Journal of Machine Learning Research 27 (2026) 1-61 Submitted 4/22; Revised 3/26; Published 3/26

The Role of Contextual Information in Best Arm Identification

Masahiro Kato* MKATO-CSECON@G.ECC.U-TOKYO.AC.JP
AI Lab
CyberAgent, Inc.
Shibuya, Tokyo 150-6121, Japan

Kaito Ariu* ARIU@KTH.SE
AI Lab
CyberAgent, Inc.
Shibuya, Tokyo 150-6121, Japan

Editor: Aurelien Garivier

Abstract

We study the best-arm identification problem with fixed confidence when contextual (covariate) information is available in stochastic bandits. In each round, we observe contextual information before selecting an arm. The distribution of the reward associated with the selected arm depends on the observed contextual information. We are interested in finding the arm with the maximum mean reward marginalized over the contextual distribution and not the mean reward conditioned on contexts. Our goal is to identify the best arm with a minimal number of samples under a given error probability. First, we derive the instance-specific sample-complexity lower bounds under the contextual information. Then, we propose a context-aware version of the *Track-and-Stop strategy*, wherein the proportions of arm draws track the set of optimal allocations, and prove that the expected number of arm draws asymptotically matches the lower bound. We demonstrate that the contextual information can be used to improve the efficiency of the identification of the best marginalized mean reward when compared with the results of Garivier and Kaufmann (2016). Furthermore, we experimentally confirm that contextual information contributes to faster best-arm identification.

Keywords: multi-armed bandit problem, best arm identification, contextual information, information theoretic analysis, adaptive sequential testing

1. Introduction

This paper studies best-arm identification (BAI) with contextual information in stochastic multi-armed bandit (MAB) problems. We define the best arm as the arm with the maximum marginalized mean reward, where the expectation is taken over the context distribution and is not conditional on a specific context. We call this setting contextual BAI. The goal is to identify the best arm with a fixed confidence level and a smaller sample complexity (the expected stopping time) defined by the probably approximately correct (PAC) framework. The instance-specific sample complexity of BAI without contextual information is now well understood. There exists an instance-specific lower bound (Kaufmann et al., 2016; Gariv-

*Authors contributed equally.

ier and Kaufmann, 2016) and optimal algorithms whose performance guarantee matches the lower bound (Kaufmann et al., 2016; Garivier and Kaufmann, 2016; Degenne et al., 2019); however, the corresponding characterization for contextual BAI has not been fully elucidated.

Formally, we consider the following setting. At each time $t = 1, 2, \dots$, an agent observes a context (covariate) $X_t \in \mathcal{X}$ and chooses an arm $A_t \in [K] = \{1, \dots, K\}$, where $\mathcal{X}$ denotes the context space. Then, the agent immediately receives a reward (or outcome) R_t linked to the arm A_t. This setting is called the bandit feedback or Rubin causal model (Neyman, 1923; Rubin, 1974); that is, a reward in round t is $R_t = \sum_{a=1}^{K} \mathbb{1}[A_t = a] R_{t,a}$, where $R_{t,a}$ is a potential independent (random) reward. We assume that X_t is independent and identically distributed (i.i.d.) over $[T]$ and denote the distribution of X_t by ζ. Given the context $x \in \mathcal{X}$, we denote the reward distributions of the potential outcomes as $\boldsymbol{p} = (p_{1,x}, p_{2,x}, \dots, p_{K,x})$ and their means as $\boldsymbol{\mu} = (\mu_{1,x}, \mu_{2,x}, \dots, \mu_{K,x})$. Let $\mathcal{V} = (\boldsymbol{p}, \zeta)$ (this can be written as $\nu = ((\mu_{a,x}), (\zeta_x))$ when the rewards follow a distribution that belongs to a single parameter exponential family, and the contexts are finite) be a bandit problem. Let $\mathbb{P}_{\mathcal{V}}$ (resp. $\mathbb{P}_{\nu}$) and $\mathbb{E}_{\mathcal{V}}$ (resp. $(\mathbb{E}_{\nu})$) be the probability and expectations under model $\mathcal{V}$ (resp. ν), respectively. Then, $\mu_a = \mathbb{E}_{X\sim\zeta}[\mu_{a,X}] = \mathbb{E}_{X\sim\zeta}[\mathbb{E}_{\mathcal{V}}[R_{t,a}|X]] = \mathbb{E}_{\mathcal{V}}[R_{t,a}]$ is the average reward marginalized over $\mathcal{X}$. We assume that $\mathcal{V}$ belongs to a class $\Omega = \{(\boldsymbol{p}, \zeta) : \exists a^* \in [K] \ s.t. \ \forall a \neq a^*, \mu_{a^*} > \mu_a\}$; that is, the best arm $a^*(\mathcal{V}) = \arg\max_a \mu_a$ is uniquely defined. Let $p_{a,x}$ and $q_{a,x}$ be two absolutely continuous probability distributions (w.r.t. the Lebesgue measure) of $R_{t,a}$, given $X_t = x$. We define the Kullback–Leibler (KL) divergence from $p_{a,x}$ to $q_{a,x}$ as

$$\mathrm{KL}(p_{a,x}, q_{a,x}) := \begin{cases} \int_{\mathbb{R}} \log\left(\frac{p_{a,x}}{q_{a,x}}(r)\right) \mathrm{d}p_{a,x}(r) \text{ if } q_{a,x} \ll p_{a,x}, \\ +\infty \quad \text{otherwise.} \end{cases}$$

We assume that for all $(\boldsymbol{p}, \zeta), (\boldsymbol{q}, \zeta) \in \Omega$, if $p_{a,x} \neq q_{a,x}$, then $0 < \mathrm{KL}(p_{a,x}, q_{a,x}) < +\infty$. For distributions that belong to the single parameter exponential family, we introduce the KL divergence from the distribution with mean μ to the distribution with mean ν as $\mathrm{kl}(\mu, \nu)$. Furthermore, for the Bernoulli distributions, we denote the KL divergence by $d(\mu, \nu) = \mu \log(\mu/\nu) + (1-\mu)\log((1-\mu)/(1-\nu))$ with the convention that $d(0,0) = d(1,1) = 0$.

Let $\mathcal{F}_t = \sigma(X_1, A_1, R_1, \dots, X_t, A_t, R_t, X_{t+1})$ and $\mathcal{G}_t = \sigma(X_1, A_1, R_1, \dots, X_t, A_t, R_t)$ be the sigma-algebras generated by the observations up to immediately before the selection of the arm at time $t+1$ and all observations up to time t, respectively. The strategy or algorithm of the best arm identification consists of the following three elements: a sampling, stopping, and decision rules. A sampling rule selects from which arm we collect the sample each time based on past observation (A_t is $\mathcal{F}_{t-1}$-measurable). The stopping rule determines when to stop sampling based on the past observation. We denote τ as this time; τ is the stopping time with respect to the filtration $(\mathcal{G}_t)_{t\geq 1}$. The decision rule estimates the best arm $\hat{a}_\tau$ based on observation up to time τ ($\hat{a}_\tau$ is $\mathcal{G}_\tau$-measurable). We measure the performance of the decision rule using the sample complexity, the expected stopping time, defined as $\mathbb{E}_{\mathcal{V}}[\tau]$.

We focus on the fixed confidence setting; that is, with a given admissible failure probability $\delta \in (0,1)$, the algorithm is guaranteed to have $\mathbb{P}(\hat{a}_\tau \neq \arg\max_a \mu_a) \leq \delta$. We define δ-PAC to formalize this property:

Definition 1 *An algorithm is δ-PAC if for all $\mathcal{V} \in \Omega$, $\mathbb{P}_{\mathcal{V}}(\hat{a}_\tau \neq a^*(\mathcal{V})) \leq \delta$ and $\mathbb{P}_{\mathcal{V}}(\tau < \infty) = 1$.*

Later, we propose algorithms that are δ-PAC.

We reemphasize that although we can use contextual information, our primary interest is not in the mean reward conditioned on each context. Similar problems are frequently considered in the literature on causal inference that mainly discusses the efficient estimation of causal parameters. The assigned treatment (chosen arm) and observed outcomes for each treatment (reward) and covariate (context) are given therein. Here, we are not interested in the distribution of the covariate; rather we are interested in the estimation of the expected value of the outcome of the treatment marginalized over the covariate distribution; that is, the average treatment effect (ATE) (Imbens and Rubin, 2015). For this setting, van der Laan (2008) and Hahn et al. (2011) proposed experimental design methods to estimate the ATE more efficiently by assigning treatments based on the covariate. According to their results, even if the covariates are marginalized, the variance of the estimator can be reduced with the help of the covariate information. Karlan and Wood (2014) applied the method of Hahn et al. (2011) to test how donors respond to new information about the effectiveness of charity. These studies were extended by Tabord-Meehan (2018) and Kato et al. (2020).

For each $x \in \mathcal{X}$, we define allocations over arms given context x as $\Sigma_x^K = \{(w_{a,x})_{a\in[K]} \in \mathbb{R}_+^K : w_{1,x} + \cdots + w_{K,x} = 1\}$ and define the set of allocations over all contexts as $\mathcal{W} = \{(w_{a,x})_{a\in[K],x\in\mathcal{X}} : \forall x \in \mathcal{X}\ (w_{a,x})_{a\in[K]} \in \Sigma_x^K\}$. We denote by $N_x(t)$ and $N_{a,x}(t)$ the number of times we observe context x, and we choose arm a given context x; that is, $N_x(t) = \sum_{s=1}^t \mathbb{1}\{X_s = x\}$ and $N_{a,x}(t) = \sum_{s=1}^t \mathbb{1}\{X_s = x, A_s = a\}$, respectively.

Main results. We briefly summarize our contributions.

First, we establish the instance-specific lower bound on contextual BAI for both continuous and finite context cases. The derived lower bound formula has smaller sample complexity than that of lower bound formula in Garivier and Kaufmann (2016), suggesting that a faster BAI may be possible.

Then, we propose optimal algorithms for two cases: (i) two-armed Gaussian bandits where the arms and context jointly follow the multivariate Gaussian distribution; and (ii) MAB with reward distributions belonging to the single parameter exponential family and finite contexts. We prove that the sample complexity upper bounds of the proposed algorithms asymptotically match the lower bounds.

Organization. This paper is organized as follows. In Section 2, we derive the general instance-specific lower bounds for contextual BAI for a case with continuous contexts. Then, in Section 3, we discuss an optimal algorithm for two-armed Gaussian bandits with continuous contexts. Section 4 focuses on the lower bound when the number of contexts is finite, and the reward distributions are from the single parameter exponential family. In Section 5, for the finite context case, we obtain the optimal allocations for each pair of contexts and actions by simplifying the lower bound formula. In Section 6, in the same setting of Section 4, we show an optimal algorithm. We describe details of the sampling, stopping, and decision rules that are the core of the proposed algorithm and demonstrate that the algorithm is δ-PAC. We further confirm that the sample complexity of the proposed algorithm is asymptotically optimal. Section 7 presents the results of our numerical experiments.

Related work. The stochastic MAB problem is a classical abstraction of the sequential decision-making problem (Thompson, 1933; Robbins, 1952; Lai and Robbins, 1985). BAI

is a paradigm of the MAB problem, where we consider pure exploration to find the best arm. Several strategies and efficiency metrics have been proposed for BAI (Bechhofer et al., 1968; Paulson, 1964; Mannor and Tsitsiklis, 2004; Even-Dar et al., 2006; Bubeck et al., 2011; Gabillon et al., 2012; Karnin et al., 2013; Garivier and Kaufmann, 2016; Jamieson et al., 2014). BAI with linear bandits (Soare et al., 2014; Xu et al., 2018; Tao et al., 2018; Fiez et al., 2019; Jedra and Proutiere, 2020), BAI with multiple queries, and the partition identification problem studied by Juneja and Krishnasamy (2019) are different directions for generalizing BAI.

Our setting is a generalization of BAI without contextual information. We can use the side information (explicitly or implicitly) at each round. There have been limited studies that address pure exploration in contextual bandits. Tekin and van der Schaar (2015), Guan and Jiang (2018), and Deshmukh et al. (2018) also consider BAI with contextual information; however, they do not discuss the instance-specific optimality. After this study, Qin and Russo (2022) also considers a related topic.

From the causal inference perspective, contextual BAI is closely related to a (semiparametric) experimental design for efficient ATE estimation (van der Laan, 2008; Hahn et al., 2011; Karlan and Wood, 2014; Athey and Imbens, 2016; Tabord-Meehan, 2018; Kato et al., 2024a). The goal of efficient ATE estimation by adaptive experimentation is often in choosing the best treatment (arm) via hypothesis testing. Therefore, it can be considered as a case where the proposed method should be applied, especially when there are multiple treatments (arms).

Russac et al. (2021) also addresses a similar problem independently of us. Their problem setting is the same as ours in that they can observe discrete contexts. However, they consider a problem that is slightly different from BAI, namely A/B/n testing, where the comparison is made with a designated control arm. In that problem setting, optimal allocation is uniquely obtained, and they do not have to consider multiple candidates of optimal allocations as we do. Besides, we also derive the result for the case of continuous contexts, which they do not address. On the other hand, they discuss the problem more generally by considering four situations, (a) active mode, (b) proportional mode, (c) agnostic mode, and (d) oblivious mode, depending on how the decision is made. The (b) proportional mode discussed by them is closer to the setting discussed in this paper. In these senses, our results and theirs, while similar, are independent and parallel, and correspond to complementary studies.

Readers might be interested in identifying an arm with the highest context-conditional expected reward instead of an arm with the highest context-marginalized expected reward. However, to identify the context-conditional best arm, we usually need to introduce a regression model to approximate the context-conditional expected reward. In real-world applications, there are many cases where we want to avoid using such models. First, there is a risk of model misspecification; that is, an incorrect model cannot approximate the context-conditional expected reward well. Second, if we consider a large class of models, training such models requires large samples, which is costly in practice. These issues are the reasons why we are still interested in estimating the ATE (the difference in the context-marginalized expected reward) in causal inference, although various methods have been proposed to estimate the context-conditional ATE. Note that in BAI, to the best of our knowledge, it is still under investigation how to identify the context-conditional best arm, unlike the setting of regret minimization. Recently, Kato et al. (2024b) proposes a method for identifying the

context-conditional best arm in the setting of fixed-budget BAI, but they only show a minimax rate optimality of their proposed algorithm, unlike our exact optimality with matching upper and lower bounds.

2. General Non-Asymptotic Lower Bounds

In this section, we provide the instance-specific sample complexity lower bounds for general contextual BAI. The proof is based on standard change-of-measure arguments (Kaufmann et al., 2016). However, the derivations must consider the possibly continuous context distributions, which are non-trivial.

Based on this lower bound, we find that contextual information either helps or does not harm BAI. Our result is the same as those of existing studies on fixed-confidence BAI without contextual information, except that we can obtain help from the existence of the contextual information. At first glance, it does not necessarily seem advantageous to use contextual information as the marginalized mean reward is not directly related to the contextual information. However, the lower bound with contextual information (see Section 2) is strictly lower than the sample complexity derived by Kaufmann et al. (2016) and Garivier and Kaufmann (2016).

Assume $\mathcal{X} = \mathbb{R}$. Then, we present the non-asymptotic sample complexity lower bound.

Theorem 2 *Let $\delta \in (0, 1/2)$. Assume that for all $x \in \mathbb{R}$, distributions $p_{1,x}, \dots, p_{K,x}$ are absolutely continuous with respect to the Lebesgue measure. Then, for any δ-PAC strategy and any $\mathcal{V} = (\boldsymbol{p}, \zeta) \in \Omega$,*

$$\mathbb{E}_{\mathcal{V}}[\tau_\delta] \geq T^\star(\mathcal{V}) d(\delta, 1-\delta),$$

where

$$T^\star(\mathcal{V}) := \left(\sup_{\boldsymbol{w} \in \mathcal{W}} \inf_{(\boldsymbol{q}, \zeta) \in \mathrm{Alt}(\mathcal{V})} \sum_{a=1}^{K} \int_{\mathbb{R}} w_{a,x} \mathrm{KL}(p_{a,x}, q_{a,x}) \zeta(x) \mathrm{d}x \right)^{-1},$$

and $\mathrm{Alt}(\mathcal{V}) := \{(\boldsymbol{q}, \zeta) \in \Omega : a^*((\boldsymbol{q}, \zeta)) \neq a^*((\boldsymbol{p}, \zeta))\}$ is the set of alternative problems.

We provide the proof of Theorem 2 in Appendix D. Based on this lower bound, we develop optimal algorithms for (i) two-armed Gaussian bandits where the arms and context jointly follow the multivariate Gaussian distribution; and (ii) MAB with reward distributions belonging to the single parameter exponential family and finite contexts.

Setting practical considerations aside, if we assume that the contextual distribution is known, we can develop optimal algorithms for more general cases, such as multi-armed bandits following a distribution belonging to the exponential family. To develop such algorithms, we derive an optimal allocation from the lower bound in Theorem 2 and then apply the Track-and-Stop algorithm in Garivier and Kaufmann (2016). However, computing an optimal allocation can be a challenging task since we need to solve the optimization problem in the lower bound of Theorem 2. For example, if we consider infinite contexts, we may need to use some model for $\boldsymbol{w}$, such as linear models and neural networks. The existence of the integral also complicates the problem. Furthermore, if the contextual distribution is

unknown and needs to be estimated during a trial, the estimation error generally affects the performance.

As a similar problem, in fixed-confidence BAI without contexts, Jourdan et al. (2023) discusses Gaussian bandits with unknown variances and finds that a specific algorithm is required for that case. In this study, we find that we can obtain practical algorithms for two cases. This is because, in these cases, we can model the reward and contextual distribution with a simple distribution that is easy to handle. In two-armed Gaussian bandits with Gaussian contexts, the whole distribution follows the multivariate Gaussian distribution. In the MAB with reward distributions belonging to the single-parameter exponential family and finite contexts, the whole distribution becomes a combination of the exponential family and multinomial distribution. This property allows us to avoid the technical issues in this problem. If we consider a more general case with an unknown distribution, we may need to develop different lower bounds and algorithms to take the estimation error into account, as in Jourdan et al. (2023). Additionally, if the contexts are infinite, we may need to use models such as linear models and neural networks to approximate the allocation ratio $\boldsymbol{w}$.

Efficiency gains from the context use. In Figure 1, we illustrate the efficiency gain by using contextual information. We consider a two-armed, one-dimensional context $X_t \in \mathbb{R}$. Suppose that $(R_{t,1}, R_{t,2}, X_t)^\top$ follows a multivariate normal distribution with mean vector $(1,0,0)^\top$. We assume that the variances of $R_{t,1}$, $R_{t,2}$, and X_t are 1. We investigate the variation in the theoretical sample complexity by varying the correlation coefficients between X_t and $R_{t,1}$ and X_t and $R_{t,2}$, which are denoted as $\rho_{1\mathcal{X}} \in [-1,1]$ and $\rho_{2\mathcal{X}} \in [0,1]$, respectively. Note that we omit the other domains due to symmetry with the current domain. Note that when ignoring (marginalizing) the context, arm 1 follows $\mathcal{N}(1,1)$ and arm 2 follows $\mathcal{N}(0,1)$, where $\mathcal{N}(\mu, \sigma^2)$ denotes a normal distribution with a mean μ and variance σ^2. Here, for $\delta = 0.05$, we calculate the sample complexity lower bounds of the standard setting of BAI from the result of Garivier and Kaufmann (2016) and those of the contextual case from our results. We denote the former as ℓ and the latter as $\widetilde{\ell}$. Then, we compute the sample complexity gain $(1-\widetilde{\ell}/\ell)$ for different pairs of $(\rho_{1\mathcal{X}}, \rho_{2\mathcal{X}})$ and illustrate it in Figure 1. Here, the efficiency gain is about 67.5% at $(\rho_{1\mathcal{X}}, \rho_{2\mathcal{X}}) = (-1,0)$, $(\rho_{1\mathcal{X}}, \rho_{2\mathcal{X}}) = (1,0)$, and $(\rho_{1\mathcal{X}}, \rho_{2\mathcal{X}}) = (0,1)$, while it is 100% at $(\rho_{1\mathcal{X}}, \rho_{2\mathcal{X}}) = (-1,1)$ and $(\rho_{1\mathcal{X}}, \rho_{2\mathcal{X}}) = (1,1)$. That is, at $(\rho_{1\mathcal{X}}, \rho_{2\mathcal{X}}) = (-1,1)$ and $(\rho_{1\mathcal{X}}, \rho_{2\mathcal{X}}) = (1,1)$, the lower-bound value becomes zero. This is because, at these points, the variance of each arm's reward is zero. The details are explained in the following section.

