward vector $\boldsymbol{x}_t$, which are combined in the table of potential rewards $\underline{d} = (\boldsymbol{x}_1, \dots)$. For local DP, we denote by $\mathcal{M}$ the perturbation mechanism that each user $u_t$ uses to send their noisy rewards. While Table 2 gathers problem-specific notation, Table 3 groups notation for the algorithms.

| Notation | Type | Description |
|---|---|---|
| $K$ | $\mathbb{N}$ | Number of arms |
| $\mathcal{F}$ | $\subseteq \mathcal{P}([0,1])$ | Class of distributions, e.g., Bernoulli |
| $\sigma$ | $\mathbb{R}_+^\star$ | $\sigma$-sub-Gaussian, e.g., $\sigma = 1/2$ for Bernoulli |
| $a$ | $[K]$ | Arm |
| $\nu_a$ | $\mathcal{F}$ | Distribution of arm $a \in [K]$ |
| $\boldsymbol{\nu}$ | $\mathcal{F}^K$ | Vector of distributions, $\nu \triangleq (\nu_a)_{a\in[K]}$ |
| $\mathcal{M}$ | $\subseteq \mathcal{F}^K$ | Class of bandit instances |
| $\mu_a$ | $(0,1)$ | Mean of arm $a \in [K]$, $\mu_a \triangleq \mathbb{E}_{X\sim\nu_a}[X]$ |
| $\mu$ | $(0,1)^K$ | Vector of means, $\mu \triangleq (\mu_a)_{a\in[K]}$ |
| $\boldsymbol{\nu}_G$ | | Gaussian instance $\mathcal{N}(\mu, I_K)$ |
| $a^\star(\boldsymbol{\nu})$ | $\subseteq [K]$ | Set of best arms, $a^\star(\boldsymbol{\nu}) \triangleq \arg\max_{a\in[K]} \mu_a$ |
| $a^\star$ | $[K]$ | Unique best arm, i.e., $a^\star(\mu) = \{a^\star\}$ |
| $\Delta_a$ | $(0,1)$ | Mean gap arm $a \in [K]$, i.e., $\Delta_a \triangleq \mu_{a^\star} - \mu_a$, $\Delta_{a^\star} = \Delta_{\min} \triangleq \min_{a\neq a^\star} \Delta_a$ and $\Delta_{\max} \triangleq \max_{a\neq a^\star} \Delta_a$ |
| $\epsilon$ | $\mathbb{R}_+^\star$ | Privacy budget, e.g., $\epsilon$-global or $\epsilon$-local DP |
| $\delta$ | $(0,1)$ | Risk for $\delta$-correctness, i.e., confidence $1-\delta$ |
| $\mathrm{Alt}(\boldsymbol{\nu})$ | $\subseteq \mathcal{M}$ | Alternative instances with different best arms |
| $\Sigma_K$ | $\subseteq \mathbb{R}_+^K$ | $(K-1)$-dimensional probability simplex |
| $\mathbf{d}$ | | Measure between distributions, e.g., KL, TV or $\mathrm{TV}^2$ |
| $\beta$ | $(0,1)$ | Fixed proportion, e.g., $\beta = 1/2$ |
| $T^\star_{\mathbf{d}}(\boldsymbol{\nu})$, $T^\star_{\mathbf{d},\beta}(\boldsymbol{\nu})$ | $\mathbb{R}_+^\star$ | ($\beta$-)Characteristic time for $\boldsymbol{\nu}$ given $\mathbf{d}$, see Eq. (1) |
| $\omega^\star_{\mathbf{d}}(\boldsymbol{\nu})$, $\omega^\star_{\mathbf{d},\beta}(\boldsymbol{\nu})$ | $\Sigma_K$ | ($\beta$-)Optimal allocation, maximiser of Eq. (1) |
| $T^\star_\ell(\boldsymbol{\nu};\epsilon)$ | $\mathbb{R}_+^\star$ | $\epsilon$-Local lower bound for $\boldsymbol{\nu}$, see Theorem 9 |
| $T^\star_g(\boldsymbol{\nu};\epsilon)$ | $\mathbb{R}_+^\star$ | $\epsilon$-Global lower bound for $\boldsymbol{\nu}$, see Theorem 16 |

Table 2: Notation for the setting.

| Notation | Type | Description |
|---|---|---|
| $n$ | $\mathbb{N}$ | Time |
| $a_n$ | $[K]$ | Arm sampled at time $n$ |
| $r_n$ | $[0,1]$ | Sample observed at the end of time $n$, i.e. $r_n \sim \nu_{a_n}$ |
| $\mathcal{H}_n$ | | History after $n$, $\mathcal{H}_n = \sigma(a_1, r_1, \cdots, a_n, r_n)$ |
| $\top$ | | Stopping action |
| $\tau_\delta$ | $\mathbb{N}$ | Sample complexity (i.e., stopping time) |
| $\hat{a}_n$ | $[K]$ | Arm recommended before time $n$ |
| $\widehat{a}$ | $[K]$ | Arm recommended when stopping |
| $B_n$ | $[K]$ | (UCB) Leader at time $n$ |
| $C_n$ | $[K]$ | (TC) Challenger at time $n$ |
| $b_a$ | $\mathbb{N}^K \to \mathbb{R}_+^\star$ | Confidence bonuses for arm $a$, e.g., $b_a^G$ and $b_a^{G,\epsilon}$ as in Eq. (3) and (11) |
| $W_{a,b}$ | $\mathbb{R}^K \times \mathbb{N}^K \to \mathbb{R}_+^\star$ | Transportation costs for $(a,b)$, e.g., $W_{a,b}^G$ and $W_{a,b}^{G,\epsilon}$ as in Eq. (3) and (13) |
| $c_{a,b}$ | $\mathbb{N}^K \times (0,1) \to \mathbb{R}_+^\star$ | Stopping threshold function for $(a,b)$, e.g., $c_{a,b}^G$, $c_{a,b}^{G,\epsilon}$ and $\tilde{c}_{a,b}^{G,\epsilon}$ as in Eq. (4), (12) and (14) |
| $\tilde{r}_n$ | $[0,1]$ | Randomised Response mechanism ($\epsilon$-local DP) |
| $N_{n,a}$ | $\mathbb{N}$ | Number of pulls of arm $a$ before time $n$ |
| $k_{n,a}$ | $\mathbb{N}$ | Current phase of arm $a$ at time $n$ |
| $\tilde{N}_{k_{n,a},a}$ | $\mathbb{N}$ | Local count of arm $a$ before time $n$ |
| $\hat{\mu}_{n,a}$ | $\mathbb{N}$ | Non-private estimator based on $\tilde{N}_{n,a}$ observations |
| $\tilde{\mu}_{n,a}$ | $\mathbb{N}$ | Private estimator based on $\tilde{N}_{n,a}$ observations |
| $T_k(a)$ | $\mathbb{N}$ | Time $n$ where the arm $a$ changes to phase $k$ |
| $Y_{k,a}$ | $\mathbb{R}$ | Laplace "noise" at phase $k$ for arm $a$ ($\epsilon$-global DP) |
| $L_{n,a}$ | $\mathbb{N}$ | Counts of $B_t = a$ before time $n$ |
| $N_{n,a}^a$ | $\mathbb{N}$ | Counts of $(B_t, a_t) = (a,a)$ before time $n$ |

Table 3: Notation for the algorithm.

## Appendix C. Algorithm Standalone Pseudocodes

For clarity, we state the algorithms as standalone pseudocodes: TTUCB algorithm (Jourdan and Degenne, 2024) in Algorithm 6, CTB-TT algorithm in Algorithm 7, $(\epsilon, \gamma)$-DP TTUCB algorithm in Algorithm 8, AdaP-TT algorithm in Algorithm 9 and AdaP-TT$^\star$ algorithm in Algorithm 10.

**Algorithm 6** TTUCB Algorithm (Jourdan and Degenne, 2024)

**Input:** setting parameter $\delta \in (0,1)$, algorithmic hyperparameters $(\beta, s, \alpha) \in (0,1)$, e.g., $(\beta, s, \alpha) = (1/2, 1.2, 1.2)$, stopping threshold $(c^G_{a,b})_{(a,b)\in[K]^2}$ as in Eq. (4), confidence bonuses $(b^G_a)_{a\in[K]}$ and transportation costs $(W^G_{a,b})_{(a,b)\in[K]^2}$ as in Eq. (3).
**Initialisation:** Observe $r_a \sim \nu_a$ for all $a \in [K]$; Initialise the estimators $\hat{\mu}_{n,a} = r_a$, $N_{n,a} = 1$, $L_{n,a} = N^a_{n,a} = 0$ where $n = K+1$ ;
**for** $n > K$ **do**
  Set arm $\hat{a}_n = \arg\max_{a\in[K]} \hat{\mu}_{n,a}$ ; ▷ Recommendation rule
  **if** $W^G_{\hat{a}_n,a}(\hat{\mu}_n, N_n) \geq c^G_{\hat{a}_n,a}(N_n, \delta)$ for all $a \neq \hat{a}_n$ **then** ▷ GLR stopping rule
    Set $a_n = \top$ and **return** $\hat{a}_n$ ;
  **end if**
  Set arm $B_n = \arg\max_{a\in[K]} \left\{\hat{\mu}_{n,a} + b^G_a(N_n)\right\}$; ▷ UCB leader
  Set arm $C_n = \arg\min_{a\neq B_n} W^G_{B_n,a}(\hat{\mu}_n, N_n)$; ▷ TC challenger
  Set arm $a_n = B_n$ if $N^{B_n}_{n,B_n} \leq \beta L_{n+1,B_n}$, and $a_n = C_n$ otherwise; ▷ $\beta$-tracking
  Pull $a_n$ and observe $r_n \sim \nu_{a_n}$; ▷ Sampling rule
  Set $N_{n+1,a_n} \leftarrow N_{n,a_n} + 1$, $L_{n+1,B_n} \leftarrow L_{n,B_n} + 1$, $N^{B_n}_{n+1,B_n} \leftarrow N^{B_n}_{n,B_n} + \mathbb{1}(B_n = a_n)$ ;
  Set $\hat{\mu}_{n+1,a} = N^{-1}_{n+1,a} \sum_{t\in[n]} r_t \mathbb{1}\{a_t = a\}$ and $n \leftarrow n+1$ ; ▷ Update rule
**end for**

**Algorithm 7** CTB-TT Algorithm

**Input:** setting parameters $(\epsilon, \delta) \in \mathbb{R}_+^\star \times (0,1)$, algorithmic hyperparameters $(\beta, s, \alpha) \in (0,1)$, e.g., $(\beta, s, \alpha) = (1/2, 1.2, 1.2)$, stopping threshold $(c^G_{a,b})_{(a,b)\in[K]^2}$ as in Eq. (4), confidence bonuses $(b^G_a)_{a\in[K]}$ and transportation costs $(W^G_{a,b})_{(a,b)\in[K]^2}$ as in Eq. (3) with $\sigma = 1/2$.
**Initialisation:** Observe $\tilde{r}_a \sim \mathrm{Ber}\left(\frac{r_a(e^\epsilon-1)+1}{e^\epsilon+1}\right)$ where $r_a \sim \nu_a$ for all $a \in [K]$; Initialise the estimators $\tilde{\mu}_{n,a} = \tilde{r}_a$, $N_{n,a} = 1$, $L_{n,a} = N^a_{n,a} = 0$ where $n = K+1$ ;
**for** $n > K$ **do**
  Set $\tilde{a}_n = \arg\max_{a\in[K]} \tilde{\mu}_{n,a}$ ; ▷ Recommendation rule
  **if** $W^G_{\tilde{a}_n,a}(\tilde{\mu}_n, N_n) \geq c^G_{\tilde{a}_n,a}(N_n, \delta)$ for all $a \neq \tilde{a}_n$ **then** ▷ GLR stopping rule
    Set arm $a_n = \top$ and **return** $\tilde{a}_n$ ;
  **end if**
  Set arm $B_n = \arg\max_{a\in[K]} \left\{\tilde{\mu}_{n,a} + b^G_a(N_n)\right\}$; ▷ UCB leader
  Set arm $C_n = \arg\min_{a\neq B_n} W^G_{B_n,a}(\tilde{\mu}_n, N_n)$; ▷ TC challenger
  Set arm $a_n = B_n$ if $N^{B_n}_{n,B_n} \leq \beta L_{n+1,B_n}$, and $a_n = C_n$ otherwise; ▷ $\beta$-tracking
  Pull $a_n$, observe and store $\tilde{r}_n \sim \mathrm{Ber}\left(\frac{r_n(e^\epsilon-1)+1}{e^\epsilon+1}\right)$ where $r_n \sim \nu_{a_n}$; ▷ Sampling rule
  Set $N_{n+1,a_n} \leftarrow N_{n,a_n} + 1$, $L_{n+1,B_n} \leftarrow L_{n,B_n} + 1$, $N^{B_n}_{n+1,B_n} \leftarrow N^{B_n}_{n,B_n} + \mathbb{1}(B_n = a_n)$ ;
  Set $\tilde{\mu}_{n+1,a} = N^{-1}_{n+1,a} \sum_{t\in[n]} \tilde{r}_t \mathbb{1}\{a_t = a\}$ and $n \leftarrow n+1$; ▷ Update rule
**end for**

**Algorithm 8** $(\epsilon, \gamma)$-DP TTUCB Algorithm, based on Algorithm 1 in Zheng et al. (2020)

**Input:** setting parameters $(\epsilon, \gamma, \delta) \in \mathbb{R}_+^\star \times (0,1)$, algorithmic hyperparameters $(\beta, s, \alpha) \in (0,1)$, e.g., $(\beta, s, \alpha) = (1/2, 1.2, 1.2)$, stopping threshold $(c^G_{a,b})_{(a,b)\in[K]^2}$ as in Eq. (4), confidence bonuses $(b^G_a)_{a\in[K]}$ and transportation costs $(W^G_{a,b})_{(a,b)\in[K]^2}$ as in Eq. (3) with $\sigma = 1/2$.
**Initialisation:** Set $\sigma_{\epsilon,\gamma} \triangleq \frac{\sqrt{2\log(1.25/\gamma)}}{\epsilon}$; Observe $\tilde{r}_a \sim \mathcal{N}(r_a, \sigma^2_{\epsilon,\gamma})$ where $r_a \sim \nu_a$ for all $a \in [K]$; Initialise the estimators $\tilde{\mu}_{n,a} = \tilde{r}_a$, $N_{n,a} = 1$, $L_{n,a} = N^a_{n,a} = 0$ where $n = K+1$;
**for** $n > K$ **do**
  Set $\tilde{a}_n = \arg\max_{a\in[K]} \tilde{\mu}_{n,a}$ ; ▷ Recommendation rule
  **if** $W^G_{\tilde{a}_n,a}(\tilde{\mu}_n, N_n) \geq c^G_{\tilde{a}_n,a}(N_n, \delta)$ for all $a \neq \tilde{a}_n$ **then** ▷ GLR stopping rule
    Set arm $a_n = \top$ and **return** $\tilde{a}_n$ ;
  **end if**
  Set arm $B_n = \arg\max_{a\in[K]} \left\{\tilde{\mu}_{n,a} + b^G_a(N_n)\right\}$; ▷ UCB leader
  Set arm $C_n = \arg\min_{a\neq B_n} W^G_{B_n,a}(\tilde{\mu}_n, N_n)$; ▷ TC challenger
  Set arm $a_n = B_n$ if $N^{B_n}_{n,B_n} \leq \beta L_{n+1,B_n}$, and $a_n = C_n$ otherwise; ▷ $\beta$-tracking
  Pull $a_n$, observe and store $\tilde{r}_n \sim \mathcal{N}(r_n, \sigma^2_{\epsilon,\gamma})$ where $r_n \sim \nu_{a_n}$; ▷ Sampling rule
  Set $N_{n+1,a_n} \leftarrow N_{n,a_n} + 1$, $L_{n+1,B_n} \leftarrow L_{n,B_n} + 1$, $N^{B_n}_{n+1,B_n} \leftarrow N^{B_n}_{n,B_n} + \mathbb{1}(B_n = a_n)$ ;
  Set $\tilde{\mu}_{n+1,a} = N^{-1}_{n+1,a} \sum_{t\in[n]} \tilde{r}_t \mathbb{1}\{a_t = a\}$ and $n \leftarrow n+1$ ; ▷ Update rule
**end for**

**Algorithm 9** AdaP-TT Algorithm

**Input:** setting parameters $(\epsilon, \delta) \in \mathbb{R}_+^\star \times (0,1)$, algorithmic hyperparameter $\beta \in (0,1)$, e.g., $\beta = 1/2$, stopping threshold $(c_{a,b}^{G,\epsilon})_{(a,b)\in[K]^2}$ as in Eq. (12), confidence bonuses $(b_a^{G,\epsilon})_{a\in[K]}$ as in Eq. (11) and transportation costs $(W_{a,b}^G)_{(a,b)\in[K]^2}$ as in Eq. (3) with $\sigma = 1/2$.

**Initialisation:** Observe $r_a \sim \nu_a$ for all $a \in [K]$ and draw $Y_{1,a} \sim \mathrm{Lap}\,(1/\epsilon)$; Initialise the estimators $\tilde{\mu}_{n,a} = r_a + Y_{1,a}$, $k_a = 1$, $T_1(a) = K+1$, $\tilde{N}_{n,a} = N_{n,a} = 1$, $L_{n,a} = N_{n,a}^a = 0$ where $n = K+1$ ;

**for** $n > K$ **do**

  **if** $N_{n,a} \geq 2N_{T_{k_{n,a}}(a),a}$ **then** ▷ Per-arm doubling update grid

    Change phase $k_{n,a} \leftarrow k_{n,a} + 1$ for arm $a$ ;

    Set $T_{k_{n,a}}(a) = n$ and $\tilde{N}_{k_{n,a},a} = N_{T_{k_{n,a}}(a),a} - N_{T_{k_{n,a}-1}(a),a}$ ;

    Set $\hat{\mu}_{k_{n,a},a} = \tilde{N}_{k_{n,a},a}^{-1} \sum_{t=T_{k_{n,a}-1}(a)}^{T_{k_{n,a}}(a)-1} r_t \mathbb{1}\{a_t = a\}$ ; ▷ Forgetting past observations

    Set $\tilde{\mu}_{k_{n,a},a} = \hat{\mu}_{k_{n,a},a} + Y_{k_{n,a},a}$ where $Y_{k_{n,a},a} \sim \mathrm{Lap}((\epsilon\tilde{N}_{k_{n,a},a})^{-1})$ ▷ Private estimator

  **end if**

  Set arm $\tilde{a}_n = \arg\max_{a\in[K]} \tilde{\mu}_{n,a}$ ; ▷ Recommendation rule

  **if** $W_{\tilde{a}_n,a}^G(\tilde{\mu}_n, \tilde{N}_n) \geq c_{\tilde{a}_n,a}^{G,\epsilon}(\tilde{N}_n, \delta)$ for all $a \neq \tilde{a}_n$ **then** ▷ GLR stopping rule

    Set $a_n = \top$ and **return** $\tilde{a}_n$ ;

  **end if**

  Set arm $B_n = \arg\max_{a\in[K]} \left\{\tilde{\mu}_{n,a} + b_a^{G,\epsilon}(\tilde{N}_n)\right\}$; ▷ UCB leader

  Set arm $C_n = \arg\min_{a\neq B_n} W_{B_n,a}^G(\tilde{\mu}_n, N_n)$ ; ▷ TC challenger

  Set arm $a_n = B_n$ if $N_{n,B_n}^{B_n} \leq \beta L_{n+1,B_n}$, and $a_n = C_n$ otherwise; ▷ $\beta$-tracking

  Pull $a_n$, observe and store $r_n \sim \nu_{a_n}$; ▷ Sampling rule

  Set $N_{n+1,a_n} \leftarrow N_{n,a_n} + 1$, $L_{n+1,B_n} \leftarrow L_{n,B_n} + 1$, $N_{n+1,B_n}^{B_n} \leftarrow N_{n,B_n}^{B_n} + \mathbb{1}(B_n = a_n)$ and $n \leftarrow n+1$ ;

**end for**

**Algorithm 10** AdaP-TT$^\star$ Algorithm

**Input:** setting parameters $(\epsilon, \delta) \in \mathbb{R}_+^\star \times (0,1)$, algorithmic hyperparameter $\beta \in (0,1)$, e.g., $\beta = 1/2$, stopping threshold $(\tilde{c}_{a,b}^{G,\epsilon})_{(a,b)\in[K]^2}$ as in Eq. (14), confidence bonuses $(b_a^{G,\epsilon})_{a\in[K]}$ as in Eq. (11) and transportation costs $(W_{a,b}^{G,\epsilon})_{(a,b)\in[K]^2}$ as in Eq. (13) with $\sigma = 1/2$.
**Initialisation:** Observe $r_a \sim \nu_a$ for all $a \in [K]$ and draw $Y_{1,a} \sim \mathrm{Lap}\,(1/\epsilon)$; Initialise the estimators $\tilde{\mu}_{n,a} = r_a + Y_{1,a}$, $k_a = 1$, $T_1(a) = K+1$, $\tilde{N}_{n,a} = N_{n,a} = 1$, $L_{n,a} = N_{n,a}^a = 0$ where $n = K+1$ ;
**for** $n > K$ **do**
  **if** $N_{n,a} \geq 2N_{T_{k_{n,a}}(a),a}$ **then** ▷ Per-arm doubling update grid
    Change phase $k_{n,a} \leftarrow k_{n,a} + 1$ for arm $a$ ;
    Set $T_{k_{n,a}}(a) = n$ and $\tilde{N}_{k_{n,a},a} = N_{T_{k_{n,a}}(a),a} - N_{T_{k_{n,a}-1}(a),a}$ ;
    Set $\hat{\mu}_{k_{n,a},a} = \tilde{N}_{k_{n,a},a}^{-1} \sum_{t=T_{k_{n,a}-1}(a)}^{T_{k_{n,a}}(a)-1} r_t \mathbb{1}\{a_t = a\}$ ; ▷ Forgetting past observations
    Set $\tilde{\mu}_{k_{n,a},a} = \hat{\mu}_{k_{n,a},a} + Y_{k_{n,a},a}$ where $Y_{k_{n,a},a} \sim \mathrm{Lap}((\epsilon \tilde{N}_{k_{n,a},a})^{-1})$ ▷ Private estimator
  **end if**
  Set arm $\tilde{a}_n = \arg\max_{a\in[K]} \tilde{\mu}_{n,a}$ ; ▷ Recommendation rule
  **if** $W_{\tilde{a}_n,a}^{G,\epsilon}(\tilde{\mu}_n, \tilde{N}_n) \geq \tilde{c}_{\tilde{a}_n,a}^{G,\epsilon}(\tilde{N}_n, \delta)$ for all $a \neq \tilde{a}_n$ **then** ▷ GLR stopping rule
    Set $a_n = \top$ and **return** $\tilde{a}_n$ ;
  **end if**
  Set arm $B_n = \arg\max_{a\in[K]} \left\{\tilde{\mu}_{n,a} + b_a^{G,\epsilon}(\tilde{N}_n)\right\}$; ▷ UCB leader
  Set arm $C_n = \arg\min_{a\neq B_n} W_{B_n,a}^{G,\epsilon}(\tilde{\mu}_n, N_n)$ ; ▷ TC challenger
  Set arm $a_n = B_n$ if $N_{n,B_n}^{B_n} \leq \beta L_{n+1,B_n}$, and $a_n = C_n$ otherwise; ▷ $\beta$-tracking
  Pull $a_n$, observe and store $r_n \sim \nu_{a_n}$; ▷ Sampling rule
  Set $N_{n+1,a_n} \leftarrow N_{n,a_n} + 1$, $L_{n+1,B_n} \leftarrow L_{n,B_n} + 1$, $N_{n+1,B_n}^{B_n} \leftarrow N_{n,B_n}^{B_n} + \mathbb{1}(B_n = a_n)$ and $n \leftarrow n+1$ ;
**end for**

## Appendix D. Lower Bounds on the Expected Sample Complexity

In this section, we provide the proofs for the sample complexity lower bounds. First, we present the canonical model for BAI to introduce the relevant quantities. Then, we prove the $\epsilon$-local DP sample complexity lower bound. Finally, for global-DP, we first prove an $\epsilon$-global version of the transportation lemma, i.e. Lemma 15. Using this lemma, we prove the $\epsilon$-global DP sample complexity lower bound of Theorem 16. We also prove the formula expressing the TV characteristic time for Bernoulli instances.

### D.1 Canonical Model for BAI

Let $\boldsymbol{\nu} \triangleq \{\nu_a : a \in [K]\}$ be a bandit instance, consisting of $K$ arms with finite means $\{\mu_a\}_{a\in[K]}$. Now, we recall the interaction between a BAI strategy $\pi$ and the bandit instance $\nu$ in the Protocol 3. The BAI strategy $\pi$ halts at $\tau$, samples a sequence of actions $\underline{A}^\tau$,

and recommends the action $\hat{A}$. Let $\mathbb{P}_{\boldsymbol{\nu},\pi}$ be the probability distribution over the triplets $(\tau, \underline{A}^\tau, \hat{A})$, when the BAI strategy $\pi$ interacts with the bandit instance $\nu$.

For a fixed $T > 1$, a sequence of actions $\underline{a}^T = (a_1, \dots, a_T) \in [K]^T$ and a recommendation $\hat{a} \in [K]$, we define the event $E = \{\tau = T, \underline{A}^\tau = \underline{a}^T, \hat{A} = \hat{a}\}$. We have that

$$\mathbb{P}_{\boldsymbol{\nu},\pi}(E) = \int_{\underline{r}^T=(r_1,\dots,r_T)\in\mathbb{R}^T} \pi(\underline{a}^T, \hat{a}, T \mid \underline{r}^T) \prod_{t=1}^{T} \mathrm{d}\nu_{a_t}(r_t) dr_t$$

where

$$\pi(\underline{a}^T, \widehat{a}, T \mid \underline{r}^T) \triangleq \mathrm{Rec}_{T+1}\left(\widehat{a} \mid \mathcal{H}_T\right) \mathrm{S}_{T+1}\left(\top \mid \mathcal{H}_T\right) \prod_{t=1}^{T} \mathrm{S}_t\left(a_t \mid \mathcal{H}_{t-1}\right)$$

and $\mathcal{H}_t = (a_1, r_1, \dots, a_t, r_t)$.

*Remark on the bandit feedback for global DP.* Let $\pi$ be an $\epsilon$-DP BAI strategy. Let $T \geq 1$, $\underline{a}^T \in [K]^T$ a sequence sampled actions and $\widehat{a} \in [K]$ a recommended actions. This time, let $\underline{r}^T = \{r_1, \dots, r_T\} \in \mathbb{R}^T$ and $\underline{r'}^T \in \mathbb{R}^T$ two neighbouring sequence of rewards, i.e. $d_{\mathrm{Ham}}(\underline{r}^T, \underline{r'}^T) \triangleq \sum_{t=1}^{T} \mathbb{1}\{r_t \neq r'_t\} = 1$ . Consider the table of rewards $\underline{d}^T$ consisting of concatenating $\underline{r}^T$ colon-wise $K$ times, i.e. $\underline{d}^T_{t,i} = \underline{r}^T_t$ for all $i \in [K]$ and all $t \in [T]$ . Define $\underline{d'}^T$ similarly with respect to $\underline{r'}^T$.

In this case, by definition of $\pi$, $\underline{d}^T$ and $\underline{d'}^T$, it is direct that

$$\pi(\underline{a}^T, \widehat{a}, T \mid \underline{r}^T) = \pi(\underline{a}^T, \widehat{a}, T \mid \underline{d}^T)$$

and $d_{\mathrm{Ham}}(\underline{d}^T, \underline{d'}^T) = 1$.

Which means that

$$\pi(\underline{a}^T, \widehat{a}, T \mid \underline{r}^T) \leq e^\epsilon \pi(\underline{a}^T, \widehat{a}, T \mid \underline{r'}^T).$$

In other words, *if $\pi$ is $\epsilon$-pure DP for neighbouring table of rewards $\underline{d}^T$, then $\pi$ is also $\epsilon$-pure DP for neighbouring sequence of observed rewards $\underline{r}^T$.*

*Remark on the local DP canonical model.* Let $(\mathcal{M}, \pi)$ be a pair of perturbation mechanism and BAI satisfying $\epsilon$-local DP. Let $\boldsymbol{\nu} \triangleq \{\nu_a : a \in [K]\}$ be a bandit instance. In the local DP interaction protocol, the BAI strategy $\pi$ only accesses the noisy rewards from the perturbation mechanism, i.e. $z_t \sim \mathcal{M}(r_t)$, where $r_t \sim \nu_{a_t}$. Thus, we can define an environment $\boldsymbol{\nu}^{\mathcal{M}} \triangleq \{\nu_a^{\mathcal{M}} : a \in [K]\}$ *induced* by the perturbation mechanism, where

$$\nu_a^{\mathcal{M}}(Z) = \int_{r\in\mathbb{R}} \mathcal{M}(Z \mid r)\, \mathrm{d}\nu_a(r) dr$$

is the marginal over the noisy rewards of arm $a$.

Thus, the local DP canonical model of the interaction between $(\mathcal{M}, \pi)$ and an environment $\boldsymbol{\nu}$ is equivalent to the "classical" canonical model between $\pi$ and the induced environment $\boldsymbol{\nu}^{\mathcal{M}}$.

### D.2 Expected Sample Complexity Lower Bound under $\epsilon$-local DP

**Theorem 9 1 (Sample complexity lower bound for $\epsilon$-local DP FC-BAI)** *Let $\delta \in (0,1)$ and $\epsilon > 0$. For any $\delta$-correct $\epsilon$-local DP pair $(\mathcal{M}, \pi)$ of perturbation mechanism and BAI strategy, we have that $\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta] \geq T^\star_\ell(\boldsymbol{\nu}; \epsilon) \log(1/(2.4\delta))$ with*

$$T^\star_\ell(\boldsymbol{\nu}; \epsilon)^{-1} \triangleq \sup_{\omega \in \Sigma_K} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\nu})} \sum_{a \in [K]} \omega_a \min \left\{ \mathrm{KL}\left(\nu_a \parallel \lambda_a\right), c(\epsilon) \left(\mathrm{TV}\left(\nu_a \parallel \lambda_a\right)\right)^2 \right\},$$

*where $c(\epsilon) = \min\{4, e^{2\epsilon}\}(e^\epsilon - 1)^2$ is a privacy term. For two probability distributions $\mathbb{P}, \mathbb{Q}$ on the measurable space $(\Omega, \mathcal{F})$, the TV divergence is $\mathrm{TV}(\mathbb{P} \parallel \mathbb{Q}) \triangleq \sup_{A \in \mathcal{F}} \{\mathbb{P}(A) - \mathbb{Q}(A)\}$.*

**Proof** Let $(\mathcal{M}, \pi)$ a perturbation mechanism and BAI strategy pair that it $\epsilon$-local DP.

We suppose that $\pi$ is $\delta$-correct.

Using the remark in the local DP canonical model, $\pi$ is $\delta$-correct with respect to the environment $\boldsymbol{\nu}^{\mathcal{M}} \triangleq \{\nu_a^{\mathcal{M}} : a \in [K]\}$ induced by the perturbation mechanism $\mathcal{M}$, where

$$\nu_a^{\mathcal{M}}(Z) = \int_{r \in \mathbb{R}} \mathcal{M}(Z \mid r) \, \mathrm{d}\nu_a(r) dr$$

is the marginal over the noisy rewards of arm $a$.

Thus using Lemma 1 in Kaufmann et al. (2016), we get that

$$\sum_{a=1}^{K} \mathbb{E}\left[N_a(\tau)\right] \mathrm{KL}\left(\nu_a^{\mathcal{M}} \parallel \lambda_a^{\mathcal{M}}\right) \geq \mathrm{kl}(1-\delta, \delta)$$

for any alternative environment $\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\nu})$.

Using Theorem 1 in Duchi et al. (2013), we have that

$$\begin{aligned}
\mathrm{KL}\left(\nu_a^{\mathcal{M}} \parallel \lambda_a^{\mathcal{M}}\right) &\leq \mathrm{KL}\left(\nu_a^{\mathcal{M}} \parallel \lambda_a^{\mathcal{M}}\right) + \mathrm{KL}\left(\lambda_a^{\mathcal{M}} \parallel \nu_a^{\mathcal{M}}\right) \\
&\leq \min\{4, e^{2\epsilon}\}(e^\epsilon - 1)^2 (\mathrm{TV}\left(\nu_a \parallel \lambda_a\right))^2 \\
&= c(\epsilon)(\mathrm{TV}\left(\nu_a \parallel \lambda_a\right))^2
\end{aligned}$$

where $c(\epsilon) \triangleq \min\{4, e^{2\epsilon}\}(e^\epsilon - 1)^2$.

On the other hand, using the data-processing inequality, we also have that

$$\mathrm{KL}\left(\nu_a^{\mathcal{M}} \parallel \lambda_a^{\mathcal{M}}\right) \leq \mathrm{KL}\left(\nu_a \parallel \lambda_a\right)$$

Thus, combining the two inequalities gives that

$$\begin{aligned}
\mathrm{kl}(1-\delta, \delta) &\leq \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\nu})} \sum_{a=1}^{K} \mathbb{E}\left[N_a(\tau)\right] \min\left\{\mathrm{KL}\left(\nu_a \parallel \lambda_a\right), c(\epsilon)(\mathrm{TV}\left(\nu_a \parallel \lambda_a\right))^2\right\} \\
&= \mathbb{E}[\tau] \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\nu})} \sum_{a=1}^{K} \frac{\mathbb{E}\left[N_a(\tau)\right]}{\mathbb{E}[\tau]} \min\left\{\mathrm{KL}\left(\nu_a \parallel \lambda_a\right), c(\epsilon)(\mathrm{TV}\left(\nu_a \parallel \lambda_a\right))^2\right\} \\
&\leq \mathbb{E}[\tau] \left( \sup_{\omega \in \Sigma_K} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\nu})} \sum_{a=1}^{K} \omega_a \min\left\{\mathrm{KL}\left(\nu_a \parallel \lambda_a\right), c(\epsilon)(\mathrm{TV}\left(\nu_a \parallel \lambda_a\right))^2\right\} \right)
\end{aligned}$$

$$= \mathbb{E}[\tau](T_\ell^\star(\boldsymbol{\nu}, \epsilon))^{-1}$$

The theorem follows by noting that for $\delta \in (0,1), \mathrm{kl}(1-\delta, \delta) \geq \log(1/2.4\delta)$. ■

### D.3 Expected Sample Complexity Lower Bound under $\epsilon$-global DP

#### D.3.1 Upper Bound on the KL as a Transport Problem

First, we recall the result that conditioning increases the KL.

**Theorem 25 (Conditioning Increases the KL)** *Let $P_X \overset{P_{Y|X}}{\longrightarrow} P_Y$ and $P_X \overset{Q_{Y|X}}{\longrightarrow} Q_Y$. Then*

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

We apply this result to our problem of upper-bounding the marginals of DP mechanisms.

**Theorem 13 2 (KL Upper Bound as a Transport Problem)** *If $\mathcal{M}$ is $\epsilon$-pure DP,*

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

*where $\Pi(\mathcal{P}_1, \mathcal{P}_2)$ is the set of all couplings between $\mathcal{P}_1$ and $\mathcal{P}_2$.*

**Proof** Let $\mathcal{C} \in \Pi(\mathcal{P}_1, \mathcal{P}_2)$ a coupling between $\mathcal{P}_1$ and $\mathcal{P}_2$.
Then, we re-write:

$$\begin{aligned} M_1(A) &\triangleq \int_{D \in \mathcal{X}^n} \mathcal{M}_D(A) \, \mathrm{d}\mathcal{P}_1(D) \\ &= \int_{D, D' \in \mathcal{X}^n} \mathcal{M}_D(A) \, \mathrm{d}\mathcal{C}\left(D, D'\right) \end{aligned}$$

and

$$\begin{aligned} M_2(A) &\triangleq \int_{D' \in \mathcal{X}^n} \mathcal{M}_{D'}(A) \, \mathrm{d}\mathcal{P}_2\left(D'\right) \\ &= \int_{D, D' \in \mathcal{X}^n} \mathcal{M}_{D'}(A) \, \mathrm{d}\mathcal{C}\left(D, D'\right) \end{aligned}$$

by the definition of the coupling.
Then, using Theorem 25, we get

$$\mathrm{KL}\left(M_1 \parallel M_2\right) \leq \mathbb{E}_{(D,D') \sim \mathcal{C}}\left[\mathrm{KL}\left(\mathcal{M}_D \parallel \mathcal{M}_{D'}\right)\right]$$

Finally, using group privacy, we have

$$\mathrm{KL}\left(\mathcal{M}_D \parallel \mathcal{M}_{D'}\right) \leq \epsilon d_{\mathrm{Ham}}(D, D')$$

Combining the last two inequalities gives

$$\mathrm{KL}\left(M_1 \parallel M_2\right) \leq \epsilon \, \mathbb{E}_{(D,D') \sim \mathcal{C}}[d_{\mathrm{Ham}}(D, D')]$$

And since this is true for any coupling $\mathcal{C}$, taking the infimum over couplings concludes the proof. ■

### D.3.2 Sequential KL Decomposition for Bandits under DP

Now, we adapt Theorem 14 for the bandit marginals. First, for simplicity, we start with the setting where $T$ the number of rounds in the bandit interaction is fixed.

Let $\nu = \{P_a, a \in [K]\}$ and $\nu' = \{P'_a, a \in [K]\}$ be two bandit instances. We recall that, when the policy $\pi$ interacts with the bandit instance $\nu$, it induces a marginal distribution $\mathbb{M}_{\nu\pi}$ over the sequence of actions, where

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

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

$$\mathbb{M}_{\nu\pi}(C) \triangleq \sum_{(a_1,\dots,a_T)\in C} m_{\nu\pi}(a_1, a_2, \dots, a_T).$$

We define $\mathbb{M}_{\nu'\pi}$ similarly.

The goal is to upper bound the quantity $\mathrm{KL}\left(\mathbb{M}_{\nu\pi} \parallel \mathbb{M}_{\nu'\pi}\right)$. The marginals $\mathbb{M}_{\nu\pi}$ and $\mathbb{M}_{\nu\pi}$ in the sequential setting "look like" marginals generated by "product distributions". However, the hardness of the sequential setting resides in the fact that the data-generating distributions depend on the actions chosen, which are stochastic. Thus, the results of the previous section cannot directly be applied. To adapt the proof ideas of the previous section to the bandit case, we introduce the idea of a coupled bandit instance.

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

Here, we build a coupled environment $\gamma$ of $\nu$ and $\nu'$. The policy $\pi$ interacts with the coupled environment $\gamma$ up to a given time horizon $T$ to produce an augmented history $\{(a_t, r_t, r'_t)\}_{t=1}^T$. The iterative steps of this interaction process are:

1. The probability of choosing an action $a_t = a$ at time $t$ is dictated only by the policy $\pi_t$ and $a_1, r_1, a_2, r_2, \dots, a_{t-1}, r_{t-1}$, *i.e.* the policy ignores $\{r'_s\}_{s=1}^{t-1}$.

2. The distribution of rewards $(r_t, r'_t)$ is $c_{a_t}$ and is conditionally independent of the previous observed history $\{(a_s, r_s, r'_s)\}_{t=1}^{t-1}$.

This interaction is similar to the interaction process of policy $\pi$ with the first bandit instance $\nu$, with the addition of sampling an extra $r'_t$ from the coupling of $P_{a_t}$ and $P'_{a_t}$. This, in essence, corresponds to the "up" branch in Theorem 25.

The distribution of the augmented history induced by the interaction of $\pi$ and the coupled environment can be defined as

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

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

With the new notation

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

Similarly, we define

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

which corresponds to the "down" branch in Theorem 25, where the policy ignores the rewards $r_1, \dots, r_T$ in the interaction.

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

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

and

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

By the data-processing inequality, we get that

$$\mathrm{KL}\left(\mathbb{M}_{\nu\pi} \parallel \mathbb{M}_{\nu'\pi}\right) \leq \mathrm{KL}\left(p_{\gamma\pi} \parallel q_{\gamma\pi}\right)$$

We have that

$$\begin{aligned}
\mathrm{KL}\left(p_{\gamma\pi} \parallel q_{\gamma\pi}\right) &\overset{(a)}{=} \mathbb{E}_{\mathbf{h} \triangleq (\mathbf{a},\mathbf{r},\mathbf{r'}) \sim p_{\gamma\pi}} \left[ \log \left( \frac{\pi(\mathbf{a} \mid \mathbf{r}) c_{\mathbf{a}}(\mathbf{r}, \mathbf{r'})}{\pi(\mathbf{a} \mid \mathbf{r'}) c_{\mathbf{a}}(\mathbf{r}, \mathbf{r'})} \right) \right] \\
&= \mathbb{E}_{\mathbf{h} \triangleq (\mathbf{a},\mathbf{r},\mathbf{r'}) \sim p_{\gamma\pi}} \left[ \log \left( \frac{\pi(\mathbf{a} \mid \mathbf{r}}{\pi(\mathbf{a} \mid \mathbf{r'}} \right) \right]
\end{aligned}$$

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

**Global DP.** Now, if the policy $\pi$ is $\epsilon$-global DP, then by group privacy $\pi(\mathbf{a} \mid \mathbf{r}) \leq e^{\epsilon d_{\mathrm{Ham}}(\mathbf{r},\mathbf{r'})} \pi(\mathbf{a} \mid \mathbf{r'})$ for any sequence of actions, and any two sequence of rewards $\mathbf{r}$ and $\mathbf{r'}$. Thus, computing the expectation of $d_{\mathrm{Ham}}(\mathbf{r}, \mathbf{r'})$ when $\mathbf{r}$ and $\mathbf{r'}$ are generated through the coupled environment provides the following theorem.

**Theorem 26 (KL Decomposition for $\epsilon$-global DP)** *If $\pi$ is $\epsilon$-global DP, then*

$$\mathrm{KL}\left(\mathbb{M}_{\nu\pi} \parallel \mathbb{M}_{\nu'\pi}\right) \leq \epsilon \, \mathbb{E}_{\nu\pi} \left( \sum_{t=1}^{T} t_{a_t} \right),$$

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

**Proof** The proof follows by computing

$$\begin{aligned}
\mathbb{E}_{\mathbf{h} \triangleq (\mathbf{a},\mathbf{r},\mathbf{r'}) \sim p_{\gamma\pi}} \left[ \log \left( \frac{\mathcal{V}^{\pi}_{\mathbf{r}}(\mathbf{a})}{\mathcal{V}_{\mathbf{r'}}(\mathbf{a})} \right) \right] &\overset{(a)}{\leq} \mathbb{E}_{\mathbf{h} \triangleq (\mathbf{a},\mathbf{r},\mathbf{r'}) \sim p_{\gamma\pi}} \left[ \epsilon d_{\mathrm{Ham}}(\mathbf{r}, \mathbf{r'}) \right] \\
&\overset{(b)}{=} \epsilon \sum_{t=1}^{T} \mathbb{E}_{\mathbf{h} \triangleq (\mathbf{a},\mathbf{r},\mathbf{r'}) \sim p_{\gamma\pi}} \left[ \mathbb{1} \left\{ r_t \neq r'_t \right\} \right]
\end{aligned}$$

$$\overset{(c)}{=} \epsilon \sum_{t=1}^{T} \mathbb{E}_{\mathbf{h}\triangleq(\mathbf{a},\mathbf{r},\mathbf{r'})\sim p_{\gamma\pi}} \left[ \mathbb{E}_{\mathbf{h}\triangleq(\mathbf{a},\mathbf{r},\mathbf{r'})\sim p_{\gamma\pi}} \left[ \mathbb{1}\left\{ r_t \neq r'_t \right\} | a_t \right] \right]$$
$$\overset{(d)}{=} \epsilon \sum_{t=1}^{T} \mathbb{E}_{\mathbf{h}\triangleq(\mathbf{a},\mathbf{r},\mathbf{r'})\sim p_{\gamma\pi}} \left[ t_{a_t} \right]$$
$$\overset{(e)}{=} \epsilon \sum_{t=1}^{T} \mathbb{E}_{\mathbf{a}\sim m_{\nu\pi}} \left[ t_{a_t} \right]$$

where: (a) is by group privacy, (b) is by the definition of the hamming distance, (c) is by the towering property of the expectation, (d) is by the definition of the maximal coupling and (e) is because the sum only depends on the sequence of actions, with marginal distribution $m_{\nu\pi}$. ■

**Comparison to the product distribution setting:** The result of Theorem 26 generalises the result of Theorem 14 to the sequential setting under pure DP. Since the actions are stochastic, there is an additional expectation over the generation process of the sequence of actions, sampled from $\mathbb{M}_{\nu\pi}$. Also, Theorem 26 can be seen as an $\epsilon$-DP version of the general KL decomposition lemma (Exercice 15.8, (b) in Lattimore and Szepesvári (2020)), which recall states that $\mathrm{KL}\left(\mathbb{P}_{\nu\pi} \parallel \mathbb{P}_{\nu'\pi}\right) = \mathbb{E}_{\nu\pi}\left(\sum_{t=1}^{T} \mathrm{KL}\left(P_{a_t} \parallel P'_{a_t}\right)\right)$.

**Remark 27 (Improvement of a factor of 6 for Azize and Basu 2022)** *Theorem of Azize and Basu (2022) states*

$$\mathrm{KL}\left(\mathbb{M}_{\nu\pi} \parallel \mathbb{M}_{\nu'\pi}\right) \leq 6\epsilon\, \mathbb{E}_{\nu\pi}\left(\sum_{t=1}^{T} t_{a_t}\right).$$

*We used a generalisation of Karwa Vadhan lemma to prove this result for product distributions. Using the coupled environment idea in the new proof, we eliminate the extra 6 factor in the upper bound. This improves all the regret and sample complexity lower bounds in this manuscript by a factor of 6 compared to the results in Azize and Basu (2022); Azize et al. (2023).*

**Remark 28 (Stopping time version of the KL decomposition under DP)** *Let $\pi$ be a DP BAI strategy. Let $\nu$ and $\lambda$ be two bandit instances. Denote by $\mathbb{M}_{\boldsymbol{\nu},\pi}$ the marginal distribution of $(\underline{A}, \widehat{A}, \tau)$ , i.e. the sequence of action, final recommendation and stopping time when the BAI strategy $\pi$ interacts with $\nu$. By adapting the proof technique seen before for the canonical bandit setting under FC-BAI, we get that*

$$\mathrm{KL}\left(\mathbb{M}_{\nu\pi} \parallel \mathbb{M}_{\lambda\pi}\right) \leq \epsilon \sum_{a=1}^{K} \mathbb{E}_{\boldsymbol{\nu},\pi}\left[N_a(\tau)\right] \mathrm{TV}\left(P_a \parallel P'_a\right)$$

*where $\tau$ is the stopping time, $t_{a_t} \triangleq \mathrm{TV}\left(P_{a_t} \parallel P'_{a_t}\right)$ and $\mathbb{E}_{\nu\pi}$ is the expectation under $m_{\nu\pi}$.*

**Proof** For completeness, we recall the canonical bandit setting under FC-BAI, using the notation of Remark 28, and then adapt the coupling technique to FC-BAI. The main difference with the previous section is that the number of interactions $\tau$ is now a random variable, and we use Wald's lemma to deal with that.

**Step 1: Canonical Model under FC-BAI:** Let $\boldsymbol{\nu} \triangleq \{P_a : a \in [K]\}$ be a bandit instance, consisting of $K$ arms with finite means $\{\mu_a\}_{a\in[K]}$. Now, we recall the interaction between a BAI strategy $\pi$ and the bandit instance $\nu$ in the Protocol 3. The BAI strategy $\pi$ halts at $\tau$, samples a sequence of actions $\underline{A}^\tau \triangleq (a_1, \dots, a_\tau)$, and recommends the action $\hat{A}$. Let $\mathbb{M}_{\boldsymbol{\nu},\pi}$ be the marginal distribution over the output $(\tau, \underline{A}^\tau, \hat{A})$, when the BAI strategy $\pi$ interacts with the bandit instance $\nu$. Then,

$$m_{\boldsymbol{\nu},\pi}(\tau = T, \underline{A}^\tau = \underline{a}^T, \hat{A} = \hat{a}) = \int_{\underline{r}^T=(r_1,\dots,r_T)\in\mathbb{R}^T} \pi(\underline{a}^T, \hat{a}, T \mid \underline{r}^T) \prod_{t=1}^{T} p_{a_t}(r_t) dr_t$$

where

$$\pi(\underline{a}^T, \widehat{a}, T \mid \underline{r}^T) \triangleq \mathrm{Rec}_{T+1}\left(\widehat{a} \mid \mathcal{H}_T\right) \mathrm{S}_{T+1}\left(\top \mid \mathcal{H}_T\right) \prod_{t=1}^{T} \mathrm{S}_t\left(a_t \mid \mathcal{H}_{t-1}\right)$$

and $\mathcal{H}_t = (a_1, r_1, \dots, a_t, r_t)$.

**Step 2: Coupling technique applied to FC-BAI marginals.** Let $\nu = \{P_a : a \in [K]\}$ and $\lambda = \{P'_a : a \in [K]\}$ be two bandit instances. Define $c_a$ as the maximal coupling between $P_a$ and $P'_a$. Here, we build again a coupled environment $\gamma$ of $\nu$ and $\lambda$. The BAI strategy $\pi$ interacts with the coupled environment $\gamma$, decides to halt at step $\tau$ to produce an augmented history $(\tau, a_1, r_1, r'_1, \dots, a_\tau, r_\tau, r'_\tau, \hat{a})$. 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 sampling rule $S_t$ and $\mathcal{H}_{t-1} \triangleq (a_1, r_1, a_2, r_2, \dots, a_{t-1}, r_{t-1})$, *i.e.* the sampling rule ignores $\{r'_s\}_{s=1}^{t-1}$.

2. The distribution of rewards $(r_t, r'_t)$ is $c_{a_t}$ and is conditionally independent of the previous observed history $\{(a_s, r_s, r'_s)\}_{t=1}^{t-1}$.

3. The stopping rule and recommendation rule only depend on the history $\mathcal{H}_{t-1}$, the stopping and recommendation rules ignore $\{r'_s\}_{s=1}^{t-1}$

The distribution of the augmented history induced by the interaction of $\pi$ and the coupled environment can be defined as

$$\begin{aligned} p_{\gamma\pi}(\tau, a_1, r_1, r'_1 \dots, a_\tau, r_\tau, r'_\tau, \hat{a}) &\triangleq \mathrm{Rec}_{\tau+1}\left(\widehat{a} \mid \mathcal{H}_\tau\right) \mathrm{S}_{\tau+1}\left(\top \mid \mathcal{H}_\tau\right) \prod_{t=1}^{\tau} \mathrm{S}_t\left(a_t \mid \mathcal{H}_{t-1}\right) c_{a_t}(r_t, r'_t) \\ &= \pi(\underline{a}^\tau, \hat{a}, \tau \mid \underline{r}^\tau) \prod_{t=1}^{\tau} c_{a_t}(r_t, r'_t) \end{aligned}$$

where $\tau \in \mathbb{N}$, $a_t \in [K]$ and $r_t \in \mathbb{R}$ are fixed.

To simplify the notation, let $\underline{r}^\tau \triangleq (r_1, \dots, r_\tau)$, $\underline{r'}^\tau \triangleq (r'_1, \dots, r'_\tau)$ and $c_{\underline{a}^\tau}(\underline{r}^\tau, \underline{r'}^\tau) \triangleq \prod_{t=1}^\tau c_{a_t}(r_t, r'_t)$. We put $\mathbf{h} \triangleq (\tau, \underline{a}^\tau, \underline{r}^\tau, \underline{r'}^\tau, \hat{a})$.

With the new notation

$$p_{\gamma\pi}(\tau, \underline{a}^\tau, \underline{r}^\tau, \underline{r'}^\tau, \hat{a}) \triangleq \pi(\underline{a}^\tau, \hat{a}, \tau \mid \underline{r}^\tau) c_{\underline{a}^\tau}(\underline{r}^\tau, \underline{r'}^\tau)$$

Similarly, we define

$$q_{\gamma\pi}(\tau, \underline{a}^\tau, \underline{r}^\tau, \underline{r'}^\tau, \hat{a}) \triangleq \pi(\underline{a}^\tau, \hat{a}, \tau \mid \underline{r'}^\tau) c_{\underline{a}^\tau}(\underline{r}^\tau, \underline{r'}^\tau)$$

which corresponds to the augmented history interaction, where the policy ignores the rewards $(r_1, \dots, r_t, \dots)$ in the interaction.

It follows that $m_{\nu\pi}$ and $m_{\lambda\pi}$ are the marginals of $p_{\gamma\pi}$ and $q_{\gamma\pi}$ when integrated over the rewards, *i.e.*

$$m_{\nu\pi}(\tau, \underline{a}^\tau, \hat{a}) = \int_{\underline{r}^\tau, \underline{r'}^\tau} p_{\gamma\pi}(\underline{r}^\tau, \underline{r'}^\tau) \prod_{t=1}^\tau \mathrm{d}r_t \, \mathrm{d}r'_t$$

and

$$m_{\lambda\pi}(\tau, \underline{a}^\tau, \hat{a}) = \int_{\underline{r}^\tau, \underline{r'}^\tau} q_{\gamma\pi}(\underline{r}^\tau, \underline{r'}^\tau) \prod_{t=1}^\tau \mathrm{d}r_t \, \mathrm{d}r'_t$$

By the data-processing inequality, we get that

$$\mathrm{KL}\left(\mathbb{M}_{\nu\pi} \parallel \mathbb{M}_{\lambda\pi}\right) \leq \mathrm{KL}\left(p_{\gamma\pi} \parallel q_{\gamma\pi}\right)$$

On the other hand, we have

$$\begin{aligned}
\mathrm{KL}\left(p_{\gamma\pi} \parallel q_{\gamma\pi}\right) &\overset{(a)}{=} \mathbb{E}_{\mathbf{h} \triangleq (\tau, \underline{a}^\tau, \underline{r}^\tau, \underline{r'}^\tau, \hat{a}) \sim p_{\gamma\pi}} \left[ \log \left( \frac{\pi(\underline{a}^\tau, \hat{a}, \tau \mid \underline{r}^\tau) c_{\underline{a}^\tau}(\underline{r}^\tau, \underline{r'}^\tau)}{\pi(\underline{a}^\tau, \hat{a}, \tau \mid \underline{r'}^\tau) c_{\underline{a}^\tau}(\underline{r}^\tau, \underline{r'}^\tau)} \right) \right] \\
&= \mathbb{E}_{\mathbf{h} \triangleq (\tau, \underline{a}^\tau, \underline{r}^\tau, \underline{r'}^\tau, \hat{a}) \sim p_{\gamma\pi}} \left[ \log \left( \frac{\pi(\underline{a}^\tau, \hat{a}, \tau \mid \underline{r}^\tau)}{\pi(\underline{a}^\tau, \hat{a}, \tau \mid \underline{r'}^\tau)} \right) \right] \\
&\overset{(b)}{\leq} \mathbb{E}_{\mathbf{h} \triangleq (\tau, \underline{a}^\tau, \underline{r}^\tau, \underline{r'}^\tau, \hat{a}) \sim p_{\gamma\pi}} \left[ \epsilon d_{\mathrm{Ham}}(\underline{r}^\tau, \underline{r'}^\tau) \right] \\
&\overset{(c)}{=} \epsilon \, \mathbb{E}_{\mathbf{h} \triangleq (\tau, \underline{a}^\tau, \underline{r}^\tau, \underline{r'}^\tau, \hat{a}) \sim p_{\gamma\pi}} \left[ \sum_{t=1}^\tau \mathbb{1}\left\{ r_t \neq r'_t \right\} \right] \\
&= \epsilon \, \mathbb{E}_{\mathbf{h} \triangleq (\tau, \underline{a}^\tau, \underline{r}^\tau, \underline{r'}^\tau, \hat{a}) \sim p_{\gamma\pi}} \left[ \sum_{a=1}^K \sum_{s=1}^{N_a(\tau)} \mathbb{1}\left\{ r_{s,a} \neq r'_{s,a} \right\} \right] \\
&\overset{(d)}{=} \epsilon \sum_{a=1}^K \mathbb{E}_{\mathbf{h} \triangleq (\tau, \underline{a}^\tau, \underline{r}^\tau, \underline{r'}^\tau, \hat{a}) \sim p_{\gamma\pi}} [N_a(\tau)] \, \mathrm{TV}\left(P_a \parallel P'_a\right) \\
&\overset{(e)}{=} \epsilon \sum_{a=1}^K \mathbb{E}_{\boldsymbol{\nu}, \pi} \left[ N_a(\tau) \right] \mathrm{TV}\left(P_a \parallel P'_a\right)
\end{aligned}$$

where: (a): by definition of $p_{\gamma\pi}$, $q_{\gamma\pi}$ and the KL divergence (b) is by group privacy, (c) is by the definition of the hamming distance, (d) is using Wald's lemma and the definition

of the maximal coupling, (e) is because $N_a(\tau)$ does not depend on the sequence of rewards $(r'_t)$, when $\mathbf{h}$ is generated through $p_{\gamma,\pi}$.

■

#### D.3.3 Transportation Lemma under $\epsilon$-global DP: Proof of Lemma 15

**Lemma 15 3 (Transportation lemma under $\epsilon$-global DP)** *Let $\delta \in (0,1)$ and $\epsilon > 0$. Let $\boldsymbol{\nu}$ be a bandit instance and $\lambda \in \mathrm{Alt}(\boldsymbol{\nu})$. For any $\delta$-correct $\epsilon$-global DP BAI strategy, we have that*

$$\epsilon \sum_{a=1}^{K} \mathbb{E}_{\boldsymbol{\nu},\pi}\left[N_a(\tau)\right] \mathrm{TV}\left(\nu_a \parallel \lambda_a\right) \geq \mathrm{kl}(1-\delta,\delta),$$

*where* $\mathrm{kl}(x,y) \triangleq x \log \frac{x}{y} + (1-x)\log\frac{1-x}{1-y}$ *for* $x, y \in (0,1)$.

**Proof** *Step 1: Distinguishability due to $\delta$-correctness.* Let $\pi$ be a $\delta$-correct $\epsilon$-global DP BAI strategy. Let $\boldsymbol{\nu}$ be a bandit instance and $\lambda \in \mathrm{Alt}(\boldsymbol{\nu})$.

Let $\mathbb{M}_{\boldsymbol{\nu},\pi}$ denote the probability distribution of $(\underline{A}, \widehat{A}, \tau)$ when the BAI strategy $\pi$ interacts with $\nu$. For any alternative instance $\lambda \in \mathrm{Alt}(\boldsymbol{\nu})$, the data-processing inequality gives that

$$\begin{aligned}\mathrm{KL}\left(\mathbb{M}_{\boldsymbol{\nu},\pi} \parallel \mathbb{M}_{\boldsymbol{\lambda},\pi}\right) &\geq \mathrm{kl}\left(\mathbb{M}_{\boldsymbol{\nu},\pi}\left(\widehat{A} = a^\star(\boldsymbol{\nu})\right), \mathbb{M}_{\boldsymbol{\lambda},\pi}\left(\widehat{A} = a^\star(\boldsymbol{\nu})\right)\right) \\ &\geq \mathrm{kl}(1-\delta,\delta). \end{aligned} \tag{16}$$

where the second inequality is because $\pi$ is $\delta$-correct i.e. $\mathbb{M}_{\boldsymbol{\nu},\pi}\left(\widehat{A} = a^\star(\boldsymbol{\nu})\right) \geq 1-\delta$ and $\mathbb{M}_{\boldsymbol{\lambda},\pi}\left(\widehat{A} = a^\star(\boldsymbol{\nu})\right) \leq \delta$, and the monotonicity of the kl.