3. Two-armed Gaussian Bandits with Continuous Context

In this section, we provide an example for the case of continuous contexts and prove the upper bound of the sample complexity. We consider the following two-armed bandit problem. For each t, $R_{t,1}$, $R_{t,2}$, and $X_t \in \mathbb{R}$ are drawn from the following Gaussian distributions $\mathcal{N}(\mu_1, \sigma_1^2)$, $\mathcal{N}(\mu_2, \sigma_2^2)$, and $\mathcal{N}(\mu_{\mathcal{X}}, \sigma_{\mathcal{X}}^2)$, respectively ($\mu_1 > \mu_2$). Assume that the vector $(R_{t,1}, R_{t,2}, X_t)$ forms a multivariate Gaussian distribution. We denote $\mathrm{Cov}(R_{t,1}, X_t) = \sigma_{1\mathcal{X}}$ and $\mathrm{Cov}(R_{t,2}, X_t) = \sigma_{2\mathcal{X}}$. Suppose the algorithm knows that $(R_{t,1}, R_{t,2}, X_t)$ form a multivariate Gaussian distribution, knows the values of σ_1^2, σ_2^2, $\mu_{\mathcal{X}}$, $\sigma_{\mathcal{X}}$, $\sigma_{1\mathcal{X}}$, and $\sigma_{2\mathcal{X}}$, and does not know the values of μ_1 and μ_2. Let $\widetilde{\Omega}$ be a set of all such problems. Given an observa-

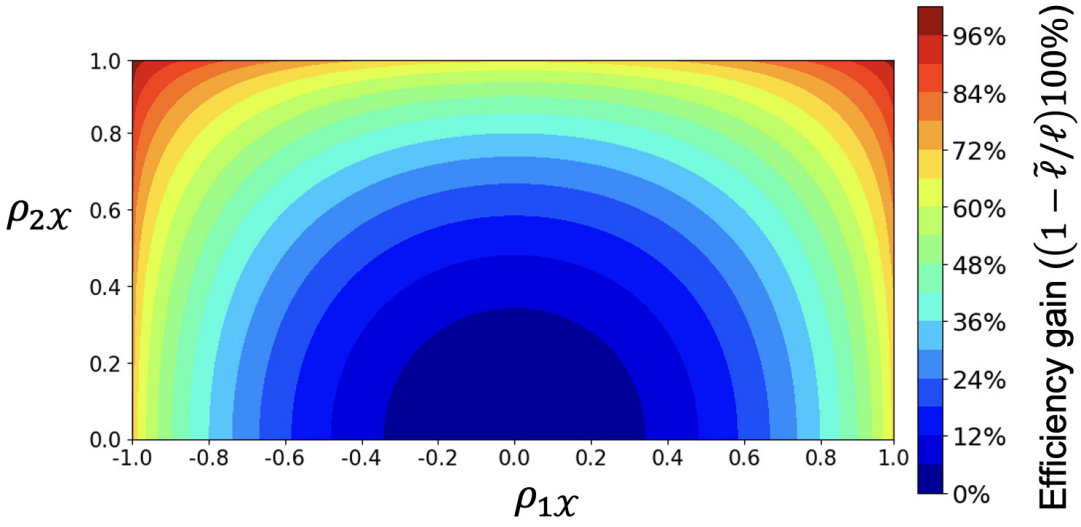

Figure 1: Sample complexity gains through context. The x axis denotes $\rho_{1\mathcal{X}} \in [-1, 1]$ and the y axis denotes $\rho_{2\mathcal{X}} \in [0, 1]$. The contour lines indicate the sample complexity gains: $(1 - \widetilde{\ell}/\ell)100\%$. The efficiency gain is about 67.5% at $(\rho_{1\mathcal{X}}, \rho_{2\mathcal{X}}) = (-1, 0)$, $(\rho_{1\mathcal{X}}, \rho_{2\mathcal{X}}) = (1, 0)$, and $(\rho_{1\mathcal{X}}, \rho_{2\mathcal{X}}) = (0, 1)$, while it is 100% at $(\rho_{1\mathcal{X}}, \rho_{2\mathcal{X}}) = (-1, 1)$ and $(\rho_{1\mathcal{X}}, \rho_{2\mathcal{X}}) = (1, 1)$.

tion $X_t = x$, we have conditional distributions of $R_{t,1}$ and $R_{t,2}$ where for each $a \in \{1, 2\}$, $R_{t,a} \sim \mathcal{N}\left(\mu_a + \frac{\sigma_{a\mathcal{X}}}{\sigma_{\mathcal{X}}^2}(x - \mu_{\mathcal{X}}), \sigma_a^2 - \frac{\sigma_{1\mathcal{X}}^2}{\sigma_{\mathcal{X}}^2}\right) = \mathcal{N}\left(\mu_a + \frac{\rho_{a\mathcal{X}}\sigma_a}{\sigma_{\mathcal{X}}}(x - \mu_{\mathcal{X}}), \sigma_a^2(1 - \rho_{a\mathcal{X}}^2)\right)$. Here, $\rho_{a\mathcal{X}}$ is the correlation coefficient between the context and arm $a \in \{1, 2\}$. We denote $\sigma_1'^2 = \sigma_1^2 - \frac{\sigma_{1\mathcal{X}}^2}{\sigma_{\mathcal{X}}^2}$ and $\sigma_2'^2 = \sigma_2^2 - \frac{\sigma_{2\mathcal{X}}^2}{\sigma_{\mathcal{X}}^2}$. From our lower bound in Theorem 2, we can derive the following lower bound for this specific problem. We give the proof in Appendix E.2.

Theorem 3 *Let $\delta \in (0, 1/2)$. For any δ-PAC strategy and $\mathcal{V} \in \widetilde{\Omega}$, we have*

$$\mathbb{E}_{\mathcal{V}}[\tau_\delta] \geq \frac{2(\sigma_1' + \sigma_2')^2}{(\mu_1 - \mu_2)^2} d(\delta, 1 - \delta).$$

Note that when $\sigma_{1\mathcal{X}}^2 > 0$ or $\sigma_{2\mathcal{X}}^2 > 0$, $\sigma_1 + \sigma_2 > \sigma_1' + \sigma_2'$; that is, the value of the lower bound derived in Theorem 3 is strictly smaller than that of the lower bound derived by Kaufmann et al. (2016), $\frac{2(\sigma_1+\sigma_2)^2}{(\mu_1-\mu_2)^2}\mathrm{kl}(\delta, 1 - \delta)$. Let $\alpha = \sigma_1'/(\sigma_1' + \sigma_2')$. We also note that the simple α-elimination algorithm by Kaufmann et al. (2016) with α achieves the lower bound as well as a strictly better sample complexity than that given in Kaufmann et al. (2016). We give the proof in Appendix E.3.

Theorem 4 *If $\alpha = \sigma_1'/(\sigma_1' + \sigma_2')$, then the α-elimination strategy using the exploration rate $\beta(t, \delta) = \log \frac{t}{\delta} + 2 \log \log(6t)$ is δ-PAC on $\widetilde{\Omega}$ and for every $\mathcal{V} \in \widetilde{\Omega}$ and $\epsilon > 0$, satisfies*

$$\mathbb{E}_{\mathcal{V}}[\tau_\delta] \leq (1 + \epsilon)\frac{2(\sigma_1' + \sigma_2')^2}{(\mu_1 - \mu_2)^2} \log\left(\frac{1}{\delta}\right) + o\left(\log\left(\frac{1}{\delta}\right)\right).$$

Algorithm 1: α-elimination with contextual information

Input: Confidence level δ, threshold $\beta(t,\delta)$, σ_a, $\sigma_{\mathcal{X}}$, $\rho_{a\mathcal{X}}$, $\mu_{\mathcal{X}}$.

Initialization: $t=0$. $\hat{\mu}_1(0)=\hat{\mu}_2(0)=0$, $\sigma_0^2(\alpha)=1$.

$\sigma_1'^2 \leftarrow \sigma_1^2 - \frac{\sigma_{1\mathcal{X}}^2}{\sigma_{\mathcal{X}}^2}$, $\sigma_2'^2 \leftarrow \sigma_2^2 - \frac{\sigma_{2\mathcal{X}}^2}{\sigma_{\mathcal{X}}^2}$.

$\alpha \leftarrow \sigma_1'/(\sigma_1'+\sigma_2')$

while $|\hat{\mu}_1(t)-\hat{\mu}_2(t)| \leq \sqrt{2\sigma_t^2(\alpha)\beta(t,\delta)}$ **do**

  $t \leftarrow t+1$.

  Observe X_t.

  if $\lceil \alpha t \rceil = \lceil \alpha(t-1) \rceil$ **then**

    $A_t \leftarrow 2$

  else

    $A_t \leftarrow 1$

  end

  Observe R_t.

  $\hat{\mu}_1(t) = \frac{1}{\sum_{s=1}^t \mathbb{1}[A_s=1]} \sum_{s=1}^t \left(R_{s,1} - \frac{\rho_{1\mathcal{X}}\sigma_1}{\sigma_{\mathcal{X}}}(X_s - \mu_{\mathcal{X}}) \right) \mathbb{1}[A_s=1]$.

  $\hat{\mu}_2(t) = \frac{1}{\sum_{s=1}^t \mathbb{1}[A_s=2]} \sum_{s=1}^t \left(R_{s,2} - \frac{\rho_{2\mathcal{X}}\sigma_2}{\sigma_{\mathcal{X}}}(X_s - \mu_{\mathcal{X}}) \right) \mathbb{1}[A_s=2]$.

  Compute $\sigma_t^2(\alpha) = \sigma_1'^2/\lceil \alpha t \rceil + \sigma_2'^2/(t - \lceil \alpha t \rceil)$.

end

return $\arg\max_{a=1,2} \hat{\mu}_a(t)$

Hence, α-elimination is optimal for this problem. The details of α-elimination with contextual information is shown in Appendix E.1. The pseudo-code is shown in Algorithm 1. Thus, apparently irrelevant contextual information improves optimal sample complexity.

Gaussian bandits with unknown variances. When variances are unknown, our results cannot be directly applied. After the initial public draft of this study, Jourdan et al. (2023) considered Gaussian distributions with unknown variances, while Garivier and Kaufmann (2016) focused on Gaussian distributions with known variances. In Jourdan et al. (2023), they assume a class of distributions with unknown variances and develop corresponding lower and upper bounds. Unlike our setting and that of Garivier and Kaufmann (2016), the variances are not fixed for the distributions in their class and vary across different instances. They then propose a strategy and demonstrate its asymptotic optimality. Their findings suggest that specific algorithms and theoretical analyses are needed to handle unknown variances. While combining our approach with the results of Jourdan et al. (2023) is a promising direction for future work, it is beyond the scope of this study and remains an avenue for future investigation.

4. Lower Bound with Finite Contexts

Although we derived the optimal algorithm for BAI with continuous contexts in the previous sections, it requires some assumptions that may not be practical, e.g., multivariate normal distribution and known variance. We also consider a more practical algorithm by considering

BAI with finite contexts. In this section, we consider a lower bound when the number of contexts is finite. For $\nu = ((\mu_{a,x})_{a,x}, (\zeta_x))$, we suppose that $\mathcal{X}$ is finite (ζ follows the multinomial distribution), and for each arm a and context x, arm distribution belongs to the canonical one-parameter exponential family (Cappé et al., 2013; Kaufmann et al., 2016; Garivier and Kaufmann, 2016; Juneja and Krishnasamy, 2019):

$$\mathcal{P} = \left\{ (p_\pi)_{\pi \in \Pi} : \frac{\mathrm{d}p_\pi}{\mathrm{d}\lambda} = \exp(\pi u - b(\pi)) \right\}, \tag{1}$$

where λ is some reference measure on $\mathbb{R}$, $b : \Pi \mapsto \mathbb{R}$ is a convex, twice differentiable function, and $\Pi \subset \mathbb{R}$ is a parameter space. Note that a distribution $p_\pi \in \mathcal{P}$ can be parameterized by its mean $\dot{b}(\pi)$. As discussed in Cappé et al. (2013); Garivier and Kaufmann (2016), the KL divergence from p_π to $p_{\pi'}$ is given by

$$\mathrm{KL}(p_\pi, p_{\pi'}) = \mathrm{kl}(\dot{b}(\pi), \dot{b}(\pi')) = b(\pi') - b(\pi) - \dot{b}(\pi)(\pi' - \pi).$$

For each arm a and context x pair, we represent the unique distribution in $\mathcal{P}$ by $(\mu_{a,x})$. We further write the multinomial contextual distribution by (ζ_x).

We denote by Θ a set of BAI problems with finite contexts and the single parameter (canonical) exponential family. The lower bound is given in the following theorem.

Theorem 5 *Let $\delta \in (0, 1/2)$. For any δ-PAC strategy and any $\mathcal{V} = ((\mu_{a,x}), (\zeta_x)) \in \Theta$, it holds that*

$$\mathbb{E}_\nu[\tau_\delta] \geq T^\star(\nu) d(\delta, 1-\delta),$$

where

$$T^\star(\nu)^{-1} \coloneqq \sup_{\boldsymbol{w} \in \mathcal{W}} \inf_{((\lambda_{a,x}),(\zeta_x)) \in \mathrm{Alt}(\nu)} \sum_{x \in \mathcal{X}} \zeta_x \sum_{a=1}^{K} w_{a,x} \mathrm{kl}(\mu_{a,x}, \lambda_{a,x}).$$

As an intuition behind $T^\star(\nu)$, the probability of misidentification is roughly $\exp(-\tau \left(T^\star(\nu)\right)^{-1})$; that is, larger $\left(T^\star(\nu)\right)^{-1}$ means a strategy with smaller sample complexity.

We note specific properties of this lower bound. From the results in Garivier and Kaufmann (2016), we know that when the optimal arm is unique, the expected value of the sampling budget of the optimal BAI algorithm does not diverge; rather it is less than or equal to the order of $\log(1/\delta)$. Therefore, from the assumption of the proposed model, $T^\star(\nu)$ is finite under certain regularity conditions; for example, the context marginalized distribution of the reward R_t is sub-Gaussian.

To derive the lower bound, we show the following lemma, which is an extension of Lemma 1 of Kaufmann et al. (2016).

Lemma 6 *Let $N_{a,x}(\tau) = \sum_{t=1}^{\tau} \mathbb{1}\{X_t = x, A_t = a\}$. Let $\nu = ((\mu_{a,x}), \zeta), \nu' = ((\lambda_{a,x}), \zeta) \in \Theta$. For any almost surely finite stopping time τ with respect to $(\mathcal{G}_t)_{t \geq 1}$, it holds that*

$$\sum_{x \in \mathcal{X}} \sum_{a \in [K]} \mathbb{E}_\nu[N_{a,x}(\tau)] \mathrm{kl}(\mu_{a,x}, \lambda_{a,x}) \geq \sup_{\mathcal{E} \in \mathcal{G}_\tau} d(\mathbb{P}_\nu(\mathcal{E}), \mathbb{P}_{\nu'}(\mathcal{E})).$$

The proof is provided in Appendix B. Here, we offer the proof sketch of Theorem 5 as follows.

Proof sketch. From Lemma 6 with $\mathcal{E} = \{\hat{a}_\tau = a^*(\nu)\}$, for each $\nu \in \Theta$ and $\nu' \in \mathrm{Alt}(\nu)$, we have

$$\sum_{x\in\mathcal{X}}\sum_{a\in[K]} \mathbb{E}_\nu[N_{a,x}(\tau)]\mathrm{kl}(\mu_{a,x},\lambda_{a,x}) \geq d(\mathbb{P}_\nu(\mathcal{E}),\mathbb{P}_{\nu'}(\mathcal{E})) \geq d(\delta,1-\delta),$$

where, for the last inequality, we use the definition of the δ-PAC algorithm and monotonicity of the KL divergence. Then, for each $\nu \in \Theta$, for some $(w_{a,x})_{a\in[K],x\in\mathcal{X}} \in \mathcal{W}$, we can obtain $d(\delta,1-\delta) \leq \mathbb{E}_\nu[\tau_\delta] \sup_{\boldsymbol{w}\in\mathcal{W}} \inf_{((\lambda_{a,x}),\zeta)\in\mathrm{Alt}(\nu)} \sum_{x\in\mathcal{X}} \zeta_x \sum_{a\in[K]} w_{a,x}\mathrm{kl}(\mu_{a,x},\lambda_{a,x})$.

In Section 5.3, we explain that the lower bound with contextual information is smaller than or equal to the lower bound without contextual information shown by Garivier and Kaufmann (2016).

5. Optimal Allocation in Contextual BAI with Finite Contexts

In this section, we first provide a simplification of the lower bound derived in Section 2. Then, we examine the characteristics of the optimal allocations used in the proof. It becomes apparent that the set of optimal allocations is, in general, not unique. Therefore, we define the notion of convergence to the set and prove that the estimated optimal allocations converge to the set of optimal allocations (even though they might not converge to a point).

5.1 Simplification of the Lower Bound

Without loss of generality, let $a^*(\nu) = 1$. First, we show a simpler equivalence form for the optimization problem $T^\star(\nu)^{-1}$ in the following theorem.

Lemma 7 *For each $\boldsymbol{w} \in \mathcal{W}$, we have*

$$\begin{aligned}
&\inf_{((\lambda_{a,x}),\zeta)\in\mathrm{Alt}(\nu)} \sum_{x\in\mathcal{X}} \zeta_x \sum_{a=1}^{K} w_{a,x}\mathrm{kl}(\mu_{a,x},\lambda_{a,x}) \\
&= \min_{a\neq 1} \inf_{\sum_{x\in\mathcal{X}}\zeta_x\lambda_{a,x} > \sum_{x\in\mathcal{X}}\zeta_x\lambda_{1,x}} \sum_{x\in\mathcal{X}} \zeta_x \Big(w_{1,x}\mathrm{kl}(\mu_{1,x},\lambda_{1,x}) + w_{a,x}\mathrm{kl}(\mu_{a,x},\lambda_{a,x}) \Big) \qquad (2)
\end{aligned}$$

We provide the proof in Appendix F.1.

Moreover, we can further simplify the constraint in the minimization problem. We define

$$f_a((\lambda_{a,x})) = \sum_{x\in\mathcal{X}} \zeta_x \left\{ w_{1,x}\mathrm{kl}(\mu_{1,x},\lambda_{1,x}) + w_{a,x}\mathrm{kl}(\mu_{a,x},\lambda_{a,x}) \right\}.$$

Then, we show the following lemma.