*Step 2: Using the KL decomposition under global DP.* By Remark 28, we have

$$\mathrm{KL}\left(\mathbb{M}_{\boldsymbol{\nu},\pi} \parallel \mathbb{M}_{\boldsymbol{\lambda},\pi}\right) \leq \epsilon \sum_{a=1}^{K} \mathbb{E}_{\boldsymbol{\nu},\pi}\left[N_a(\tau)\right] \mathrm{TV}\left(\nu_a \parallel \lambda_a\right). \tag{17}$$

Combining Inequalities 16 and 17 concludes the proof. ■

#### D.3.4 Proof of Theorem 16

**Theorem 16 4** *Let $\delta \in (0,1)$ and $\epsilon > 0$. For any $\delta$-correct and $\epsilon$-global DP FC-BAI algorithm, we have that $\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta] \geq T^\star_g(\boldsymbol{\nu};\epsilon)\log(1/(2.4\delta))$ with*

$$T^\star_g(\boldsymbol{\nu};\epsilon)^{-1} \triangleq \sup_{\omega \in \Sigma_K} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\nu})} \min\left\{ \sum_{a\in[K]} \omega_a \mathrm{KL}\left(\nu_a \parallel \lambda_a\right), \epsilon \sum_{a\in[K]} \omega_a \mathrm{TV}\left(\nu_a \parallel \lambda_a\right) \right\}.$$

**Proof** Let $\pi$ be a $\delta$-correct $\epsilon$-global DP BAI strategy. Let $\boldsymbol{\nu}$ be a bandit instance and $\lambda \in \mathrm{Alt}(\boldsymbol{\nu})$.

Let $\mathbb{E}$ denote the expectation under $\mathbb{P}_{\boldsymbol{\nu},\pi}$, ie $\mathbb{E} \triangleq \mathbb{E}_{\boldsymbol{\nu},\pi}$.

By Lemma 15, we have that $\epsilon \sum_{a=1}^{K} \mathbb{E}\left[N_a(\tau)\right] \mathrm{TV}\left(\nu_a \parallel \lambda_a\right) \geq \mathrm{kl}(1-\delta, \delta)$.

Lemma 1 from Kaufmann et al. (2016) gives that $\sum_{a=1}^{K} \mathbb{E}\left[N_a(\tau)\right] \mathrm{KL}\left(\nu_a \parallel \lambda_a\right) \geq \mathrm{kl}(1-\delta, \delta)$.

Since these two inequalities hold for all $\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\nu})$, we get

$$\begin{aligned}
\mathrm{kl}(1-\delta, \delta) &\leq \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\nu})} \min \left( \epsilon \sum_{a=1}^{K} \mathbb{E}\left[N_a(\tau)\right] \mathrm{TV}\left(\nu_a \parallel \lambda_a\right), \sum_{a=1}^{K} \mathbb{E}\left[N_a(\tau)\right] \mathrm{KL}\left(\nu_a \parallel \lambda_a\right) \right) \\
&\overset{(a)}{=} \mathbb{E}[\tau] \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\nu})} \min \left( \epsilon \sum_{a=1}^{K} \frac{\mathbb{E}\left[N_a(\tau)\right]}{\mathbb{E}[\tau]} \mathrm{TV}\left(\nu_a \parallel \lambda_a\right), \sum_{a=1}^{K} \frac{\mathbb{E}\left[N_a(\tau)\right]}{\mathbb{E}[\tau]} \mathrm{KL}\left(\nu_a \parallel \lambda_a\right) \right) \\
&\overset{(b)}{\leq} \mathbb{E}[\tau] \left( \sup_{\omega \in \Sigma_K} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\nu})} \min \left( \epsilon \sum_{a=1}^{K} \omega_a \mathrm{TV}\left(\nu_a \parallel \lambda_a\right), \sum_{a=1}^{K} \omega_a \mathrm{KL}\left(\nu_a \parallel \lambda_a\right) \right) \right) .
\end{aligned}$$

(a) is due to the fact that $\mathbb{E}[\tau]$ does not depend on $\boldsymbol{\lambda}$. (b) is obtained by noting that the vector $(\omega_a)_{a \in [K]} \triangleq \left( \frac{\mathbb{E}_{\nu,\pi}[N_a(\tau)]}{\mathbb{E}_{\nu,\pi}[\tau]} \right)_{a \in [K]}$ belongs to the simplex $\Sigma_K$.

The theorem follows by noting that for $\delta \in (0,1), \mathrm{kl}(1-\delta, \delta) \geq \log(1/(2.4\delta))$. ■

### D.3.5 TV Characteristic Time for Bernoulli Instances: Proof of Corollary 17

**Proposition 29 (TV characteristic time for Bernoulli instances)** *Let $\nu$ be a bandit instance, i.e. such that $\nu_a = \mathit{Bernoulli}(\mu_a)$ and $\mu_1 > \mu_2 \geq \cdots \geq \mu_K$. Let $\Delta_a \triangleq \mu_1 - \mu_a$ and $\Delta_{min} \triangleq \min_{a \neq 1} \Delta_a$. We have that*

$$T^{\star}_{\mathrm{TV}}(\boldsymbol{\nu}) = \frac{1}{\Delta_{min}} + \sum_{a=2}^{K} \frac{1}{\Delta_a}, \qquad \textit{and} \qquad \frac{1}{\Delta_{min}} \leq T^{\star}_{\mathrm{TV}}(\boldsymbol{\nu}) \leq \frac{K}{\Delta_{min}}.$$

**Proof** *Step 1:* Let $\nu$ be a bandit instance, i.e. such that $\nu_a \triangleq \mathrm{Bernoulli}(\mu_a)$ and $\mu_1 > \mu_2 \geq \cdots \geq \mu_K$.

For the alternative bandit instance $\boldsymbol{\lambda}$, we refer to the mean of arm $a$ as $\rho_a$, i.e. $\lambda_a \triangleq \mathrm{Bernoulli}(\rho_a)$.

By the definition of $T^{\star}_{\mathrm{TV}}$, we have that

$$\begin{aligned}
\left(T^{\star}_{\mathrm{TV}}(\boldsymbol{\nu})\right)^{-1} &= \sup_{\omega \in \Sigma_K} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\nu})} \sum_{a=1}^{K} \omega_a \mathrm{TV}\left(\nu_a \parallel \lambda_a\right) \\
&\overset{(a)}{=} \sup_{\omega \in \Sigma_K} \min_{a \neq 1} \inf_{\boldsymbol{\lambda} : \rho_a > \rho_1} \omega_1 \left|\mu_1 - \rho_1\right| + \omega_a \left|\mu_a - \rho_a\right| \\
&\overset{(b)}{=} \sup_{\omega \in \Sigma_K} \min_{a \neq 1} \min(\omega_1, \omega_a) \Delta_a \\
&\overset{(c)}{=} \sup_{\omega \in \Sigma_K} \omega_1 \min_{a \neq 1} \min(1, \frac{\omega_a}{\omega_1}) \Delta_a
\end{aligned}$$

$$\overset{(d)}{=} \sup_{(x_2,\dots,x_K)\in(\mathbb{R}^+)^{K-1}} \frac{\min_{a\neq 1} g_a(x_a)}{1+x_2+\cdots+x_K}\,,$$

where $g_a(x_a) \triangleq \min(1, x_a)\Delta_a$.

Equality (a) is obtained due to the fact that $\mathrm{Alt}(\boldsymbol{\nu}) = \bigcup_{a\neq 1}\{\boldsymbol{\lambda} : \rho_a > \rho_1\}$, and for Bernoullis, $\mathrm{TV}(\nu_a \parallel \lambda_a) = |\mu_a - \rho_a|$.

Equality (b) is true, since $\inf_{\boldsymbol{\lambda}:\rho_a>\rho_1} \omega_1 |\mu_1 - \rho_1| + \omega_a |\mu_a - \rho_a| = \min(\omega_1, \omega_a)\Delta_a$.

Equality (c) holds true, since $\omega_1 \neq 1$ (if $\omega_1 = 0$, the value of the objective is 0).

Equality (d) is obtained by the change of variable $x_a \triangleq \frac{\omega_a}{\omega_1}$

*Step 2:* Let $(x_2, \dots, x_K) \in (\mathbb{R}^+)^{K-1}$. By the definition of $g_a$, we have that

$$g_a(x_a) \leq x_a\Delta_a \quad \text{and} \quad g_a(x_a) \leq \Delta_a.$$

This leads to the inequalities

$$\min_{a\neq 1} g_a(x_a) \leq g_a(x_a) \leq x_a\Delta_a \quad \text{and} \quad \min_{a\neq 1} g_a(x_a) \leq \Delta_{\min}.$$

Thus,

$$\begin{aligned}\left(\min_{a\neq 1} g_a(x_a)\right)\left(\frac{1}{\Delta_{\min}} + \sum_{a=2}^{K}\frac{1}{\Delta_a}\right) &= \frac{\min_{a\neq 1} g_a(x_a)}{\Delta_{\min}} + \sum_{a=2}^{K}\frac{\min_{a\neq 1} g_a(x_a)}{\Delta_a} \\ &\leq 1 + \sum_{a=2}^{K} x_a\,.\end{aligned}$$

This means that for every $(x_2, \dots, x_K) \in (\mathbb{R}^+)^{K-1}$,

$$\frac{\min_{a\neq 1} g_a(x_a)}{1+x_2+\cdots+x_K} \leq \frac{1}{\frac{1}{\Delta_{\min}} + \sum_{a=2}^{K}\frac{1}{\Delta_a}}.$$

Here, the upper bound is achievable for $x_a^\star = \frac{\Delta_{\min}}{\Delta_a}$, since $g_a(x_a^\star) = \Delta_{\min}$ for all $a \neq 1$. This concludes that

$$T^\star_{\mathrm{TV}}(\boldsymbol{\nu})^{-1} = \frac{1}{\frac{1}{\Delta_{\min}} + \sum_{a=2}^{K}\frac{1}{\Delta_a}} \qquad \Longrightarrow\ T^\star_{\mathrm{TV}}(\boldsymbol{\nu}) = \frac{1}{\Delta_{\min}} + \sum_{a=2}^{K}\frac{1}{\Delta_a}\,.$$

*Step 3:* The lower and upper bounds on $T^\star_{\mathrm{TV}}(\boldsymbol{\nu})$ follow from the fact that $\frac{1}{\Delta_a} \geq 0$ for all $a$, and $\frac{1}{\Delta_a} \leq \frac{1}{\Delta_{\min}}$ for all $a \neq 1$.

Hence, we conclude the proof. ■

#### D.3.6 On the Total Variation Distance and the Hardness of Privacy

Our lower bound suggests that the hardness of the DP-FC-BAI problem is characterized by $T^\star_{\mathrm{TV}}(\boldsymbol{\nu})$, which is a total variation counterpart of the classic KL-based characteristic time $T^\star_{\mathrm{KL}}(\boldsymbol{\nu})$ in FC-BAI Garivier and Kaufmann (2016). The total variation distance appears

to be the natural measure to quantify the hardness of privacy in other settings such as regret minimization Azize and Basu (2022), Karwa-Vadhan lemma Karwa and Vadhan (2018) and Differentially Private Assouad, Fano, and Le Cam Acharya et al. (2021). The high-level intuition is that: Pure DP can be seen as a multiplicative stability constraint of $e^\epsilon$ when one data point changes. With group privacy, if two data sets differ in $d_{ham}$ points, then one incurs a factor $e^{d_{ham}\ \epsilon}$. Now, by sampling $n$ i.i.d points from a distribution $P$ and $n$ i.i.d points from a distribution $Q$, the Karwa-Vadhan lemma states that the incurred factor is $e^{(nTV(P,Q))\ \epsilon}$. This is proved by building a maximal coupling, which is the coupling that minimizes the Hamming distance in expectation. In brief, *the total variation naturally appears in lower bounds since it is the quantity that characterises the hardness of the optimal transport problem minimizing the hamming distance*, i.e $\mathrm{TV}(P,Q) = \inf_{(X,Y)\sim(P,Q)} E(1_{X\neq Y})$. However, it is possible that the problem can be characterized by other f-divergences. Finally, one can always go from TV to KL using Pinsker's inequality, though that would always be less tight than the TV-based lower bound.

*On the relation between $T^\star_{\mathrm{TV}}(\boldsymbol{\nu})$ and $T^\star_{\mathrm{KL}}(\boldsymbol{\nu})$.* A direct application of Pinsker's inequality gives that $T^\star_{\mathrm{TV}}(\boldsymbol{\nu}) \geq \sqrt{2T^\star_{\mathrm{KL}}(\boldsymbol{\nu})}$. For completeness, we present here the exact calculations:

For every alternative mean parameter $\lambda$ and every arm $a$, using Pinkser's inequality, we have that $d_{TV}(\mu_a, \lambda_a) \leq \sqrt{\frac{1}{2} d_{KL}(\mu_a, \lambda_a)}$. Therefore, for every allocation over arms $\omega$, we have

$$\sum_a \omega_a d_{TV}(\mu_a, \lambda_a) \leq \sum_a \omega_a \sqrt{\frac{1}{2} d_{KL}(\mu_a, \lambda_a)} \leq \sqrt{\frac{1}{2} \sum_a \omega_a d_{KL}(\mu_a, \lambda_a)}\,.$$

Taking the supremum over the simplex and the infimum over the set of alternative mean parameters yields $T^\star_{\mathrm{TV}}(\boldsymbol{\nu})^{-1} \leq \sqrt{\frac{1}{2} T^\star_{\mathrm{KL}}(\boldsymbol{\nu})^{-1}}$. This concludes the proof.

## Appendix E. Privacy analysis

We prove that AdaP-TT and AdaP-TT$^\star$ satisfy $\epsilon$-global DP. We first provide the privacy lemma that justifies using doubling and forgetting. Using the privacy lemma and the post-processing property of DP, we conclude the privacy analysis of AdaP-TT and AdaP-TT$^\star$.

### E.1 Privacy Lemma for Non-overlapping Sequences

**Lemma 30 (Privacy of non-overlapping sequence of empirical means)** *Let $\mathcal{M}$ be a mechanism that takes a **set** as input and outputs the private empirical mean, i.e.,*

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

*Let $\ell < T$ and $t_1, \dots t_\ell, t_{\ell+1}$ be in $[1, T]$ such that $1 = t_1 < \dots < t_\ell < t_{\ell+1} - 1 = T$.*
*Let's define the following mechanism*

$$\mathcal{G} : \{r_1, \dots, r_T\} \to \bigotimes_{i=1}^{\ell} \mathcal{M}_{\{r_{t_i}, \dots, r_{t_{i+1}-1}\}} \tag{18}$$

*In other words, $\mathcal{G}$ is the mechanism we get by applying $\mathcal{M}$ to the non-overlapping partition of the sequence $\{r_1, \dots, r_T\}$ according to $t_1 < \cdots < t_\ell < t_{\ell+1}$, i.e.*

$$\begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_T \end{pmatrix} \xrightarrow{\mathcal{G}} \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_\ell \end{pmatrix}$$

*where $\mu_i \sim \mathcal{M}_{\{r_{t_i}, \dots, r_{t_{i+1}-1}\}}$.*

*For $r_t \in [0,1]$, the mechanism $\mathcal{G}$ is $\epsilon$-DP.*

**Proof** Let $r^T \triangleq (r_1, \dots, r_T)$ and $r'^T \triangleq (r'_1, \dots, r'_T)$ 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 $\ell'$ 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 $\mu \triangleq (\mu_1, \dots, \mu_\ell)$ a fixed sequence of outcomes. Then,

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

where the last inequality holds true because $\mathcal{M}$ satisfies $\epsilon$-DP following Theorem 2. ■

### E.2 Privacy Analysis of AdaP-TT and AdaP-TT$^\star$

**Theorem 31 (Privacy analysis)** *For rewards in $[0,1]$, AdaP-TT and AdaP-TT$^\star$ satisfy $\epsilon$-global DP.*

**Remark 32** *The following proof is valid for any BAI strategy that only uses the DAF($\epsilon$) to estimate the means.*

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

We also fix a sequence of sampled actions $\underline{a}^T = \{a_1, \dots, a_T\} \in [K]^T$ and a recommended action $\hat{a} \in K$.

Let $\pi$ be a BAI strategy that only uses DAF($\epsilon$) to estimate the means, i.e. either AdaP-TT or AdaP-TT$^\star$.

We want to show that: $\pi(\underline{a}^T, \widehat{a}, T \mid \underline{\mathbf{d}}^T) \leq e^\epsilon \pi(\underline{a}^T, \widehat{a}, T \mid \underline{\mathbf{d}}'^T)$.

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 30.

*Step 1. Sequential decomposition of the output probability*

We observe that due to the sequential nature of the interaction, the output probability can be decomposed to a part that depends on $\underline{\mathbf{d}}^{j-1} \triangleq \{\mathbf{x}_1, \dots, \mathbf{x}_{j-1}\}$, which is identical for both $\underline{\mathbf{d}}^T$ and $\underline{\mathbf{d}}'^T$ and a second conditional part on the history.

Specifically, we have that

$$\pi(\underline{a}^T, \widehat{a}, T \mid \underline{\mathbf{d}}^T) \triangleq \mathrm{Rec}_{T+1}\left(\widehat{a} \mid \mathcal{H}_T\right) \mathrm{S}_{T+1}\left(\top \mid \mathcal{H}_T\right) \prod_{t=1}^{T} \mathrm{S}_t\left(a_t \mid \mathcal{H}_{t-1}\right)$$
$$\triangleq \mathcal{P}^{\pi}_{\underline{\mathbf{d}}^{j-1}}(\underline{a}^j) \mathcal{P}^{\pi}_{\underline{\mathbf{d}}}(a_{>j}, \hat{a}, T \mid \underline{a}^j)$$

where

- $a_{>j} \triangleq (a_{j+1}, \ldots, a_T)$
- $\mathcal{P}^{\pi}_{\underline{\mathbf{d}}^{j-1}}(\underline{a}^j) \triangleq \prod_{t=1}^{j} \mathrm{S}_t\left(a_t \mid \mathcal{H}_{t-1}\right)$
- $\mathcal{P}^{\pi}_{\underline{\mathbf{d}}}(a_{>j}, \hat{a}, T \mid \underline{a}^j) \triangleq \mathrm{Rec}_{T+1}\left(\widehat{a} \mid \mathcal{H}_T\right) \mathrm{S}_{T+1}\left(\top \mid \mathcal{H}_T\right) \prod_{t=j+1}^{T} \mathrm{S}_t\left(a_t \mid \mathcal{H}_{t-1}\right)$

Similarly

$$\pi(\underline{a}^T, \widehat{a}, T \mid \underline{\mathbf{d}'}^T) \triangleq \mathcal{P}^{\pi}_{\underline{\mathbf{d}}^{j-1}}(\underline{a}^j) \mathcal{P}^{\pi}_{\underline{\mathbf{d}'}}(a_{>j}, \hat{a}, T \mid \underline{a}^j)$$

since $\underline{\mathbf{d}'}^{j-1} = \underline{\mathbf{d}}^{j-1}$.

Which means that

$$\frac{\pi(\underline{a}^T, \widehat{a}, T \mid \underline{\mathbf{d}}^T)}{\pi(\underline{a}^T, \widehat{a}, T \mid \underline{\mathbf{d}'}^T)} = \frac{\mathcal{P}^{\pi}_{\underline{\mathbf{d}}}(a_{>j}, \hat{a}, T \mid \underline{a}^j)}{\mathcal{P}^{\pi}_{\underline{\mathbf{d}'}}(a_{>j}, \hat{a}, T \mid \underline{a}^j)} \tag{19}$$

*Step 2. The adaptive episodes are the same, before step $j$*

Let $\ell$ such that $t_\ell \leq j < t_{\ell+1}$ when $\pi$ interacts with $\underline{\mathbf{d}}^T$. Let us call it $\psi^{\pi}_{\underline{\mathbf{d}}^T}(j) \triangleq \ell$. Similarly, let $\ell'$ such that $t_{\ell'} \leq j < t_{\ell'+1}$ when $\pi$ interacts with $\underline{\mathbf{d}'}^T$. Let us call it $\psi^{\pi}_{\underline{\mathbf{d}'}^T}(j) \triangleq \ell'$.

Since $\psi^{\pi}_{\underline{\mathbf{d}}^T}(j)$ only depends on $\underline{\mathbf{d}}^{j-1}$, which is identical for $\underline{\mathbf{d}}^T$ and $\underline{\mathbf{d}'}^T$, we have that $\psi^{\pi}_{\underline{\mathbf{d}}^T}(j) = \psi^{\pi}_{\underline{\mathbf{d}'}^T}(j)$ with probability 1.

We call $\xi_j$ the last **time-step** of the episode $\psi^{\pi}_{\underline{\mathbf{d}}^T}(j)$, i.e $\xi_j \triangleq t_{\psi^{\pi}_{\underline{\mathbf{d}}^T}(j)+1} - 1$.

*Step 3. Private sufficient statistics*

Let $r_t \triangleq \underline{\mathbf{d}}^T_{t,a_t}$, be the reward corresponding to the action $a_t$ in the table $\underline{\mathbf{d}}^T$. Similarly, $r'_t \triangleq \underline{\mathbf{d}'}^T_{t,a_t}$ for $\underline{\mathbf{d}'}^T$.

Let us define $L_j \triangleq \mathcal{G}_{\{r_1,\ldots,r_{\xi_j}\}}$ and $L'_j \triangleq \mathcal{G}_{\{r'_1,\ldots,r'_{\xi_j}\}}$, where $\mathcal{G}$ is defined as in Eq. 18, using the same episodes for $d$ and $d'$. In other words, $L_j$ is the list of private empirical means computed on a non-overlapping sequence of rewards before step $\xi_j$.

Using the forgetting structure of $\pi$, there exists a randomised mapping $f_{\underline{\mathbf{d}}_{>\xi_j}}$ such that $\mathcal{P}^{\pi}_{\underline{\mathbf{d}}}(. \mid \underline{a}^j) = f_{\underline{\mathbf{d}}_{>\xi_j}} \circ L_j$ and $\mathcal{P}^{\pi}_{\underline{\mathbf{d}'}}(. \mid \underline{a}^j) = f_{\underline{\mathbf{d}}_{>\xi_j}} \circ L'_j$.

In other words, the interaction of $\pi$ with $\underline{\mathbf{d}}$ and $\underline{\mathbf{d}'}$ from step $\xi_j + 1$ until $T$ only depends on the sufficient statistics $L_j$, which summarises what happened before $\xi_j$, and the new inputs $\underline{\mathbf{d}}_{>\xi_j}$, which are the same for $\underline{\mathbf{d}}$ and $\underline{\mathbf{d}'}$.

*Step 4. Concluding with Lemma 30 and the post-processing lemma*

Since rewards are in $[0, 1]$, using Lemma 30, we have that $\mathcal{G}$ is $\epsilon$-DP.

Since $\mathcal{P}^{\pi}_{\underline{\mathbf{d}}}(. \mid \underline{a}^j)$ is just a post-processing of the output of $\mathcal{G}$, we have that

$$\frac{\mathcal{P}^{\pi}_{\underline{\mathbf{d}}}(a_{>j}, \hat{a}, T \mid \underline{a}^j)}{\mathcal{P}^{\pi}_{\underline{\mathbf{d}}'}(a_{>j}, \hat{a}, T \mid \underline{a}^j)} \leq e^{\epsilon} ,$$

and Eq. (19) concludes the proof. ■

## Appendix F. Globally Differentially Private GLR Stopping Rules

After studying the non-private GLR stopping rule with phases (Appendix F.1), we study the private GLR stopping rule with non-private transportation costs $W^G_{a,b}$ in Appendix F.2 (Lemma 20) and with adapted transportation costs $W^{G,\epsilon}_{a,b}$ in Appendix F.3 (Lemma 22).

### F.1 Non-private GLR Stopping Rule with Per-arm Phases

Before accounting for the privacy (i.e. Laplace noise), we first highlight the price of DAF($\epsilon$) for $\epsilon = +\infty$. This stopping condition is only evaluated at the beginning of each phase for each arm since it involves quantities that are fixed until we switch phase again, and it recommends $\hat{a}_n = \arg\max_{a\in[K]} \hat{\mu}_{k_{n,a},a}$ which is the best arm for the non-private empirical means. Lemma 33 yields a threshold function ensuring $\delta$-correctness.

**Lemma 33** *Let $\delta \in (0,1)$. Let $s > 1$, $\zeta$ be the Riemann $\zeta$ function, $c^G_{a,b}$ as in Eq. (4) and $k(x) = \log_2 x + 2$. Combining the DAF($\epsilon$) estimator for $\epsilon = +\infty$ with the GLR stopping rule with $W^G_{a,b}$ as in Eq. (3) and the stopping threshold $c^G_{a,b}(\omega, \delta(\zeta(s)^2 k(\omega_a)^s k(\omega_b)^s)^{-1})$ yields a $\delta$-correct algorithm for $\sigma$-sub-Gaussian distributions regardless of the sampling rule.*

**Proof** The non-private GLR stopping rule matches the one used for Gaussian bandits. Proving $\delta$-correctness of a GLR stopping rule is done by leveraging concentration results.

**Lemma 34 (Theorem 9 in Kaufmann and Koolen 2021)** *Let $\boldsymbol{\nu}$ be a sub-Gaussian bandit with means $\boldsymbol{\mu} \in \mathbb{R}^K$ and variance proxy $\sigma$. Let $S \subseteq [K]$ and $x > 0$.*

$$\mathbb{P}_{\boldsymbol{\nu}}\left(\exists n \in \mathbb{N}, \sum_{a\in S} \frac{N_{n,a}}{2\sigma^2}(\mu_{n,a} - \mu_a)^2 > \sum_{a\in S} 2\log\left(4 + \log\left(N_{n,a}\right)\right) + |S|\mathcal{C}_G\left(\frac{x}{|S|}\right)\right) \leq e^{-x} ,$$

*where $\mathcal{C}_G$ is defined in Kaufmann and Koolen (2021) as*

$$\mathcal{C}_G(x) \triangleq \min_{\lambda\in]1/2,1]} \frac{g_G(\lambda) + x}{\lambda} \text{ and } g_G(\lambda) \triangleq 2\lambda - 2\lambda\log(4\lambda) + \log\zeta(2\lambda) - \frac{1}{2}\log(1-\lambda) . \quad (20)$$

*Here, $\zeta$ is the Riemann $\zeta$ function and $\mathcal{C}_G(x) \approx x + \log(x)$.*

We consider the concentration event $\mathcal{E}^{(1)}_\delta = \bigcap_{a\neq a^\star} \bigcap_{n\in\mathbb{N}} \mathcal{E}^{(1)}_\delta(a,n)$ with $\mathcal{E}^{(1)}_\delta(a,n) =$

$$\left\{ \frac{\tilde{N}_{k_{n,a},a}}{2\sigma^2}(\hat{\mu}_{k_{n,a},a} - \mu_a)^2 + \frac{\tilde{N}_{k_{n,a^\star},a^\star}}{2\sigma^2}(\hat{\mu}_{k_{n,a^\star},a^\star} - \mu_{a^\star})^2 < c^G_{a,a^\star}(\tilde{N}_{k_n}, \frac{\delta}{\zeta(s)^2 k^s_{n,a} k^s_{n,a^\star}}) \right\} . \quad (21)$$

For all $a \neq a^\star$ and all $(k_a, k_{a^\star}) \in \mathbb{N}^2$, the estimators $(\hat{\mu}_{k_c,c})_{c\in\{a,a^\star\}}$ are based solely on the observations collected for arm $a$ (resp. arm $a^\star$) between times $n \in \{T_{k_a-1}(a), \cdots, T_{k_a}(a)-1\}$ (resp. $n \in \{T_{k_{a^\star}-1}(a^\star), \cdots, T_{k_{a^\star}}(a^\star)-1\}$) with local counts $(\tilde{N}_{k_c,c})_{c\in\{a,a^\star\}}$, i.e. dropping past observations. Using a direct union bound, we obtain that $\mathbb{P}_\nu((\mathcal{E}_\delta^{(1)})^\complement)$ is smaller than

$$\sum_{a\neq a^\star}\sum_{k_a,k_{a^\star}\in\mathbb{N}} \mathbb{P}_{\boldsymbol{\nu}}\left(\frac{\tilde{N}_{k_a,a}}{2\sigma^2}(\hat{\mu}_{k_a,a}-\mu_a)^2 + \frac{\tilde{N}_{k_{a^\star},a^\star}}{2\sigma^2}(\hat{\mu}_{k_{a^\star},a^\star}-\mu_{a^\star})^2 \geq c^G_{a,a^\star}(\tilde{N}_{k_n}, \frac{\delta}{\zeta(s)^2 k_a^s k_{a^\star}^s})\right)$$
$$\leq \frac{\delta}{K-1}\frac{1}{\zeta(s)^2}\sum_{a\neq a^\star}\sum_{(k_a,k_{a^\star})\in\mathbb{N}^2}\frac{1}{(k_a k_{a^\star})^s} = \delta\,.$$

where the last inequality uses Lemma 34 for all $a \neq a^\star$ and all $(k_a, k_{a^\star}) \in \mathbb{N}^2$. Therefore,

$$\mathbb{P}_\nu(\tau_\delta < +\infty,\, \hat{a}_{\tau_\delta} \neq a^\star) \leq \delta + \mathbb{P}_\nu(\mathcal{E}_\delta^{(1)} \cap \{\tau_\delta < +\infty,\, \hat{a}_{\tau_\delta} \neq a^\star\})\,.$$

Under $\mathcal{E}_\delta^{(1)} \cap \{\tau_\delta < +\infty,\, \hat{a}_{\tau_\delta} \neq a^\star\}$, we have $\hat{a}_{\tau_\delta} = \arg\max_{b\in[K]} \hat{\mu}_{k_{\tau_\delta,a},a} \neq a^\star$ and

$$\begin{aligned}
c^G_{\hat{a}_{\tau_\delta},a^\star}(\tilde{N}_{k_{\tau_\delta}}, \frac{\delta}{\zeta(s)^2 k_{\tau_\delta,a}^s k_{\tau_\delta,a^\star}^s}) &\leq \frac{(\hat{\mu}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}} - \hat{\mu}_{k_{\tau_\delta,a^\star},a^\star})^2}{2\sigma^2(1/\tilde{N}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}} + 1/\tilde{N}_{k_{\tau_\delta,a^\star},a^\star})} \\
&= \inf_{y\geq x}\left\{\frac{\tilde{N}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}}}{2\sigma^2}(\hat{\mu}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}} - x)^2 + \frac{\tilde{N}_{k_{\tau_\delta,a^\star},a^\star}}{2\sigma^2}(\hat{\mu}_{k_{\tau_\delta,a^\star},a^\star} - y)^2\right\} \\
&\leq \frac{\tilde{N}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}}}{2\sigma^2}(\hat{\mu}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}} - \mu_{\hat{a}_{\tau_\delta}})^2 + \frac{\tilde{N}_{k_{\tau_\delta,a^\star},a^\star}}{2\sigma^2}(\hat{\mu}_{k_{\tau_\delta,a^\star},a^\star} - \mu_{a^\star})^2 \\
&< c^G_{\hat{a}_{\tau_\delta},a^\star}(\tilde{N}_{k_{\tau_\delta}}, \frac{\delta}{\zeta(s)^2 k_{\tau_\delta,a}^s k_{\tau_\delta,a^\star}^s})\,.
\end{aligned}$$

This is a contradiction, hence $\mathcal{E}_\delta^{(1)} \cap \{\tau_\delta < +\infty,\, \hat{a}_{\tau_\delta} \neq a^\star\} = \emptyset$. This concludes the proof. ■

### F.2 Private GLR with Non-private Transportation Cost: Proof of Lemma 20

The proof of Lemma 20 is similar as the one detailed in Appendix F.1, with the added difficulty of controlling the Laplace noise. We consider the concentration event $\mathcal{E}_\delta = \mathcal{E}_{\delta/2}^{(1)} \cap \mathcal{E}_{\delta/2}^{(2)}$ where $\mathcal{E}_\delta^{(1)}$ as in Eq. (21) and $\mathcal{E}_\delta^{(2)} = \bigcap_{a\in[K]}\bigcap_{n\in\mathbb{N}} \mathcal{E}_\delta^{(2)}(a,n)$ with

$$\mathcal{E}_\delta^{(2)}(a,n) = \left\{\epsilon\tilde{N}_{k_{n,a},a}|Y_{k_{n,a},a}| < \log\left(\frac{K\zeta(s)k_{n,a}^s}{\delta}\right)\right\}\,. \tag{22}$$

Since $Y_{k_{n,a},a} \sim \mathrm{Lap}\left((\epsilon\tilde{N}_{k_{n,a},a})^{-1}\right)$, we have that $\tilde{N}_{k_{n,a},a}|Y_{k_{n,a},a}| \sim \mathcal{E}(\epsilon)$ for all $a \in [K]$ and all $n \in \mathbb{N}$, where $\mathcal{E}(\cdot)$ denotes the exponential distribution. Using concentration results for exponential distribution, a direct union bound yields that $\mathbb{P}_\nu((\mathcal{E}_\delta^{(2)})^\complement) \leq \delta$, hence

$$\mathbb{P}_\nu(\tau_\delta < +\infty,\, \hat{a}_{\tau_\delta} \neq a^\star) \leq \delta + \mathbb{P}_\nu(\mathcal{E}_\delta \cap \{\tau_\delta < +\infty,\, \hat{a}_{\tau_\delta} \neq a^\star\})\,.$$

Under $\mathcal{E}_\delta \cap \{\tau_\delta < +\infty,\, \hat{a}_{\tau_\delta} \neq a^\star\}$, we have $\hat{a}_{\tau_\delta} = \arg\max_{b\in[K]} \tilde{\mu}_{k_{\tau_\delta,a},a} \neq a^\star$ and

$$\begin{aligned}
c^{G,\epsilon}_{\hat{a}_{\tau_\delta},a^\star}(\tilde{N}_{k_{\tau_\delta}},\delta) &\leq \frac{\tilde{N}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}}}{2\sigma^2}(\tilde{\mu}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}} - \mu_{\hat{a}_{\tau_\delta}})^2 + \frac{\tilde{N}_{k_{\tau_\delta,a^\star},a^\star}}{2\sigma^2}(\tilde{\mu}_{k_{\tau_\delta,a^\star},a^\star} - \mu_{a^\star})^2 \\
&\leq \frac{\tilde{N}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}}}{\sigma^2}(\hat{\mu}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}} - \mu_{\hat{a}_{\tau_\delta}})^2 + \frac{\tilde{N}_{k_{\tau_\delta,a^\star},a^\star}}{\sigma^2}(\hat{\mu}_{k_{\tau_\delta,a^\star},a^\star} - \mu_{a^\star})^2 \\
&\quad + \frac{\tilde{N}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}}}{\sigma^2} Y^2_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}} + \frac{\tilde{N}_{k_{\tau_\delta,a^\star},a^\star}}{\sigma^2} Y^2_{k_{\tau_\delta,a^\star},a^\star} \\
&< 2c^{G}_{\hat{a}_{\tau_\delta},a^\star}(\tilde{N}_{k_{\tau_\delta}}, \frac{\delta}{2\zeta(s)^2 k^s_{\tau_\delta,a} k^s_{\tau_\delta,a^\star}}) + \frac{1}{\epsilon^2\sigma^2} \sum_{c\in\{a,a^\star\}} \frac{1}{\tilde{N}_{k_{\tau_\delta,c},c}} \left(\log \frac{2K\zeta(s)k^s_{\tau_\delta,c}}{\delta}\right)^2 .
\end{aligned}$$

where we used that $\tilde{\mu}_{k_{\tau_\delta,a},a} = \tilde{\mu}_{k_{\tau_\delta,a},a} + Y_{k_{\tau_\delta,a},a}$ and $(x-y)^2 \leq 2x^2 + 2y^2$. This is a contradiction, hence $\mathcal{E}_\delta \cap \{\tau_\delta < +\infty,\, \hat{a}_{\tau_\delta} \neq a^\star\} = \emptyset$. To conclude the proof, we take as a specific parameter $s = 2$, hence $\zeta(2) = \pi^2/6$.

### F.3 Private GLR with Adapted Transportation Cost: Proof of Lemma 22

The proof of Lemma 22 is similar as the one detailed in Appendix F.2. The main difference lies in the considered transportation costs, namely $W^{G,\epsilon}_{a,b}$ instead of $W^G_{a,b}$.

**Lemma 35 (Lemma 28 in Jourdan et al. 2023)** *Let $\delta \in (0,1)$. For all $x \geq 1$, let $\overline{W}_{-1}(x) = -W_{-1}(-e^{-x})$ (see Lemma 39), where $W_{-1}$ is the negative branch of the Lambert W function. Let $c(x,\delta) = \frac{1}{2}\overline{W}_{-1}\left(2\log(K/\delta) + 4\log(4+\log x) + 1/2\right)$. Consider $\sigma$-sub-Gaussian bandits with means $\mu \in \mathbb{R}^K$. Then,*

$$\mathbb{P}\left(\exists n\in\mathbb{N},\, \exists a \in [K],\, \frac{N_{n,a}}{2\sigma^2}(\mu_{n,a} - \mu_a)^2 > c(N_{n,a},\delta)\right) \leq \delta\,.$$

Recall that $h(\tilde{N}_{k_{n,a},a},\delta) = c(\tilde{N}_{k_{n,a},a}, \frac{\delta}{3\zeta(s)k^s_{n,a}})$. We use the concentration event $\mathcal{E}_\delta = \mathcal{E}^{(1)}_{\delta/3} \cap \mathcal{E}^{(2)}_{\delta/3} \cap \mathcal{E}^{(3)}_{\delta/3}$ where $\mathcal{E}^{(1)}_\delta$ as in Eq. (21), $\mathcal{E}^{(2)}_\delta$ as in Eq. (22) and $\mathcal{E}^{(3)}_\delta = \bigcap_{a\in[K]}\bigcap_{n\in\mathbb{N}} \mathcal{E}^{(3)}_\delta(a,n)$ with

$$\mathcal{E}^{(3)}_\delta(a,n) = \left\{ \frac{\tilde{N}_{k_{n,a},a}}{2\sigma^2}(\hat{\mu}_{k_{n,a},a} - \mu_a)^2 < h(\tilde{N}_{k_{n,a},a}, 3\delta) \right\}\,.$$

Using Lemma 35, a direct union bound yields that $\mathbb{P}_\nu((\mathcal{E}^{(3)}_\delta)^\complement) \leq \delta$, hence

$$\mathbb{P}_\nu(\tau_\delta < +\infty,\, \hat{a}_{\tau_\delta} \neq a^\star) \leq \delta + \mathbb{P}_\nu(\mathcal{E}_\delta \cap \{\tau_\delta < +\infty,\, \hat{a}_{\tau_\delta} \neq a^\star\})\,.$$

Under $\mathcal{E}_\delta \cap \{\tau_\delta < +\infty,\, \hat{a}_{\tau_\delta} \neq a^\star\}$, we have $\hat{a}_{\tau_\delta} = \arg\max_{b\in[K]} \tilde{\mu}_{k_{\tau_\delta,a},a} \neq a^\star$.
**Case 1.** Under $\mathcal{E}_\delta \cap \{\tau_\delta < +\infty,\, \hat{a}_{\tau_\delta} \neq a^\star,\, (\hat{\mu}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}} - \hat{\mu}_{k_{\tau_\delta,a^\star},a^\star})_+ < \epsilon/2\}$, we have

$$\frac{1}{2}c^{G,\epsilon}_{\hat{a}_{\tau_\delta},a^\star}(\tilde{N}_{k_{\tau_\delta}},\delta) + \frac{\sqrt{2}}{\epsilon\sigma} \sum_{c\in\{\hat{a}_{\tau_\delta},a^\star\}} \sqrt{\frac{h(\tilde{N}_{k_{\tau_\delta,c}},\delta)}{\tilde{N}_{k_{\tau_\delta,c}}}} \log\left(\frac{3K\zeta(s)k^s_{\tau_\delta,c}}{\delta}\right)$$

$$
\begin{aligned}
&\leq \frac{\tilde{N}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}}}{2\sigma^2}(\tilde{\mu}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}} - \mu_{\hat{a}_{\tau_\delta}})^2 + \frac{\tilde{N}_{k_{\tau_\delta,a^\star},a^\star}}{2\sigma^2}(\tilde{\mu}_{k_{\tau_\delta,a^\star},a^\star} - \mu_{a^\star})^2 \\
&= \frac{\tilde{N}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}}}{2\sigma^2}(\hat{\mu}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}} - \mu_{\hat{a}_{\tau_\delta}})^2 + \frac{\tilde{N}_{k_{\tau_\delta,a^\star},a^\star}}{2\sigma^2}(\hat{\mu}_{k_{\tau_\delta,a^\star},a^\star} - \mu_{a^\star})^2 \\
&\quad + \frac{\tilde{N}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}}}{2\sigma^2}Y^2_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}} + \frac{\tilde{N}_{k_{\tau_\delta,a^\star},a^\star}}{2\sigma^2}Y^2_{k_{\tau_\delta,a^\star},a^\star} \\
&\quad + \frac{\tilde{N}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}}}{\sigma^2}Y_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}}(\hat{\mu}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}} - \mu_{\hat{a}_{\tau_\delta}}) + \frac{\tilde{N}_{k_{\tau_\delta,a^\star},a^\star}}{\sigma^2}Y_{k_{\tau_\delta,a^\star},a^\star}(\hat{\mu}_{k_{\tau_\delta,a^\star},a^\star} - \mu_{a^\star}) \\
&< \frac{1}{2}c^{G,\epsilon}_{\hat{a}_{\tau_\delta},a^\star}(\tilde{N}_{k_{\tau_\delta}},\delta) + \frac{\sqrt{2}}{\epsilon\sigma}\sum_{c\in\{\hat{a}_{\tau_\delta},a^\star\}}\sqrt{\frac{h(\tilde{N}_{k_{\tau_\delta,c}},\delta)}{\tilde{N}_{k_{\tau_\delta,c}}}\log\left(\frac{3K\zeta(s)k^s_{\tau_\delta,c}}{\delta}\right)}\,.
\end{aligned}
$$

This is a contradiction, hence $\mathcal{E}_\delta \cap \{\tau_\delta < +\infty,\, \hat{a}_{\tau_\delta} \neq a^\star\,(\hat{\mu}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}} - \hat{\mu}_{k_{\tau_\delta,a^\star},a^\star})_+ < \epsilon/2\} = \emptyset$.

**Case 2.** Under $\mathcal{E}_\delta \cap \{\tau_\delta < +\infty,\, \hat{a}_{\tau_\delta} \neq a^\star,\, (\hat{\mu}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}} - \hat{\mu}_{k_{\tau_\delta,a^\star},a^\star})_+ \geq \epsilon/2\}$, we have

$$
\begin{aligned}
&\frac{1}{2\sigma^2}\log\left(\frac{3K\zeta(s)\max_{c\in\{\hat{a}_{\tau_\delta},a^\star\}}k_{\tau_\delta,c}}{\delta}\right) + \frac{\epsilon}{2\sqrt{2}\sigma}\sum_{c\in\{\hat{a}_{\tau_\delta},a^\star\}}\sqrt{\tilde{N}_{k_{\tau_\delta,c}}h(\tilde{N}_{k_{\tau_\delta,c}},\delta)} \\
&\quad \leq \frac{\epsilon(\tilde{\mu}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}} - \tilde{\mu}_{k_{\tau_\delta,a^\star},a^\star})}{4\sigma^2(1/\tilde{N}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}} + 1/\tilde{N}_{k_{\tau_\delta,a^\star},a^\star})} \\
&\quad \leq \frac{\epsilon}{4\sigma^2}\min\{\tilde{N}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}}, \tilde{N}_{k_{\tau_\delta,a^\star},a^\star})\}(\tilde{\mu}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}} - \tilde{\mu}_{k_{\tau_\delta,a^\star},a^\star}) \\
&\quad = \frac{\epsilon}{4\sigma^2}\inf_{y\geq x}\left\{\tilde{N}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}}|\tilde{\mu}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}} - x| + \tilde{N}_{k_{\tau_\delta,a^\star},a^\star}|\tilde{\mu}_{k_{\tau_\delta,a^\star},a^\star} - y|\right\} \\
&\quad \leq \frac{\epsilon}{4\sigma^2}\tilde{N}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}}|\tilde{\mu}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}} - \mu_{\hat{a}_{\tau_\delta}}| + \frac{\epsilon}{4\sigma^2}\tilde{N}_{k_{\tau_\delta,a^\star},a^\star}|\tilde{\mu}_{k_{\tau_\delta,a^\star},a^\star} - \mu_{a^\star}| \\
&\quad \leq \frac{\epsilon}{4\sigma^2}\tilde{N}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}}|\hat{\mu}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}} - \mu_{\hat{a}_{\tau_\delta}}| + \frac{\epsilon}{4\sigma^2}\tilde{N}_{k_{\tau_\delta,a^\star},a^\star}|\hat{\mu}_{k_{\tau_\delta,a^\star},a^\star} - \mu_{a^\star}| \\
&\qquad + \frac{\epsilon}{4\sigma^2}\tilde{N}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}}|Y_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}}| + \frac{\epsilon}{4\sigma^2}\tilde{N}_{k_{\tau_\delta,a^\star},a^\star}|Y_{k_{\tau_\delta,a^\star},a^\star}| \\
&\quad < \frac{1}{2\sigma^2}\log\left(\frac{3K\zeta(s)\max_{c\in\{\hat{a}_{\tau_\delta},a^\star\}}k_{\tau_\delta,c}}{\delta}\right) + \frac{\epsilon}{2\sqrt{2}\sigma}\sum_{c\in\{\hat{a}_{\tau_\delta},a^\star\}}\sqrt{\tilde{N}_{k_{\tau_\delta,c}}h(\tilde{N}_{k_{\tau_\delta,c}},\delta)}\,.
\end{aligned}
$$

This is a contradiction, hence $\mathcal{E}_\delta \cap \{\tau_\delta < +\infty,\, \hat{a}_{\tau_\delta} \neq a^\star\,(\hat{\mu}_{k_{\tau_\delta,\hat{a}_{\tau_\delta}},\hat{a}_{\tau_\delta}} - \hat{\mu}_{k_{\tau_\delta,a^\star},a^\star})_+ \geq \epsilon/2\} = \emptyset$.

**Summary.** Putting both cases together and take as a specific parameter $s = 2$ yields the result.

## Appendix G. Expected Sample Complexity of AdaP-TT and AdaP-TT$^\star$

Let $\beta \in (0,1)$, $\epsilon \in \mathbb{R}^\star_+$, and $\boldsymbol{\nu}$ be a bandit instance consisting of $\sigma$-sub-Gaussian distributions with distinct means $\boldsymbol{\mu} \in \mathbb{R}^K$, i.e. $\min_{a\neq b}|\mu_a - \mu_b| > 0$. For conciseness, we denote $\Delta_a \triangleq \mu_{a^\star} - \mu_a$, $\Delta_{\min} \triangleq \min_{a\neq a^\star}\Delta_a$, and $\Delta_{\max} \triangleq \max_{a\neq a^\star}\Delta_a$. For Gaussian, we define the unique

$\beta$-optimal allocations $\omega^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}) = \{(\omega^\star_{\beta,a})_{a\in[K]}\}$ and $\omega^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}_{G,\epsilon}) = \{(\omega^\star_{\epsilon,\beta,a})_{a\in[K]}\}$ as

$$\omega^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}) \triangleq \underset{\omega\in\Sigma_K,\omega_{a^\star}=\beta}{\arg\max} \min_{a\neq a^\star} \frac{\Delta_a^2}{1/\beta + 1/\omega_a} \ , \ \omega^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}_{G,\epsilon}) \triangleq \underset{\omega\in\Sigma_K,\omega_{a^\star}=\beta}{\arg\max} \min_{a\neq a^\star} \frac{\Delta_a \min\{\epsilon/2,\Delta_a\}}{1/\beta + 1/\omega_a}.$$

At equilibrium, we have equality of the transportation costs (see Jourdan and Degenne (2024) for example), namely

$$\forall a \neq a^\star, \quad \frac{\Delta_a^2}{1/\beta + 1/\omega^\star_{\beta,a}} = 2\sigma^2 T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu})^{-1} \quad , \quad \frac{\Delta_a \min\{\epsilon/2,\Delta_a\}}{1/\beta + 1/\omega^\star_{\epsilon,\beta,a}} = 2\sigma^2 T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}_{G,\epsilon})^{-1}. \tag{23}$$

Our proof follows the unified sample complexity analysis of Top Two algorithms from Jourdan et al. (2022).

Let $\gamma > 0$. Let $\omega \in \Sigma_K$ be any allocation over arms such that $\min_a \omega_a > 0$. We denote by $T_{\boldsymbol{\mu},\gamma}(\omega)$ the *convergence time* towards $\omega$, which is a random variable quantifying the number of samples required for the global empirical allocations $N_n/(n-1)$ to be $\gamma$-close to $\omega$ for any subsequent time, namely

$$T_{\boldsymbol{\mu},\gamma}(\omega) \triangleq \inf\left\{ T \geq 1 \mid \forall n \geq T, \ \left\| \frac{N_n}{n-1} - \omega \right\|_\infty \leq \gamma \right\}. \tag{24}$$

As the AdaP-TT and AdaP-TT$^\star$ algorithms share the same leader, all results solely on the leader applies to both of them. As the AdaP-TT and AdaP-TT$^\star$ algorithms consider a TC challenger with different transportation costs, all results involving the challenger should be slightly modified. Except if specified otherwise, all the results presented in the following hold for both algorithms.

The rest of Appendix G is organised as follows. After recalling some technical results (Appendix G.1), we prove sufficient exploration (Appendix G.2) Second, we prove that convergence towards the $\beta$-optimal allocation (Appendix G.3) in finite time. Third, we explicit the cost of doubling and forgetting (Appendix G.4). Finally, we conclude the proof of Theorems 21 and 23 (Appendix G.5).

### G.1 Technical Results

Before delving into the proofs, we first recall some useful technical results.

*Doubling trick.* Due to the doubling, the growth of the counts is exponential (Lemma 36).

**Lemma 36** *For all $(a,k) \in [K]\times\mathbb{N}$ s.t. $\mathbb{E}_{\boldsymbol{\nu}}[T_k(a)] < +\infty$, $N_{T_k(a),a} = 2^{k-1}$ and $\tilde{N}_{k,a} = 2^{k-2}$.*

**Proof** Let $a \in [K]$. After initialisation, we have $k = 1$, $T_1(a) = K+1$ and $N_{T_1(a),a} = 1$. Using the definition of the phase switch, it is direct to see that $N_{T_2(a),a} = 2$ and $\tilde{N}_{2,a} = 1$ when $\mathbb{E}_{\boldsymbol{\nu}}[T_2(a)] < +\infty$.

Now, we proceed by recurrence. Suppose that $N_{T_k(a),a} = 2^{k-1}$ and $\tilde{N}_{k,a} = 2^{k-2}$ when $\mathbb{E}_{\boldsymbol{\nu}}[T_k(a)] < +\infty$. If $\mathbb{E}_{\boldsymbol{\nu}}[T_{k+1}(a)] < +\infty$, then it means that the phase $k$ ends for arm $a$ almost surely. Since we sample only one arm at each round, at the beginning of phase $k+1$ for arm $a$, we have $N_{T_{k+1}(a),a} = 2N_{T_k(a),a} = 2^k$ by using the definition of the phase switch.

Then, we have directly that $\tilde{N}_{k+1,a} = N_{T_{k+1}(a),a} - N_{T_k(a),a} = 2^k - 2^{k-1} = 2^{k-1}$. ■

*Tracking.* We denote by $N^a_{n,b} \triangleq \sum_{t\in[n-1]} \mathbb{1}\left(B_t = a,\, a_t = C_t = b\right)$ the number of times the arm $b$ was pulled while the arm $a$ was the leader, and by $L_{n,a} \triangleq \sum_{t\in[n-1]} \mathbb{1}\left(B_t = a\right)$ the number of times arm $a$ was the leader.

**Lemma 37 (Lemma 2.2 in Jourdan and Degenne 2024)** *For all $n > K$ and all $a \in [K]$, we have $-1/2 \leq N^a_{n,a} - \beta L_{n,a} \leq 1$.*

*Concentration results.* In order to control the randomness of $(\tilde{\mu}_{k_a,a})_{a\in[K]}$, we use a standard concentration result on the empirical mean of sub-Gaussian random variables and on sub-exponential observations (Lemma 38). Since Bernoulli distributions are 1/2-sub-Gaussian and the absolute value of a Laplace is an exponential distribution, Lemma 38 applies to our setting.

**Lemma 38** *There exists a sub-Gaussian random variable $W_\mu$ such that, almost surely,*

$$\forall a \in [K],\, \forall k_a \in \mathbb{N}, \quad |\hat{\mu}_{k_a,a} - \mu_a| \leq W_\mu \sqrt{\frac{\log(e + \tilde{N}_{k_a,a})}{\tilde{N}_{k_a,a}}}\,.$$

*There exists a sub-exponential random variable $W_\epsilon$ such that, almost surely,*

$$\forall a \in [K],\, \forall k_a \in \mathbb{N}, \quad |Y_{k_a,a}| \leq W_\epsilon \frac{\log(e + k_a)}{\tilde{N}_{k_a,a}}\,.$$

*In particular, any random variable which is polynomial in $(W_\epsilon, W_\mu)$ has a finite expectation.*

**Proof** The first part is a known result, e.g. Appendix E.2 in Jourdan et al. (2022). Let

$$W_\epsilon \triangleq \sup_{a\in[K]} \sup_{k_a\in\mathbb{N}} \frac{\tilde{N}_{k_a,a}|Y_{k_a,a}|}{\log(e + k_a)}\,.$$

By definition, we have that, almost surely,

$$\forall a \in [K],\, \forall k_a \in \mathbb{N}, \quad |Y_{k_a,a}| \leq W_\epsilon \frac{\log(e + k_a)}{\tilde{N}_{k_a,a}}\,.$$

Since $\tilde{N}_{k,i}|Y_{k,i}| \sim \mathcal{E}(\epsilon)$, Lemma 72 in Jourdan et al. (2022) yields that $W_\epsilon$ is a sub-exponential random variable. Since $W_\mu$ is sub-Gaussian and $W_\epsilon$ is a sub-exponential, any random variable which is polynomial in $(W_\epsilon, W_\mu)$ has a finite expectation. ■

*Inversion results.* Lemma 39 gathers properties on the function $\overline{W}_{-1}$, which is used in the literature to obtain concentration results.

**Lemma 39 (Jourdan et al. 2023)** *Let $\overline{W}_{-1}(x) \triangleq -W_{-1}(-e^{-x})$ for all $x \geq 1$, where $W_{-1}$ is the negative branch of the Lambert W function. The function $\overline{W}_{-1}$ is increasing on*

$(1, +\infty)$ *and strictly concave on* $(1, +\infty)$*. In particular,* $\overline{W}'_{-1}(x) = \left(1 - \frac{1}{\overline{W}_{-1}(x)}\right)^{-1}$ *for all* $x > 1$*. Then, for all* $y \geq 1$ *and* $x \geq 1$*,*

$$\overline{W}_{-1}(y) \leq x \quad \iff \quad y \leq x - \log(x)\,.$$

*Moreover, for all* $x > 1$*,*

$$x + \log(x) \leq \overline{W}_{-1}(x) \leq x + \log(x) + \min\left\{\frac{1}{2}, \frac{1}{\sqrt{x}}\right\}\,.$$

Lemma 40 is an inversion result to upper bound a time, which is implicitly defined. It is a direct consequence of Lemma 39.

**Lemma 40** *Let* $\overline{W}_{-1}$ *defined in Lemma 39. Let* $A > 0$*,* $B > 0$ *such that* $B/A + \log A > 1$ *and*

$$C(A, B) = \sup\{x \mid\ x < A\log x + B\}\,.$$

*Then,* $C(A, B) < h_1(A, B)$ *with* $h_1(z, y) = z\overline{W}_{-1}\left(y/z + \log z\right)$*.*

**Proof** Since $B/A + \log A > 1$, we have $C(A, B) \geq A$, hence

$$C(A, B) = \sup\{x \mid\ x < A\log(x) + B\} = \sup\{x \geq A \mid\ x < A\log(x) + B\}\,.$$

Using Lemma 39 yields that

$$x \geq A\log x + B \iff \frac{x}{A} - \log\left(\frac{x}{A}\right) \geq \frac{B}{A} + \log A \iff x \geq A\overline{W}_{-1}\left(\frac{B}{A} + \log A\right)\,.$$

■

### G.2 Sufficient Exploration

The first step of in the generic analysis of Top Two algorithms Jourdan et al. (2022) consists in showing sufficient exploration. The main idea is that, if there are still undersampled arms, either the leader or the challenger will be among them. Therefore, after a long enough time, no arm can still be undersampled. We emphasise that there are multiple ways to select the leader/challenger pair in order to ensure sufficient exploration. Therefore, other choices of leader/challenger pair would yield similar results.

Given an arbitrary phase $p \in \mathbb{N}$, we define the sampled enough set, i.e. the arms having reached phase $p$, and the arm with highest mean in this set (when not empty) as

$$S_n^p = \{a \in [K] \mid N_{n,a} \geq 2^{p-1}\} \quad \text{and} \quad a_n^\star = \arg\max_{a \in S_n^p} \mu_a\,. \tag{25}$$

Since $\min_{a \neq b} |\mu_a - \mu_b| > 0$, $a_n^\star$ is unique. Let $p \in \mathbb{N}$ such that $(p-1)/4 \in \mathbb{N}$. We define the highly and the mildly under-sampled sets as

$$U_n^p \triangleq \{a \in [K] \mid N_{n,a} < 2^{(p-1)/2}\} \quad \text{and} \quad V_n^p \triangleq \{a \in [K] \mid N_{n,a} < 2^{3(p-1)/4}\}\,. \tag{26}$$

Those arms have not reached phase $(p-1)/2$ and phase $3(p-1)/4$, respectively.