Lemma 8 *For each $\boldsymbol{w} \in \mathcal{W}$, suppose that $(\lambda^*_{a,x})$ satisfies*

$$\min_{a\neq 1} \inf_{\sum_{x\in\mathcal{X}}\zeta_x\lambda_{a,x} > \sum_{x\in\mathcal{X}}\zeta_x\lambda_{1,x}} f_a((\lambda_{a,x})) = \min_{a\neq 1} f_a((\lambda^*_{a,x})).$$

For all $a^* \in \arg\min_{a\in[K]} \inf_{\sum_{x\in\mathcal{X}}\zeta_x\lambda_{a,x} > \sum_{x\in\mathcal{X}}\zeta_x\lambda_{1,x}} f_a((\lambda_{a,x}))$, we have

$$\sum_{x\in\mathcal{X}} \zeta_x \lambda^*_{a,x} = \sum_{x\in\mathcal{X}} \zeta_x \lambda^*_{1,x}.$$

Consequently, we can equivalently write the optimization problem as

$$T^\star(\nu)^{-1} = \max_{\boldsymbol{w}\in\mathcal{W}} \min_{a\neq 1} L_{1,a}((\mu_{1,x}, \mu_{a,x}, \zeta_x, w_{1,x}, w_{a,x})_{x\in\mathcal{X}}), \tag{3}$$

where for $a, b \in [K]$,

$$\begin{aligned} &L_{a,b}((\mu_{a,x}, \mu_{b,x}, \zeta_x, w_{a,x}, w_{b,x})_{x\in\mathcal{X}}) \\ &= \min_{\sum_{x\in\mathcal{X}} \zeta_x \lambda_{b,x} = \sum_{x\in\mathcal{X}} \zeta_x \lambda_{a,x}} \sum_{x\in\mathcal{X}} \zeta_x \left\{ w_{a,x}\mathrm{kl}(\mu_{a,x}, \lambda_{a,x}) + w_{b,x}\mathrm{kl}(\mu_{b,x}, \lambda_{b,x}) \right\} \end{aligned}$$

We provide the proof in Appendix F.2.

5.2 Characteristics of the Lower Bound

Let $2^{\mathcal{W}}$ be a power set of $\mathcal{W}$. We define a point-to-set map $\Phi : \Theta \to 2^{\mathcal{W}}$; that is, the set of all optimal allocations for the bandit problem ν as

$$\Phi(\nu) = \left\{ \boldsymbol{w} \in \mathcal{W} \mid m(\boldsymbol{w}, \nu) = \max_{\boldsymbol{w}'\in\mathcal{W}} m(\boldsymbol{w}', \nu) \right\},$$

where

$$m(\boldsymbol{w}, \nu) = \min_{a\neq 1} \min_{\sum_{x\in\mathcal{X}} \zeta_x \lambda_{a,x} = \sum_{x\in\mathcal{X}} \zeta_x \lambda_{1,x}} \sum_{x\in\mathcal{X}} \zeta_x \left\{ w_{1,x}\mathrm{kl}(\mu_{1,x}, \lambda_{1,x}) + w_{a,x}\mathrm{kl}(\mu_{a,x}, \lambda_{a,x}) \right\}.$$

The interpretation of $m(\boldsymbol{w}, \nu)$ is that, unlike the corresponding part in Garivier and Kaufmann (2016), we can further minimize the lower bound by choosing an optimal allocation from a wider domain than the case without contextual information as long as the constraints are satisfied. For example, let us consider a case where two arms a and b, and two contexts 1 and 2 are given. Here, under certain circumstances, one needs to think about saving the allocations to arm a in context 1, allocating more to arm a in context 2, and get more budget to arm b in context 1. Thus, solving $m(\boldsymbol{w}, \nu)$ is inherently different from optimizing the allocations separately for each context; that is, a case where we apply a BAI algorithm without contextual information for each discrete context such as Garivier and Kaufmann (2016).

From this simplified formula of the lower bound, we obtain the following lemmas. We provide the proofs in Appendix F.3-F.4.

Lemma 9 *Fix* $\boldsymbol{w} \in \mathcal{W}$. *We regard* ν *as a point in* $\mathbb{R}^{|\mathcal{X}|(K+1)}$: $\nu = ((\mu_{a,x}), \zeta) \in \mathbb{R}^{|\mathcal{X}|(K+1)}$. *Then,* $m(\boldsymbol{w}, \nu)$ *is continuous at every* $\nu \in \Theta$.

Note that the reason why ν is in $\mathbb{R}^{|\mathcal{X}|(K+1)}$ is that we include $\zeta \in \mathbb{R}^{|\mathcal{X}|}$ in ν with $(\mu_{a,x}) \in \mathbb{R}^{|\mathcal{X}|K}$.

Lemma 10 *Fix* $\nu \in \Theta$. *Then,* $m(\boldsymbol{w}, \nu)$ *is continuous at every* $\boldsymbol{w} \in \mathcal{W}$.

The set of the optimal allocations is not, in general, unique. Therefore, we introduce the notion of convergence, where the metric is defined as the minimum distance from the point to the set.

Definition 11 *Let $(\boldsymbol{w}_k)_{k\geq 1} = ((w_{a,x}^{(k)}))_{k\geq 1}$ be a sequence of points in $\mathcal{W}$. Let $\bar{\mathcal{W}} \subset \mathcal{W}$. We say $(\boldsymbol{w}_k)_{k\geq 1}$ converges to $\bar{\mathcal{W}}$ if for any $\varepsilon > 0$, there exists $n(\varepsilon) \in \mathbb{N}$ such that for all $k \geq n(\varepsilon)$, it holds that*

$$\inf_{(w_{a,x})\in\bar{\mathcal{W}}} \max_{a,x} |w_{a,x}^{(k)} - w_{a,x}| < \varepsilon.$$

Using this definition of convergence, we obtain the following lemmas. We provide the proofs in Appendix F.5–F.6.

Lemma 12 *Let $(\nu^k = ((\mu_{a,x}^{(k)}), \zeta_x^{(k)}))_{k\geq 1}$ be a sequence converging to ν. Construct a sequence $(\boldsymbol{w}_k)_{k\geq 1}$ such that $\boldsymbol{w}_k \in \Phi(\nu^k)$. Then $\boldsymbol{w}_k$ converges to $\Phi(\nu)$.*

Lemma 13 *The set of all optimal allocations for the bandit problem ν, $\Phi(\nu)$, is convex.*

5.3 Efficiency Gain

Here, we show that the lower bound with contextual information is smaller than or equal to the lower bound without contextual information shown by Garivier and Kaufmann (2016). For simplicity of discussion, we consider a two-armed bandit case. Let us denote the lower bound without contextual information by $\Gamma^\star(\nu)\mathrm{kl}(\mathbb{P}_\nu(\mathcal{E}), \mathbb{P}_{\nu'}(\mathcal{E}))$, where $\Gamma^\star(\nu)$ is defined as the same quantity as $T^\star(\boldsymbol{\mu})$ in Garivier and Kaufmann (2016). Let us also denote the optimal allocation in Garivier and Kaufmann (2016) by γ_1^* and γ_2^* and one of the optimal allocations in ours by $(w_{1,x}^*)$ and $(w_{2,x}^*)$. Then, $\Gamma^\star(\nu)^{-1} \leq T^\star(\nu)^{-1}$ holds as follows:

$$\begin{aligned}
&\Gamma^\star(\nu)^{-1} \\
&= \min_{\lambda_1=\lambda_2} \left\{ \gamma_1^* \mathrm{kl}(\mu_1, \lambda_1) + \gamma_2^* \mathrm{kl}(\mu_2, \lambda_2) \right\} \\
&= \min_{\lambda_1=\lambda_2} \left\{ \gamma_1^* \mathrm{kl}\left(\sum_{x\in\mathcal{X}} \zeta_x \mu_{1,x}, \lambda_1 \right) + \gamma_2^* \mathrm{kl}\left(\sum_{x\in\mathcal{X}} \zeta_x \mu_{2,x}, \lambda_2 \right) \right\} \\
&= \min_{\sum_{x\in\mathcal{X}} \zeta_x \lambda_{1,x} = \sum_{x\in\mathcal{X}} \zeta_x \lambda_{2,x}} \left\{ \gamma_1^* \mathrm{kl}\left(\sum_{x\in\mathcal{X}} \zeta_x \mu_{1,x}, \sum_{x\in\mathcal{X}} \zeta_x \lambda_{1,x} \right) + \gamma_2^* \mathrm{kl}\left(\sum_{x\in\mathcal{X}} \zeta_x \mu_{2,x}, \sum_{x\in\mathcal{X}} \zeta_x \lambda_{2,x} \right) \right\} \\
&\overset{(a)}{\leq} \min_{\sum_{x\in\mathcal{X}} \zeta_x \lambda_{1,x} = \sum_{x\in\mathcal{X}} \zeta_x \lambda_{2,x}} \sum_{x\in\mathcal{X}} \zeta_x \left\{ \gamma_1^* \mathrm{kl}\left(\mu_{1,x}, \lambda_{1,x}\right) + \gamma_2^* \mathrm{kl}\left(\mu_{2,x}, \lambda_{2,x}\right) \right\} \\
&\leq \min_{\sum_{x\in\mathcal{X}} \zeta_x \lambda_{1,x} = \sum_{x\in\mathcal{X}} \zeta_x \lambda_{2,x}} \sum_{x\in\mathcal{X}} \zeta_x \left\{ w_{1,x}^* \mathrm{kl}\left(\mu_{1,x}, \lambda_{1,x}\right) + w_{2,x}^* \mathrm{kl}\left(\mu_{2,x}, \lambda_{2,x}\right) \right\} \\
&= T^\star(\nu)^{-1},
\end{aligned}$$

where for (a), we use the convexity of the KL divergence. Next, we discuss when the equality holds. For brevity, we consider a case with only two contexts. Let us denote the optimal λ_1 in the case without contextual information by λ_1^* (note that $\lambda_1 = \lambda_2$) and the optimal $(\lambda_{1,1}, \lambda_{1,2})$ and $(\lambda_{2,1}, \lambda_{2,2})$ in the case with contextual information by $(\lambda_{1,1}^*, \lambda_{1,2}^*)$ and $(\lambda_{2,1}^*, \lambda_{2,2}^*)$. Then, the equality holds only if the following three conditions simultaneously hold:

- $\lambda_1^* = \zeta_1 \lambda_{1,1}^* = \zeta_2 \lambda_{1,2}^* = \zeta_1 \lambda_{2,1}^* = \zeta_2 \lambda_{2,2}^*$;
- $\frac{\mu_{1,1}}{\lambda_{1,1}^*} = \frac{\mu_{1,1}}{\lambda_{1,2}^*}$ and $\frac{\mu_{2,1}}{\lambda_{2,1}^*} = \frac{\mu_{2,2}}{\lambda_{2,2}^*}$;
- $\gamma_1^* = \gamma_{1,1}^* = \gamma_{1,2}^*$.

We believe that it is difficult to summarize these conditions in a simpler form, but except for cases where the expected reward does not change among contexts, situations satisfying these conditions are extremely limited.

6. Contextual Track-and-Stop Algorithm

In this section, we propose an optimal algorithm for contextual BAI, called the Contextual Track-and-Stop (CTS) algorithm for the case of finite context. The strategy is an extension of the Track-and-Stop (TS) algorithm by Garivier and Kaufmann (2016) for contextual BAI. We further prove that the proposed algorithm is δ-PAC.

Recall that the optimal algorithm of BAI with fixed confidence Garivier and Kaufmann (2016) consists of sampling, stopping, and decision rules. We follow the same path for the contextual BAI. We show the pseudo-code of the proposed CTS algorithm in Algorithm 2. There, the empirical averages $\hat{\mu}_{a,x}(t)$ and $\hat{\zeta}_x(t)$ are defined as follows: for each $a \in [K]$ and $x \in \mathcal{X}$, $\hat{\mu}_{a,x}(t) = \left(\sum_{s=1}^t R_s \mathbb{1}\{A_s = a, X_s = x\}\right) / N_{a,x}(t)$ and $\hat{\zeta}_x(t) = \left(\sum_{s=1}^t \mathbb{1}\{X_s = x\}\right) / t$. Our procedure is similar to TS with D-tracking, proposed by Garivier and Kaufmann (2016). However, incorporating contextual information is a non-trivial extension of their method. The algorithm consists of sampling, stopping, and decision rules. The details of the sampling rule are described in Section 6.1. The stopping rule, in particular, for determining the threshold $\beta(t, \delta)$, is described in Section 6.2 when the reward distributions are Bernoulli and in Section 6.3 when the reward distributions belong to the canonical one-parameter exponential family.

Our proposed algorithm consists of sampling, stopping, and recommendation rules. In the sampling rule, we use forced exploration, which is an extension of D-tracking of Garivier and Kaufmann (2016) and is known to be empirically superior to their C-tracking. To estimate the optimal weights, we solve an empirically approximated optimization problem (3) by applying optimization solvers directly. Several methods are proposed to solve the maximin problem more efficiently, such as the application of no-regret learning algorithms in Degenne et al. (2019). However, we cannot use them directly for solving contextual BAI, in which we have a different form of the maximin problem than that of BAI without context. Jedra and Proutiere (2020) (BAI with linear models) and Russac et al. (2021) (A/B/n testing with contextual information) also directly solve the maximin problem. In the stopping rule, we use the criterion proposed by Kaufmann and Koolen (2021), which refines the stopping rule of Garivier and Kaufmann (2016). Then, we recommend an arm with the maximum sample average of the reward.

6.1 Sampling Rule

To design an algorithm with minimal sample complexity, the sampling rule should match the optimal proportions of the arm draws; that is, an allocation in the set $\Phi(\nu)$. Because

Algorithm 2: CTS algorithm

Input: Confidence level δ and threshold $\beta(t,\delta)$.
Initialization: $t=0$, $N_x(0)=0$, $N_{a,x}(0)=0$.
while $(Z(t) := \max_{a\in[K]} \min_{b\in[K]\setminus\{a\}} Z_{a,b}(t) < \beta(t,\delta))$ **do**
  $t \leftarrow t+1$.
  Observe X_t.
  if $\varphi^g_{X_t,t} = \{a : N_{a,X_t}(t) < \sqrt{N_x(t)} - K/2\} \neq \emptyset$ **then**
    $a \leftarrow \arg\min_{a\in\varphi^g_{X_t,t}} N_{a,X_t}(t)$
  else
    $a \leftarrow \arg\max_{a\in[K]} \sum_{s=1}^t \mathbb{1}[X_s = X_t] w_{a,X_t}(t) - N_{a,X_t}(t)$.
  end
  Sample arm a and update $N_x(t)$, $N_{a,x}(t)$, $\hat{\zeta}_x(t)$, $\hat{\mu}_{a,x}(t)$, $Z(t)$.
  $\hat{a}_t = \arg\max_{a\in[K]} \sum_{x\in\mathcal{X}} \hat{\zeta}_x(t)\hat{\mu}_{a,x}(t)$.
  $\boldsymbol{w}(t) \leftarrow \arg\max_{\boldsymbol{w}\in\mathcal{W}} \min_{a\neq\hat{a}_t} L_{\hat{a}_t,a}((\hat{\mu}_{\hat{a}_t,x}(t), \hat{\mu}_{a,x}(t), \hat{\zeta}_x(t), w_{\hat{a}_t,x}, w_{a,x})_{x\in\mathcal{X}})$.
end
return $\hat{a}_\tau = \hat{a}_t$

$\mu_{a,x}$ and ζ_x are unknown, our sampling rule tracks, in round t, the optimal allocations in the plug-in estimate $\Phi(\hat{\nu}(t))$, where $\hat{\nu}(t) = ((\hat{\mu}_{a,x}(t))_{a\in[K],x\in\mathcal{X}}, (\hat{\zeta}_x(t))_{x\in\mathcal{X}})$.

The design of our tracking rule is equivalent to computing a sequence of allocations $(w_{a,x}(t))_{t\geq 1}$. The only requirement we actually impose on this sequence is the following condition:

$$\lim_{t\to\infty} \min_{\boldsymbol{w}'\in\Phi(\hat{\nu}(t))} \max_{a\in[K],x\in\mathcal{X}} \left| w_{a,x}(t) - w'_{a,x} \right| = 0. \qquad \text{a.s.} \tag{4}$$

This condition is sufficient to guarantee the asymptotic optimality of the algorithm. We introduce a set $\varphi^g_{x,t} = \{a : N_{a,x}(t) < \sqrt{N_x(t)} - K/2\}$ consisting of the context-action pairs that are poorly explored. Then, in round t, after observing a context $X_t \in \mathcal{X}$, our sampling rule (A_t) is sequentially defined as

$$A_t \in \begin{cases} \arg\min_{a\in\varphi^g_{X_t,t}} N_{a,X_t} & \text{if } \varphi^g_{X_t,t} \neq \emptyset \\ \arg\max_{1\leq a\leq K} \sum_{s=0}^t \mathbb{1}[X_s = X_t] w_{a,X_t}(t) - N_{a,X_t}(t). \end{cases} \tag{5}$$

We offer the following lemma under this sampling rule. The proof is provided in Appendix G.1.

Lemma 14 *Under any sampling rule (5) that satisfies the condition (4), it holds that*

$$\mathbb{P}_\nu \left(\inf_{\boldsymbol{w}^*\in\Phi(\nu)} \lim_{t\to\infty} \max_{a\in[K],x\in\mathcal{X}} \left| \frac{N_{a,x}(t)}{t} - \zeta_x w^*_{a,x} \right| = 0 \right) = 1.$$

This lemma shows that the sampling rule can keep the allocation close to the optimal allocations. Thus, together with any sequence of allocations satisfying (4), the sampling rule defined by (5) keeps the empirical allocation close to the set of optimal allocations.

To compute $w_{a,x}(t)$ in (5), we need to solve the minimax problem defined in (3) with the estimated parameters. If the number of contexts and arms is very large, it may be difficult to solve. However, except for such an extreme case, we can solve the problem by using minimax optimization based on the convex optimization algorithm in a short time. The computation is similar to that in Jedra and Proutiere (2020).

We remark that the application of the original TS algorithm (Garivier and Kaufmann, 2016) for each context separately is not optimal for contextual BAI. Our problem setting makes finding the best allocations difficult, which is quite different from running BAI in parallel for each context. It is necessary to find good allocations of each arm to the right context, and the allocations among contexts are entangled. For example, to achieve our derived lower bound, one needs to think about saving the allocations to an arm a in context 1, then allocating more to arm a in context 2, and getting more budget to another arm b in context 1. In contrast, when separately applying the original TS algorithm, we cannot attain such an optimal allocation.

6.2 Threshold in the Stopping Rule

In this subsection, we present the stopping rule, in particular the threshold for the Bernoulli bandit model. We aim to design an algorithm that stops as early as possible while maintaining the failure probability at most δ. We demonstrate that the stopping rule using the generalized likelihood ratio test (GLRT) for contextual BAI is δ-PAC when the exploration ratio is properly tuned. Such a stopping rule is also known as Chernoff's stopping rule (Chernoff, 1959). Although the approach for deriving the threshold is inspired by and similar to that of Garivier and Kaufmann (2016), our computation with the contextual information is more involved.