*Lemma 41 shows that, when the leader is sampled enough, it is the arm with highest true mean among the sampled enough arms.*

**Lemma 41** *Let $S_n^p$ and $a_n^\star$ as in (25). There exists $p_0$ with $\mathbb{E}_{\boldsymbol{\nu}}[\exp(\alpha p_0)] < +\infty$ for all $\alpha > 0$ such that if $p \geq p_0$, for all $n$ such that $S_n^p \neq \emptyset$, $B_n \in S_n^p$ implies that $B_n = a_n^\star = \arg\max_{a \in S_n^p} \tilde{\mu}_{k_{n,a},a}$.*

**Proof** Let $p_0$ to be specified later. Let $p \geq p_0$. Let $n \in \mathbb{N}$ such that $S_n^p \neq \emptyset$, where $S_n^p$ and $a_n^\star$ as in Equation (25). Let $(k_{n,a})_{a\in[K]}$ be the phases indices for all arms. Since $N_{n,a} \geq 2^{p-1}$ for all $a \in S_n^p$, we have $k_{n,a} \geq p$ and $\tilde{N}_{k_{n,a},a} \geq 2^{p-2}$ by using Lemma 36. Using Lemma 38, we obtain that

$$\tilde{\mu}_{k_{n,a_n^\star},a_n^\star} \geq \mu_{a_n^\star} - W_\mu \sqrt{\frac{\log(e+2^{p-2})}{2^{p-2}}} - W_\epsilon \frac{\log(e+p)}{2^{p-2}}\,,$$
$$\tilde{\mu}_{k_{n,a},a} \leq \mu_a + W_\mu \sqrt{\frac{\log(e+2^{p-2})}{2^{p-2}}} + W_\epsilon \frac{\log(e+p)}{2^{p-2}}\,, \quad \forall a \in S_n^p \setminus \{a_n^\star\}.$$

Here, we use that $x \to \log(e+x)/x$ is decreasing.

Let $\overline{\Delta}_{\min} = \min_{a\neq b} |\mu_a - \mu_b|$. By assumption on the considered instances, we know that $\overline{\Delta}_{\min} > 0$. Let $p_1 = \lceil \log_2(X_1 - e) \rceil + 2$ and $p_2 = \lceil \log_2((X_2 - e - 2)\log 2 + 1) \rceil + 2$ with

$$X_1 = \sup\left\{ x > 1 \mid\ x \leq 64\overline{\Delta}_{\min}^{-2} W_\mu^2 \log x + e \right\} \leq h_1(64\overline{\Delta}_{\min}^{-2} W_\mu^2,\ e)\,,$$
$$X_2 = \sup\left\{ x > 1 \mid\ x \leq \frac{8}{\log 2}\overline{\Delta}_{\min}^{-1} W_\epsilon \log x + e + 2 - 1/\log 2 \right\} \leq h_1(8\overline{\Delta}_{\min}^{-1} W_\epsilon/\log 2,\ 4)\,,$$

where we used Lemma 40, and $h_1$ defined therein. Then, for all $p \in \mathbb{N}$ such that $p \geq \max\{p_1, p_2\} + 1$ and all $n \in \mathbb{N}$ such that $S_n^p \neq \emptyset$, we have $\tilde{\mu}_{k_{n,a_n^\star},a_n^\star} \geq \mu_{a_n^\star} - \overline{\Delta}_{\min}/4$ and $\tilde{\mu}_{k_{n,a},a} \leq \mu_a + \overline{\Delta}_{\min}/4$ for all $a \in S_n^p \setminus \{a_n^\star\}$, hence $a_n^\star = \arg\max_{a\in[K]} \tilde{\mu}_{k_{n,a},a}$.

We have, for all $\alpha \in \mathbb{R}_+$,

$$\exp(\alpha p_1) \leq e^{3\alpha}(X_1 - e)^{\alpha/\log 2} \quad \text{hence} \quad \mathbb{E}_{\boldsymbol{\nu}}[\exp(\alpha p_1)] < +\infty\,,$$

where we used Lemma 38 and $h_1(x,e) \sim_{x\to+\infty} x\log x$ to obtain that $\exp(\alpha p_1)$ is at most polynomial in $W_\mu$. Likewise, we obtain that $\mathbb{E}_{\boldsymbol{\nu}}[\exp(\alpha p_2)] < +\infty$ for all $\alpha \in \mathbb{R}_+$.

Let us define the UCB indices by $I_{k_{n,a},a} = \tilde{\mu}_{k_{n,a},a} + \sqrt{k_{n,a}/\tilde{N}_{k_{n,a},a}} + k_{n,a}/(\epsilon\tilde{N}_{k_{n,a},a})$. Using the above, we have

$$I_{k_{n,a_n^\star},a_n^\star} \geq \mu_{a_n^\star} - W_\mu \sqrt{\frac{\log(e+2^{p-2})}{2^{p-2}}} - W_\epsilon \frac{\log(e+p)}{2^{p-2}}\,,$$
$$\forall a \in S_n^p \setminus \{a_n^\star\}, \quad I_{k_{n,a},a} \leq \mu_a + W_\mu \sqrt{\frac{\log(e+2^{p-2})}{2^{p-2}}} + W_\epsilon \frac{\log(e+p)}{2^{p-2}} + \sqrt{\frac{p}{2^{p-2}}} + \frac{p}{\epsilon 2^{p-2}}\,,$$

where we used Lemma 36 and the fact that $x \to \log(e+x)/x$ and $x \to x2^{2-x}$ are decreasing function for $x \geq 2$. Let $p_3 = \lceil \log_2 X_3 \rceil + 2$ and $p_4 = \lceil \log_2 X_4 \rceil + 2$ with

$$X_3 = \sup\left\{ x > 1 \mid\ x \leq 64\overline{\Delta}_{\min}^{-2}(\log_2 x + 2) \right\} \leq h_1(64\overline{\Delta}_{\min}^{-2}/\log 2,\ 128\overline{\Delta}_{\min}^{-2})\,,$$
$$X_4 = \sup\left\{ x > 1 \mid\ x \leq 8\epsilon^{-1}\overline{\Delta}_{\min}^{-1}(\log_2 x + 2) \right\} \leq h_1(8\epsilon^{-1}\overline{\Delta}_{\min}^{-1}/\log 2,\ 16\overline{\Delta}_{\min}^{-1}\epsilon^{-1})\,,$$

where we used Lemma 40, and $h_1$ defined therein. We highlight that $(p_3, p_4)$ are deterministic values, hence their expectation is finite. Then, for all $p \in \mathbb{N}$ such that $p \geq p_0 = \max\{p_1, p_2, p_3, p_4\} + 1$ and all $n \in \mathbb{N}$ such that $S_n^p \neq \emptyset$, we have $I_{k_{n,a_n^\star},a_n^\star} \geq \mu_{a_n^\star} - \overline{\Delta}_{\min}/4$ and $I_{k_{n,a},a} \leq \mu_a + \overline{\Delta}_{\min}/2$ for all $a \in S_n^p \setminus \{a_n^\star\}$, hence $a_n^\star = B_n$ since we have $B_n = \arg\max_{a\in[K]} I_{k_{n,a},a}$.

Since we have $\mathbb{E}_{\boldsymbol{\nu}}[\exp(\alpha p_0)] < +\infty$ for all $\alpha \in \mathbb{R}_+$, this concludes the proof. ■

*Lemma 42 shows that the transportation costs between the sampled enough arms with largest true means and the other sampled enough arms are increasing fast enough.*

**Lemma 42** *Let $S_n^p$ and $a_n^\star$ as in Eq. (25). There exists $p_1$ with $\mathbb{E}_{\boldsymbol{\nu}}[\exp(\alpha p_1)] < +\infty$ for all $\alpha > 0$ such that if $p \geq p_1$, for all $n$ such that $S_n^p \neq \emptyset$, for all $b \in S_n^p \setminus \{a_n^\star\}$, we have*

$$\text{[AdaP-TT]} \quad \frac{\tilde{\mu}_{k_{n,a_n^\star},a_n^\star} - \tilde{\mu}_{k_{n,b},b}}{\sqrt{1/\tilde{N}_{k_{n,a_n^\star},a_n^\star} + 1/\tilde{N}_{k_{n,b},b}}} \geq 2^{p/2} C_\mu \,,$$

$$\text{[AdaP-TT}^\star\text{]} \quad \frac{(\tilde{\mu}_{k_{n,a_n^\star},a_n^\star} - \tilde{\mu}_{k_{n,b},b}) \min\{\epsilon/2, \tilde{\mu}_{k_{n,a_n^\star},a_n^\star} - \tilde{\mu}_{k_{n,b},b}\}}{1/\tilde{N}_{k_{n,a_n^\star},a_n^\star} + 1/\tilde{N}_{k_{n,b},b}} \geq 2^p C_\mu \,,$$

*where $C_\mu > 0$ is a problem dependent constant.*

**Proof** Let $p_1$ to be specified later. Let $p \geq p_1$. Let $n \in \mathbb{N}$ such that $S_n^p \neq \emptyset$, where $S_n^p$ and $a_n^\star$ as in Equation (25). Let $(k_{n,a})_{a\in[K]}$ be the phases indices for all arms. Since $N_{n,a} \geq 2^{p-1}$ for all $a \in S_n^p$, we have $k_{n,a} \geq p$ and $\tilde{N}_{k_{n,a},a} \geq 2^{p-2}$ by using Lemma 36. Let $\overline{\Delta}_{\min} = \min_{a\neq b} |\mu_a - \mu_b|$, which satisfies $\overline{\Delta}_{\min} > 0$ by assumption on the instance considered.

Using Lemma 38, for all $b \in S_n^p \setminus \{a_n^\star\}$, we obtain

$$\tilde{\mu}_{k_{n,a_n^\star},a_n^\star} - \tilde{\mu}_{k_{n,b},b} \geq \overline{\Delta}_{\min} - W_\mu \sqrt{\frac{\log(e + 2^{p-2})}{2^{p-4}}} - W_\epsilon \frac{\log(e+p)}{2^{p-3}} \,.$$

Let $p_3 = \lceil \log_2((X_3 - e)/4) \rceil + 4$ and $p_2 = \lceil \log_2((X_2 - e - 3)\log 2 + 1) \rceil + 3$ with

$$X_3 = \sup\left\{x > 1 \mid\ x \leq 64\overline{\Delta}_{\min}^{-2} W_\mu^2 \log x + e\right\} \leq h_1(64\overline{\Delta}_{\min}^{-2} W_\mu^2,\, e) \,,$$
$$X_2 = \sup\left\{x > 1 \mid\ x \leq 4\overline{\Delta}_{\min}^{-1} W_\epsilon \log x + e + 3 - 1/\log 2\right\} \leq h_1(4\overline{\Delta}_{\min}^{-1} W_\epsilon,\, 5) \,,$$

where we used Lemma 40, and $h_1$ defined therein. Then, for all $p \in \mathbb{N}$ such that $p \geq p_1 = \max\{p_3, p_2\} + 1$ and all $n \in \mathbb{N}$ such that $S_n^p \neq \emptyset$, we have, for all $b \in S_n^p \setminus \{a_n^\star\}$,

$$\tilde{\mu}_{k_{n,a_n^\star},a_n^\star} - \tilde{\mu}_{k_{n,b},b} \geq \overline{\Delta}_{\min}/2 \,.$$

As in the proof of Lemma 41, we obtain that $\mathbb{E}_{\boldsymbol{\nu}}[\exp(\alpha p_1)] < +\infty$ for all $\alpha \in \mathbb{R}_+$.

Then, for all $b \in S_n^p \setminus \{a_n^\star\}$, we have

$$\frac{\tilde{\mu}_{k_{n,a_n^\star},a_n^\star} - \tilde{\mu}_{k_{n,b},b}}{\sqrt{1/\tilde{N}_{k_{n,a_n^\star},a_n^\star} + 1/\tilde{N}_{k_{n,b},b}}} \geq 2^{p/2} \frac{\overline{\Delta}_{\min}}{2^{5/2}} \,,$$

where we used that $\min\{\tilde{N}_{k_{n,a_n^\star},a}, \tilde{N}_{k_{n,b},b}\} \geq 2^{p-2}$. Setting $C_\mu = \overline{\Delta}_{\min}/2^{5/2}$ yields the first result.

The second result is obtained similarly by taking $C_\mu = \frac{\overline{\Delta}_{\min}}{16} \min\{\epsilon/2, \frac{\overline{\Delta}_{\min}}{2}\}$ ■

*Lemma 43 shows that the transportation costs between sampled enough arms and under-sampled arms are not increasing too fast.*

**Lemma 43** *Let $S_n^p$ be as in Eq. (25). For all $p \geq 1$ and all $n$ such that $S_n^p \neq \emptyset$, for all $a \in S_n^p$ and $b \notin S_n^p$,*

$$\textit{[AdaP-TT]} \quad \frac{\tilde{\mu}_{k_{n,a},a} - \tilde{\mu}_{k_{n,b},b}}{\sqrt{1/\tilde{N}_{k_{n,a},a} + 1/\tilde{N}_{k_{n,b},b}}} \leq 2^{p/2} D_\mu + 2W_\mu \sqrt{\log(e + 2^{p-2})} + 2W_\epsilon \log(e+p)\,,$$

$$\textit{[AdaP-TT}^\star\textit{]} \quad \frac{(\tilde{\mu}_{k_{n,a},a} - \tilde{\mu}_{k_{n,b},b}) \min\{\epsilon/2, \tilde{\mu}_{k_{n,a},a} - \tilde{\mu}_{k_{n,b},b}\}}{1/\tilde{N}_{k_{n,a},a} + 1/\tilde{N}_{k_{n,b},b}} \leq 2^p D_\mu + 8W_\mu^2 \log(e + 2^{p-2}) + 8W_\epsilon^2 \log(e+p)^2\,,$$

*where $D_\mu > 0$ is a problem dependent constant and $(W_\mu, W_\epsilon)$ are the random variables defined in Lemma 38.*

**Proof** Let $p \geq 1$. Let $n \in \mathbb{N}$ such that $S_n^p \neq \emptyset$, where $S_n^p$ as in Equation (25). Let $(k_{n,a})_{a\in[K]}$ be the phases indices for all arms. Since $N_{n,a} \geq 2^{p-1}$ for all $a \in S_n^p$, we have $k_{n,a} \geq p$ and $\tilde{N}_{k_{n,a},a} \geq 2^{p-2}$ by using Lemma 36. Likewise, $N_{n,a} < 2^{p-1}$ for all $a \notin S_n^p$, we have $k_{n,a} < p$ and $\tilde{N}_{k_{n,a},a} < 2^{p-2}$. Let $\overline{\Delta}_{\max} = \min_{a\neq b} |\mu_a - \mu_b|$, which satisfies $\overline{\Delta}_{\max} > 0$ by assumption on the instance considered. Using Lemma 38, for all $a \in S_n^p$ and $b \notin S_n^p$, we obtain

$$\begin{aligned}
\frac{\tilde{\mu}_{k_{n,a},a} - \tilde{\mu}_{k_{n,b},b}}{\sqrt{1/\tilde{N}_{k_{n,a},a} + 1/\tilde{N}_{k_{n,b},b}}} &\leq \sqrt{\tilde{N}_{k_{n,b},b}}(\tilde{\mu}_{k_{n,a},a} - \tilde{\mu}_{k_{n,b},b}) \\
&\leq \sqrt{\tilde{N}_{k_{n,b},b}}(\mu_a - \mu_b) + 2W_\mu \sqrt{\log(e + \tilde{N}_{k_{n,b},b})} + 2W_\epsilon \frac{\log(e+k_{n,b})}{\sqrt{\tilde{N}_{k_{n,b},b}}} \\
&\leq 2^{(p-2)/2}\overline{\Delta}_{\max} + 2W_\mu \sqrt{\log(e+2^{p-2})} + 2W_\epsilon \log(e+p)
\end{aligned}$$

where we used that $\tilde{N}_{k_{n,b},b} \geq 1$, $k_{n,b} < p$, $\tilde{N}_{k_{n,b},b} < 2^{p-2} \leq \tilde{N}_{k_{n,a},a}$ and $x \to \log(e+x)/x$ is decreasing. Taking $D_\mu = \overline{\Delta}_{\max}/2$ yields the first result.

The proof of the second result follows along the same line by noting that this transportation cost is lower than the other:

$$\begin{aligned}
&\frac{(\tilde{\mu}_{k_{n,a},a} - \tilde{\mu}_{k_{n,b},b}) \min\{\epsilon/2, \tilde{\mu}_{k_{n,a},a} - \tilde{\mu}_{k_{n,b},b}\}}{1/\tilde{N}_{k_{n,a},a} + 1/\tilde{N}_{k_{n,b},b}} \\
&\leq 2\tilde{N}_{k_{n,b},b}(\hat{\mu}_{k_{n,a},a} - \hat{\mu}_{k_{n,b},b})^2 + 2\tilde{N}_{k_{n,b},b}(Y_{k_{n,a},a} - Y_{k_{n,b},b})^2 \\
&\leq 2^p \overline{\Delta}_{\max}^2 + 8W_\mu^2 \log(e+2^{p-2}) + 8W_\epsilon^2 \log(e+p)^2\,.
\end{aligned}$$

Taking $D_\mu = \overline{\Delta}_{\max}^2$ yields the result. ■

*Lemma 44 shows that the challenger is mildly undersampled if the leader is not mildly undersampled.*

**Lemma 44** *Let $V_n^p$ be as in Equation (26). There exists $p_2$ with $\mathbb{E}_{\boldsymbol{\nu}}[\exp(\alpha p_2)] < +\infty$ for all $\alpha > 0$ such that if $p \geq p_2$, for all $n$ such that $U_n^p \neq \emptyset$, $B_n \notin V_n^p$ implies $C_n \in V_n^p$.*

**Proof** Let $p_2$ to be specified later. Let $p \geq p_2$. Let $n \in \mathbb{N}$ such that $U_n^p \neq \emptyset$ and $V_n^p \neq [K]$, where $U_n^p \subseteq V_n^p$ are defined in Equation (26). In the following, we suppose that $B_n \notin V_n^p$.

Let $(k_{n,a})_{a\in[K]}$ be the phases indices for all arms. Let $p_0$ as in Lemma 41. Let $b_n^\star = \arg\max_{b\notin V_n^p} \mu_b$. Then, for all $p \geq 4p_0/3 - 1/3$ and all $n$ such that $B_n \notin V_n^p$, Lemma 41 yields that $B_n = b_n^\star = \arg\max_{a\notin V_n^p} \tilde{\mu}_{k_{n,a},a}$.

Let $p_1$ and $C_\mu$ as in Lemma 42, and $D_\mu$ as in Lemma 43. Then, for all $p \geq \frac{4}{3}\max\{p_0, p_1\} - 1/3$ and all $n$ such that $B_n \notin V_n^p$, we have $B_n = b_n^\star$ and

$$\begin{aligned}
\forall b \notin V_n^p, \quad & \frac{\tilde{\mu}_{k_{n,b_n^\star},b_n^\star} - \tilde{\mu}_{k_{n,b},b}}{\sqrt{1/\tilde{N}_{k_{n,b_n^\star},b_n^\star} + 1/\tilde{N}_{k_{n,b},b}}} \geq 2^{(3p+1)/8} C_\mu \,, \\
\forall b \in U_n^p, \quad & \frac{\tilde{\mu}_{k_{n,b_n^\star},b_n^\star} - \tilde{\mu}_{k_{n,b},b}}{\sqrt{1/\tilde{N}_{k_{n,b_n^\star},b_n^\star} + 1/\tilde{N}_{k_{n,b},b}}} \leq 2^{(p+1)/4} D_\mu + 2W_\mu \sqrt{\log(e + 2^{(p+1)/2-2})} \\
& \qquad\qquad + 2W_\epsilon \log(e + (p+1)/2) \,,
\end{aligned}$$

where we used the first results of Lemmas 42 and 43. Let $p_3 = 16\lceil \log_2(2D_\mu/C_\mu)\rceil + 1$, then we have $2^{(p-1)/16} > \frac{D_\mu}{C_\mu}$ for all $p \geq p_3$. Let $p_4 = \frac{16}{9}\lceil \log_2 X_4\rceil + 25$ and $p_5 = \frac{32}{9}\lceil \log_2 X_5\rceil + 7$ where

$$\begin{aligned}
X_4 &= \sup\left\{ x > 1 \mid x \leq \frac{W_\mu^2}{C_\mu^2} \log(e + x^{8/9} 2^{25/18 - 3/4}) \right\} \,, \\
X_5 &= \sup\left\{ x > 1 \mid x \leq \frac{2W_\epsilon}{C_\mu} \log(e + 4 + 32\log_2(x)/18) \right\} \,.
\end{aligned}$$

As in the proof of Lemma 41, using Lemma 38 yields that $\mathbb{E}_{\boldsymbol{\nu}}[\exp(\alpha p_4)] < +\infty$ and $\mathbb{E}_{\boldsymbol{\nu}}[\exp(\alpha p_5)] < +\infty$ for all $\alpha \in \mathbb{R}_+$. Let $p_2 = \max\{p_3, p_4, p_5, 4\max\{p_0, p_1\}/3 - 1/3\} + 1$. Then, we have shown that for all $p \geq p_2$, for all $n$ such that $B_n \notin V_n^p$, we have $B_n = b_n^\star$ and

$$\min_{b\notin V_n^p} \frac{\tilde{\mu}_{k_{n,b_n^\star},b_n^\star} - \tilde{\mu}_{k_{n,b},b}}{\sqrt{1/\tilde{N}_{k_{n,b_n^\star},b_n^\star} + 1/\tilde{N}_{k_{n,b},b}}} > \max_{b\in U_n^p} \frac{\tilde{\mu}_{k_{n,b_n^\star},b_n^\star} - \tilde{\mu}_{k_{n,b},b}}{\sqrt{1/\tilde{N}_{k_{n,b_n^\star},b_n^\star} + 1/\tilde{N}_{k_{n,b},b}}} \,,$$

Therefore, by definition of the TC challenger $C_n = \arg\min_{b\neq b_n^\star} \frac{\tilde{\mu}_{k_{n,b_n^\star},b_n^\star} - \tilde{\mu}_{k_{n,b},b}}{\sqrt{1/\tilde{N}_{k_{n,b_n^\star},b_n^\star} + 1/\tilde{N}_{k_{n,b},b}}}$, we obtain that $C_n \in V_n^p$. Otherwise, there would be a contradiction given that we assumed that $U_n^p \neq \emptyset$. Given all the condition exhibited above, it is direct to see that $\mathbb{E}_{\boldsymbol{\nu}}[\exp(\alpha p_2)] < +\infty$ for all $\alpha > 0$. This concludes the proof for the AdaP-TT algorithm.

For the AdaP-TT$^\star$ algorithm, the proof is done similarly based on the second results of Lemmas 42 and 43. As above, we can construct $\tilde{p}_3$, with $\mathbb{E}_\nu[\exp(\alpha\tilde{p}_3)] < +\infty$ for all $\alpha \in \mathbb{R}_+$, such that for all $p \geq \tilde{p}_3$, we have

$$2^{(3p+1)/4} C_\mu > 2^{(p+1)/2} D_\mu + 8W_\mu^2 \log(e + 2^{(p+1)/2-2}) + 8W_\epsilon^2 \log(e + (p+1)/2)^2 \,.$$

Let $\tilde{p}_2 = \max\{\tilde{p}_3, 4\max\{p_0, p_1\}/3 - 1/3\} + 1$. Then, we have shown that for all $p \geq \tilde{p}_2$, for all $n$ such that $B_n \notin V_n^p$, we have $B_n = b_n^\star$ and

$$\begin{aligned}
&\min_{b \notin V_n^p} \frac{(\tilde{\mu}_{k_{n,b_n^\star},b_n^\star} - \tilde{\mu}_{k_{n,b},b}) \min\{\epsilon/2, \tilde{\mu}_{k_{n,b_n^\star},b_n^\star} - \tilde{\mu}_{k_{n,b},b}\}}{1/\tilde{N}_{k_{n,b_n^\star},b_n^\star} + 1/\tilde{N}_{k_{n,b},b}} \\
&> \max_{b \in U_n^p} \frac{(\tilde{\mu}_{k_{n,b_n^\star},b_n^\star} - \tilde{\mu}_{k_{n,b},b}) \min\{\epsilon/2, \tilde{\mu}_{k_{n,b_n^\star},b_n^\star} - \tilde{\mu}_{k_{n,b},b}\}}{1/\tilde{N}_{k_{n,b_n^\star},b_n^\star} + 1/\tilde{N}_{k_{n,b},b}} \,.
\end{aligned}$$

Then, we conclude similarly by using the definition of the TC challenger. ■

*Lemma 45 shows that all the arms are sufficient explored for large enough $n$.*

**Lemma 45** *There exists $N_0$ with $\mathbb{E}_{\boldsymbol{\nu}}[N_0] < +\infty$ such that for all $n \geq N_0$ and all $a \in [K]$,*

$$N_{n,a} \geq \sqrt{n/K} \quad \textit{and} \quad k_{n,a} \geq \frac{\log(n/K)}{2\log 2} + 1 \,.$$

**Proof** Let $p_0$ and $p_2$ as in Lemmas 41 and 44. Combining Lemmas 41 and 44 yields that, for all $p \geq p_3 = \max\{p_2, 4p_0/3 - 1/3\}$ and all $n$ such that $U_n^p \neq \emptyset$, we have $B_n \in V_n^p$ or $C_n \in V_n^p$. We have $\mathbb{E}_{\boldsymbol{\nu}}[2^{p_2}] < +\infty$. We have $2^{p-1} \geq K 2^{3(p-1)/4}$ for all $p \geq p_4 = 4\lceil \log_2 K \rceil + 1$. Let $p \geq \max\{p_3, p_4\}$.

Suppose towards contradiction that $U_{K2^{p-1}}^p$ is not empty. Then, for any $1 \leq t \leq K2^{p-1}$, $U_t^p$ and $V_t^p$ are non empty as well. Using the pigeonhole principle, there exists some $a \in [K]$ such that $N_{2^{p-1},a} \geq 2^{3(p-1)/4}$. Thus, we have $\left|V_{2^{p-1}}^p\right| \leq K - 1$. Our goal is to show that $|V_{2^p}^p| \leq K - 2$. A sufficient condition is that one arm in $V_{2^{p-1}}^p$ is pulled at least $2^{3(p-1)/4}$ times between $2^{p-1}$ and $2^p - 1$.

*Case 1.* Suppose there exists $a \in V_{2^{p-1}}^p$ such that $L_{2^p,a} - L_{2^{p-1},a} \geq \frac{2^{3(p-1)/4}}{\beta} + 3/(2\beta)$. Using Lemma 37, we obtain

$$N_{2^p,a}^a - N_{2^{p-1},a}^a \geq \beta(L_{2^p,a} - L_{2^{p-1},a}) - 3/2 \geq 2^{3(p-1)/4} \,,$$

hence $a$ is sampled $2^{3(p-1)/4}$ times between $2^{p-1}$ and $2^p - 1$.

*Case 2.* Suppose that for all $a \in V_{2^{p-1}}^p$, we have $L_{2^p,a} - L_{2^{p-1},a} < 2^{3(p-1)/4}/\beta + 3/(2\beta)$. Then,

$$\sum_{a \notin V_{2^{p-1}}^p} (L_{2^p,a} - L_{2^{p-1},a}) \geq 2^{p-1} - K\left(2^{3(p-1)/4}/\beta + 3/(2\beta)\right)$$

Using Lemma 37, we obtain

$$\left| \sum_{a \notin V_{2^{p-1}}^p} (N_{2^p,a}^a - N_{2^{p-1},a}^a) - \beta \sum_{a \notin V_{2^{p-1}}^p} (L_{2^p,a} - L_{2^{p-1},a}) \right| \leq 3(K-1)/2 \,.$$

Combining all the above, we obtain

$$\begin{aligned}
&\sum_{a \notin V^p_{2^{p-1}}} (L_{2^p,a} - L_{2^{p-1},a}) - \sum_{a \notin V^p_{2^{p-1}}} (N^a_{2^p,a} - N^a_{2^{p-1},a}) \\
&\geq (1-\beta) \sum_{a \notin V^p_{2^{p-1}}} (L_{2^p,a} - L_{2^{p-1},a}) - 3(K-1)/2 \\
&\geq (1-\beta) \left(2^{p-1} - K\left(2^{3(p-1)/4}/\beta + 3/(2\beta)\right)\right) - 3(K-1)/2 \geq K 2^{3(p-1)/4} \,,
\end{aligned}$$

where the last inequality is obtained for $p \geq p_5$ with

$$p_5 = \sup \left\{ p \in \mathbb{N} \mid (1-\beta)\left(2^{p-1} - K\left(2^{3(p-1)/4}/\beta + \frac{3}{2\beta}\right)\right) - \frac{3}{2}(K-1) < K 2^{3(p-1)/4} \right\} .$$

The LHS summation is exactly the number of times where an arm $a \notin V^p_{2^{p-1}}$ was leader but wasn't sampled, hence

$$\sum_{t=2^{p-1}}^{2^p-1} \mathbb{1}\left(B_t \notin V^p_{2^{p-1}},\, a_t = C_t\right) \geq K 2^{3(p-1)/4}$$

For any $2^{p-1} \leq t \leq 2^p - 1$, $U^p_t$ is non-empty, hence we have $B_t \notin V^p_{2^{p-1}}$ (hence $B_t \notin V^p_t$) implies $C_t \in V^p_t \subseteq V^p_{2^{p-1}}$. Therefore, we have shown that

$$\sum_{t=2^{p-1}}^{2^p-1} \mathbb{1}\left(a_t \in V^p_{2^{p-1}}\right) \geq \sum_{t=2^{p-1}}^{2^p-1} \mathbb{1}\left(B_t \notin V^p_{2^{p-1}},\, a_t = C_t\right) \geq K 2^{3(p-1)/4} \,.$$

Therefore, there is at least one arm in $V^p_{2^{p-1}}$ that is sampled $2^{3(p-1)/4}$ times between $2^{p-1}$ and $2^p - 1$.

In summary, we have shown $|V^p_{2^p}| \leq K - 2$ for all $p \geq p_6 = \max\{p_3, p_4, p_5\}$. By induction, for any $1 \leq k \leq K$, we have $\left|V^p_{k2^{p-1}}\right| \leq K - k$, and finally $U^p_{K2^{p-1}} = \emptyset$ for all $p \geq p_6$. Defining $N_0 = K 2^{p_6 - 1}$, we have $\mathbb{E}_{\boldsymbol{\nu}}[N_0] < +\infty$ by using Lemmas 41 and 44 for $p_3 = \max\{p_2, 4p_0/3 - 1/3\}$ and $p_4$ and $p_5$ are deterministic. For all $n \geq N_0$, we let $2^{p-1} = \frac{n}{K}$. Then, by applying the above, we have $U^p_{K2^{p-1}} = U^{\log_2(n/K)+1}_n$ is empty, which shows that $N_{n,a} \geq \sqrt{n/K}$ for all $a \in [K]$. Using Lemma 36, we obtain that $k_{n,a} \geq \frac{\log(n/K)}{2\log 2} + 1$ for all $a \in [K]$. This concludes the proof. ■

### G.3 Convergence Towards $\beta$-optimal Allocation

The second step of in the generic analysis of Top Two algorithms Jourdan et al. (2022) is to show the convergence of the empirical proportions towards the $\beta$-optimal allocation. First, we show that the leader coincides with the best arm. Hence, the tracking procedure will ensure that the empirical proportion of time we sample it is exactly $\beta$. Second, we show that a sub-optimal arm whose empirical proportion overshoots its $\beta$-optimal allocation will not be sampled next as challenger. Therefore, this "overshoots implies not sampled" mechanism

will ensure the convergence towards the $\beta$-optimal allocation. We emphasise that there are multiple ways to select the leader/challenger pair in order to ensure convergence towards the $\beta$-optimal allocation. Therefore, other choices of leader/challenger pair would yield similar results. Note that our results heavily rely on having obtained sufficient exploration first.

*Convergence for the best arm.* Lemma 46 exhibits a random phase which ensures that the leader and the candidate answer are equal to the best arm for large enough $n$.

**Lemma 46** *Let $N_0$ be as in Lemma 45. There exists $N_1 \geq N_0$ with $\mathbb{E}_{\boldsymbol{\nu}}[N_1] < +\infty$ such that, for all $n \geq N_1$, we have $\hat{a}_n = B_n = a^\star$.*

**Proof** Let $k \geq 1$. Suppose that $\mathbb{E}_{\boldsymbol{\nu}}[\max_{a\in[K]} T_k(a)] < +\infty$. Then, Lemma 36 yields that $N_{T_k(a),a} = 2^{k-1}$ and $\tilde{N}_{k,a} = 2^{k-2}$. Using Lemma 38, we obtain that

$$\begin{aligned}
\tilde{\mu}_{k,a^\star} &\geq \mu_{a^\star} - W_\mu \sqrt{\frac{\log(e+2^{k-2})}{2^{k-2}}} - W_\epsilon \frac{\log(e+k)}{2^{k-2}} \,, \\
\forall a \neq a^\star, \quad \tilde{\mu}_{k,a} &\leq \mu_a + W_\mu \sqrt{\frac{\log(e+2^{k-2})}{2^{k-2}}} + W_\epsilon \frac{\log(e+k)}{2^{k-2}} \,.
\end{aligned}$$

Let $p_1 = \lceil \log_2(X_1 - e) \rceil + 2$ and $p_2 = \lceil \log_2(X_2 - e - 1) \rceil + 2$ with

$$\begin{aligned}
X_1 &= \sup\left\{x > 1 \mid\ x \leq 64\Delta_{\min}^{-2} W_\mu^2 \log x + e\right\} \leq h_1(64\Delta_{\min}^{-2} W_\mu^2,\, e) \,, \\
X_2 &= \sup\left\{x > 1 \mid\ x \leq 8\Delta_{\min}^{-1} W_\epsilon \log x + e + 1\right\} \leq h_1(8\Delta_{\min}^{-1} W_\epsilon,\, e+1) \,, \\
X_2 &\geq \sup\left\{x > 1 \mid\ x \leq 8\Delta_{\min}^{-1} W_\epsilon \log(e + 2 + \log x)\right\} \,,
\end{aligned}$$

where we used Lemma 40, and $h_1$ defined therein. Then, for all $k \in \mathbb{N}^K$ such that $\min_{a\in[K]} k_a > p_0 = \max\{p_1, p_2\}$ such that $\mathbb{E}_{\boldsymbol{\nu}}[\max_{a\in[K]} T_{k_a}(a)] < +\infty$, we have $\tilde{\mu}_{k,a^\star} \geq \mu_{a^\star} - \Delta_{\min}/4$ and $\tilde{\mu}_{k,a} \leq \mu_a + \Delta_{\min}/4$ for all $a \neq a^\star$, hence $a^\star = \arg\max_{a\in[K]} \tilde{\mu}_{k,a}$. We have, for all $\alpha \in \mathbb{R}_+$,

$$\exp(\alpha p_1) \leq e^{3\alpha}(X_1 - e)^{\alpha/\log 2} \quad \text{hence} \quad \mathbb{E}_{\boldsymbol{\nu}}[\exp(\alpha p_1)] < +\infty \,,$$

where we used Lemma 38 and $h_1(x, e) \sim_{x\to+\infty} x \log x$ to obtain that $\exp(\alpha p_1)$ is at most polynomial in $W_\mu$. Likewise, we obtain that $\mathbb{E}_{\boldsymbol{\nu}}[\exp(\alpha p_2)] < +\infty$ for all $\alpha \in \mathbb{R}_+$. Therefore, we have $\mathbb{E}_{\boldsymbol{\nu}}[\exp(\alpha p_0)] < +\infty$ for all $\alpha \in \mathbb{R}_+$.

Let us define the UCB indices by $I_{k,a} = \tilde{\mu}_{k,a} + \sqrt{k/\tilde{N}_{k,a}} + k/(\epsilon\tilde{N}_{k,a})$. Using the above, we have

$$\begin{aligned}
I_{k,a^\star} &\geq \mu_{a^\star} - W_\mu \sqrt{\frac{\log(e+2^{k-2})}{2^{k-2}}} - W_\epsilon \frac{\log(e+k)}{2^{k-2}} + \frac{k}{\epsilon 2^{k-2}} \,, \\
\forall a \neq a^\star, \quad I_{k,a} &\leq \mu_a + W_\mu \sqrt{\frac{\log(e+2^{k-2})}{2^{k-2}}} + W_\epsilon \frac{\log(e+k)}{2^{k-2}} + \frac{k}{\epsilon 2^{k-2}} \,.
\end{aligned}$$

Therefore, we have $a^\star = \arg\max_{a\in[K]} I_{k,a}$ for all $k \in \mathbb{N}^K$ such that $\min_a k_a > \max\{p_1, p_2\}$ such that $\mathbb{E}_{\boldsymbol{\nu}}[\max_{a\in[K]} T_{k_a}(a)] < +\infty$.

Let $N_0$ as in Lemma 45. Using Lemma 45, we obtain that, for all $n \geq N_0$ and all $a \in [K]$, $k_{n,a} \geq \log_2(n/K)/2 + 1$. Therefore, we obtain $\min_{a\in[K]} k_{n,a} > \max\{p_1, p_2\}$ is implied by

$n \geq N_1 = \max\{K4^{\max\{p_1,p_2\}}, N_0\}$. Using the above, we conclude that $\mathbb{E}_{\boldsymbol{\nu}}[N_1] < +\infty$ and $\hat{a}_n = B_n = a^\star$ for all $n \geq N_1$. ■

*Lemma 47 shows that that the pulling proportion of the best arm converges towards $\beta$, provided the phase defined in Lemma 46 is reached in finite time for all arms.*

**Lemma 47** *Let $\gamma > 0$, and $N_1$ be as in Lemma 46. There exists a deterministic constant $C_0 \geq 1$ such that, for all $n \geq C_0 N_1$,*

$$\left|\frac{N_{n,a^\star}}{n-1} - \beta\right| \leq \gamma\,.$$

**Proof** Let $\gamma > 0$. Let $N_1$ as in Lemma 46. Let $M \geq N_1$. Using Lemma 46, we obtain $B_n = a^\star$ for all $n \geq M$. Therefore, we obtain $L_{n,a^\star} \geq n - M$ and $\sum_{a \neq a^\star} N^a_{n,a^\star} \leq M$ for all $n \geq M$. Using Lemma 37 yields that

$$\begin{aligned}\left|\frac{N_{n,a^\star}}{n-1} - \beta\right| &\leq \frac{|N^{a^\star}_{n,a^\star} - \beta L_{n,a^\star}|}{n-1} + \beta\left|\frac{L_{n,a^\star}}{n-1} - 1\right| + \frac{1}{n-1}\sum_{a \neq a^\star} N^a_{n,a^\star} \\ &\leq \frac{1}{2(n-1)} + \beta\frac{2(M-1)}{n-1} \leq \gamma\,,\end{aligned}$$

where the last inequality is obtained by taking $n \geq \max\{M, (1/2 + 2\beta(M-1))/\gamma + 1\}$. ■

*Convergence for the sub-optimal arms.* Lemma 48 exhibits a random phase which ensures that if a sub-optimal arm overshoots its $\beta$-optimal allocation then it cannot be selected as challenger for large enough $n$.

**Lemma 48** *Let $\gamma > 0$. Let $N_1$ and $C_0$ be as in Lemma 46 and 47. There exists $N_2 \geq C_0 N_1$ with $\mathbb{E}_{\boldsymbol{\nu}}[N_2] < +\infty$ such that, for all $n \geq N_2$,*

$$\exists a \neq a^\star, \quad \frac{N_{n,a}}{n-1} \geq \gamma + \begin{cases}\omega^\star_{\beta,a} & \text{[AdaP-TT]} \\ \omega^\star_{\epsilon,\beta,a} & \text{[AdaP-TT}^\star\text{]}\end{cases} \quad \Longrightarrow \quad C_n \neq a\,,$$

**Proof** Let $\gamma > 0$ and $\tilde{\gamma} > 0$. Let $N_1$ as in Lemma 46 and $C_0$ as in Lemma 47 for $\tilde{\gamma}$. Let $n \geq C_0 N_1$.

Let $a \neq a^\star$ such that $\frac{N_{n,a}}{n-1} \geq \omega^\star_{\beta,a} + \gamma$. Suppose towards contradiction that $\frac{N_{n,b}}{n-1} > \omega^\star_{\beta,a}$ for all $b \notin \{a^\star, a\}$. Then, for all $n \geq C_0 N_1$, we have

$$1 - \beta + \tilde{\gamma} \geq 1 - \frac{N_{n,a^\star}}{n-1} = \sum_{b \neq a^\star}\frac{N_{n,b}}{n-1} > \gamma + \sum_{b \neq a^\star}\omega^\star_{\beta,b} = 1 - \beta + \gamma\,,$$

which yields a contradiction for $\tilde{\gamma} \leq \gamma$. Therefore, for all $n \geq C_0 N_1$, we have

$$\exists a \neq a^\star, \quad \frac{N_{n,a}}{n-1} \geq \omega^\star_{\beta,a} + \gamma \quad \Longrightarrow \quad \exists b \notin \{a^\star, a\}, \quad \frac{N_{n,b}}{n-1} \leq \omega^\star_{\beta,b}\,.$$

Then, we have

$$\sqrt{\frac{1+N_{n,a^\star}/N_{n,b}}{1+N_{n,a^\star}/N_{n,a}}} \geq \sqrt{\frac{1+(\beta-\tilde{\gamma})/\omega^\star_{\beta,b}}{1+(\beta+\tilde{\gamma})/(\omega^\star_{\beta,a}+\gamma)}}\,.$$

In the following, we use Lemma 38 and similar manipulations as in the proof of Lemma 46. Therefore, we obtain that, for all $c \neq a^\star$,

$$\begin{aligned}\left|\tilde{\mu}_{k_{n,a^\star},a^\star} - \tilde{\mu}_{k_{n,c},c} - \Delta_c\right| \leq W_\mu &\left(\sqrt{\frac{\log(e+2^{k_{n,a^\star}-2})}{2^{k_{n,a^\star}-2}}} + \sqrt{\frac{\log(e+2^{k_{n,c}-2})}{2^{k_{n,c}-2}}}\right) \\ &+ W_\epsilon\left(\frac{\log(e+k_{n,a^\star})}{2^{k_{n,a^\star}-2}} + \frac{\log(e+k_{n,c})}{2^{k_{n,c}-2}}\right)\,.\end{aligned}$$

Let $p_3 = \lceil \log_2(X_1 - e)\rceil + 2$ and $p_2 = \lceil \log_2(X_2 - e - 1)\rceil + 2$ with

$$\begin{aligned}X_3 &= \sup\left\{x>1 \mid\ x \leq 16\eta^{-2}W_\mu^2 \log x + e\right\} \leq h_1(16\eta^{-2}W_\mu^2,\, e)\,,\\ X_2 &= \sup\left\{x>1 \mid\ x \leq 4\eta^{-1}W_\epsilon \log x + e + 1\right\} \leq h_1(4\eta^{-1}W_\epsilon,\, e+1)\,,\\ X_2 &\geq \sup\left\{x>1 \mid\ x \leq 4\eta^{-1}W_\epsilon \log(e+2+\log x)\right\}\,,\end{aligned}$$

where we used Lemma 40, and $h_1$ defined therein. We have, for all $\alpha \in \mathbb{R}_+$,

$$\exp(\alpha p_3) \leq e^{3\alpha}(X_3 - e)^{\alpha/\log 2} \quad \text{hence} \quad \mathbb{E}_{\boldsymbol{\nu}}[\exp(\alpha p_3)] < +\infty\,,$$

where we used Lemma 38 and $h_1(x,e) \sim_{x\to+\infty} x\log x$ to obtain that $\exp(\alpha p_3)$ is at most polynomial in $W_\mu$. Likewise, we obtain that $\mathbb{E}_{\boldsymbol{\nu}}[\exp(\alpha p_2)] < +\infty$ for all $\alpha \in \mathbb{R}_+$.

Using Lemma 45 (with $C_0N_1 \geq N_1 \geq N_0$), we obtain that, for all $n \geq C_0N_1$ and all $a \in [K]$, $k_{n,a} \geq \log_2(n/K)/2 + 1$. Therefore, we obtain $\min_{a\in[K]} k_{n,a} > \max\{p_2, p_3\}$ is implied by $n \geq N_2 = \max\{K4^{\max\{p_3,p_2\}}, C_0N_1\}$. Using the above, we conclude that $\mathbb{E}_{\boldsymbol{\nu}}[N_2] < +\infty$ and $\max_{c\neq a^\star} |\tilde{\mu}_{k_{n,a^\star},a^\star} - \tilde{\mu}_{k_{n,c},c} - \Delta_c| \leq \eta$ for all $n \geq N_2$.

Then, for all $n \geq N_2$, we have $B_n = a^\star$ and

$$\frac{\tilde{\mu}_{k_{n,a^\star},a^\star} - \tilde{\mu}_{k_{n,a},a}}{\tilde{\mu}_{k_{n,a^\star},a^\star} - \tilde{\mu}_{k_{n,b},b}}\sqrt{\frac{1+N_{n,a^\star}/N_{n,b}}{1+N_{n,a^\star}/N_{n,a}}} \geq \frac{\Delta_a - \eta}{\Delta_b + \eta}\sqrt{\frac{1+(\beta-\tilde{\gamma})/\omega^\star_{\beta,b}}{1+(\beta+\tilde{\gamma})/(\omega^\star_{\beta,a}+\gamma)}} > 1\,,$$

where the last inequality is obtained by taking $\eta$ and $\tilde{\gamma}$ sufficiently small and by using (23)

$$\frac{\Delta_a}{\Delta_b}\sqrt{\frac{1+\beta/\omega^\star_{\beta,b}}{1+\beta/\omega^\star_{\beta,a}}} = 1\,.$$

Therefore, we have shown that $B_n = a^\star$ and

$$\frac{\tilde{\mu}_{k_{n,a^\star},a^\star} - \tilde{\mu}_{k_{n,a},a}}{\sqrt{1/N_{n,a^\star} + 1/N_{n,a}}} > \frac{\tilde{\mu}_{k_{n,a^\star},a^\star} - \tilde{\mu}_{k_{n,b},b}}{\sqrt{1/N_{n,a^\star} + 1/N_{n,b}}} \quad \text{hence} \quad C_n \neq a\,.$$

This concludes the proof of the first result.

For the AdaP-TT$^\star$ algorithm, the proof is done similarly. As above, we can construct $\tilde{N}_2$ with $\mathbb{E}_{\boldsymbol{\nu}}[\tilde{N}_2] < +\infty$ such that, for all $n \geq \tilde{N}_2$, we have $B_n = a^\star$ and

$$\frac{(\tilde{\mu}_{k_{n,a^\star},a^\star} - \tilde{\mu}_{k_{n,a},a}) \min\{\epsilon/2, \tilde{\mu}_{k_{n,a^\star},a^\star} - \tilde{\mu}_{k_{n,a},a}\}}{(\tilde{\mu}_{k_{n,a^\star},a^\star} - \tilde{\mu}_{k_{n,b},b}) \min\{\epsilon/2, \tilde{\mu}_{k_{n,a^\star},a^\star} - \tilde{\mu}_{k_{n,b},b}\}} \frac{1 + N_{n,a^\star}/N_{n,b}}{1 + N_{n,a^\star}/N_{n,a}}$$
$$\geq \frac{(\Delta_a - \eta) \min\{\epsilon/2, \Delta_a - \eta\}}{(\Delta_b + \eta) \min\{\epsilon/2, \Delta_b + \eta\}} \frac{1 + (\beta - \tilde{\gamma})/\omega^\star_{\epsilon,\beta,b}}{1 + (\beta + \tilde{\gamma})/(\omega^\star_{\epsilon,\beta,a} + \gamma)} > 1 \,,$$

where the last inequality is obtained by taking $\eta$ and $\tilde{\gamma}$ sufficiently small and by using (23)

$$\frac{\Delta_a \min\{\epsilon/2, \Delta_a\}}{\Delta_b \min\{\epsilon/2, \Delta_b\}} \sqrt{\frac{1 + \beta/\omega^\star_{\epsilon,\beta,b}}{1 + \beta/\omega^\star_{\epsilon,\beta,a}}} = 1 \,.$$

Then, we conclude similarly by using the definition of the TC challenger. ■

*Lemma 49 shows that that the pulling proportion of the best arm converges towards $\beta$ for large enough $n$.*

**Lemma 49** *Let $\gamma > 0$ and $T_{\boldsymbol{\mu},\gamma}(w)$ as in Eq. (24). Then, we have $\mathbb{E}_{\boldsymbol{\nu}}[T_{\boldsymbol{\mu},\gamma}(\omega^\star_\beta)] < +\infty$ (AdaP-TT) and $\mathbb{E}_{\boldsymbol{\nu}}[T_{\boldsymbol{\mu},\gamma}(\omega^\star_{\epsilon,\beta})] < +\infty$ (AdaP-TT$^\star$).*

**Proof** Let $\gamma > 0$ and $\tilde{\gamma} > 0$. Let $N_2$ as in Lemma 48 for $\tilde{\gamma}$. Let $M \geq N_2$. Using Lemmas 46, 47 and 48 for all $n \geq M$, we obtain that $B_n = a^\star$, $\left|\frac{N_{n,a^\star}}{n-1} - \beta\right| \leq \tilde{\gamma}$ and

$$\exists a \neq a^\star, \quad \frac{N_{n,a}}{n-1} \geq \omega^\star_{\beta,a} + \tilde{\gamma} \quad \Longrightarrow \quad C_n \neq a \,.$$

For all $a \neq a^\star$, let us define $t_{n,a}(\tilde{\gamma}) = \max\Big\{t \mid M \leq t \leq n,\ N_{t,a}/(n-1) < \omega^\star_{\beta,a} + \tilde{\gamma}\Big\}$. Since $N_{t,a}/(n-1) \leq N_{t,a}/(t-1)$ for $t \leq n$, we have

$$\begin{aligned}
\frac{N_{n,a}}{n-1} &\leq \frac{M-1}{n-1} + \frac{1}{n-1} \sum_{t=M}^{n} \mathbb{1}\left(a_t = C_t = a\right) \\
&\leq \frac{M-1}{n-1} + \frac{1}{n-1} \sum_{t=M}^{n} \mathbb{1}\left(\frac{N_{t,a}}{n-1} < \omega^\star_{\beta,a} + \tilde{\gamma},\ a_t = C_t = a\right) \\
&\leq \frac{M-1}{n-1} + \frac{N_{t_{n,a}(\tilde{\gamma}),a}}{n-1} < \frac{M-1}{n-1} + \omega^\star_{\beta,a} + \tilde{\gamma} \,.
\end{aligned}$$

The second inequality uses Lemma 48, and the two last inequalities use the definition of $t_{n,a}(\tilde{\gamma})$. Using that $\sum_{a\in[K]} \frac{N_{n,a}}{n-1} = \sum_{a\in[K]} \omega^\star_{\beta,a} = 1$, we obtain

$$\frac{N_{n,a}}{n-1} = 1 - \sum_{b\neq a} \frac{N_{n,a}}{n-1} \geq 1 - \sum_{b\neq a} \left(\omega^\star_{\beta,b} + \tilde{\gamma} + \frac{M-1}{n-1}\right) = \omega^\star_{\beta,a} - (K-1)\left(\tilde{\gamma} + \frac{M-1}{n-1}\right) \,.$$

Taking $\tilde{\gamma} \leq \gamma/(2(K-1))$ and $n \geq \max\{M, 2(K-1)(M-1)/\gamma + 1\}$ yields that

$$\left\|\frac{N_n}{n-1} - \omega^\star_\beta\right\|_\infty \leq \gamma \,.$$

Let $T_{\boldsymbol{\mu},\gamma}(w)$ as in Eq. (24). Then, we showed that $T_{\boldsymbol{\mu},\gamma}(\omega^\star_\beta) \leq \max\{M, 2(K-1)(M-1)/\gamma + 1\}$. Therefore, we have

$$\mathbb{E}_{\boldsymbol{\nu}}[T_{\boldsymbol{\mu},\gamma}(\omega^\star_\beta)] \leq \mathbb{E}_{\boldsymbol{\nu}}[\max\{M, 2(K-1)(M-1)/\gamma + 1\}] < +\infty\,,$$

which concludes the proof of the first result.

For the AdaP-TT$^\star$ algorithm, the proof is exactly the same by replacing $\omega^\star_\beta$ by $\omega^\star_{\epsilon,\beta}$. ■

### G.4 Cost of Doubling and Forgetting

Compared to the generic analysis of Top Two algorithms Jourdan et al. (2022), we need to control the sample complexity cost of the DAF($\epsilon$) update (Algorithm 5). Due to this reason, we have to pay a multiplicative four-factor: one two-factor due to doubling, and another two-factor due to forgetting. It is possible to show that this cost exists when adapting any BAI algorithm in which the empirical proportions are converging towards an allocation $\omega$ such that $\min_a \omega_a > 0$, i.e. there exists $\omega$ such that $\mathbb{E}_{\boldsymbol{\nu}}[T_{\boldsymbol{\mu},\gamma}(\omega)] < +\infty$. As shown in Lemma 49, this is the case for the AdaP-TT and AdaP-TT$^\star$ algorithms.

*Lemma 50 shows that the phase switches of the arms happen in a round-robin fashion, which means that an arm switches phase for a second time after all other arms first switch their own phases.*

**Lemma 50** *Let $\omega \in \Sigma_K$ such that $\min_a \omega_a > 0$. Assume that there exists $\gamma_{\boldsymbol{\mu}} > 0$ such that for $\mathbb{E}_{\boldsymbol{\nu}}[T_{\boldsymbol{\mu},\gamma}(\omega)] < +\infty$ for all $\gamma \in (0, \gamma_{\boldsymbol{\mu}})$, where $T_{\boldsymbol{\mu},\gamma}(\omega)$ is defined in Equation (24). Let $\eta > 0$. There exists $\tilde{\gamma}_{\boldsymbol{\mu}} \in (0, \gamma_{\boldsymbol{\mu}})$ such that, for all $\gamma \in (0, \tilde{\gamma}_{\boldsymbol{\mu}})$, there exists $N_3 \geq T_{\boldsymbol{\mu},\gamma}(\omega)$ with $\mathbb{E}_{\boldsymbol{\nu}}[N_3] < +\infty$ which satisfies*

$$\forall n \geq N_3, \quad \frac{\max_{a\in[K]} T_{k_{n,a}}(a) - 1}{\min_{a\in[K]} T_{k_{n,a}}(a) - 1} \leq 2 + \eta\,.$$

**Proof** Let $\eta > 0$. Let $\tilde{\gamma}_{\boldsymbol{\mu}} \in (0, \gamma_{\boldsymbol{\mu}})$ such that $2\max_{a\in[K]}(\omega_a + \gamma)/(\omega_a - \gamma) \leq 2 + \eta$, which is possible since $\min_a \omega_a > 0$. Let $\gamma \in (0, \tilde{\gamma}_{\boldsymbol{\mu}})$. By assumption, we have $\mathbb{E}_{\boldsymbol{\nu}}[T_{\boldsymbol{\mu},\gamma}(\omega)] < +\infty$. Then, for all $n \geq T_{\boldsymbol{\mu},\gamma}(\omega)$,

$$\left\| \frac{N_n}{n-1} - \omega \right\|_\infty \leq \gamma\,.$$

Let $M \geq T_{\boldsymbol{\mu},\gamma}(\omega)$. Let use denote by $k_M = (k_{M,a})_{a\in[K]}$ the current phases for all arms $a \in [K]$ at time $M$. Then, for all $n \geq M$ and all $a \in [K]$, we have $N_{n,a} \geq (n-1)(\omega_a - \gamma)$. Therefore, taking $n \geq \max_{a\in[K]} 2^{k_{M,a}}(\omega_a - \gamma)^{-1} + 1$, we obtain that $N_{n,a} \geq 2^{k_{M,a}}$ for all $a \in [K]$, hence we have $\max_{a\in[K]} T_{k_{M,a}+1}(a) \leq n$. Since $\min_{a\in[K]} T_{k_{M,a}+1}(a) \geq M$, we have

$$\max_{a\in[K]} \left| \frac{N_{T_{k_{M,a}+1}(a),a}}{n-1} - \omega_a \right| \leq \gamma\,.$$

Likewise, taking $n \geq \max_{a\in[K]} 2^{k_{M,a}+1}(\omega_a - \gamma)^{-1} + 1$, we obtain that $N_{n,a} \geq 2^{k_{M,a}+1}$ for all $a \in [K]$, hence we have $\max_{a\in[K]} T_{k_{M,a}+2}(a) \leq n$. Let $a_1 = \arg\min_{a\in[K]} T_{k_{M,a}+2}(a)$. By definition and using Lemma 36, we have

$$2^{k_{M,a_1}+1} = N_{T_{k_{M,a_1}+2}(a_1),a_1} \leq (T_{k_{M,a_1}+2}(a_1) - 1)(\omega_{a_1} + \gamma)\,,$$

$$\forall a \neq a_1, \quad 2^{k_{M,a}} \leq N_{T_{k_{M,a_1}+2}(a_1),a} \leq (T_{k_{M,a_1}+2}(a_1) - 1)(\omega_a + \gamma) \,.$$

Let $a_2 = \arg\max_{a\in[K]} T_{k_{M,a}+2}(a)$. By definition and using Lemma 36, we have

$$2^{k_{M,a_2}+1} = N_{T_{k_{M,a_2}+2}(a_2),a_2} \geq (T_{k_{M,a_2}+2}(a_2) - 1)(\omega_{a_2} - \gamma) \,,$$

Therefore, combining the above yields

$$(T_{k_{M,a_2}+2}(a_2) - 1) \leq (T_{k_{M,a_1}+2}(a_1) - 1) 2\frac{\omega_{a_2} + \gamma}{\omega_{a_2} - \gamma} \leq (T_{k_{M,a_2}+2}(a_2) - 1)(2 + \eta) \,,$$

where the last inequality uses that $\gamma \in (0, \tilde{\gamma}_{\boldsymbol{\mu}})$ and $\tilde{\gamma}_{\boldsymbol{\mu}} \in (0, \gamma_{\boldsymbol{\mu}})$ is such that $2\max_{a\in[K]}(\omega_a + \gamma)/(\omega_a - \gamma) \leq 2 + \eta$. We take $n \geq N_3 = \max_{a\in[K]} T_{k_{M,a}+2}(a)$, hence we have $k_{n,a} \geq k_{M,a} + 2$ for all $a \in [K]$. Since $\mathbb{E}_{\boldsymbol{\nu}}[T_{\boldsymbol{\mu},\gamma}(\omega)] < +\infty$ (i.e. arms are sampled linearly), it is direct to see that $\mathbb{E}_{\boldsymbol{\nu}}[\max_{a\in[K]} T_{k_{M,a}+2}(a)] < +\infty$. This concludes the proof. ■

### G.5 Asymptotic Upper Bound on the Expected Sample Complexity

The final step of the generic analysis of Top Two algorithms (Jourdan et al., 2022) is to invert the private GLR stopping rule by leveraging the convergence of the empirical proportions towards the $\beta$-optimal allocation. Provided this convergence is shown, the asymptotic upper bound on the expected sample complexity only depends on the dependence in $\log(1/\delta)$ of the threshold that ensures $\delta$-correctness. Compared to the non-private GLR stopping rule, the private GLR stopping rules pay an extra cost to ensure privacy. In Section 4.3, the stopping threshold is adapted with an additive term in $\mathcal{O}(\log(1/\delta)^2)$. In Section 4.4, both the stopping threshold and the transportation costs are modified.