We consider a case where the reward $R_{t,a}$ follows a Bernoulli distribution conditioned on $X_t = x$. Here, the likelihood is given as

$$p_{\mu_a}\big((\underline{R}_{a,x}(t)), \underline{X}(t)\big) = \prod_{x\in\mathcal{X}} \big(\zeta_x \mu_{a,x}\big)^{\sum_{s=1}^t \mathbb{1}[A_s=a, X_s=x, R_s=1]} \big(\zeta_x(1-\mu_{a,x})\big)^{\sum_{s=1}^t \mathbb{1}[A_s=a, X_s=x, R_s=0]}.$$

Then, for all pairs of the arms, $a, b \in [K]$, the GLRT statistic is given as

$$Z_{a,b}(t) = \log \frac{\max_{\overline{\xi}_a(t)\geq\overline{\xi}_b(t)} p_{\xi_a}\big((\underline{R}_{a,x}(t)), \underline{X}(t)\big) p_{\xi_b}\big((\underline{R}_{b,x}(t)), \underline{X}(t)\big)}{\max_{\overline{\xi}_a(t)\leq\overline{\xi}_b(t)} p_{\xi_a}\big((\underline{R}_{a,x}(t)), \underline{X}(t)\big) p_{\xi_b}\big((\underline{R}_{b,x}(t)), \underline{X}(t)\big)},$$

where $\overline{\xi}_a(t) = \sum_{x\in\mathcal{X}} \hat{\zeta}_x(t)\xi_{a,x}$. Note that the maximizer of

$$\max_{\overline{\xi}_a(t)\geq\overline{\xi}_b(t)} p_{\xi_a}\big((\underline{R}_{a,x}(t)), \underline{X}(t)\big) p_{\xi_b}\big((\underline{R}_{b,x}(t)), \underline{X}(t)\big)$$

is equivalent to that of

$$\max_{\substack{(\xi_{a,x},\xi_{b,x})_{x\in\mathcal{X}}\in[0,1]^{|\mathcal{X}|\times 2} \\ \sum_{x\in\mathcal{X}} \hat{\zeta}_x(t)\xi_{a,x} \geq \sum_{x\in\mathcal{X}} \hat{\zeta}_x(t)\xi_{b,x}}} t \sum_{x\in\mathcal{X}} \sum_{c\in\{a,b\}} \left\{ \frac{N_{c,x}}{t} \left\{ \hat{\mu}_{c,x}(t) \log \frac{\xi_{c,x}}{1-\xi_{c,x}} + \log(1-\xi_{c,x}) \right\} \right\}.$$

We denote the maximizers by $(\widetilde{\xi}_{a,x}(t))$ and $(\widetilde{\xi}_{b,x}(t))$. Similarly, we denote the solution of the maximization problem in the denominator by $(\widetilde{\xi}^{\dagger}_{a,x}(t))$ and $(\widetilde{\xi}^{\dagger}_{b,x}(t))$.

In the numerator, if $\sum_{x\in\mathcal{X}} \hat{\zeta}_x(t)\hat{\mu}_{a,x}(t) \geq \sum_{x\in\mathcal{X}} \hat{\zeta}_x(t)\hat{\mu}_{b,x}(t)$, then the maximum likelihood estimator falls within the optimization constraint; that is, $\widetilde{\xi}_{a,x}(t) = \hat{\mu}_{a,x}(t)$ and $\widetilde{\xi}_{b,x}(t) = \hat{\mu}_{b,x}(t)$. Therefore, our remaining problem is to compute the denominator. Because $\sum_{x\in\mathcal{X}} \hat{\zeta}_x(t)\hat{\mu}_{a,x}(t) \geq \sum_{x\in\mathcal{X}} \hat{\zeta}_x(t)\hat{\mu}_{b,x}(t)$ does not satisfy the constraint condition in the denominator, it is hard to obtain the closed-form expression of the denominator and we need to solve the optimization problem numerically. Given the solutions, $(\widetilde{\xi}^{\dagger}_{a,x}(t))$ and $(\widetilde{\xi}^{\dagger}_{b,x}(t))$, the GLRT statistic $Z_{a,b}(t)$ is equal to

$$
\begin{aligned}
& t\sum_{x\in\mathcal{X}}\sum_{c\in\{a,b\}} \left\{ \frac{N_{c,x}(t)}{t}\Big\{ \hat{\mu}_{c,x}(t)\log\frac{\hat{\mu}_{c,x}(t)}{1-\hat{\mu}_{c,x}(t)} + \log(1-\hat{\mu}_{c,x}(t)) \right. \\
& \qquad\qquad \left. - \hat{\mu}_{c,x}(t)\log\frac{\widetilde{\xi}^{\dagger}_{c,x}(t)}{1-\widetilde{\xi}^{\dagger}_{c,x}(t)} - \log(1-\widetilde{\xi}^{\dagger}_{c,x}(t))\Big\} \right\} \\
&= \max_{\substack{(\xi_{a,x},\xi_{b,x})_{x\in\mathcal{X}} \\ \sum_{x\in\mathcal{X}}\hat{\zeta}_x(t)\xi_{a,x}\leq\sum_{x\in\mathcal{X}}\hat{\zeta}_x(t)\xi_{b,x}}} t\sum_{x\in\mathcal{X}}\left(\frac{N_{a,x}(t)}{t} d(\hat{\mu}_{a,x}(t),\xi_{a,x}(t)) + \frac{N_{b,x}(t)}{t} d(\hat{\mu}_{b,x}(t),\xi_{b,x}(t)) \right).
\end{aligned}
$$

By multiplying $Z_{a,b}(t)$ by $-1/t$, we can find that solving the maximization problem is equal to solving the inner minimization problem of (2), or equivalently the problem defined in (3), by letting $\zeta_x w_{1,x} = \frac{N_{a,x}}{t}$, $\zeta_x w_{b,x} = \frac{N_{a,x}}{t}$, $\mu_{1,x} = \hat{\mu}_{a,x}(t)$, and $\mu_{a,x} = \hat{\mu}_{b,x}(t)$. From Lemma 8, the constraint $\sum_{x\in\mathcal{X}} \zeta_x\xi_{a,x} \leq \sum_{x\in\mathcal{X}} \zeta_x\xi_{b,x}$ holds with equality; that is,

$$
Z_{a,b}(t) = tL_{a,b}\left(\left(\hat{\mu}_{a,x}(t), \hat{\mu}_{b,x}(t), \hat{\zeta}_x(t), N_{a,x}(t)/N_x(t), N_{b,x}(t)/N_x(t) \right)_{x\in\mathcal{X}} \right).
$$

It is also easy to observe that when $\sum_{x\in\mathcal{X}} \zeta_x\hat{\mu}_{a,x}(t) \leq \sum_{x\in\mathcal{X}} \zeta_x\hat{\mu}_{b,x}(t)$, then $Z_{a,b}(t) = -Z_{b,a}(t)$.

Using the GLRT statistic, we use the following stopping rule:

$$
\tau_\delta = \inf\Big\{ t\in\mathbb{N} : Z(t) := \max_{a\in[K]}\min_{b\in[K]\setminus\{a\}} Z_{a,b}(t) > \beta(t,\delta) \Big\}, \tag{6}
$$

where $\beta(t,\delta)$ is the threshold of the GLRT statistic $Z_{a,b}(t)$ (exploration rate), which controls the failure probability under the stopping rule.

Next, we determine $\beta(t,\delta)$ such that the proposed algorithm is δ-PAC. We present the following theorem to decide the threshold $\beta(t,\delta)$ in the stopping rule.

Theorem 15 *Let $\delta\in(0,1)$. For a Bernoulli bandit model, if $\beta(t,\delta) = \log\left(\frac{2t(K-1)}{\delta}\right)$, then for all $\nu\in\Theta$, it holds that*

$$
\mathbb{P}_\nu\left(\tau_\delta < \infty,\, \hat{a}_{\tau_\delta} \neq a^*\right) \leq \delta.
$$

The proof is provided in Appendix G.2. The proof with contextual information is accomplished by using the fact that joint distribution of the contexts and the rewards is the Multinomial distribution. This theorem confirms that the proposed algorithm is δ-PAC when $\beta(t,\delta) = \log\left((2t(K-1))/\delta\right)$. We note that this threshold does not depend on the cardinality of $\mathcal{X}$.

6.3 Stopping Rule for a Canonical One-parameter Exponential Family and Known Contextual Distribution

For the Bernoulli bandit, we derive the stopping and recommendation rule by using the fact that the rewards and finite contexts jointly follow a multinomial distribution.We cannot use this property when the conditional rewards follow different distributions such as a Gaussian distribution. For example, when the rewards follow a Gaussian distribution, the rewards and contexts jointly follow a Gaussian mixture model, not a Gaussian distribution. This fact makes derivation of the δ-PAC threshold difficult. However, if the contextual distribution is known, we can extend the existing results, such as Garivier and Kaufmann (2016) and Kaufmann and Koolen (2021), to derive the threshold.

We consider a case where for each $a \in [K]$ and $x \in \mathcal{X}$, the reward $R_{t,a}$ follows a distribution that belongs to the canonical one-parameter exponential family (1) conditioned on $X_t = x$ and the context X_t follows a multinomial distribution with known parameters; that is, we treat the estimator $\hat{\zeta}_x$ as the true value ζ_x in our proposed CTS algorithm. Similarly to the Bernoulli case, the likelihood of the observations $(\underline{R}_{a,x}(t))_{x\in\mathcal{X}}, \forall a \in [K]$ and $\underline{X}(t)$ regarding arm a is given as follows.

$$p_{\mu_a}\big((\underline{R}_{a,x}(t))_{x\in\mathcal{X}}, \underline{X}(t)\big) = \prod_{s=1}^{t}\prod_{x\in\mathcal{X}}\left(\zeta_x \exp\left(\dot{b}^{-1}(\mu_{a,x})R_s - b(\dot{b}^{-1}(\mu_{a,x}))\right)\right)^{\mathbb{1}[A_s=a, X_s=x]}.$$

Then, for all pairs of the arms, $a, b \in [K]$, the GLRT statistic is given as

$$Z_{a,b}(t) = \log \frac{\max_{\overline{\xi}_a(t)\geq\overline{\xi}_b(t)} p_{\xi_a}\big((\underline{R}_{a,x}(t)), \underline{X}(t)\big)p_{\xi_b}\big((\underline{R}_{b,x}(t)), \underline{X}(t)\big)}{\max_{\overline{\xi}_a(t)\leq\overline{\xi}_b(t)} p_{\xi_a}\big((\underline{R}_{a,x}(t)), \underline{X}(t)\big)p_{\xi_b}\big((\underline{R}_{b,x}(t)), \underline{X}(t)\big)},$$

where $\overline{\xi}_a(t) = \sum_{x\in\mathcal{X}} \zeta_x \xi_{a,x}$. For the numerator optimization problem, from the definition of the single parameter exponential family, the maximizer of

$$\max_{\overline{\xi}_a(t)\geq\overline{\xi}_b(t)} p_{\xi_a}\big((\underline{R}_{a,x}(t)), \underline{X}(t)\big)p_{\xi_b}\big((\underline{R}_{b,x}(t)), \underline{X}(t)\big)$$

is equivalent to the maximizer of the optimization problem

$$\max_{\substack{(\xi_{a,x},\xi_{b,x})_{x\in\mathcal{X}} \\ \sum_{x\in\mathcal{X}}\zeta_x\xi_{a,x}\geq\sum_{x\in\mathcal{X}}\zeta_x\xi_{b,x}}} t\sum_{x\in\mathcal{X}}\sum_{c\in\{a,b\}}\frac{N_{c,x}(t)}{t}\left\{\dot{b}^{-1}(\xi_{c,x})\hat{\mu}_{c,x}(t) - b(\dot{b}^{-1}(\xi_{c,x}))\right\}.$$

As for the case of a Bernoulli bandit model, using the notation, we compute the GLRT statistic $Z_{a,b}(t)$ as follows. Now, suppose that $\sum_{x\in\mathcal{X}}\zeta_x\hat{\mu}_{a,x}(t) \geq \sum_{x\in\mathcal{X}}\zeta_x\hat{\mu}_{b,x}(t)$. Then, $\widetilde{\xi}_{a,x}(t) = \hat{\mu}_{a,x}(t)$ in the numerator and $\widetilde{\xi}_{b,x}(t) = \hat{\mu}_{b,x}(t)$. We numerically solve the optimization problem in the denominator and obtain the solutions, $(\widetilde{\xi}^{\dagger}_{a,x}(t))$ and $(\widetilde{\xi}^{\dagger}_{b,x}(t))$. Then, $Z_{a,b}(t)$ is equal to

$$t\sum_{x\in\mathcal{X}}\sum_{c\in\{a,b\}}\frac{N_{c,x}(t)}{t}\Big\{\dot{b}^{-1}(\hat{\mu}_{c,x}(t))\hat{\mu}_{c,x}(t) - b(\dot{b}^{-1}(\hat{\mu}_{c,x}(t))) - \dot{b}^{-1}(\widetilde{\xi}^{\dagger}_{c,x}(t))\hat{\mu}_{c,x}(t) + b\left(\dot{b}^{-1}(\widetilde{\xi}^{\dagger}_{c,x}(t))\right)\Big\},$$

$$= \max_{\substack{(\xi_{a,x},\xi_{b,x})_{x\in\mathcal{X}} \\ \sum_{x\in\mathcal{X}}\zeta_x\xi_{a,x}\leq\sum_{x\in\mathcal{X}}\zeta_x\xi_{b,x}}} t\sum_{x\in\mathcal{X}}\left(\frac{N_{a,x}(t)}{t}\mathrm{kl}(\hat{\mu}_{a,x}(t),\xi_{a,x}(t))+\frac{N_{b,x}(t)}{t}\mathrm{kl}(\hat{\mu}_{b,x}(t),\xi_{b,x}(t))\right).$$

A similar argument can be made when $\sum_{x\in\mathcal{X}}\zeta_x\hat{\mu}_{a,x}(t) < \sum_{x\in\mathcal{X}}\zeta_x\hat{\mu}_{b,x}(t)$ by reversing the sign of the constraint.

Next, we define the stopping rule using the GLRT statistic $Z_{a,b}(t)$ as follows.

$$\tau_\delta = \inf\left\{t\in\mathbb{N} : Z(t) := \max_{a\in[K]}\min_{b\in[K]\setminus\{a\}} Z_{a,b}(t) > \beta(t,\delta)\right\},$$

where we decide the threshold $\beta(t,\delta)$ later. Let $\hat{\mu}_c(t) = \sum_{x\in\mathcal{X}}\zeta_x\hat{\mu}_{c,x}(t)$ for $c\in\mathcal{A}$. If $\sum_{x\in\mathcal{X}}\zeta_x\mu_{a,x} = \mu_a \leq \mu_b = \sum_{x\in\mathcal{X}}\zeta_x\mu_{b,x}$ and $\hat{\mu}_a > \hat{\mu}_b$, then

$$\begin{aligned}
&Z_{a,b}(t)\\
&= \min_{\substack{(\xi_{a,x},\xi_{b,x})_{x\in\mathcal{X}} \\ \sum_{x\in\mathcal{X}}\zeta_x\xi_{a,x}\leq\sum_{x\in\mathcal{X}}\zeta_x\xi_{b,x}}} t\sum_{x\in\mathcal{X}}\left(N_{a,x}(t)\mathrm{kl}(\hat{\mu}_{a,x}(t),\xi_{a,x}(t))+N_{b,x}(t)\mathrm{kl}(\hat{\mu}_{b,x}(t),\xi_{b,x}(t))\right)\\
&\leq \sum_{x\in\mathcal{X}}\left(N_{a,x}(t)\mathrm{kl}\left(\hat{\mu}_{a,x}(t),\mu_{a,x}\right)+N_{b,x}(t)\mathrm{kl}\left(\hat{\mu}_{b,x}(t),\mu_{b,x}\right)\right).
\end{aligned}$$

Then, we decompose the probability $\mathbb{P}_\nu\left(\tau_\delta<\infty,\, \hat{a}_{\tau_\delta}\neq a^*\right)$ as

$$\begin{aligned}
&\mathbb{P}_\nu\left(\tau_\delta<\infty,\, \hat{a}_{\tau_\delta}\neq a^*\right)\\
&\leq \mathbb{P}_\nu\left(\exists a\neq a^*, \exists t\in\mathbb{N} : \hat{\mu}_a(t) > \hat{\mu}_{a^*}(t), Z_{a,a^*}(t) > \beta(t,\delta)\right)\\
&\leq \mathbb{P}_\nu\left(\exists a\neq a^*, \exists t\in\mathbb{N} : \sum_{c\in\{a,a^*\}}\sum_{x\in\mathcal{X}} N_{c,x}(t)\mathrm{kl}\left(\hat{\mu}_{c,x}(t),\mu_{c,x}\right) > \beta(t,\delta)\right). \qquad (7)
\end{aligned}$$

Thus, if we choose a threshold $\beta(t,\delta)$ such that the upper bound of the last equation (7) is δ, we can guarantee that the algorithm is δ-PAC.

Using the results of Kaufmann and Koolen (2021), which refines existing deviation bounds and the threshold in Garivier and Kaufmann (2016), we can guarantee that the algorithm is δ-PAC with a tight threshold. We use the following theorem from Kaufmann and Koolen (2021).

Theorem 16 (From Theorem 7 of Kaufmann and Koolen (2021)) *Let us define $h(u) = u - \ln u$, $\forall u\geq 1$ and $h^{-1}(u)$ (the inverse of $h(u)$). For each $z\in[1,e]$ and for all $x\geq 0$, let*

$$\widetilde{h}_z(x) = \begin{cases} e^{1/h^{-1}(x)}h^{-1}(x) & \text{if } x\geq h(1/\ln z)\\ z(x-\ln\ln z) & \text{otherwise.}\end{cases}$$

We further define the function $\mathcal{C}_{\exp}:\mathbb{R}^+\mapsto\mathbb{R}^+$ as

$$\mathcal{C}_{\exp}(x) = 2\widetilde{h}_{3/2}\left(\frac{h^{-1}(1+x)+\ln(2\zeta(2))}{2}\right),$$

where $\zeta(s) = \sum_{n=1}^{\infty} n^{-s}$. *For each subset* $\mathcal{S}$ *of the context arm pairs* $(x, a) \in \mathcal{X} \times [K]$, *for all* $x > 0$, *the following holds:*

$$\mathbb{P}_\nu \left(\exists t \in \mathbb{N} : \sum_{(x,a)\in\mathcal{S}} N_{a,x}(t) \mathrm{kl}(\hat{\mu}_{a,x}, \mu_{a,x}) \geq \sum_{(x,a)\in\mathcal{S}} 3\ln(1 + \ln(N_{a,x}(t))) + |\mathcal{S}| \mathcal{C}_{\exp}\left(\frac{x}{|\mathcal{S}|}\right) \right) \leq \exp(-x).$$

Let us define the threshold $\beta(t, \delta)$ as

$$\beta(t, \delta) = 6|\mathcal{X}| \ln\left(\ln\left(\frac{t}{2}\right) + 1\right) + 2|\mathcal{X}| \mathcal{C}_{\exp}\left(\frac{\ln \frac{K-1}{\delta}}{2|\mathcal{X}|}\right).$$

Using this threshold, the following guarantee can be obtained.

Corollary 17 *Assume the context distribution is known; that is, we set* $\hat{\zeta}_x(t) = \zeta_x, \forall x \in \mathcal{X}, \forall t \in \mathbb{N}$ *in the GLRT statistics. Let* $\delta \in (0, 1)$. *For any sampling rule, using the stopping rule (6) with the threshold, we have*

$$\beta(t, \delta) = 6|\mathcal{X}| \ln\left(\ln\left(\frac{t}{2}\right) + 1\right) + 2|\mathcal{X}| \mathcal{C}_{\exp}\left(\frac{\ln \frac{K-1}{\delta}}{2|\mathcal{X}|}\right),$$

for all $\nu \in \Theta$, $\mathbb{P}_\nu(\tau_\delta < \infty, \hat{a}_{\tau_\delta} \neq a^*) \leq \delta$.