**Lemma 51** *Let $(\delta, \beta) \in (0,1)^2$. Assume that there exists $\gamma_{\boldsymbol{\mu}} > 0$ such that $\mathbb{E}_{\boldsymbol{\nu}}[T_{\boldsymbol{\mu},\gamma}(\omega^\star_\beta)] < +\infty$ for all $\gamma \in (0, \gamma_{\boldsymbol{\mu}})$, where $T_{\boldsymbol{\mu},\gamma}(w)$ is defined in Eq. (24). Combining such a sampling rule, using the DAF($\epsilon$) update, with the GLR stopping rule with $W^G_{a,b}$ as in Eq. (3) and the stopping threshold $c^{G,\epsilon}_{a,b}$ as in Eq. (12) yields a $\delta$-correct algorithm which satisfies that, for all $\boldsymbol{\nu}$ with mean $\boldsymbol{\mu}$ such that $|a^\star(\boldsymbol{\mu})| = 1$,*

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta]}{\log(1/\delta)} \leq 4T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}) \left(1 + \sqrt{1 + \frac{\Delta^2_{\max}}{2\epsilon^2\sigma^4}}\right) \,.$$

*where $T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu})$ as in Eq. (5) with $\sigma = 1/2$.*

*Assume that there exists $\gamma_{\boldsymbol{\mu}} > 0$ such that $\mathbb{E}_{\boldsymbol{\nu}}[T_{\boldsymbol{\mu},\gamma}(\omega^\star_{\epsilon,\beta})] < +\infty$. Combining such a sampling rule, using the DAF($\epsilon$) update, with the GLR stopping rule with $W^{G,\epsilon}_{a,b}$ as in Eq. (13) and the stopping threshold $\tilde{c}^{G,\epsilon}_{a,b}$ as in Eq. (14) yields a $\delta$-correct algorithm which satisfies that, for all $\boldsymbol{\nu}$ with mean $\boldsymbol{\mu}$ such that $|a^\star(\boldsymbol{\mu})| = 1$,*

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta]}{\log(1/\delta)} \leq \begin{cases} 4T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}) g_1\left(\Delta_{\max}/(\sigma^2\epsilon)\right) & \text{if } \Delta_{\max} < 3\epsilon \\ 12T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}_{G,\epsilon}) g_2(3\epsilon^2 T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}_{G,\epsilon}) \max\{\beta, 1-\beta\}/2)/\sigma^2 & \text{otherwise} \end{cases} ,$$

*where* $T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}_{G,\epsilon})$ *as in Eq.* (15). *The function* $g_1(y) = \sup\left\{x \mid x^2 < x + y\sqrt{2x} + \frac{y^2}{4}\right\}$ *is increasing on* $[0, 12]$ *and satisfies that* $g_1(0) = 1$ *and* $g_1(12) \leq 10$. *The function* $g_2(y) = 1 + 2(\sqrt{1 + 1/y} - 1)^{-1}$ *is increasing on* $\mathbb{R}_+^\star$ *and satisfies that* $\lim_{y\to 0} g_2(y) = 1$.

**Proof** Lemma 20 and Lemma 22 yields the $\delta$-correctness of both algorithms.

AdaP-TT *algorithm.* Let $\zeta > 0$, $a^\star$ be the unique best arm. Using (23) and the continuity of

$$(\boldsymbol{\mu}, w) \mapsto \min_{a \neq a^\star(\boldsymbol{\mu})} \frac{(\mu_{a^\star(\boldsymbol{\mu})} - \mu_a)^2}{2\sigma^2(1/w_{a^\star(\boldsymbol{\mu})} + 1/w_a)}$$

yields that there exists $\gamma_\zeta > 0$ such that $\left\|\frac{N_n}{n-1} - \omega^\star_\beta\right\|_\infty \leq \gamma_\zeta$ and $\max_{a\in[K]} |\tilde{\mu}_{k_{n,a}+1,a} - \mu_a| \leq \gamma_\zeta$ implies that

$$\forall a \neq a^\star, \quad \frac{(\tilde{\mu}_{k_{n,a^\star}+1,a^\star} - \tilde{\mu}_{k_{n,a}+1,a})^2}{(n-1)/N_{n,a^\star} + (n-1)/N_{n,a}} \geq \frac{2\sigma^2(1-\zeta)}{T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu})}\,,$$
$$\frac{n-1}{N_{n,a^\star}} + \frac{n-1}{N_{n,a}} \leq \frac{\Delta_a^2}{2\sigma^2}(1+\zeta)T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu})\,.$$

We choose such a $\gamma_\zeta$. Let $\gamma_{\boldsymbol{\mu}} > 0$ be such that for $\mathbb{E}_{\boldsymbol{\nu}}[T_{\boldsymbol{\mu},\gamma}(\omega^\star_\beta)] < +\infty$ for all $\gamma \in (0, \gamma_{\boldsymbol{\mu}})$, where $T_{\boldsymbol{\mu},\gamma}(\omega)$ is defined in Eq. (24). Let $\eta > 0$. Let $\tilde{\gamma}_{\boldsymbol{\mu}} \in (0, \gamma_{\boldsymbol{\mu}})$ as in Lemma 50 for this $\eta$. In the following, let us consider $\gamma \in (0, \min\{\tilde{\gamma}_{\boldsymbol{\mu}}, \gamma_\zeta, \beta/4, \Delta_{\min}/4\})$.

Let $N_3 \geq T_{\boldsymbol{\mu},\gamma}(\omega^\star_\beta)$ with $\mathbb{E}_{\boldsymbol{\nu}}[N_3] < +\infty$ as Lemma 50 for those $(\gamma, \eta)$. Then, we have $\mathbb{E}_{\boldsymbol{\nu}}[T_{\boldsymbol{\mu},\gamma}(\omega^\star_\beta)] < +\infty$ and

$$\forall n \geq N_3, \quad \frac{\max_{a\in[K]} T_{k_{n,a}}(a) - 1}{\min_{a\in[K]} T_{k_{n,a}}(a) - 1} \leq 2 + \eta\,.$$

Since arms are sampled linearly, it is direct to construct $N_4 \geq N_3$ with $\mathbb{E}_{\boldsymbol{\nu}}[N_4] < +\infty$ such that, for all $n \geq N_4$, we have $\max_{a\in[K]} \max_{k\in\{k_{n,a}, k_{n,a}+1\}} |\tilde{\mu}_{k,a} - \mu_a| \leq \gamma$, Therefore, we have $\hat{a}_n = a^\star$.

Let $\kappa \in (0, 1)$. Let $n \geq N_4/\kappa$ and $(k_{n,a})_{a\in[K]}$ be the current phases at time $n$. Combining the above, we have $\hat{a}_n = a^\star$ and

$$\max_{a\in[K]} |\tilde{\mu}_{k_{n,a}+1,a} - \mu_a| \leq \gamma \quad , \quad \left\|\frac{N_n}{n-1} - \omega^\star_\beta\right\|_\infty \leq \gamma \quad \text{and} \quad \frac{\max_{a\in[K]} T_{k_{n,a}}(a) - 1}{\min_{a\in[K]} T_{k_{n,a}}(a) - 1} \leq 2 + \eta\,.$$

Let $a_1 = \arg\min_{a\in[K]} T_{k_{n,a}}(a)$ and $a_2 = \arg\max_{a\in[K]} T_{k_{n,a}}(a)$. Therefore, we obtain

$$\begin{aligned}
\forall a \neq a^\star, \quad \frac{(\tilde{\mu}_{k_{n,\hat{a}_n}+1,\hat{a}_n} - \tilde{\mu}_{k_{n,a}+1,a})^2}{1/\tilde{N}_{k_{n,\hat{a}_n}+1,\hat{a}_n} + 1/\tilde{N}_{k_{n,a}+1,a}} &= \frac{(\tilde{\mu}_{k_{n,a^\star}+1,a^\star} - \tilde{\mu}_{k_{n,a}+1,a})^2}{1/N_{T_{k_{n,a^\star}}(a^\star),a^\star} + 1/N_{T_{k_{n,a}}(a),a}} \\
&\geq \frac{(\tilde{\mu}_{k_{n,a^\star}+1,a^\star} - \tilde{\mu}_{k_{n,a}+1,a})^2}{1/N_{T_{k_{n,a_1}}(a_1),a^\star} + 1/N_{T_{k_{n,a_1}}(a_1),a}} \\
&\geq (\min_{a\in[K]} T_{k_{n,a}}(a) - 1)\frac{2\sigma^2(1-\zeta)}{T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu})}\,.
\end{aligned}$$

Similarly, we can show that, for all $a \neq a^\star$,

$$\begin{aligned}\frac{1}{\tilde{N}_{k_{n,a^\star}+1,a^\star}} + \frac{1}{\tilde{N}_{k_{n,a}+1,a}} &= \frac{1}{N_{T_{k_{n,a^\star}}(a^\star),a^\star}} + \frac{1}{N_{T_{k_{n,a}}(a),a}} \\ &\leq \frac{1}{N_{T_{k_{n,a_1}}(a_1),a^\star}} + \frac{1}{N_{T_{k_{n,a_1}}(a_1),a}} \\ &\leq \frac{1}{\min_{a\in[K]} T_{k_{n,a}}(a) - 1} \frac{\Delta_a^2}{2\sigma^2}(1+\zeta) T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}) \\ &\leq \frac{1}{\min_{a\in[K]} T_{k_{n,a}}(a) - 1} \frac{\Delta_{\max}^2}{2\sigma^2}(1+\zeta) T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}) \,.\end{aligned}$$

Let $c^G_{a,b}$ as in Eq. (4). Using Lemma 36, we obtain, for all $a \neq a^\star$,

$$\begin{aligned}c^G_{a^\star,a}(\tilde{N}_{k_n+1}, \delta(2\zeta(s)^2(k_{n,a^\star}+1)^s(k_{n,a}+1)^s)^{-1}) &\leq 4\log(4 + (\max_{b\in[K]} k_{n,b} - 1)\log 2) \\ + 2\mathcal{C}_G\left(\log(1/\delta)/2 + s\log(\max_{b\in[K]} k_{n,b} - 1) + \log(2(K-1)\zeta(s)^2)/2\right)&\end{aligned}$$

Likewise, we obtain, for all $a \in [K]$,

$$\begin{aligned}&\frac{1}{\epsilon^2\sigma^2} \sum_{c\in\{a^\star,a\}} \frac{1}{\tilde{N}_{k_{n,c},c}} \left(\log \frac{2K\zeta(s)(k_{n,c}+1)^s}{\delta}\right)^2 \\ &\leq \frac{\Delta_{\max}^2}{2\epsilon^2\sigma^4} \frac{(1+\zeta)T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu})}{\min_{b\in[K]} T_{k_{n,b}}(b) - 1} \left(\log(1/\delta) + s\log(\max_{b\in[K]} k_{n,b} + 1) + \log(2K\zeta(s))\right)^2\end{aligned}$$

Let us denote by $T^+_{k_n+1} = \max_{b\in[K]} T_{k_{n,b}+1}(b)$, $T^+_{k_n+2} = \max_{b\in[K]} T_{k_{n,b}+2}(b)$, $T^-_{k_n+1} = \min_{b\in[K]} T_{k_{n,b}+1}(b)$, $T^-_{k_n} = \min_{b\in[K]} T_{k_{n,b}}(b)$. Let $T$ be a time such that $T \geq T^+_{k_n+1} \geq \kappa T$. Using Lemmas 36 and 50, we have

$$(k_{n,b}-1)\log 2 = \log N_{T_{k_{n,b}}(b),b} \leq \log T_{k_{n,b}}(b) \leq \log T^+_{k_n} \leq \log T^-_{k_n} + \log(2+\eta) \,.$$

Using the DAF($\epsilon$) update with the GLR stopping rule with $W^G_{a,b}$ as in Eq. (3) and the stopping threshold $c^{G,\epsilon}_{a,b}$ as in Eq. (12), we have

$$\begin{aligned}&\min\{\tau_\delta, T\} - \kappa T \leq \sum_{T\geq T^+_{k_n}\geq\kappa T} (T^+_{k_n+2} - T^+_{k_n+1}) \mathbb{1}\left(\tau_\delta > T^+_{k_n+1}\right) \\ &\leq \sum_{T^+_{k_n}=\kappa T}^{T} (T^+_{k_n+2} - T^+_{k_n+1}) \mathbb{1}\left(\exists a \neq a^\star, \frac{(\tilde{\mu}_{k_{n,a^\star}+1,a^\star} - \tilde{\mu}_{k_{n,a}+1,a})^2}{2\sigma^2\left(\frac{1}{\tilde{N}_{k_{n,a^\star}+1,a^\star}} + \frac{1}{\tilde{N}_{k_{n,a}+1,a}}\right)} < c^{G,\epsilon}_{a^\star,a}(\tilde{N}_{k_n+1},\delta)\right) \\ &\leq \sum_{T\geq T^+_{k_n}\geq\kappa T} (T^+_{k_n+2} - T^+_{k_n+1}) \mathbb{1}\left((T^-_{k_n}-1)\frac{1-\zeta}{T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu})} < 8\log(4 + \log T^-_{k_n} + \log(2+\eta))\right.\end{aligned}$$

$$+ 4\mathcal{C}_G\left(\log(1/\delta)/2 + s\log(2+\log_2 T^-_{k_n} + \log_2(2+\eta)) + \log(2(K-1)\zeta(s)^2)/2\right)$$
$$+\frac{\Delta^2_{\max}}{2\epsilon^2\sigma^4}\frac{(1+\zeta)T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu})}{T^-_{k_n}-1}\left(\log(1/\delta) + s\log(2+\log_2 T^-_{k_n} + \log_2(2+\eta)) + \log(2K\zeta(s))\right)^2\Bigg) ,$$

Let $T_\zeta(\delta)$ defined as the largest deterministic time such that the above condition is satisfied when replacing $T^-_{k_n}$ by $(1-\kappa)T$. Let $k_\delta$ be the largest random vector of phases such that that $T^+_{k_\delta+1} \le T_\zeta(\delta)$ almost surely, hence $T^+_{k_\delta+2} > T_\zeta(\delta)$ almost surely. Then, using the above yields that $\tau_\delta \le T^+_{k_\delta+2}$ almost surely, hence

$$\limsup_{\delta\to 0}\frac{\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta]}{\log(1/\delta)} \le \limsup_{\delta\to 0}\frac{\mathbb{E}_{\boldsymbol{\nu}}[T^+_{k_\delta+2}]}{\log(1/\delta)} \le (2+\eta)^2\limsup_{\delta\to 0}\frac{\mathbb{E}_{\boldsymbol{\nu}}[T^+_{k_\delta+1}]}{\log(1/\delta)} \le (2+\eta)^2\limsup_{\delta\to 0}\frac{T_\zeta(\delta)}{\log(1/\delta)}$$

where the second inequality uses Lemma 50 twice, i.e. $T^+_{k_\delta+2} \le (2+\eta)T^-_{k_\delta+2} \le (2+\eta)^2 T^+_{k_\delta+1}$, and the last one used the definition of $k_\delta$ and that $T_\zeta(\delta)$ is deterministic.

Since we are only interested in upper bounding $\limsup_{\delta\to 0}\frac{T_\zeta(\delta)}{\log(1/\delta)}$, we can safely drop the second orders terms in $T$ and $\log(1/\delta)$. This allows us to remove the terms in $\mathcal{O}(\log\log T)$ and in $\mathcal{O}(\log\log(1/\delta))$. Using that $\mathcal{C}_G(x) = x + \mathcal{O}(\log x)$, tedious manipulations yields that

$$\limsup_{\delta\to 0}\frac{T_\zeta(\delta)}{\log(1/\delta)} \le \frac{T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu})}{1-\kappa}D_\zeta(\mu,\epsilon) ,$$

where

$$D_\zeta(\mu,\epsilon) = \sup\left\{x \mid x^2 < \frac{2}{1-\zeta}x + \frac{1+\zeta}{1-\zeta}\frac{\Delta^2_{\max}}{2\epsilon^2\sigma^4}\right\} \le \frac{1}{1-\zeta}\left(1+\sqrt{1+(1-\zeta^2)\frac{\Delta^2_{\max}}{2\epsilon^2\sigma^4}}\right) .$$

The last inequality uses that $x^2 - 2bx - c < 0$ for all $x \in [0, b(1+\sqrt{1+c/b^2}))$. Therefore, we have shown that

$$\limsup_{\delta\to 0}\frac{\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta]}{\log(1/\delta)} \le (2+\eta)^2\frac{T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu})}{(1-\kappa)(1-\zeta)}\left(1+\sqrt{1+(1-\zeta^2)\frac{\Delta^2_{\max}}{2\epsilon^2\sigma^4}}\right) .$$

Letting $\kappa$, $\eta$ and $\zeta$ goes to zero concludes the proof of the first result.

AdaP-TT$^\star$ *algorithm.* For the AdaP-TT$^\star$ algorithm, the proof is done with similar arguments. Using (23) and the continuity of $(\boldsymbol{\mu}, w) \to \min_{a\neq a^\star} W^{G,\epsilon}_{a^\star,a}(\boldsymbol{\mu}, w)$, defined in Eq. (13), we obtain another $\gamma_\zeta > 0$ such that $\left\|\frac{N_n}{n-1} - \omega^\star_{\epsilon,\beta}\right\|_\infty \le \gamma_\zeta$ and $\max_{a\in[K]}|\tilde{\mu}_{k_{n,a}+1,a} - \mu_a| \le \gamma_\zeta$ implies that

$$\forall a \neq a^\star, \quad \frac{(\tilde{\mu}_{k_{n,a^\star}+1,a^\star} - \tilde{\mu}_{k_{n,a}+1,a})\min\{\epsilon/2, \tilde{\mu}_{k_{n,a^\star}+1,a^\star} - \tilde{\mu}_{k_{n,a}+1,a}\}}{(n-1)/N_{n,a^\star} + (n-1)/N_{n,a}} \ge \frac{2\sigma^2(1-\zeta)}{T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}_{G,\epsilon})} ,$$
$$\frac{n-1}{N_{n,a^\star}} + \frac{n-1}{N_{n,a}} \le \frac{\Delta_a\min\{\epsilon/2,\Delta_a\}}{2\sigma^2}(1+\zeta)T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}_{G,\epsilon}) .$$

We choose such a $\gamma_\zeta$. Let $\gamma_{\boldsymbol{\mu}} > 0$ be such that for $\mathbb{E}_{\boldsymbol{\nu}}[T_{\boldsymbol{\mu},\gamma}(\omega^\star_{\epsilon,\beta})] < +\infty$ for all $\gamma \in (0,\gamma_{\boldsymbol{\mu}})$. Let $\eta > 0$. Let $\tilde{\gamma}_{\boldsymbol{\mu}} \in (0,\gamma_{\boldsymbol{\mu}})$ as in Lemma 50 for this $\eta$. In the following, let us consider $\gamma \in (0, \min\{\tilde{\gamma}_\mu, \gamma_\zeta, \beta/4, \Delta_{\min}/4, (\epsilon/2 - \max_{a,\Delta_a<\epsilon/2}\Delta_a)/2\})$.

Let $\kappa \in (0,1)$. As above, we can construct $N_3$ with Lemma 50 and $N_4 \geq N_3$ such that $\mathbb{E}_\nu[N_4] < +\infty$. Let $n \geq N_4/\kappa$ and $(k_{n,a})_{a\in[K]}$ the current phases. Then, we have $\hat{a}_n = a^\star$,

$$\max_{a\in[K]} |\tilde{\mu}_{k_{n,a}+1,a} - \mu_a| \leq \gamma\,, \quad \left\| \frac{N_n}{n-1} - \omega^\star_{\epsilon,\beta} \right\|_\infty \leq \gamma \text{ and } \frac{\max_{a\in[K]} T_{k_{n,a}}(a) - 1}{\min_{a\in[K]} T_{k_{n,a}}(a) - 1} \leq 2 + \eta\,.$$

Depending on the value of the private empirical gap, the stopping condition that is checked is different. For all $a \neq a^\star$ such that $\Delta_a < \epsilon/2$, we have $\tilde{\mu}_{k_{n,a^\star}+1,a^\star} - \tilde{\mu}_{k_{n,a}+1,a} \leq \Delta_a + 2\gamma < \epsilon/2$. For all $a \neq a^\star$ such that $\Delta_a \geq \epsilon/2$, we have either $\tilde{\mu}_{k_{n,a^\star}+1,a^\star} - \tilde{\mu}_{k_{n,a}+1,a} \geq \epsilon/2$ or $\tilde{\mu}_{k_{n,a^\star}+1,a^\star} - \tilde{\mu}_{k_{n,a}+1,a} \leq \epsilon/2$ and $\Delta_a \leq \epsilon/2 + 2\gamma$. Let $a_1 = \arg\min_{a\in[K]} T_{k_{n,a}}(a)$ and $a_2 = \arg\max_{a\in[K]} T_{k_{n,a}}(a)$. Therefore, we obtain similarly that, for all $a \neq a^\star$,

$$\frac{(\tilde{\mu}_{k_{n,a^\star}+1,a^\star} - \tilde{\mu}_{k_{n,a}+1,a}) \min\{\epsilon/2, \tilde{\mu}_{k_{n,a^\star}+1,a^\star} - \tilde{\mu}_{k_{n,a}+1,a}\}}{1/\tilde{N}_{k_{n,a^\star}+1,a^\star} + 1/\tilde{N}_{k_{n,a}+1,a}} \geq (\min_{b\in[K]} T_{k_{n,b}}(b) - 1) \frac{2\sigma^2(1-\zeta)}{T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}_{G,\epsilon})}\,.$$

Similarly, for all $a \neq a^\star$ such that $\tilde{\mu}_{k_{n,a^\star}+1,a^\star} - \tilde{\mu}_{k_{n,a}+1,a} \leq \epsilon/2$, hence $\Delta_a \leq \epsilon/2 + 2\gamma$, we have

$$\begin{aligned}
&\frac{1}{\tilde{N}_{k_{n,a^\star}+1,a^\star}} + \frac{1}{\tilde{N}_{k_{n,a}+1,a}} \leq \frac{1}{\min_{b\in[K]} T_{k_{n,b}}(b) - 1} \frac{\Delta_a \min\{\epsilon/2, \Delta_a\}}{2\sigma^2} (1+\zeta) T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}_{G,\epsilon})\,,\\
&\frac{1}{\sqrt{\tilde{N}_{k_{n,a^\star}+1,a^\star}}} + \frac{1}{\sqrt{\tilde{N}_{k_{n,a}+1,a}}}\\
&\leq \sqrt{\frac{1}{(\min_{b\in[K]} T_{k_{n,b}}(b) - 1)} \frac{\Delta_a \min\{\epsilon/2, \Delta_a\}}{\sigma^2} (1+\zeta) T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}_{G,\epsilon})},\\
&\frac{1}{2\epsilon^2\sigma^2} \sum_{c\in\{a^\star,a\}} \frac{1}{\tilde{N}_{k_{n,c}+1,c}} \left( \log \frac{3K(k_{n,c}+1)^s \zeta(s)}{\delta} \right)^2\\
&\leq \frac{\Delta_a \min\{\epsilon/2, \Delta_a\}(1+\zeta) T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}_{G,\epsilon})}{4\epsilon^2\sigma^4 (\min_{b\in[K]} T_{k_{n,b}}(b) - 1)} \left( \log(1/\delta) + s \log(\max_{b\in[K]} k_{n,b} + 1) + \log(3K\zeta(s)) \right)^2\,,\\
&\frac{\sqrt{2}}{\epsilon\sigma} \sum_{c\in\{a^\star,a\}} \sqrt{\frac{h(\tilde{N}_{k_{n,c}+1,c}, \delta)}{\tilde{N}_{k_{n,c}+1,c}}} \log \left( \frac{3K\zeta(s)(k_{n,c}+1)^s}{\delta} \right)\\
&\leq \sqrt{\frac{2\Delta_a \min\{\epsilon/2, \Delta_a\}(1+\zeta) T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}_{G,\epsilon})}{\epsilon^2\sigma^4 (\min_{b\in[K]} T_{k_{n,b}}(b) - 1)}}\\
&\quad \sqrt{h(2^{\max_{b\in[K]} k_{n,b}-1}, \delta)} \left( \log(1/\delta) + s \log(\max_{b\in[K]} k_{n,b} + 1) + \log(3K\zeta(s)) \right)\,.
\end{aligned}$$

Moreover, for all $a \neq a^\star$ such that $\tilde{\mu}_{k_{n,a^\star}+1,a^\star} - \tilde{\mu}_{k_{n,a}+1,a} \geq 3\epsilon$, hence $\Delta_a \geq 3\epsilon$, we have

$$\begin{aligned}
&\sqrt{\tilde{N}_{k_{n,a}+1,a}} + \sqrt{\tilde{N}_{k_{n,a^\star}+1,a^\star}} \leq \sqrt{T_{k_{n,a_2}}(a_2) - 1} \left( \sqrt{\beta+\gamma} + \sqrt{\omega^\star_{\epsilon,\beta,a} + \gamma} \right)\\
&\quad \leq \sqrt{\min_{b\in[K]} T_{k_{n,b}}(b) - 1} \sqrt{2(2+\eta)(\max\{\beta, 1-\beta\} + \gamma)}\,,
\end{aligned}$$

$$\frac{\epsilon}{2\sqrt{2\sigma^2}} \sum_{c\in\{a^\star,a\}} \sqrt{\tilde{N}_{k_{n,c}+1,c} h(\tilde{N}_{k_{n,c}+1,c},\delta)}$$
$$\leq \sqrt{h(2^{\max_{b\in[K]} k_{n,b}-1},\delta)} \frac{\epsilon}{2\sigma} \sqrt{\min_{b\in[K]} T_{k_{n,b}}(b)-1} \sqrt{(2+\eta)(\max\{\beta,1-\beta\}+\gamma)}\,.$$

Let $T^+_{k_n+1} = \max_b T_{k_{n,b}+1}(b)$, $T^+_{k_n+2} = \max_b T_{k_{n,b}+2}(b)$, $T^-_{k_n+1} = \min_b T_{k_{n,b}+1}(b)$, $T^-_{k_n} = \min_b T_{k_{n,b}}(b)$. Let $T$ be a time such that $T \geq T^+_{k_n+1} \geq \kappa T$. Then, $(\max_b k_{n,b}-1)\log 2 \leq \log T^-_{k_n} + \log(2+\eta)$. As above, using the DAF($\epsilon$) update with the GLR stopping rule with $W^{G,\epsilon}_{a,b}$ as in Eq. (13) and the stopping threshold $\tilde{c}^{G,\epsilon}_{a,b}$ as in Eq. (14), we have

$$\min\{\tau_\delta, T\} - \kappa T \leq \sum_{T\geq T^+_{k_n}\geq \kappa T} (T^+_{k_n+2} - T^+_{k_n+1}) \mathbb{1}\left(\tau_\delta > T^+_{k_n+1}\right)$$
$$\leq \sum_{T\geq T^+_{k_n}\geq \kappa T} (T^+_{k_n+2} - T^+_{k_n+1}) \mathbb{1}\left(\exists a \neq a^\star,\right.$$
$$\left(\tilde{\mu}_{k_{n,a^\star}+1,a^\star} - \tilde{\mu}_{k_{n,a}+1,a} < \epsilon/2, \frac{(\tilde{\mu}_{k_{n,a^\star}+1,a^\star} - \tilde{\mu}_{k_{n,a}+1,a})^2}{2\sigma^2(1/\tilde{N}_{k_{n,a^\star}+1,a^\star} + 1/\tilde{N}_{k_{n,a}+1,a})} < \tilde{c}^{G,\epsilon}_{a^\star,a}(\tilde{N}_{k_n+1},\delta)\right) \vee$$
$$\left.\left(\tilde{\mu}_{k_{n,a^\star}+1,a^\star} - \tilde{\mu}_{k_{n,a}+1,a} \geq \epsilon/2, \frac{\epsilon(\tilde{\mu}_{k_{n,a^\star}+1,a^\star} - \tilde{\mu}_{k_{n,a}+1,a})}{4\sigma^2(1/\tilde{N}_{k_{n,a^\star}+1,a^\star} + 1/\tilde{N}_{k_{n,a}+1,a})} < \tilde{c}^{G,\epsilon}_{a^\star,a}(\tilde{N}_{k_n+1},\delta)\right)\right)\,.$$

Leveraging the inequalities explicited above, we can upper bound it by a condition which only involves $T^-_{k_n}$ and problem dependent quantities (in a highly convoluted fashion). As above, we define $T_\zeta(\delta)$ as the largest deterministic time such that the above condition is satisfied when replacing $T^-_{k_n}$ by $(1-\kappa)T$. Then, we obtain similarly that

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_\nu[\tau_\delta]}{\log(1/\delta)} \leq (2+\eta)^2 \limsup_{\delta\to 0} \frac{T_\zeta(\delta)}{\log(1/\delta)}\,.$$

Droping the second orders terms in $T$ and $\log(1/\delta)$ and using that $\mathcal{C}_G(x) = x + \mathcal{O}(\log x)$ and $\overline{W}_{-1}(x) = x + \mathcal{O}(\log x)$, tedious manipulations yields that

$$\limsup_{\delta\to 0} \frac{T_\zeta(\delta)}{\log(1/\delta)} \leq \frac{T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}_{G,\epsilon})}{1-\kappa} \max\{D^{(1)}_{\gamma,\zeta}(\mu,\epsilon), \mathbb{1}\left(\Delta_{\max} \geq \epsilon/2\right) D^{(2)}_{\gamma,\zeta,\eta}(\mu,\epsilon)\}\,,$$

where $g(x,y) = \sqrt{2xy} + y/4$ and

$$D^{(1)}_{\gamma,\zeta}(\mu,\epsilon) = \sup\left\{x \mid x^2(1-\zeta) < x + g\left(x, \frac{1+\zeta}{\epsilon^2\sigma^4} \max_{\Delta_a \leq \epsilon/2+2\gamma} \Delta_a \min\{\Delta_a, \epsilon/2\}\right)\right\}\,,$$
$$D^{(2)}_{\gamma,\zeta,\eta}(\mu,\epsilon) = \sup\left\{x \mid 2x\sigma(1-\zeta) < \epsilon\sqrt{x}\sqrt{(2+\eta)T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}_{G,\epsilon})(\max\{\beta,1-\beta\}+\gamma)} + \frac{1}{\sigma}\right\}\,.$$

Letting $\kappa$, $\gamma$, $\eta$ and $\zeta$ goes to zero yields that

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_\nu\left[\tau_\delta\right]}{\log(1/\delta)} \leq 4T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}_{G,\epsilon}) \max\{D^{(1)}_{0,0}(\mu,\epsilon), \mathbb{1}\left(\Delta_{\max} \geq \epsilon/2\right) D^{(2)}_{0,0,0}(\mu,\epsilon)\}\,.$$

Using that $x^2 - 2bx - c < 0$ for all $x \in [0, b(1+\sqrt{1+c/b^2}))$, we obtain that

$$D^{(2)}_{0,0,0}(\mu,\epsilon) = \left(\sup\left\{x \mid x^2 < \sqrt{\epsilon^2\sigma^{-2}T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}_{G,\epsilon})\max\{\beta,1-\beta\}/2}x + \sigma^{-2}/2\right\}\right)^2$$
$$\leq \frac{\epsilon^2}{8\sigma^2}T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}_{G,\epsilon})\max\{\beta,1-\beta\}\left(1+\sqrt{1+\frac{4}{\epsilon^2 T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}_{G,\epsilon})\max\{\beta,1-\beta\}}}\right)^2,$$
$$D^{(1)}_{0,0}(\mu,\epsilon) = g_1\left(\max_{\Delta_a \leq 3\epsilon}\frac{\Delta_a}{\sigma^2\epsilon}\right) \quad \text{with} \quad g_1(y) = \sup\left\{x \mid x^2 < x + y\sqrt{2x} + y^2/4\right\}.$$

In more details, we have

$$g_1(0) = \sup\left\{x \mid x^2 < x\right\} = 1 \quad \text{and} \quad g_1(12) = \sup\left\{x \mid x^2 < x + 12\sqrt{2x} + 36\right\} \leq 10\,,$$

where the last inequality is obtained by numerical analysis. The function $g_2$ is obtained by noting that $y(1+\sqrt{1+1/y})^2 = 1 + 2(\sqrt{1+1/y}-1)^{-1}$. When $\Delta_{\max} < \epsilon/2$, we have $T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}_{G,\epsilon}) = T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu})$ where $T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu})$ as in Eq. (5) with $\sigma = 1/2$. This concludes the proof of the second result. ■

*Concluding the proof of Theorems 21 and 23.* Combining Lemmas 45, 49, 50 and 51 concludes the proof of Theorems 21 and 23. We restrict the result to instances such that $\min_{a\neq b}|\mu_a - \mu_b| > 0$ in order for Lemma 45 to hold. Note that this is an artifact of the asymptotic proof which could be alleviated with more careful considerations. ■

## Appendix H. On the Number of Rounds of Adaptivity

Due to its generality, using the DAF update yields a batched version of any existing FC-BAI algorithm, which satisfies $\epsilon$-global DP. At the end of the episode of arm $a$ (after updating its mean), it is possible to compute the sequence of all the arms to be pulled before the end of the next episode (for another arm), without taking the collected observations into account. In contrast to the classical batched setting where the batch size is fixed, the size of the resulting batches is adaptive and data-dependent.