Proof

With Theorem 16 and the union bound over the set of $(K-1)$ pairs: $(a, a^*), a \neq a^*$, we bound $\mathbb{P}_\nu(\tau_\delta < \infty, \hat{a}_{\tau_\delta} \neq a^*)$ as

$$\begin{aligned}(7) \leq \mathbb{P}_\nu \Bigg(\exists a \neq a^*, \exists t \in \mathbb{N} : \sum_{c\in\{a,a^*\}} \sum_{x\in\mathcal{X}} N_{c,x}(t) \mathrm{kl}\left(\hat{\mu}_{c,x}(t), \mu_{c,x}\right) > \\ \sum_{c\in\{a,a^*\}} \sum_{x\in\mathcal{X}} 3\ln(1 + \ln(N_{c,x}(t))) + 2|\mathcal{X}| \mathcal{C}_{\exp}\left(\frac{\ln \frac{K-1}{\delta}}{2|\mathcal{X}|}\right) \Bigg) \\ \leq \delta.\end{aligned}$$

Furthermore, it is easy to check that $\mathcal{C}_{\exp}(x) = x + o(x)$ as $x \to \infty$ (Kaufmann and Koolen, 2021). ■

6.4 Sample Complexity Analysis

In this section, we address the upper bound of the sample complexity of the proposed CTS algorithm.

First, we demonstrate that the sample complexity asymptotically matches the lower bound almost surely for a case where the reward follows a Bernoulli bandit model.

Proposition 18 *Suppose that the reward follows a Bernoulli bandit model. If the sampling rule ensures that for all $a \in [K]$, for all $x \in \mathcal{X}$, $\min_{\boldsymbol{w}^* \in \Phi(\nu)} \left| \lim_{t\to\infty} \frac{N_{a,x}(t)}{t} - \zeta_x w^*_{a,x} \right| = 0$, and we follow the stopping rule defined in Section 6.2 with $\beta(t,\delta) = \log\left(\frac{2t(K-1)}{\delta}\right)$, then for all $\delta \in (0,1)$, it holds that $\mathbb{P}_\nu(\tau_\delta < \infty) = 1$ and*

$$\mathbb{P}_\nu\left(\limsup_{\delta\to 0} \frac{\tau_\delta}{\log(1/\delta)} \le T^\star(\nu)\right) = 1.$$

We provide the proof in Appendix H.1.

We now provide an upper bound on the expected stopping time $\mathbb{E}[\tau_\delta]$. The following theorem states that the proposed CTS algorithm asymptotically matches the sample complexity lower bound derived from Theorem 5. The proof of this result is provided in Appendix H.2.

Theorem 19 *Suppose that the reward follows a Bernoulli bandit model. For each $\nu \in \Theta$, if sampling rule ensures that for all $a \in [K]$, for all $x \in \mathcal{X}$, $\min_{\boldsymbol{w}^* \in \Phi(\nu)} \left| \lim_{t\to\infty} \frac{N_{a,x}(t)}{t} - \zeta_x w^*_{a,x} \right| = 0$, and we follow the stopping rule defined in Section 6.2 with $\beta(t,\delta) = \log\left(\frac{2t(K-1)}{\delta}\right)$, then*

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_\nu[\tau_\delta]}{\log(1/\delta)} \le T^\star(\nu)\ .$$

As well as the case with a Bernoulli bandit model, we can also show that an upper bound on the expected number of the stopping times $\mathbb{E}[\tau_\delta]$ matches the lower bound almost surely for a case where the reward follows a distribution that belongs to a canonical one-parameter exponential family, and the parameters of the context distribution are known.

Corollary 20 *Suppose that the reward follows a distribution that belongs to a canonical one-parameter exponential family, and $(\zeta_x)_{x\in\mathcal{X}}$ is known. For each $\nu \in \Theta$, if sampling rule ensures that for all $a \in [K]$, for all $x \in \mathcal{X}$, $\min_{\boldsymbol{w}^* \in \Phi(\nu)} \left| \lim_{t\to\infty} \frac{N_{a,x}(t)}{t} - \zeta_x w^*_{a,x} \right| = 0$, and we follow the stopping rule defined in Section 6.3 with $\beta(t,\delta) = 6|\mathcal{X}| \ln\left(\ln\left(\frac{t}{2}\right)+1\right) + 2|\mathcal{X}|\mathcal{C}_{\exp}\left(\frac{\ln\frac{K-1}{\delta}}{2|\mathcal{X}|}\right)$, then for all $\delta \in (0,1)$, it holds that $\mathbb{P}_\nu(\tau_\delta < +\infty) = 1$ and $\mathbb{P}_\nu\Big(\limsup_{\delta\to 0} \frac{\tau_\delta}{\log(1/\delta)} \le T^\star(\nu)\Big) = 1$. Moreover, $\limsup_{\delta\to 0} \frac{\mathbb{E}_\nu[\tau_\delta]}{\log(1/\delta)} \le T^\star(\nu)$ holds.*

6.5 General Context Spaces via Context Aggregation (Bernoulli Rewards)

We consider an algorithm when the context space is not finite. For Bernoulli rewards, we can extend CTS beyond finite context spaces by *aggregating* contexts.

Let $(\mathcal{X}, \mathcal{F})$ be an arbitrary measurable context space and suppose that, conditionally on $X_t = x$, arm a yields a Bernoulli reward with mean $\mu_a(x) \in (0,1)$. The best arm is defined by the marginal mean $\mu_a := \mathbb{E}[\mu_a(X)]$. Fix any measurable mapping $\varphi : \mathcal{X} \to [M]$ (e.g., binning, clustering, hashing) and define the aggregated context $U_t := \varphi(X_t) \in [M]$. For each $u \in [M]$, let $\zeta_u := \mathbb{P}(U = u)$ and

$$\mu_{a,u} := \mathbb{E}[\mu_a(X) \mid U = u]. \tag{8}$$

Because Bernoulli distributions are determined by their mean, the conditional reward distribution of arm a given $U = u$ is again Bernoulli with mean $\mu_{a,u}$. Therefore the aggregated process (U_t) exactly fits our finite-context model, and we can run CTS using U_t instead of X_t.

Theorem 21 (Context aggregation for Bernoulli rewards) *Under the Bernoulli reward model above, for any measurable $\varphi : \mathcal{X} \to [M]$, running CTS on the aggregated contexts $U_t = \varphi(X_t)$ is δ-PAC for identifying the best arm with respect to the marginal means $\mu_a = \mathbb{E}[\mu_a(X)]$.*

Moreover, context aggregation interpolates between the non-contextual BAI problem (coarsest aggregation) and increasingly fine discretizations. Since refining an aggregation only increases the degrees of freedom in the sampling proportions, the corresponding characteristic time cannot increase.

Lemma 22 (Monotonicity under refinement) *Let φ_1 and φ_2 be two aggregations such that φ_2 is a refinement of φ_1 (i.e., $\varphi_1(x) = \varphi_1(x')$ whenever $\varphi_2(x) = \varphi_2(x')$). Denote by $T^*(\nu_\varphi)$ the characteristic time of the finite-context instance induced by φ through (8). Then $T^*(\nu_{\varphi_2}) \leq T^*(\nu_{\varphi_1})$.*

Proofs are provided in Appendix I.

Relation to plug-in approaches for general context spaces. The reduction above provides a *δ-PAC* algorithm on an arbitrary context space by intentionally *coarsening* the observed contexts. Since the learner only uses $\varphi(X)$ rather than the full X, the induced instance ν_φ can be statistically harder than the original full-information model, and therefore the resulting sample complexity need not match the instance-dependent lower bound for the original model (see Section 2 for a discussion of lower bounds on general context spaces). Nonetheless, Lemma 22 shows that refining φ cannot worsen the characteristic time of the induced finite-context instance, so one may trade off statistical efficiency and computational cost by choosing φ. In principle, one may also pursue the typically less practical plug-in route: estimate the context distribution, for example via kernel density estimation when $\mathcal{X}$ is continuous, and the conditional reward model, for example via kernel regression such as Nadaraya–Watson or local-polynomial estimators, and plug these estimates into the optimal allocation prescribed by the lower bound. Note, however, that the above aggregation scheme can itself be interpreted as a simple plug-in strategy for infinite context spaces: it uses a histogram estimator of the context distribution (via the empirical bin frequencies) and, within each bin, empirical estimates of the Bernoulli means, thereby reducing the problem to a finite-context instance to which CTS applies. From the viewpoint of nonparametric estimation, histogram aggregation is a "box-kernel" method: the bin width Δ_n plays the role of the bandwidth, and the refinement regime $\Delta_n \downarrow 0$ mirrors the familiar bias–variance trade-off of kernel density estimation and kernel regression. Moreover, under mild regularity, refining the partition allows one to recover the continuous characteristic time in Theorem 2.

Proposition 23 (Refining the partition recovers the continuous characteristic time) *Assume that the context space $\mathcal{X} \subseteq \mathbb{R}$ is compact and that rewards are Bernoulli with mean functions $x \mapsto \mu_a(x) \in [\eta, 1-\eta]$ for some $\eta \in (0, 1/2)$ and all $a \in [K]$, which are uniformly*

continuous on $\mathcal{X}$. (For this proposition we restrict the alternative set in the definition of $T^\star$ to Bernoulli instances whose mean functions also take values in $[\eta, 1-\eta]$.) Assume further that the best arm $a^\star(\mathcal{V})$ is unique. Let $\{\Pi_m\}_{m\geq 1}$ be a sequence of finite measurable partitions of $\mathcal{X}$ such that Π_{m+1} refines Π_m and $\max_{B\in\Pi_m} \mathrm{diam}(B) \to 0$ as $m \to \infty$. Denote by $\widetilde{\mathcal{V}}^{(m)}$ the induced finite-context instance obtained by aggregating contexts according to Π_m (as in Theorem 21). Then, for all sufficiently large m, $a^\star(\widetilde{\mathcal{V}}^{(m)}) = a^\star(\mathcal{V})$ and

$$\lim_{m\to\infty} T^\star(\widetilde{\mathcal{V}}^{(m)}) = T^\star(\mathcal{V}).$$

Consequently, for any $\varepsilon > 0$ there exists m_ε such that running CTS *on the aggregated contexts defined by Π_{m_ε} is δ-PAC and satisfies*

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\mathcal{V}}[\tau_\delta]}{\log(1/\delta)} \leq T^\star(\mathcal{V}) + \varepsilon.$$

These results provide a simple, practical route to handle general (possibly infinite) context spaces in the Bernoulli case: pick a suitable aggregation φ and apply CTS.

7. Simulation Studies

In this section, we investigate the behavior of the proposed algorithms.

7.1 Simulation Studies of the α-elimination algorithm

First, we examine the performance of the α-elimination algorithm using contextual information. As in Section 3, we generate samples $\{(R_{t,1}, R_{t,2}, X_t)^\top\}_{t=1}^T$ from the multivariate distribution with the mean vector $(1,0,0)^\top$. We denote the variances of $R_{t,1}$, $R_{t,2}$, and X_t as σ_1^2, σ_2^2, and $\sigma_{\mathcal{X}}^2$. Let the correlation coefficient between $R_{t,1}$ and X_t be $\rho_{1\mathcal{X}}$, and the correlation coefficient between $R_{t,2}$ and X_t be $\rho_{2\mathcal{X}}$. We fix $\sigma_2^2 = 1$, $\sigma_{\mathcal{X}}^2 = 1$, and $\rho_{2\mathcal{X}} = 0.5$. We investigate the performance of the proposed method by varying the combination of the variance σ_1^2 and correlation coefficient $\rho_{1\mathcal{X}}$. We choose σ_1^2 from $\{1, 2\}$ and $\rho_{1\mathcal{X}}$ from $\{-0.9, -0.5, 0, 0.5, 0.9\}$. For the case with $\sigma_1^2 = 1$, the α-elimination without contextual information of Kaufmann et al. (2016) results in an allocation of $\alpha = 0.5$ (uniform sampling). For the case with $\sigma_1^2 = 2$, it results in an allocation of $\alpha = \sqrt{2}/(\sqrt{2}+\sqrt{1})$. Conversely, the proposed α-elimination with contextual information uses different allocations for each correlation coefficient. We conducted 1000 trials with $\delta = 0.05$ and display the realized stopping time (sample complexity) in Figure 2 using box plots, where the right figure shows the results with $\sigma_1^2 = 1$ and the left shows the results with $\sigma_1^2 = 2$. In Figure 2, we compare the proposed algorithm with different $\rho_{1\mathcal{X}}$ with the α-elimination (without context). The results demonstrate that when using contextual information, the proposed α-elimination can stop earlier than the original α-elimination. We note the fact that the proposed algorithm can stop earlier, even though the allocation is also 0.5 when $\rho_{1\mathcal{X}}$ is 0. Here, the stopping threshold β used in the proposed algorithm is less than that used in the original algorithm, while maintaining the δ-PAC property. Note that for all cases, the realized δ does not exceed 0.05.

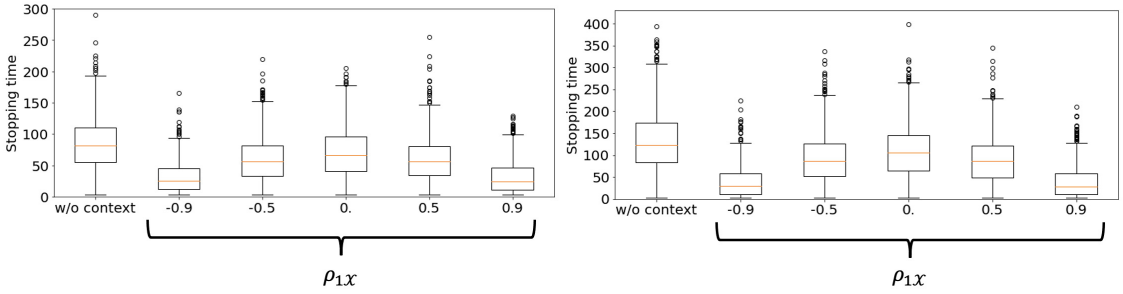

Figure 2: Results of α-elimination. The left figure displays the results with $\sigma_1^2 = 1$; the right figure displays the results with $\sigma_1^2 = 2$.

Additional experiments (plug-in variants). Additional experimental results for a plug-in variant (estimating the nuisance parameters online) are deferred to Appendix J.4. In these experiments, we compare against a *known-nuisance baseline* that uses the true nuisance parameters, which should be viewed as a convenient reference rather than an information-theoretic lower bound.

7.2 Simulation Study of the CTS algorithm

Next, we compare the performance of the proposed CTS algorithm to the TS algorithm for BAI without contextual information (Garivier and Kaufmann, 2016). For a Bernoulli bandit model, we consider a sample scenario with marginalized mean rewards $\{\mu_1, \mu_2, \mu_3, \mu_4\} = \{0.3, 0.21, 0.2, 0.19\}$, which is the same as a scenario used in Garivier and Kaufmann (2016). Suppose that there exist two contexts $X_t \in \{1, 2\}$, where the conditional mean rewards are given as $\{\mu_{1,1}, \mu_{2,1}, \mu_{3,1}, \mu_{4,1}\} = \{0.5, 0.01, 0.4, 0.01\}$ and $\{\mu_{1,2}, \mu_{2,2}, \mu_{3,2}, \mu_{4,2}\} = \{0.1, 0.41, 0, 0.37\}$.. The context 1 and 2 appear with probability 0.5, respectively.

We display the empirical values of the GLRT statistic in Figure 3. We present the GLRT statistic directly rather than the stopping time because the GLRT statistic provides more information. Note that the algorithm stops when the GLRT statistic exceeds the predefined threshold $\beta(t, \delta)$. The threshold $\beta(t, \delta)$ can be determined by us within the range suggested in Theorems 15–19, and its role does not differ significantly between the CTS and TS algorithms. The sooner this value becomes large, the smaller the sample complexity that can be achieved under a properly specified $\beta(t, \delta)$.[1]

This figure indicates that the CTS algorithm achieves a smaller sample complexity than TS, as suggested by the theoretical results. Conversely, the reason why the CTS algorithm indicates a smaller GLRT statistic compared with TS in the early rounds is likely because the number of parameters to be estimated is proportional to the number of contexts; thus it requires more time to converge in finite samples. In Appendix J, we present more details

1. In other words, the GLRT statistic and stopping time are directly related. If we were to display the stopping time, it would be necessary to show the stopping time for each threshold $\beta(t, \delta)$. This would involve drawing horizontal lines at the threshold values in Figure 3 and indicating the points of intersection between these lines and the respective GLRT statistics. By displaying the empirical GLRT statistic, as we have done, it is possible to consider the stopping times corresponding to any threshold $\beta(t, \delta)$, making the approach more general.

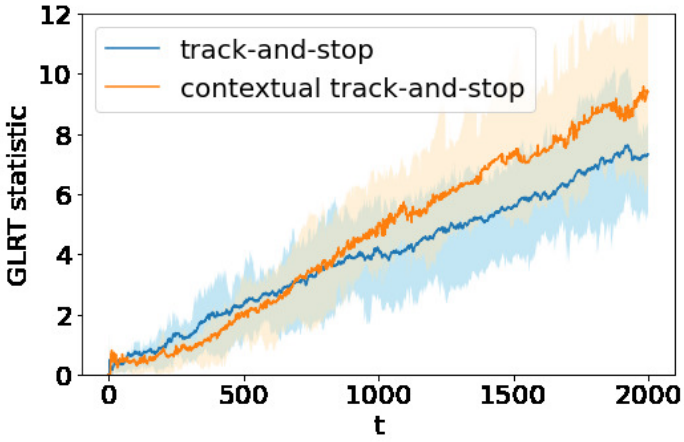

Figure 3: Graph illustrating the maximum GLRT statistic $\max_{a\in[K]} \min_{b\in[K]\setminus\{a\}} Z_{a,b}(t)$. The solid line represents the averaged value over 20 trials; the light-colored area indicates the values between the first and third quartiles.

and additional results under different settings, including additional Bernoulli instances and plug-in variants.

8. Conclusion

This paper proposed contextual BAI, in which contextual information can be used to identify the arm with the highest marginalized mean reward. We noted that even contextual information that is not immediately related to the parameter that we wish to identify could help us solve the task more efficiently. We proposed the CTS algorithm for the case in which the rewards follow Bernoulli distributions and confirmed that it performs better theoretically and experimentally when contextual information is provided. We also found that when the rewards and context follow a multivariate normal distribution in the two-armed bandit problem, we could improve the efficiency of BAI without changing the conventional algorithm. These properties have not been discussed to date.

A next step of this work is to consider a case with continuous contexts following an unknown distribution. As with the case of unknown variance in fixed-confidence BAI, the estimation error of the unknown variances would affect the performance (Jourdan et al., 2023). Therefore, for these cases, we may need to develop a lower bound for a class of distributions with unknown contextual distribution, as well as Jourdan et al. (2023), which develops lower and upper bounds for a class of Gaussian distributions with unknown variances. Another approach considers a localized instance where some parameter approaches zero (Kato et al., 2024b). Existing studies of ATE estimation, such as van der Laan (2008), Hahn et al. (2011), Kato et al. (2020) and Cook et al. (2023), employ the latter approach, which is referred to as semiparametric analysis. Considering a localized instance corresponds to evaluating the performance under the worst case, and under the worst case, we can ignore estimation errors of parameters unrelated to the parameter of interest.

Acknowledgement

The authors thank Alexandre Proutière for detailed discussions.

References

András Antos, Varun Grover, and Csaba Szepesvári. Active learning in multi-armed bandits. In *Algorithmic Learning Theory*, 2008.

Susan Athey and Guido Imbens. The econometrics of randomized experiments, 2016.

R.E. Bechhofer, J. Kiefer, and M. Sobel. *Sequential Identification and Ranking Procedures: With Special Reference to Koopman-Darmois Populations*. University of Chicago Press, 1968.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 2011.

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback–Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516 – 1541, 2013.

Herman Chernoff. Sequential Design of Experiments. *The Annals of Mathematical Statistics*, 30(3):755 – 770, 1959.

S.N. Chiu, D. Stoyan, W.S. Kendall, and J. Mecke. *Stochastic Geometry and Its Applications*. Wiley, 2013.

Thomas Cook, Alan Mishler, and Aaditya Ramdas. Semiparametric efficient inference in adaptive experiments. In *NeurIPS 2023 Workshop on Adaptive Experimental Design and Active Learning in the Real World*, 2023. URL `https://openreview.net/forum?id=xfj5jjpOaL`. a]rXiv:2311.18274.

Rémy Degenne, Wouter M Koolen, and Pierre Ménard. Non-asymptotic pure exploration by solving games. In *Advances in Neural Information Processing Systems*, 2019.

Aniket Anand Deshmukh, Srinagesh Sharma, James W. Cutler, Mark Moldwin, and Clayton Scott. Simple regret minimization for contextual bandits, 2018.

Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 2006.

Tanner Fiez, Lalit Jain, Kevin G Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. In *Advances in Neural Information Processing Systems*, 2019.

Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems*, 2012.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 2019.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, 2016.

Melody Y. Guan and Heinrich Jiang. Nonparametric stochastic contextual bandits. In *AAAI Conference on Artificial Intelligence*, 2018.

Jinyong Hahn, Keisuke Hirano, and Dean Karlan. Adaptive experimental design using the propensity score. *Journal of Business and Economic Statistics*, 2011.

William W Hogan. Point-to-set maps in mathematical programming. *SIAM review*, 1973.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press, 2015.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil' ucb : An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, 2014.

Yassir Jedra and Alexandre Proutiere. Optimal best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 2020.

Marc Jourdan, Degenne Rémy, and Kaufmann Emilie. Dealing with unknown variances in best-arm identification. In *International Conference on Algorithmic Learning Theory (ALT)*, volume 201, pages 776–849. PMLR, 2023.

Sandeep Juneja and Subhashini Krishnasamy. Sample complexity of partition identification using multi-armed bandits. In *Conference on Learning Theory*, volume 99, pages 1824–1852, 2019.

Olav Kallenberg. *Random measures, theory and applications*, volume 1. Springer, 2017.

Dean Karlan and Daniel H Wood. The effect of effectiveness: Donor response to aid effectiveness in a direct mail fundraising experiment. Working paper, National Bureau of Economic Research, 2014.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, 2013.

Masahiro Kato, Takuya Ishihara, Junya Honda, and Yusuke Narita. Adaptive experimental design for efficient treatment effect estimation, 2020.

Masahiro Kato, Akihiro Oga, Wataru Komatsubara, and Ryo Inokuchi. Active adaptive experimental design for treatment effect estimation with covariate choice. In *International Conference on Machine Learning (ICML)*, 2024a.

Masahiro Kato, Kyohei Okumura, Takuya Ishihara, and Toru Kitagawa. Adaptive experimental design for policy learning, 2024b. arXiv:2401.03756.

Emilie Kaufmann and Wouter M. Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, 22 (246):1–44, 2021.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 2016.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 1985.

Shie Mannor and John N Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 2004.

Jerzy Neyman. Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes. *Statistical Science*, 5:463–472, 1923.

Edward Paulson. A Sequential Procedure for Selecting the Population with the Largest Mean from k Normal Populations. *The Annals of Mathematical Statistics*, 1964.

Chao Qin and Daniel Russo. Adaptivity and confounding in multi-armed bandit experiments, 2022.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 1952.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 1974.

Yoan Russac, Christina Katsimerou, Dennis Bohle, Olivier Cappé, Aurélien Garivier, and Wouter M. Koolen. A/b/n testing with control in the presence of subpopulations. In *Advances in Neural Information Processing Systems*, 2021.

Marta Soare, Alessandro Lazaric, and Remi Munos. Best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems*, 2014.

Max Tabord-Meehan. Stratification trees for adaptive randomization in randomized controlled trials, 2018.

Chao Tao, Saúl Blanco, and Yuan Zhou. Best arm identification in linear bandits with linear dimension dependency. In *International Conference on Machine Learning*, 2018.

Cem Tekin and Mihaela van der Schaar. Releaf: An algorithm for learning and exploiting relevance. *IEEE Journal of Selected Topics in Signal Processing*, 2015.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 1933.

Mark J. van der Laan. The construction and analysis of adaptive group sequential designs, 2008.

Liyuan Xu, Junya Honda, and Masashi Sugiyama. A fully adaptive algorithm for pure exploration in linear bandits. In *International Conference on Artificial Intelligence and Statistics*, 2018.

Contents

Appendix A. Notations, Terms, and Abbreviations

In this section, we summarize the notations used in this paper.

Appendix B. Proof of Lemma 6

For each problem $\nu = ((\mu_{a,x}), (\zeta_x))$, for each $a \in [K]$, $x \in \mathcal{X}$, let $f^{\nu}_{a,x}$ denote the density (w.r.t. the Lebesgue measure) of the reward with the action-context pair (a, x). Let us define a log-likelihood ratio between the observation under the model $\nu = ((\mu_{a,x}), (\zeta_x))$ to the model $\nu' = ((\lambda_{a,x}), (\zeta_x))$ as

$$L_\tau = \sum_{t=1}^{\tau} \sum_{x\in\mathcal{X}} \sum_{a\in[K]} \mathbb{1}\{X_t = x, A_t = a\} \log\left(\frac{f^{\nu}_{a,x}(R_t)}{f^{\nu'}_{a,x}(R_t)}\right)$$

We have

$$\begin{aligned}
\mathbb{E}_\nu[L_\tau] &= \mathbb{E}_\nu\left[\sum_{t=1}^{\tau} \sum_{a\in[K]} \sum_{x\in\mathcal{X}} \mathbb{1}\{X_t = x, A_t = a\} \log\left(\frac{f^{\nu}_{a,x}(R_t)}{f^{\nu'}_{a,x}(R_t)}\right)\right] \\
&\overset{(a)}{=} \mathbb{E}_\nu\left[\sum_{x\in\mathcal{X}} \sum_{a\in[K]} \sum_{k=1}^{N_{a,x}(\tau)} \log\left(\frac{f^{\nu}_{a,x}(Y^{(x,a)}_k)}{f^{\nu'}_{a,x}(Y^{(x,a)}_k)}\right)\right] \\
&= \sum_{x\in\mathcal{X}} \sum_{a\in[K]} \mathbb{E}_\nu[N_{a,x}(\tau)]\mathrm{kl}(\mu_{a,x}, \lambda_{a,x}),
\end{aligned}$$

Table 1: Summary of notations

$X_t,\ A_t,\ R_t$	Context, action, and reward observed in round t
$[K],\ \mathcal{X}$	Sets of actions and contexts
$R_{t,a}$	Potential reward of arm a
ζ	Distribution of X_t
$\boldsymbol{p} = (p_{1,x}, p_{2,x}, \dots, p_{K,x})$	Reward distributions of the potential outcome given $x \in \mathcal{X}$.
$\boldsymbol{\mu} = (\mu_{1,x}, \mu_{2,x}, \dots, \mu_{K,x})$	Conditional mean rewards given $x \in \mathcal{X}$.
$\mu_a = \mathbb{E}_{X\sim\zeta}[\mu_{a,X}] = \mathbb{E}_{X\sim\zeta}[\mathbb{E}_{\mathcal{V}}[R_{t,a}\|X]]$ $= \mathbb{E}_{\mathcal{V}}[R_{t,a}]$	Marginalized mean reward of arm a
$\mathcal{V} = (\boldsymbol{p}, \zeta)$	Bandit problem.
$\nu = ((\mu_{a,x}), (\zeta_x))$	Bernoulli bandit problem with finite context.
Ω (resp. Θ)	Class of $\mathcal{V}$ (resp. ν).
$a^* = a^*(\mathcal{V}) = \arg\max_a \mu_a$	Best arm with the highest marginalized mean reward.
$\mathcal{F}_t = \sigma(X_1, A_1, R_1, \dots, X_t, A_t, R_t, X_{t+1})$	Sigma-algebras with the observations until t and X_{t+1}.
$\mathcal{G}_t = \sigma(X_1, A_1, R_1, \dots, X_t, A_t, R_t)$	Sigma-algebras with all observations up to t.
τ_δ	Stopping time under a fixed confidence $\delta > 0$.
$\hat{a}_{\tau_\delta}$	Recommended arm.
$\mathrm{Alt}(\mathcal{V}) := \{(\boldsymbol{q}, \zeta) \in \Omega : a^*((\boldsymbol{q}, \zeta)) \neq a^*((\boldsymbol{p}, \zeta))\}$	Set of alternative problems.
$N_x(t) = \sum_{s=1}^t \mathbb{1}\{X_s = x\}$	The number of times we observe context x.
$N_{a,x}(t) = \sum_{s=1}^t \mathbb{1}\{X_s = x, A_s = a\}$	The number of times we choose arm a given context x.
$\mathrm{KL}(p_{a,x}, q_{a,x})$	KL divergence from $p_{a,x}$ to $q_{a,x}$
$\mathrm{kl}(\mu, \nu)$	KL divergence of the canonical one-parameter exponential family.
$d(\mu, \nu)$ $= \mu \log(\mu/\nu) + (1-\mu)\log((1-\mu)/(1-\nu))$	KL divergence of Bernoulli distributions.
$\hat{\mu}_{a,x},\ \hat{\zeta}_x$	Estimators of $\mu_{a,x}$ and ζ_x in round t.
$w_{a,x}$	Allocation for arm a given context x.
$\mathcal{W}$	Set of allocation rule.
$p_{\mu_a}\big((\underline{R}_{a,x}(t))_{x\in\mathcal{X}}, \underline{X}(t)\big)$	Likelihood of parameters $\mu_a = (\mu_{a,x})_{a\in[K],x\in\mathcal{X}}$ given the observations $(\underline{R}_{a,x}(t))_{x\in\mathcal{X}}$ and $\underline{X}(t)$.
$Z_{a,b}(t)$	GLRT statistic.
$\beta(t, \delta)$	Threshold for stopping rule.

where for (a), we introduced random variables: $Y_k^{(x,a)}$ denotes k-th time the reward with the context x and the action a is observed and for the last equality, we used Wald's lemma for each (x, a) pair. From the data-processing inequality applied to the change-of-measure argument in Garivier et al. (2019), it holds that for any $\mathcal{E} \in \mathcal{G}_\tau$,

$$\sum_{x\in\mathcal{X}} \sum_{a\in[K]} \mathbb{E}_\nu[N_{a,x}(\tau)]\mathrm{kl}(\mu_{a,x}, \lambda_{a,x}) \geq d(\mathbb{P}_\nu(\mathcal{E}), \mathbb{P}_{\nu'}(\mathcal{E})).$$

This concludes the proof of Lemma 6.

Appendix C. Proof of Theorem 5

Proof. From Lemma 6 with $\mathcal{E} = \{\hat{a}_\tau = a^*(\nu)\}$, for each $\nu \in \Theta$ and $\nu' \in \mathrm{Alt}(\nu)$, we have

$$\sum_{x\in\mathcal{X}} \sum_{a\in[K]} \mathbb{E}_\nu[N_{a,x}(\tau)]\mathrm{kl}(\mu_{a,x}, \lambda_{a,x}) \geq d(\mathbb{P}_\nu(\mathcal{E}), \mathbb{P}_{\nu'}(\mathcal{E})) \geq \mathrm{kl}(\delta, 1-\delta),$$

where for the last inequality, we used the definition of the δ-PAC algorithm and monotonicity of the KL divergence. Let $N_x(\tau) = \sum_{t=1}^{\tau} \mathbb{1}\{X_t = x\}$. For each $\nu \in \Theta$,

$$\begin{aligned}
d(\delta, 1-\delta) &\leq \inf_{((\lambda_{a,x}),\zeta)\in \mathrm{Alt}(\nu)} \sum_{x\in\mathcal{X}} \sum_{a\in[K]} \mathbb{E}_\nu[N_{a,x}(\tau)] \mathrm{kl}(\mu_{a,x}, \lambda_{a,x}) \\
&= \inf_{((\lambda_{a,x}),\zeta)\in \mathrm{Alt}(\nu)} \sum_{x\in\mathcal{X}} \mathbb{E}_\nu[N_x(\tau)] \sum_{a\in[K]} \frac{\mathbb{E}_\nu[N_{a,x}(\tau)]}{\mathbb{E}_\nu[N_x(\tau)]} \mathrm{kl}(\mu_{a,x}, \lambda_{a,x}) \\
&= \inf_{((\lambda_{a,x}),\zeta)\in \mathrm{Alt}(\nu)} \mathbb{E}_\nu[\tau_\delta] \sum_{x\in\mathcal{X}} \frac{\mathbb{E}_\nu[N_x(\tau)]}{\mathbb{E}_\nu[\tau_\delta]} \sum_{a\in[K]} \frac{\mathbb{E}_\nu[N_{a,x}(\tau)]}{\mathbb{E}_\nu[N_x(\tau)]} \mathrm{kl}(\mu_{a,x}, \lambda_{a,x}) \\
&\overset{(a)}{=} \inf_{((\lambda_{a,x}),\zeta)\in \mathrm{Alt}(\nu)} \mathbb{E}_\nu[\tau_\delta] \sum_{x\in\mathcal{X}} \frac{\mathbb{E}_\nu[\tau_\delta]\zeta_x}{\mathbb{E}_\nu[\tau_\delta]} \sum_{a\in[K]} \frac{\mathbb{E}_\nu[N_{a,x}(\tau)]}{\mathbb{E}_\nu[N_x(\tau)]} \mathrm{kl}(\mu_{a,x}, \lambda_{a,x}) \\
&= \mathbb{E}_\nu[\tau_\delta] \inf_{((\lambda_{a,x}),\zeta)\in \mathrm{Alt}(\nu)} \sum_{x\in\mathcal{X}} \zeta_x \sum_{a\in[K]} \frac{\mathbb{E}_\nu[N_{a,x}(\tau)]}{\mathbb{E}_\nu[N_x(\tau)]} \mathrm{kl}(\mu_{a,x}, \lambda_{a,x}) \\
&\leq \mathbb{E}_\nu[\tau_\delta] \sup_{\boldsymbol{w}\in\mathcal{W}} \inf_{((\lambda_{a,x}),\zeta)\in \mathrm{Alt}(\nu)} \sum_{x\in\mathcal{X}} \zeta_x \sum_{a\in[K]} w_{a,x} \mathrm{kl}(\mu_{a,x}, \lambda_{a,x}),
\end{aligned}$$

where for (a), we used Wald's lemma for each x. This concludes the proof.

Appendix D. Proof of Theorem 2

We show Theorem 2. Let $\mathcal{B}(\mathbb{R})$ be a Borel σ-algebra on $\mathbb{R}$. Let us introduce two random counting measures on $\mathbb{R}$: (i) for each $A \in \mathcal{B}(\mathbb{R})$, $\Xi(A)$ counts the number of times contexts have arrived in A, (ii) $\Upsilon_a(A)$ counts the number of times the algorithm selected action a under the context is in A.

The intensity measure is a characteristic analogous to the mean of a real-valued random variable (Chiu et al., 2013). Let us denote the intensity measures of Ξ and Υ_a by γ and κ_a, respectively; that is, $\gamma(A) = \mathbb{E}[\Xi(A)]$ and $\kappa_a(A) = \mathbb{E}[\Upsilon_a(A)]$ for each $A \in \mathcal{B}(\mathbb{R})$. Suppose that γ and κ_a are absolutely continuous with respect to ζ (Kallenberg, 2017). Furthermore, κ_a is absolutely continuous with respect to γ. Let $\frac{d\gamma}{dx}(x)$ and $\frac{\mathrm{d}\kappa_a}{\mathrm{d}x}(x)$ be densities of γ and κ_a with respect to the Lebesgue measure.

Then, we extend our Lemma 6 to the case of continuous contexts.

Lemma 24 *Take $\mathcal{V} = (\boldsymbol{p}, \zeta), \mathcal{M} = (\boldsymbol{q}, \zeta) \in \Omega$. For any almost-surely finite stopping time τ with respect to $(\mathcal{G}_t)_{t\geq 1}$, we have*

$$\sum_{a=1}^{K} \int_{\mathbb{R}} \frac{\mathrm{d}\kappa_a}{\mathrm{d}x}(x) \mathrm{KL}(p_{a,x}, q_{a,x}) \mathrm{d}x \geq \sup_{\mathcal{E}\in\mathcal{G}_\tau} d(\mathbb{P}_\mathcal{V}(\mathcal{E}), \mathbb{P}_\mathcal{M}(\mathcal{E})),$$

where $\mathbb{E}_\mathcal{V}$ and $\mathbb{P}_\mathcal{V}$ are the expectation and probability under model $\mathcal{V}$, respectively, and $\mathbb{E}_\mathcal{M}$ and $\mathbb{P}_\mathcal{M}$ are defined analogously under model $\mathcal{M}$.

In the proof, we use Campbell's theorem.

Proposition 25 (Campbell's theorem from Theorem 4.1 in Chiu et al. (2013)) *For any nonnegative measurable function $f(x)$ and $A \in \mathcal{B}(\mathbb{R})$, we have*

$$\mathbb{E}\left[\sum_{x\in\Upsilon_a(A)} f(x)\right] = \mathbb{E}\left[\int_A f(x)\Upsilon_a(\mathrm{d}x)\right] = \int_A f(x)\kappa_a(\mathrm{d}x).$$

We show the proof of Lemma 24 as follows.