Let $C(\tau_\delta) = \sum_{a\in[K]} k_{\tau_\delta,a}$ be the number of rounds of adaptivity, where $k_{\tau_\delta,a}$ denotes the number of episodes of arm $a \in [K]$ at stopping time. Using Jensen's inequality, the number of rounds of adaptivity is upper bounded by $\mathbb{E}_{\boldsymbol{\nu}}[C(\tau_\delta)] \leq K\log_2 \mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta]$. Therefore, any upper bound on the expected sample complexity directly implies an upper bound on the number of rounds of adaptivity.

*One global episode.* The multiplicative factor $K$ is incurred because DAF maintains one episode per arm. Alternatively, one can consider one global episode $k_n$. Formally, we switch phase as soon as all the arms have doubled their empirical counts, i.e. $N_{n,a} \geq 2N_{T_{k_n},a}$ for all $a \in [K]$. This modification allows to shave the $K$ factor since $\mathbb{E}_{\boldsymbol{\nu}}[C(\tau_\delta)] \leq \log_2 \mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta]$. When using one global episode, one can show the same asymptotic upper bound as when we used one episode per arm.

Empirically, the performance is worsen by considering one global episode, hence we recommend to use one episode per arm. A sub-optimal arm $a$ might be sampled more than $a^\star$ in early stage due to unlucky first draws. When there is only one global episode,

**Algorithm 11** Doubling-Per-Arm (DPA)

**Input:** History $\mathcal{H}_n$, arm $a \in [K]$.
**Initialization:** For all $a \in [K]$, $T_1(a) = K+1$ and $k_{K+1,a} = 1$;
**if** $N_{n,a} \geq 2N_{T_{k_{n,a}}(a),a}$ **then**
  Change phase $k_{n,a} \leftarrow k_{n,a} + 1$ for this arm $a$;
  Set $T_{k_{n,a}}(a) = n$ and $\hat{\mu}_{k_{n,a},a} = N^{-1}_{T_{k_{n,a}}(a),a} \sum_{t \in [T_{k_{n,a}}(a)-1]} r_t \mathbb{1}\{a_t = a\}$;
**end if**
**Return** $(\hat{\mu}_{n,a}, N_{n,a})$;

the learner will always have to double the counts of this sub-optimal arm before updating its estimators of the other arms. After realizing that this arm is sub-optimal, it won't be sampled frequently, hence many samples should be collected before ending the episode.

*Batched best arm identification* In the non-private setting ($\epsilon = +\infty$), we recover Batched Best-Arm Identification (BBAI) in the fixed-confidence setting. One of the question arising in this setting is the following: *Can we solve the BBAI problem with asymptotically optimal sample complexity (up to a constant factor) and a small number of batches?* A slight modification of the above result provides a positive answer.

Without the privacy constraint, there is no need to forget about past observations or to add Laplacian noise. Therefore, the DPA update is better to suited for BBAI than the DAF one. Using the DPA update yields an adaptive batched version of any existing FC-BAI algorithm. It is direct to see that the same analysis can be used to study TTUCB with DPA update. Namely, it yields a $\delta$-correct algorithm such that, for all $\mu$ with distinct means,

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta]}{\log(1/\delta)} \leq 2T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu}) \,, \; \limsup_{\delta \to 0} \left(\mathbb{E}_{\boldsymbol{\nu}}[C(\tau_\delta)] - K \log_2 \log(1/\delta)\right) \leq K \log_2(2T^\star_{\mathrm{KL},\beta}(\boldsymbol{\nu})) \,.$$

For $\beta = 1/2$, the algorithm is asymptotically optimal (up to a multiplicative factor 4) with solely $\mathcal{O}\left(K \log_2(T^\star_{\mathrm{KL}}(\boldsymbol{\nu}) \log(1/\delta))\right)$ rounds of adaptivity.

There are already several works studying BBAI (Karnin et al., 2013; Jin et al., 2019, 2024), see Table 1 in Jin et al. (2024) for a detailed comparison. Building on the Exponential-Gap Elimination algorithm (Karnin et al., 2013), Jin et al. (2019) proposed an algorithm achieving an expected sample complexity of the order of $\mathcal{O}(\sum_{a \neq a^\star} \Delta_a^{-2} \log(\log(\Delta_a^{-1})/\delta))$ with $\mathcal{O}(\log^\star_{1/\delta}(K) \log(\Delta_{\min}^{-1}))$ batches, where $\log^\star_{1/\delta}$ is the iterated logarithm function with base $1/\delta$. To the best of our knowledge, existing lower bound on the number of rounds are worst-case bounds. For constant $\delta \in (0,1)$, Tao et al. (2019) proved that for certain bandit instances, any algorithm that achieves the sample complexity bound obtained in Jin et al. (2019) requires at least $\Omega(\log(\Delta_{\min}^{-1})/\log\log\Delta_{\min}^{-1})$ batches. Jin et al. (2024) proposed the Tri-BBAI algorithm which achieves asymptotic optimality with two rounds of adaptivity (i.e. three phases). An important remark here is that the analysis of Tri-BBAI is purely asymptotic, and it is only $\delta$-correct for sufficiently small $\delta$. As an improvement with similar asymptotic guarantees as well as non-asymptotic ones, they propose Opt-BBAI which uses the same first two phases as Tri-BBAI, then uses successive elimination and checks for best arm elimination.

## Appendix I. Extended Experimental Analysis

For both local DP and global DP, we perform additional experiments on six bandit environments with Bernoulli distributions, as defined by (Sajed and Sheffet, 2019), namely

$$\mu_1 = (0.95, 0.9, 0.9, 0.9, 0.5), \qquad \mu_2 = (0.75, 0.7, 0.7, 0.7, 0.7),$$
$$\mu_3 = (0, 0.25, 0.5, 0.75, 1), \qquad \mu_4 = (0.75, 0.625, 0.5, 0.375, 0.25)\},$$
$$\mu_5 = (0.75, 0.53125, 0.375, 0.28125, 0.25), \qquad \mu_6 = (0.75, 0.71875, 0.625, 0.46875, 0.25)\}.$$

For each Bernoulli instance, we implement the algorithms with

$$\epsilon \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 10, 100, 1000\},$$

for global DP, and

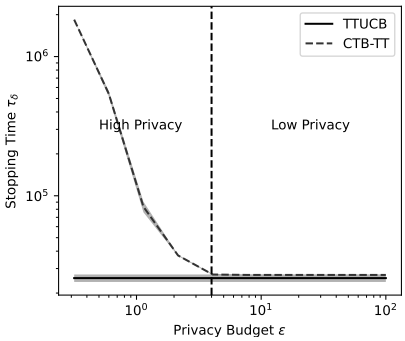
$$\epsilon \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 10, 100\},$$

for local DP.

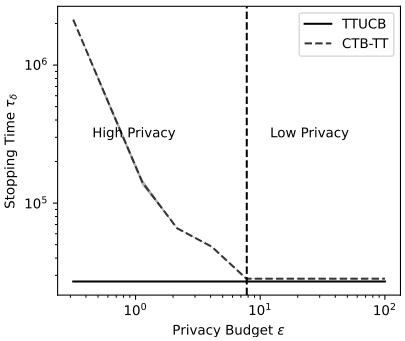
The risk level is set at $\delta = 0.01$. We verify empirically that the algorithms are $\delta$-correct by running each algorithm 1000 times.

The additional results for local DP are presented in Figure 3. For global DP, the additional results are provided in Figure 4. To show the difference between AdaP-TT and AdaP-TT$^\star$, we plot the stopping time not in a logarithmic scale in Figure 5. The additional experiments validate the same conclusions as the ones reached in Section 5.

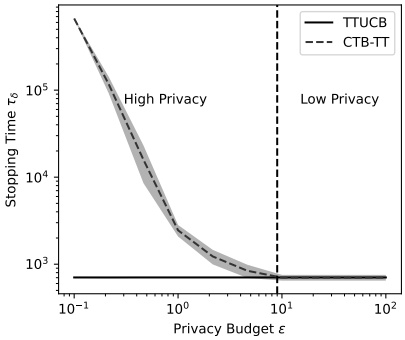
**Remark 52** *To implement the thresholds of* AdaP-TT *and* AdaP-TT$^\star$*, we use empirical thresholds that we get by approximating the theoretical thresholds. The expressions of the empirical thresholds used can be found in the code at* `https://github.com/achraf-azize/DP-BAI`.



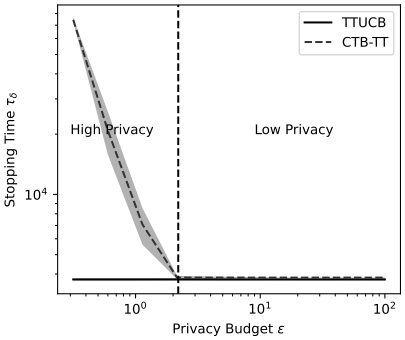

(a) $\mu_1$



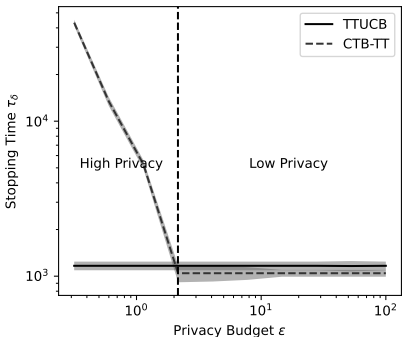

(b) $\mu_2$



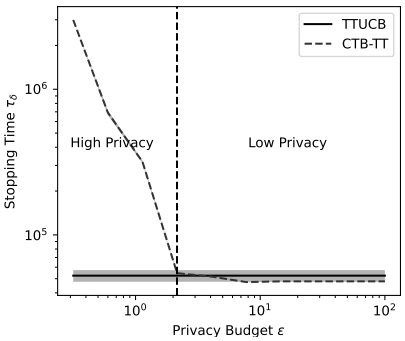

(c) $\mu_3$




(d) $\mu_4$

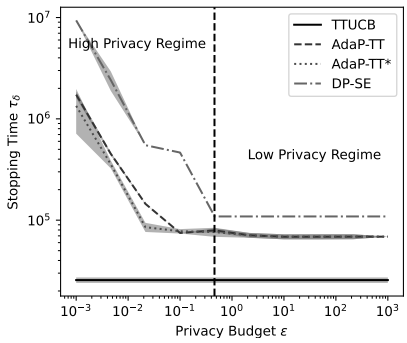



(e) $\mu_5$

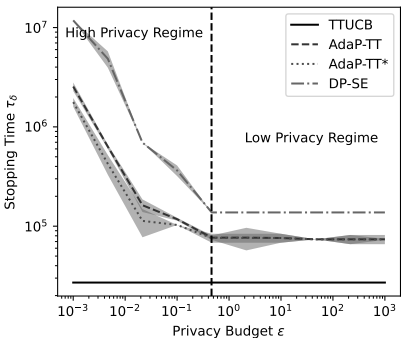



(f) $\mu_6$

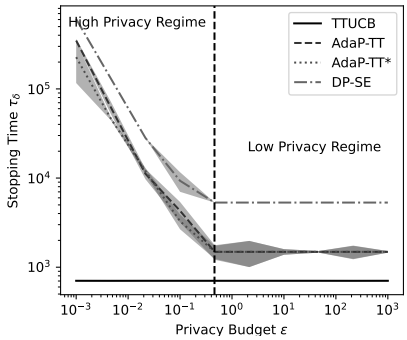
Figure 3: Evolution of the stopping time $\tau$ (mean $\pm$ std. over 1000 runs) of CTB-TT and TTUCB with respect to the privacy budget $\epsilon$ for $\delta = 10^{-2}$ on different Bernoulli instances. The shaded vertical line separates the two privacy regimes.



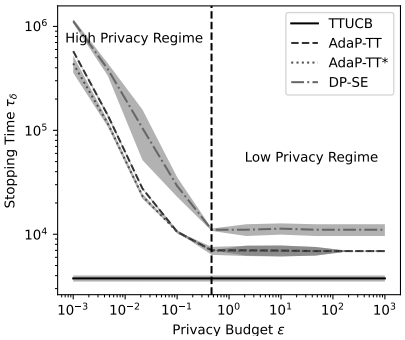

(a) $\mu_1$



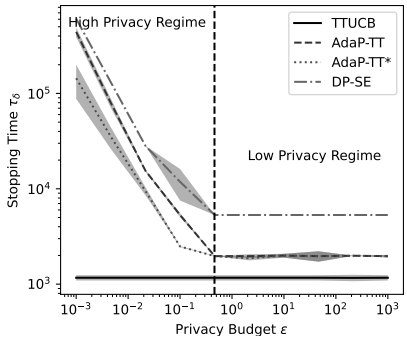

(b) $\mu_2$



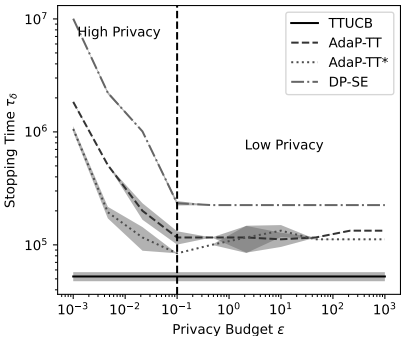

(c) $\mu_3$




(d) $\mu_4$

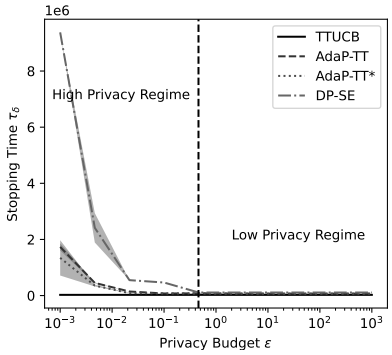



(e) $\mu_5$

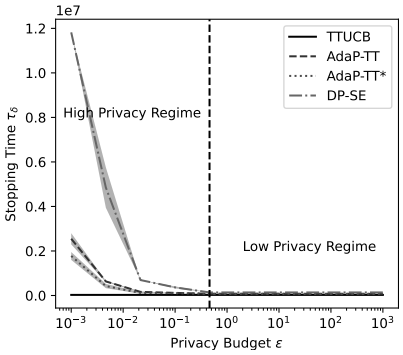



(f) $\mu_6$

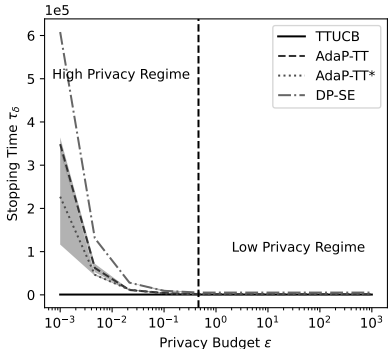
Figure 4: Evolution of the stopping time $\tau$ (mean $\pm$ std. over 1000 runs) of Imp-AdaP-TT, AdaP-TT, DP-SE, and TTUCB with respect to the privacy budget $\epsilon$ for $\delta = 10^{-2}$ on different Bernoulli instances. The shaded vertical line separates the two privacy regimes. Both the $x$-axis and $y$-axis are in logarithmic scale.



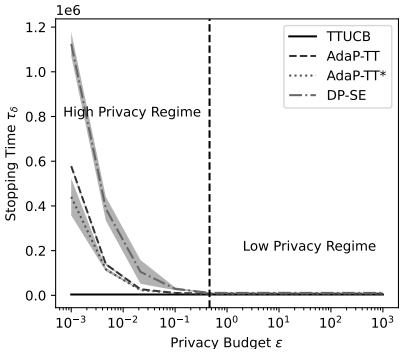

(a) $\mu_1$



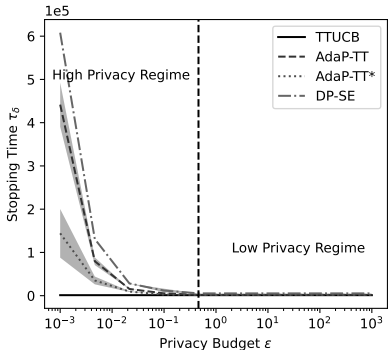

(b) $\mu_2$



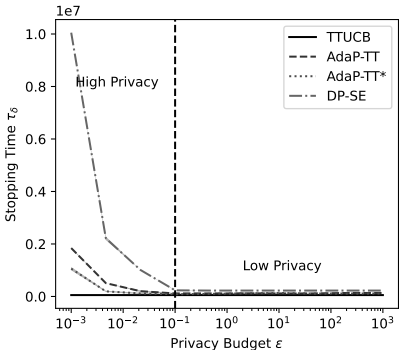

(c) $\mu_3$




(d) $\mu_4$




(e) $\mu_5$




(f) $\mu_6$

Figure 5: Evolution of the stopping time $\tau$ (mean $\pm$ std. over 1000 runs) of AdaP-TT$^\star$, AdaP-TT, DP-SE, and TTUCB with respect to the privacy budget $\epsilon$ for $\delta = 10^{-2}$ on different Bernoulli instances. The shaded vertical line separates the two privacy regimes. Only the $x$-axis is in logarithmic scale.

## References

Y. Abbasi-Yadkori, P. Bartlett, V. Gabillon, A. Malek, and M. Valko. Best of both worlds: Stochastic & adversarial best-arm identification. In *Conference on Learning Theory*, pages 918–949. PMLR, 2018.

J. Acharya, Z. Sun, and H. Zhang. Differentially private assouad, fano, and le cam. In *Algorithmic Learning Theory*, pages 48–78. PMLR, 2021.

S. Agrawal, S. Juneja, and P. W. Glynn. Optimal $\delta$-correct best-arm selection for heavy-tailed distributions. In *Algorithmic Learning Theory (ALT)*, 2020.

J.-Y. Audibert, S. Bubeck, and R. Munos. Best Arm Identification in Multi-armed Bandits. In *Conference on Learning Theory*, 2010.

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

M. Aziz, E. Kaufmann, and M.-K. Riviere. On multi-armed bandit designs for dose-finding clinical trials. *The Journal of Machine Learning Research*, 22(1):686–723, 2021.

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

A. Azize and D. Basu. Concentrated differential privacy for bandits. In *2nd IEEE Conference on Secure and Trustworthy Machine Learning*, 2024.

A. Azize, M. Jourdan, A. Al Marjani, and D. Basu. On the complexity of differentially private best-arm identification with fixed confidence. *Thirty-Seventh Conference on Neural Information Processing Systems*, 2023.

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

R. E. Bechhofer. A single-sample multiple decision procedure for ranking means of normal populations with known variances. *The Annals of Mathematical Statistics*, pages 16–39, 1954.

R. E. Bechhofer. A sequential multiple-decision procedure for selecting the best one of several normal populations with a common unknown variance, and its use with various experimental designs. *Biometrics*, 14(3):408–429, 1958.

T. T. Cai, Y. Wang, and L. Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825–2850, 2021.

A. Carpentier and A. Locatelli. Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Conference on Learning Theory*, pages 590–604. PMLR, 2016.

S. Chen, T. Lin, I. King, M. R. Lyu, and W. Chen. Combinatorial pure exploration of multi-armed bandits. *Advances in neural information processing systems*, 27, 2014.

A. Cheu. Differential privacy in the shuffle model: A survey of separations. *arXiv preprint arXiv:2107.11839*, 2021.

S. R. Chowdhury and X. Zhou. Distributed differential privacy in multi-armed bandits. In *The Eleventh International Conference on Learning Representations*, 2023.

R. Degenne, W. M. Koolen, and P. Ménard. Non-asymptotic pure exploration by solving games. *Advances in Neural Information Processing Systems*, 32, 2019.

R. Degenne, H. Shao, and W. Koolen. Structure adaptive algorithms for stochastic bandits. In *International Conference on Machine Learning*, pages 2443–2452. PMLR, 2020.

I. Dinur and K. Nissim. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 202–210, 2003.

J. Dong, A. Roth, and W. J. Su. Gaussian differential privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):3–37, 2022.

J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438. IEEE, 2013.

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

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, pages 265–284, Berlin, Heidelberg, 2006. Springer-Verlag.

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

C. Dwork, M. Naor, T. Pitassi, G. N. Rothblum, and S. Yekhanin. Pan-private streaming algorithms. In *Innovations in Computer Science*, pages 66–80, 2010b.

E. Even-Dar, S. Mannor, Y. Mansour, and S. Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6), 2006.

A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 211–222, 2003.

V. Gabillon, M. Ghavamzadeh, and A. Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. *Advances in Neural Information Processing Systems*, 25, 2012.

A. Garivier and E. Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016.

A. M. Girgis, D. Data, S. Diggavi, A. T. Suresh, and P. Kairouz. On the renyi differential privacy of the shuffle model. In *ACM SIGSAC Conference on Computer and Communications Security*, pages 2321–2341, 2021.

K. Jamieson and R. Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2014.

K. Jamieson and A. Talwalkar. Non-stochastic best arm identification and hyperparameter optimization. In *Artificial intelligence and statistics*, pages 240–248. PMLR, 2016.

K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439. PMLR, 2014.

T. Jin, J. Shi, X. Xiao, and E. Chen. Efficient pure exploration in adaptive round model. *Advances in Neural Information Processing Systems*, 32, 2019.

T. Jin, Y. Yang, J. Tang, X. Xiao, and P. Xu. Optimal batched best arm identification. *Advances in Neural Information Processing Systems*, 37:134947–134980, 2024.

M. Jourdan and R. Degenne. Non-asymptotic analysis of a ucb-based top two algorithm. *Advances in Neural Information Processing Systems*, 36, 2024.

M. Jourdan, R. Degenne, D. Baudry, R. de Heide, and E. Kaufmann. Top two algorithms revisited. *Advances in Neural Information Processing Systems*, 35:26791–26803, 2022.

M. Jourdan, R. Degenne, and E. Kaufmann. Dealing with unknown variances in best-arm identification. *International Conference on Algorithmic Learning Theory*, 2023.

M. Jourdan, R. Degenne, and E. Kaufmann. An $\varepsilon$-best-arm identification algorithm for fixed-confidence and beyond. *Advances in Neural Information Processing Systems*, 36, 2024.

P. Kairouz, K. Bonawitz, and D. Ramage. Discrete distribution estimation under local privacy. In *International Conference on Machine Learning*, pages 2436–2444. PMLR, 2016.

D. S. Kalogerias, K. E. Nikolakakis, A. D. Sarwate, and O. Sheffet. Quantile multi-armed bandits: Optimal best-arm identification and a differentially private scheme. *IEEE Journal on Selected Areas in Information Theory*, 2(2):534–548, 2021.

S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone. Pac subset selection in stochastic multi-armed bandits. In *International Conference on Machine Learning*, volume 12, pages 655–662, 2012.

Z. Karnin, T. Koren, and O. Somekh. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*. PMLR, 2013.

V. Karwa and S. Vadhan. Finite Sample Differentially Private Confidence Intervals. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*, volume 94. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2018.

E. Kaufmann and S. Kalyanakrishnan. Information complexity in bandit subset selection. In *Conference on Learning Theory*, pages 228–251. PMLR, 2013.

E. Kaufmann and W. M. Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, 22(246): 1–44, 2021.

E. Kaufmann, O. Cappé, and A. Garivier. On the complexity of best arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17:1–42, 2016.

C. Lalanne, A. Garivier, and R. Gribonval. On the statistical complexity of estimation and testing under privacy constraints. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

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

L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, 18(1):6765–6816, 2017.

P. J. Libin, T. Verstraeten, D. M. Roijers, J. Grujic, K. Theys, P. Lemey, and A. Nowé. Bayesian best-arm identification for selecting influenza mitigation strategies. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018*, 2019.

S. Lindståhl, A. Proutiere, and A. Johnsson. Measurement-based admission control in sliced networks: A best arm identification approach. In *GLOBECOM 2022-2022 IEEE Global Communications Conference*, pages 1484–1490. IEEE, 2022.

D. E. Losada, D. Elsweiler, M. Harvey, and C. Trattner. A day at the races: using best arm identification algorithms to reduce the cost of information retrieval user studies. *Applied Intelligence*, 52(5):5617–5632, 2022.

S. Mannor and J. N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648, 2004.

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

N. Mishra and A. Thakurta. (Nearly) optimal differentially private stochastic multi-arm bandits. In *Conference on Uncertainty in Artificial Intelligence*, 2015.

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

K. E. Nikolakakis, D. S. Kalogerias, and A. D. Sarwate. Optimal rates for learning hidden tree structures. *arXiv preprint arXiv:1909.09596*, 2019.

C. Qin, D. Klabjan, and D. Russo. Improving the expected improvement algorithm. *Advances in Neural Information Processing Systems*, 30, 2017.

W. Ren, X. Zhou, J. Liu, and N. B. Shroff. Multi-armed bandits with local differential privacy. *arXiv preprint arXiv:2007.03121*, 2020.

A. Rio, M. Barlier, I. Colin, and M. Soare. Multi-agent best arm identification with private communications. In *International Conference on Machine Learning*, 2023.

D. Russo. Simple bayesian algorithms for best arm identification. In *Conference on Learning Theory*, pages 1417–1418. PMLR, 2016.

T. Sajed and O. Sheffet. An optimal private stochastic-mab algorithm based on optimal private stopping rule. In *International Conference on Machine Learning*, pages 5579–5588. PMLR, 2019.

X. Shang, R. Heide, P. Menard, E. Kaufmann, and M. Valko. Fixed-confidence guarantees for bayesian best-arm identification. In *International Conference on Artificial Intelligence and Statistics*, pages 1823–1832. PMLR, 2020.

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

M. Soare, A. Lazaric, and R. Munos. Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 27, 2014.

C. Tao, Q. Zhang, and Y. Zhou. Collaborative learning with limited interaction: Tight bounds for distributed exploration in multi-armed bandits. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 126–146, 2019.

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

K. Tucker, J. Branson, M. Dilleen, S. Hollis, P. Loughlin, M. J. Nixon, and Z. Williams. Protecting patient privacy when sharing patient-level data from clinical trials. *BMC medical research methodology*, 16(1):5–14, 2016.

P.-A. Wang, R.-C. Tzeng, and A. Proutiere. Fast pure exploration via frank-wolfe. *Advances in Neural Information Processing Systems*, 34:5810–5821, 2021.

S. L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American statistical association*, pages 63–69, 1965.

W. You, C. Qin, Z. Wang, and S. Yang. Information-directed selection for top-two algorithms. In *Conference on Learning Theory*, pages 2850–2851. PMLR, 2023.

K. Zheng, T. Cai, W. Huang, Z. Li, and L. Wang. Locally differentially private (contextual) bandits learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 12300–12310, 2020.

Y. Zhou, X. Chen, and J. Li. Optimal pac multiple arm identification with applications to crowdsourcing. In *International Conference on Machine Learning*, pages 217–225. PMLR, 2014.