Proof For each $a \in [K]$, $x \in \mathcal{X}$, let us denote by $f_{a,x}$ and $f'_{a,x}$ the probability density functions of $p_{a,x}$ and $q_{a,x}$ with respect to the Lebesgue measure. We have that

$$\mathbb{E}_{\mathcal{V}}\left[\log\left(\frac{f_{a,x}(R_{t,a})}{f'_{a,x}(R_{t,a})}\right)\middle| X_t = x\right] = \mathrm{KL}(p_{a,x}, q_{a,x}).$$

Let us define a log-likelihood ratio from the observation under the model $\mathcal{V} = ((p_{a,x}), \zeta)$ to the model $\mathcal{M} = ((q_{a,x}), \zeta)$

$$L_\tau = \sum_{t=1}^{\tau}\sum_{a\in[K]} \mathbb{1}\{A_t = a\}\log\left(\frac{f_{a,X_t}(R_{t,a})}{f'_{a,X_t}(R_{t,a})}\right).$$

Let us define $\boldsymbol{x}_\infty = (x_1, x_2, \ldots)$, $\boldsymbol{X}_\infty = (X_1, X_2, \ldots)$, and $\mathcal{X}'(\boldsymbol{x}_\infty) = \cup_{t=1}^{\infty}\{x_t\}$. We have

$$\begin{aligned}
\mathbb{E}_{\mathcal{V}}[L_\tau] &= \mathbb{E}_{\mathcal{V}}[\mathbb{E}_{\mathcal{V}}[L_\tau|\boldsymbol{X}_\infty]] \\
&= \int_{\boldsymbol{x}_\infty\in\mathbb{R}^\infty} \mathbb{E}_{\mathcal{V}}[L_\tau|\boldsymbol{X}_\infty = \boldsymbol{x}_\infty]\prod_{t=1}^{\infty}\zeta(x_t)\mathrm{d}\boldsymbol{x}_\infty \\
&= \int_{\boldsymbol{x}_\infty\in\mathbb{R}^\infty} \mathbb{E}_{\mathcal{V}}\left[\sum_{t=1}^{\tau}\sum_{a\in[K]}\mathbb{1}\{A_t = a\}\log\frac{f_{a,X_t}(R_{t,a})}{f'_{a,X_t}(R_{t,a})}\middle|\boldsymbol{X}_\infty = \boldsymbol{x}_\infty\right]\prod_{t=1}^{\infty}\zeta(x_t)\mathrm{d}\boldsymbol{x}_\infty \\
&= \int_{\boldsymbol{x}_\infty\in\mathbb{R}^\infty} \mathbb{E}_{\mathcal{V}}\left[\sum_{t=1}^{\tau}\sum_{a\in[K]}\mathbb{1}\{A_t = a, X_t = x_t\}\log\frac{f_{a,x_t}(R_{t,a})}{f'_{a,x_t}(R_{t,a})}\middle|\boldsymbol{X}_\infty = \boldsymbol{x}_\infty\right]\prod_{t=1}^{\infty}\zeta(x_t)\mathrm{d}\boldsymbol{x}_\infty \\
&\overset{(a)}{=} \sum_{a\in[K]}\int_{\boldsymbol{x}_\infty\in\mathbb{R}^\infty} \mathbb{E}_{\mathcal{V}}\left[\sum_{x\in\mathcal{X}'(\boldsymbol{x}_\infty)}\sum_{k=1}^{N_{a,x}(\tau)}\log\frac{f_{a,x}(Y_k^{(a,x)})}{f'_{a,x}(Y_k^{(a,x)})}\middle|\boldsymbol{X}_\infty = \boldsymbol{x}_\infty\right]\prod_{t=1}^{\infty}\zeta(x_t)\mathrm{d}\boldsymbol{x}_\infty \\
&= \sum_{a\in[K]}\int_{\boldsymbol{x}_\infty\in\mathbb{R}^\infty}\sum_{x\in\mathcal{X}'(\boldsymbol{x}_\infty)} \mathbb{E}_{\mathcal{V}}\left[\sum_{k=1}^{N_{a,x}(\tau)}\log\frac{f_{a,x}(Y_k^{(a,x)})}{f'_{a,x}(Y_k^{(a,x)})}\middle|\boldsymbol{X}_\infty = \boldsymbol{x}_\infty\right]\prod_{t=1}^{\infty}\zeta(x_t)\mathrm{d}\boldsymbol{x}_\infty \\
&= \sum_{a\in[K]}\int_{\boldsymbol{x}_\infty\in\mathbb{R}^\infty}\sum_{x\in\mathcal{X}'(\boldsymbol{x}_\infty)} \mathbb{E}_{\mathcal{V}}\left[N_{a,x}(\tau)|\boldsymbol{X}_\infty = \boldsymbol{x}_\infty\right]\mathbb{E}_{\mathcal{V}}\left[\log\frac{f_{a,x}(Y_1^{(a,x)})}{f'_{a,x}(Y_1^{(a,x)})}\middle|\boldsymbol{X}_\infty = \boldsymbol{x}_\infty\right]\prod_{t=1}^{\infty}\zeta(x_t)\mathrm{d}\boldsymbol{x}_\infty \\
&= \sum_{a\in[K]}\int_{\boldsymbol{x}_\infty\in\mathbb{R}^\infty}\sum_{x\in\mathcal{X}'(\boldsymbol{x}_\infty)} \mathbb{E}_{\mathcal{V}}\left[N_{a,x}(\tau)|\boldsymbol{X}_\infty = \boldsymbol{x}_\infty\right]\mathrm{KL}(p_{a,x}, q_{a,x})\prod_{t=1}^{\infty}\zeta(x_t)\mathrm{d}\boldsymbol{x}_\infty \\
&= \sum_{a\in[K]}\int_{\boldsymbol{x}_\infty\in\mathbb{R}^\infty} \mathbb{E}_{\mathcal{V}}\left[\sum_{x\in\mathcal{X}'(\boldsymbol{x}_\infty)} N_{a,x}(\tau)\mathrm{KL}(p_{a,x}, q_{a,x})|\boldsymbol{X}_\infty = \boldsymbol{x}_\infty\right]\prod_{t=1}^{\infty}\zeta(x_t)\mathrm{d}\boldsymbol{x}_\infty
\end{aligned}$$

$$
\begin{aligned}
&\overset{(b)}{=} \sum_{a\in[K]} \mathbb{E}_{\mathcal{V}}\left[\int_{\mathbb{R}} \mathrm{KL}(p_{a,x}, q_{a,x})\Upsilon_a(\mathrm{d}x)\right] \\
&\overset{(c)}{=} \sum_{a\in[K]} \int_{\mathbb{R}} \mathrm{KL}(p_{a,x}, q_{a,x})\kappa_a(\mathrm{d}x) \\
&= \sum_{a\in[K]} \int_{x\in\mathbb{R}} \frac{\mathrm{d}\kappa_a}{\mathrm{d}x}(x)\mathrm{KL}(p_{a,x}, q_{a,x})\mathrm{d}x.
\end{aligned}
$$

For (a), we introduced random variable $Y_k^{(a,x)}$, denoting k-th time the reward with the context x and the action a is observed. For (b), the computation is as follows:

$$
\begin{aligned}
&\sum_{a\in[K]} \int_{\boldsymbol{x}_\infty\in\mathbb{R}^\infty} \mathbb{E}_{\mathcal{V}}\left[\sum_{x\in\mathcal{X}'(\boldsymbol{x}_\infty)} N_{a,x}(\tau)\mathrm{KL}(p_{a,x}, q_{a,x})|\boldsymbol{X}_\infty = \boldsymbol{x}_\infty\right] \prod_{t=1}^{\infty}\zeta(x_t)\mathrm{d}\boldsymbol{x}_\infty \\
&= \sum_{a\in[K]} \mathbb{E}_{\mathcal{V}}\left[\sum_{x\in\mathcal{X}'(\boldsymbol{X}_\infty)} N_{a,x}(\tau)\mathrm{KL}(p_{a,x}, q_{a,x})\right] \\
&= \sum_{a\in[K]} \mathbb{E}_{\mathcal{V}}\left[\sum_{x\in(X_1,\ldots,X_\tau)} N_{a,x}(\tau)\mathrm{KL}(p_{a,x}, q_{a,x})\right] \\
&= \sum_{a\in[K]} \mathbb{E}_{\mathcal{V}}\left[\int_{\mathbb{R}} \mathrm{KL}(p_{a,x}, q_{a,x})\Upsilon_a(\mathrm{d}x)\right],
\end{aligned}
$$

where the last equality follows from the definition of Υ_a. For (c), we used Campbell's theorem (Proposition 25).

From the data-processing inequality applied to the change-of-measure argument in Garivier et al. (2019), it holds that for any $\mathcal{E} \in \mathcal{G}_\tau$,

$$
\sum_{a\in[K]} \int_{\mathbb{R}} \frac{\mathrm{d}\kappa_a}{\mathrm{d}x}(x)\mathrm{KL}(p_{a,x}, q_{a,x})\mathrm{d}x \geq d(\mathbb{P}_{\mathcal{V}}(\mathcal{E}), \mathbb{P}_{\mathcal{M}}(\mathcal{E})).
$$

This concludes the proof of Lemma 24. ■

Then, we show the proof of Theorem 2.

Proof From Lemma 24 with $\mathcal{E} = \{\hat{a}_\tau = a^*(\mathcal{V})\}$, for each $\mathcal{V} \in \Omega$ and $\mathcal{M} \in \mathrm{Alt}(\mathcal{V})$, we have

$$
\sum_{a\in[K]} \int_{\mathbb{R}} \frac{\mathrm{d}\kappa_a}{\mathrm{d}x}(x)\mathrm{KL}(p_{a,x}, q_{a,x})\mathrm{d}x \geq \mathrm{kl}(\mathbb{P}_{\mathcal{V}}(\mathcal{E}), \mathbb{P}_{\mathcal{M}}(\mathcal{E})) \geq d(\delta, 1-\delta),
$$

where for the last inequality, we used the definition of the δ-PAC algorithm and monotonicity of the KL divergence.

For each $\mathcal{V} \in \Omega$, we have

$$
d(\delta, 1-\delta) \leq \inf_{(\boldsymbol{p},\zeta)\in\mathrm{Alt}(\mathcal{V})} \sum_{a\in[K]} \int_{\mathbb{R}} \frac{\mathrm{d}\kappa_a}{\mathrm{d}x}(x)\mathrm{KL}(p_{a,x}, q_{a,x})\mathrm{d}x
$$

$$
\begin{aligned}
&= \inf_{(\boldsymbol{p},\zeta)\in \mathrm{Alt}(\mathcal{V})} \sum_{a\in[K]} \int_{\mathbb{R}} \frac{\mathrm{d}\kappa_a}{\mathrm{d}\gamma}\frac{\mathrm{d}\gamma}{\mathrm{d}\zeta}(x)\zeta(x)\mathrm{KL}(p_{a,x},q_{a,x})\mathrm{d}x \\
&= \inf_{(\boldsymbol{p},\zeta)\in \mathrm{Alt}(\mathcal{V})} \int_{\mathbb{R}} \underbrace{\frac{\mathrm{d}\gamma}{\mathrm{d}\zeta}(x)\zeta(x)}_{\mathbb{E}_{\mathcal{V}}[\tau_\delta]\zeta(x)} \sum_{a\in[K]} \frac{\mathrm{d}\kappa_a}{\mathrm{d}\gamma}(x)\mathrm{KL}(p_{a,x},q_{a,x})\mathrm{d}x \\
&\overset{(a)}{=} \mathbb{E}_{\mathcal{V}}[\tau_\delta] \inf_{(\boldsymbol{p},\zeta)\in \mathrm{Alt}(\mathcal{V})} \int_{\mathbb{R}} \sum_{a\in[K]} \frac{\mathrm{d}\kappa_a}{\mathrm{d}\gamma}(x)\mathrm{KL}(p_{a,x},q_{a,x})\zeta(x)\mathrm{d}x \\
&\leq \mathbb{E}_{\mathcal{V}}[\tau_\delta] \sup_{\boldsymbol{w}\in\mathcal{W}} \inf_{(\boldsymbol{p},\zeta)\in \mathrm{Alt}(\mathcal{V})} \int_{\mathbb{R}} \sum_{a\in[K]} w_{a,x}\mathrm{KL}(p_{a,x},q_{a,x})\zeta(x)\mathrm{d}x,
\end{aligned}
$$

where for (a) we used the equivalence:

$$
\begin{aligned}
\mathbb{E}_{\mathcal{V}}[\tau_\delta] &= \int_{\mathbb{R}} \frac{\mathrm{d}\gamma}{\mathrm{d}x}\mathrm{d}x \\
&= \int_{\mathbb{R}} \frac{\mathrm{d}\gamma}{\mathrm{d}\zeta}\zeta(x)\mathrm{d}x \\
&\overset{(b)}{=} \frac{\mathrm{d}\gamma}{\mathrm{d}\zeta}\int_{\mathbb{R}} \zeta(x)\mathrm{d}x \\
&= \frac{\mathrm{d}\gamma}{\mathrm{d}\zeta},
\end{aligned}
$$

where for (b), we used the fact that $\frac{\mathrm{d}\gamma}{\mathrm{d}\zeta}$ is a constant does not depend on x.

■

Appendix E. Proofs for Section 3 and α-Elimination Algorithm with Contextual Information

E.1 α-Elimination Algorithm with Contextual Information

We use an algorithm that is almost identical to the α-elimination of Kaufmann et al. (2016). The only difference between the proposed α-elimination and that of Kaufmann et al. (2016) is that we construct an estimator of the marginalized mean reward in the following form:

$$
\begin{aligned}
\hat{\mu}_1(t) &= \frac{1}{\sum_{s=1}^t \mathbb{1}[A_s=1]} \sum_{s=1}^t \left(R_{s,1} - \frac{\rho_{1\mathcal{X}}\sigma_1}{\sigma_{\mathcal{X}}}(X_s-\mu_{\mathcal{X}})\right)\mathbb{1}[A_s=1], \\
\hat{\mu}_2(t) &= \frac{1}{\sum_{s=1}^t \mathbb{1}[A_s=2]} \sum_{s=1}^t \left(R_{s,2} - \frac{\rho_{2\mathcal{X}}\sigma_2}{\sigma_{\mathcal{X}}}(X_s-\mu_{\mathcal{X}})\right)\mathbb{1}[A_s=2].
\end{aligned}
$$

Here, we used that $\mu_a = \mu_{a,x} - \frac{\rho_{a\mathcal{X}}\sigma_a}{\sigma_{\mathcal{X}}}(x-\mu_{\mathcal{X}})$. This estimator is based on the form of the conditional distribution of $R_{t,a}$. We replace $\hat{\mu}_a(t)$ in the original α-elimination with these estimators.

E.2 Proof of Theorem 3

Recall that the KL divergence from $\mathcal{N}(\mu_1, \sigma^2)$ to $\mathcal{N}(\mu_2, \sigma^2)$ is given as

$$\mathrm{KL}\left(\mathcal{N}(\mu_1, \sigma^2), \mathcal{N}(\mu_2, \sigma^2)\right) = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2}.$$

If we ignore sets of measure zero, we have

$$\begin{aligned}
T^\star(\mathcal{V})^{-1} &= \sup_{\boldsymbol{w} \in \mathcal{W}} \inf_{(\boldsymbol{q}, \zeta) \in \mathrm{Alt}(\mathcal{V})} \sum_{a=1}^{2} \int_{\mathbb{R}} w_{a,x} \mathrm{KL}(p_{a,x}, q_{a,x}) \zeta(x) \mathrm{d}x \\
&= \sup_{\boldsymbol{w} \in \mathcal{W}} \inf_{(\boldsymbol{q}, \zeta) \in \mathrm{Alt}(\mathcal{V})} \sum_{a=1}^{2} \int_{\mathbb{R}} w_{a,x} \frac{\left(\mu_a + \frac{\sigma_{a\mathcal{X}}}{\sigma^2_{\mathcal{X}}}(x - \mu_{\mathcal{X}}) - \lambda_{a,x}\right)^2}{2\sigma_a'^2} \zeta(x) \mathrm{d}x \\
&= \sup_{\boldsymbol{w} \in \mathcal{W}} \inf_{\int_{\mathbb{R}} \lambda_{2,x} \zeta(x) \mathrm{d}x > \int_{\mathbb{R}} \lambda_{1,x} \zeta(x) \mathrm{d}x} \sum_{a=1}^{2} \int_{\mathbb{R}} w_{a,x} \frac{\left(\mu_a + \frac{\sigma_{a\mathcal{X}}}{\sigma^2_{\mathcal{X}}}(x - \mu_{\mathcal{X}}) - \lambda_{a,x}\right)^2}{2\sigma_a'^2} \zeta(x) \mathrm{d}x \\
&\overset{(a)}{=} \max_{\boldsymbol{w} \in \mathcal{W}} \min_{\int_{\mathbb{R}} \lambda_{2,x} \zeta(x) \mathrm{d}x = \int_{\mathbb{R}} \lambda_{1,x} \zeta(x) \mathrm{d}x} \sum_{a=1}^{2} \int_{\mathbb{R}} w_{a,x} \frac{\left(\mu_a + \frac{\sigma_{a\mathcal{X}}}{\sigma^2_{\mathcal{X}}}(x - \mu_{\mathcal{X}}) - \lambda_{a,x}\right)^2}{2\sigma_a'^2} \zeta(x) \mathrm{d}x
\end{aligned}$$

where for (a), we used the same argument as in Lemma 8. From the property of the multivariate Gaussian distribution,

$$\lambda_{1,x} = \lambda_1 + \frac{\sigma_{1\mathcal{X}}}{\sigma^2_{\mathcal{X}}}(x - \mu_{\mathcal{X}}) \quad \text{and} \quad \lambda_{2,x} = \lambda_2 + \frac{\sigma_{2\mathcal{X}}}{\sigma^2_{\mathcal{X}}}(x - \mu_{\mathcal{X}}).$$

From $\int_{\mathbb{R}} \lambda_{2,x} \zeta(x) \mathrm{d}x = \int_{\mathbb{R}} \lambda_{1,x} \zeta(x) \mathrm{d}x$, $\lambda_1 = \lambda_2 = \lambda$. Therefore, we get

$$\frac{1}{2\sigma_a'^2} \left(\mu_a + \frac{\sigma_{a\mathcal{X}}}{\sigma^2_{\mathcal{X}}}(x - \mu_{\mathcal{X}}) - \lambda_{a,x}\right)^2 = \frac{1}{2\sigma_a'^2} \left(\mu_a - \lambda\right)^2.$$

Therefore, the optimization problem can be further simplified

$$\begin{aligned}
T^\star(\mathcal{V})^{-1} &= \max_{\boldsymbol{w} \in \mathcal{W}} \min_{\int_{\mathbb{R}} \lambda_{2,x} \zeta(x) \mathrm{d}x = \int_{\mathbb{R}} \lambda_{1,x} \zeta(x) \mathrm{d}x} \sum_{a=1}^{2} \int_{\mathbb{R}} w_{a,x} \frac{\left(\mu_a + \frac{\sigma_{a\mathcal{X}}}{\sigma^2_{\mathcal{X}}}(x - \mu_{\mathcal{X}}) - \lambda_{a,x}\right)^2}{2\sigma_a'^2} \zeta(x) \mathrm{d}x \\
&= \max_{\boldsymbol{w} \in \mathcal{W}} \min_{\lambda \in \mathbb{R}} \int_{\mathbb{R}} \sum_{a=1}^{2} w_{a,x} \frac{\left(\mu_a - \lambda\right)^2}{2\sigma_a'^2} \zeta(x) \mathrm{d}x.
\end{aligned}$$

At each point $x \in \mathbb{R}$, the optimization problem

$$\max_{w_{1,x} + w_{2,x} = 1} \min_{\lambda \in \mathbb{R}} \sum_{a=1}^{2} w_{a,x} \frac{\left(\mu_a - \lambda\right)^2}{2\sigma_a'^2}$$

is an identical problem as is given in Theorem 6 in Kaufmann et al. (2016) (two arm Gaussian bandits with known variances) and we know from Theorem 9 in Kaufmann et al. (2016),

the maximum is attained when $w_{1,x} = \sigma_1'/(\sigma_1' + \sigma_2')$. Thus, we compute

$$T^\star(\mathcal{V})^{-1} = \min_{\lambda\in\mathbb{R}} \int_{\mathbb{R}} \sum_{a=1}^{2} \frac{\sigma_a'}{\sigma_1' + \sigma_2'} \frac{(\mu_a - \lambda)^2}{2\sigma_a'^2} \zeta(x)\mathrm{d}x = \frac{1}{\sigma_1' + \sigma_2'} \min_{\lambda\in\mathbb{R}} \int_{\mathbb{R}} \sum_{a=1}^{2} \frac{(\mu_a - \lambda)^2}{2\sigma_a'} \zeta(x)\mathrm{d}x,$$

When the minimum is attained,

$$-\frac{1}{\sigma_1'}(\mu_1 - \lambda) - \frac{1}{\sigma_2'}(\mu_2 - \lambda) = 0,$$

Therefore,

$$\lambda = \frac{\frac{1}{\sigma_1'}\mu_1 + \frac{1}{\sigma_2'}\mu_2}{\frac{1}{\sigma_1'} + \frac{1}{\sigma_2'}}.$$

Then,

$$\begin{aligned}
\sum_{a=1}^{2} \frac{1}{2\sigma_a'} \left(\mu_a + \frac{\sigma_{a\mathcal{X}}}{\sigma_{\mathcal{X}}^2}(x - \mu_{\mathcal{X}}) - \lambda_{a,x}\right)^2 &= \sum_{a=1}^{2} \frac{1}{2\sigma_a'} (\mu_a - \lambda)^2 \\
&= \left(\frac{\mu_1 - \mu_2}{\sigma_1' + \sigma_2'}\right)^2 \frac{\sigma_1'}{2} + \left(\frac{\mu_1 - \mu_2}{\sigma_1' + \sigma_2'}\right)^2 \frac{\sigma_2'}{2} \\
&= \frac{(\mu_1 - \mu_2)^2}{2(\sigma_1' + \sigma_2')}
\end{aligned}$$

Therefore, we have

$$T^\star(\mathcal{V})^{-1} = \frac{1}{2}\left(\frac{\mu_1 - \mu_2}{\sigma_1' + \sigma_2'}\right)^2.$$

□

E.3 Proof of Theorem 4

We note that except that the variances of the sample from the arm a is $\sigma_a'^2$, the proof is almost identical to that of Theorem 9 of Kaufmann et al. (2016). Let $\alpha = \sigma_1'/(\sigma_1' + \sigma_2')$ and $d_t = \hat{\mu}_1(t) - \hat{\mu}_2(t)$. We first prove that the strategy is δ-PAC for every $\mathcal{V} \in \widetilde{\Omega}$. Assume that $\mu_1 > \mu_2$ and recall $\tau = \inf\{t \in \mathbb{N} : |d_t| > \sqrt{2\sigma_t^2(\alpha)\beta(t,\delta)}\}$, where $d_t := \hat{\mu}_1(t) - \hat{\mu}_2(t)$. The probability of error of the α-elimination strategy is upper bounded by

$$\begin{aligned}
\mathbb{P}_{\mathcal{V}}\left(d_\tau \le -\sqrt{2\sigma_\tau^2(\alpha)\beta(\tau,\delta)}\right) &\le \mathbb{P}_{\mathcal{V}}\left(d_\tau - (\mu_1 - \mu_2) \le -\sqrt{2\sigma_\tau^2(\alpha)\beta(\tau,\delta)}\right) \\
&\le \mathbb{P}_{\mathcal{V}}\left(\exists t \in \mathbb{N}^* : d_t - (\mu_1 - \mu_2) < -\sqrt{2\sigma_t^2(\alpha)\beta(t,\delta)}\right) \\
&\le \sum_{t=1}^{\infty} \exp\left(-\beta(t,\delta)\right),
\end{aligned}$$

where we used union bound and Chernoff bound applied to $d_t - (\mu_1 - \mu_2) \sim \mathcal{N}(0, \sigma_t^2(\alpha))$ in the last inequality. We have

$$\begin{aligned}\sum_{t=1}^{\infty} \exp\left(-\beta(t,\delta)\right) &\leq \delta \sum_{t=1}^{\infty} \frac{1}{t(\log(6t))^2} \leq \delta \left(\frac{1}{(\log 6)^2} + \int_1^{\infty} \frac{dt}{t(\log(6t))^2} \right) \\ &= \delta \left(\frac{1}{(\log 6)^2} + \frac{1}{\log(6)} \right) \leq \delta.\end{aligned}$$

For the guarantee of the expected sample complexity, we first prove the probability that τ exceeds some fixed T:

$$\begin{aligned}\mathbb{P}_{\mathcal{V}}(\tau \geq T) &\leq \mathbb{P}_{\mathcal{V}} \left(\forall t \in [T],\ d_t \leq \sqrt{2\sigma_t^2(\alpha)\beta(t,\delta)} \right) \\ &\leq \mathbb{P}_{\mathcal{V}} \left(d_T \leq \sqrt{2\sigma_T^2(\alpha)\beta(T,\delta)} \right) \\ &= \mathbb{P}_{\mathcal{V}} \left(d_T - (\mu_1 - \mu_2) \leq -\left[(\mu_1 - \mu_2) - \sqrt{2\sigma_T^2(\alpha)\beta(T,\delta)} \right] \right) \\ &\leq \exp\left(-\frac{1}{2\sigma_T^2(\alpha)} \left[(\mu_1 - \mu_2) - \sqrt{2\sigma_T^2(\alpha)\beta(T,\delta)} \right]^2 \right),\end{aligned}$$

where for the last inequality we used Chernoff bound with T such that $(\mu_1 - \mu_2) > \sqrt{2\sigma_T^2(\alpha)\beta(T,\delta)}$. For $\gamma \in (0,1)$, define

$$T_\gamma^* := \inf \left\{ t_0 \in \mathbb{N} : \forall t \geq t_0, (\mu_1 - \mu_2) - \sqrt{2\sigma_t^2(\alpha)\beta(t,\delta)} > \gamma(\mu_1 - \mu_2) \right\}.$$

We have

$$\begin{aligned}\mathbb{E}_{\mathcal{V}}[\tau] &\leq T_\gamma^* + \sum_{T=T_\gamma^*+1}^{\infty} \mathbb{P}\left(\tau \geq T\right) \\ &\leq T_\gamma^* + \sum_{T=T_\gamma^*+1}^{\infty} \exp\left(-\frac{1}{2\sigma_T^2(\alpha)} \left[(\mu_1 - \mu_2) - \sqrt{2\sigma_T^2(\alpha)\beta(T,\delta)} \right]^2 \right) \\ &\leq T_\gamma^* + \sum_{T=T_\gamma^*+1}^{\infty} \exp\left(-\frac{1}{2\sigma_T^2(\alpha)} \gamma^2 (\mu_1 - \mu_2)^2 \right).\end{aligned}$$

For all t, it is easy to show that the following upper bound on $\sigma_t^2(\alpha)$ holds:

$$\sigma_t^2(\alpha) \leq \frac{(\sigma_1' + \sigma_2')^2}{t} \times \frac{t - \frac{\sigma_1'}{\sigma_2'}}{t - \frac{\sigma_1'}{\sigma_2'} - 1}. \tag{9}$$

Using the inequality (9), we have

$$\mathbb{E}_{\mathcal{V}}[\tau] \leq T_\gamma^* + \int_0^{\infty} \exp\left(-\frac{t}{2(\sigma_1' + \sigma_2')^2} \frac{t - \frac{\sigma_1'}{\sigma_2'} - 1}{t - \frac{\sigma_1'}{\sigma_2'}} \gamma^2 (\mu_1 - \mu_2)^2 \right) dt$$

$$\leq \quad T^*_\gamma + \frac{2(\sigma_1'+\sigma_2')^2}{\gamma^2(\mu_1-\mu_2)^2}\exp\left(\frac{\gamma^2(\mu_1-\mu_2)^2}{2(\sigma_1'+\sigma_2')^2}\right).$$

Next, we upper bound T^*_γ. Let $r \in [0, e/2-1]$. There exists $N_0(r)$ such that for $t \geq N_0(r)$, $\beta(t,\delta) \leq \log(t^{1+r}/\delta)$. Again, using the inequality (9), we have $T^*_\gamma = \max(N_0(r), \widetilde{T}_\gamma)$, where $\widetilde{T}_\gamma$ is defined as

$$\widetilde{T}_\gamma = \inf\left\{ t_0 \in \mathbb{N} : \forall t \geq t_0, \frac{(\mu_1-\mu_2)^2}{2(\sigma_1'+\sigma_2')^2}(1-\gamma)^2 t > \frac{t-\frac{\sigma_1'}{\sigma_2'}-1}{t-\frac{\sigma_1'}{\sigma_2'}}\log\frac{t^{1+r}}{\delta}\right\}.$$

When $t > (1+\gamma\frac{\sigma_1'}{\sigma_2'})/\gamma$, $(t-\frac{\sigma_1'}{\sigma_2'}-1)/(t-\frac{\sigma_1'}{\sigma_2'}) \leq (1-\gamma)^{-1}$. We get $\widetilde{T}_\gamma = \max((1+\gamma\frac{\sigma_1'}{\sigma_2'})/\gamma, T'_\gamma)$, with

$$T'_\gamma = \inf\left\{ t_0 \in \mathbb{N} : \forall t \geq t_0, \exp\left(\frac{(\mu_1-\mu_2)^2}{2(\sigma_1'+\sigma_2')^2}(1-\gamma)^3 t\right) \geq \frac{t^{1+r}}{\delta}\right\}.$$

We use the following algebraic Lemma by Kaufmann et al. (2016).

Lemma 26 (Lemma 22 of Kaufmann et al. (2016)) *For every $\beta, \eta > 0$ and $s \in [1, e/2]$, the following implication is true:*

$$x_0 = \frac{s}{\beta}\log\left(\frac{e\log(1/(\beta^s\eta))}{\beta^s\eta}\right) \quad \Rightarrow \quad \forall x \geq x_0, \quad e^{\beta x} \geq \frac{x^s}{\eta}.$$

Applying Lemma 26 with $\eta = \delta$, $s = 1+r$ and $\beta = (1-\gamma)^3(\mu_1-\mu_2)^2/(2(\sigma_1'+\sigma_2')^2)$ leads to

$$T'_\gamma \leq \frac{(1+r)}{(1-\gamma)^3} \times \frac{2(\sigma_1'+\sigma_2')^2}{(\mu_1-\mu_2)^2}\left[\log\frac{1}{\delta} + \log\log\frac{1}{\delta}\right] + R(\mu_1,\mu_2,\sigma_1',\sigma_2',\gamma,r),$$

with

$$R(\mu_1,\mu_2,\sigma_1',\sigma_2',\gamma,r) = \frac{1+r}{(1-\gamma)^3}\frac{2(\sigma_1'+\sigma_2')^2}{(\mu_1-\mu_2)^2}\left[1+(1+r)\log\left(\frac{2(\sigma_1'+\sigma_2')^2}{(1-\gamma)^3(\mu_1-\mu_2)^2}\right)\right].$$

For fixed $\epsilon > 0$, choosing small enough r and γ, we have

$$\mathbb{E}_\mathcal{V}[\tau] \leq (1+\epsilon)\frac{2(\sigma_1'+\sigma_2')^2}{(\mu_1-\mu_2)^2}\left[\log\frac{1}{\delta} + \log\log\frac{1}{\delta}\right] + \mathcal{C}(\mu_1,\mu_2,\sigma_1',\sigma_2',\epsilon),$$

where $\mathcal{C}$ is a constant independent of δ summarizing the terms: $R(\mu_1,\mu_2,\sigma_1',\sigma_2',\gamma,r)$, $(1+\gamma\frac{\sigma_1'}{\sigma_2'})/\gamma$, $N_0(t)$, and $\frac{2(\sigma_1'+\sigma_2')^2}{\gamma^2(\mu_1-\mu_2)^2}\exp\left(\frac{\gamma^2(\mu_1-\mu_2)^2}{2(\sigma_1'+\sigma_2')^2}\right)$. $\mathcal{C}(\mu_1,\mu_2,\sigma_1',\sigma_2',\epsilon)$ goes to infinity when ϵ goes to zero, but for a fixed $\epsilon > 0$,

$$(1+\epsilon)\frac{2(\sigma_1'+\sigma_2')^2}{(\mu_1-\mu_2)^2}\log\log\frac{1}{\delta} + \mathcal{C}(\mu_1,\mu_2,\sigma_1',\sigma_2',\epsilon) = \underset{\delta\to 0}{o_\epsilon}\left(\log\frac{1}{\delta}\right).$$

This concludes the proof.

Appendix F. Proofs for Section 5

F.1 Proof of Lemma 7

Proof We have

$$\begin{aligned}\mathrm{Alt}(\nu) &= \{((\lambda_{a,x})_{a\in[K],x\in\mathcal{X}},\zeta)\in\Theta : a^*(((\lambda_{a,x}),\zeta)) \neq a^*(\nu)=1\}\\ &= \bigcup_{a\neq 1}\Big\{((\lambda_{a,x})_{a\in[K],x\in\mathcal{X}},\zeta)\in\Theta : \sum_{x\in\mathcal{X}}\zeta_x\lambda_{a,x} > \sum_{x\in\mathcal{X}}\zeta_x\lambda_{1,x}\Big\}.\end{aligned}$$

Then, we get

$$\begin{aligned}&\inf_{((\lambda_{a,x})_{a\in[K],x\in\mathcal{X}},\zeta)\in\mathrm{Alt}(\nu)}\sum_{x\in\mathcal{X}}\zeta_x\sum_{a=1}^K w_{a,x}\mathrm{kl}(\mu_{a,x},\lambda_{a,x})\\ &= \inf_{((\lambda_{a,x})_{a\in[K],x\in\mathcal{X}},\zeta):\exists a\in[K],\sum_{x\in\mathcal{X}}\zeta_x\lambda_{a,x}>\sum_{x\in\mathcal{X}}\zeta_x\lambda_{1,x}}\sum_{x\in\mathcal{X}}\zeta_x\sum_{a=1}^K w_{a,x}\mathrm{kl}(\mu_{a,x},\lambda_{a,x})\\ &= \min_{a\neq 1}\inf_{((\lambda_{a,x}),\zeta):\sum_{x\in\mathcal{X}}\zeta_x\lambda_{a,x}>\sum_{x\in\mathcal{X}}\zeta_x\lambda_{1,x}}\sum_{x\in\mathcal{X}}\zeta_x\sum_{a=1}^K w_{a,x}\mathrm{kl}(\mu_{a,x},\lambda_{a,x})\\ &= \min_{a\neq 1}\inf_{((\lambda_{a,x}),\zeta):\sum_{x\in\mathcal{X}}\zeta_x\lambda_{a,x}>\sum_{x\in\mathcal{X}}\zeta_x\lambda_{1,x}}\sum_{x\in\mathcal{X}}\zeta_x\Big(w_{1,x}\mathrm{kl}(\mu_{1,x},\lambda_{1,x})+w_{a,x}\mathrm{kl}(\mu_{a,x},\lambda_{a,x})\Big).\end{aligned}$$

We first fixed a and then took the infimum regarding the problems given a from the second to the third line (this decomposes the infimum in the second line). We used $\lambda_{b,x}=\lambda_{1,x}$ for $b\neq 1,a$ from the third line to the last line. ■

F.2 Proof of Lemma 8

Proof Let $a\in[K]$ be one of the arguments that minimizes

$$\inf_{((\lambda_{a,x}),\zeta):\sum_{x\in\mathcal{X}}\zeta_x\lambda_{a,x}>\sum_{x\in\mathcal{X}}\zeta_x\lambda_{1,x}} f_a((\lambda_{a,x}))$$

and suppose $\sum_{x\in\mathcal{X}}\zeta_x\lambda^*_{a,x} > \sum_{x\in\mathcal{X}}\zeta_x\lambda^*_{1,x}$. For such a, from the assumption on Θ, there exists $x\in\mathcal{X}$ such that $\mu_{1,x}>\mu_{a,x}$. For such x, from the monotonicity of the KL divergence,

$$\mu_{1,x}\geq\max(\lambda_{1,x},\lambda_{a,x})\geq\min(\lambda_{1,x},\lambda_{a,x})\geq\mu_{a,x}.$$

Then, by the assumption $\sum_{x\in\mathcal{X}}\zeta_x\lambda^*_{a,x}>\sum_{x\in\mathcal{X}}\zeta_x\lambda^*_{1,x}$, one can modify the value of $\lambda^*_{1,x}$ as $\lambda^*_{1,x}+\varepsilon$ or $\lambda^*_{a,x}$ as $\lambda^*_{a,x}-\varepsilon$ (ε is some small constant) to make the value of $f_a((\lambda_{a,x}))$ strictly smaller. This is a contradiction and concludes the proof.

■

F.3 Proof of Lemma 9

Proof Let us define a function

$$f(\nu, (\lambda_{a,x})) = \sum_{x\in\mathcal{X}} \zeta_x \sum_{a=1}^{K} w_{a,x}\mathrm{kl}(\mu_{a,x}, \lambda_{a,x}).$$

We call the point-to-set mapping

$$X(\nu) = \bigcup_{a\neq 1} \Big\{((\lambda_{a,x}), \zeta_x) \in \Theta : \sum_{x\in\mathcal{X}} \zeta_x \lambda_{a,x} > \sum_{x\in\mathcal{X}} \zeta_x \lambda_{1,x}\Big\}$$

as a constraint mapping. It is easy to check that $X(\nu)$ is outer semicontinuous at every ν. Similarly, $X(\nu)$ is inner semicontinuous at every ν. Therefore, from the stability theory in optimization Hogan (1973) and the continuity of the KL divergence, $m(\boldsymbol{w}, \nu)$ is continuous at every ν when $\boldsymbol{w}$ is fixed.

■

F.4 Proof of Lemma 10

Proof The proof is similar to that of Lemma 9. The constraint $\sum_{x\in\mathcal{X}} \zeta_x \lambda_{a,x} > \sum_{x\in\mathcal{X}} \zeta_x \lambda_{1,x}$ is invariant under the changes of $\boldsymbol{w} \in \mathcal{W}$ and the KL divergence is continuous. From the stability theory of Hogan (1973), $m(\boldsymbol{w}, \nu)$ is continuous when ν is fixed. ■

F.5 Proof of Lemma 12

Proof Suppose $\boldsymbol{w}_k$ does not converge to $\Phi(\nu)$. Then, there exists $\varepsilon > 0$ such that for any $n_1 \in \mathbb{N}$, there exists $k \geq n_1$ such that

$$\inf_{(w_{a,x})\in\bar{\mathcal{W}}} \max_{a,x} |w_{a,x}^{(k)} - w_{a,x}| \geq \varepsilon.$$

Also, there exists $C(\varepsilon) > 0$ such that

$$\max_{\boldsymbol{w}\in\mathcal{W}} m(\boldsymbol{w}, \nu) - \max_{\boldsymbol{w}\in\mathcal{W}:\inf_{(w_{a,x})\in\bar{\mathcal{W}}} \max_{a,x} |w_{a,x}^{(k)} - w_{a,x}| \geq \varepsilon} m(\boldsymbol{w}, \nu) \geq C(\varepsilon). \tag{10}$$

Let $w^* \in \arg\max m(\boldsymbol{w}, \nu)$. We can find a constant $\varepsilon_2(C(\varepsilon)) > 0$ such that for any $n_2 \in \mathbb{N}, n_2 \geq n_1$, there exists $k \geq n_2$ such that

$$\begin{aligned}
&|\max_{\boldsymbol{w}\in\mathcal{W}} m(\boldsymbol{w}, \nu) - \max_{\boldsymbol{w}\in\mathcal{W}} m(\boldsymbol{w}, \nu_k)| \\
&= |m(\boldsymbol{w}^*, \nu) - m(\boldsymbol{w}^*, \nu_k) + m(\boldsymbol{w}^*, \nu_k) - \max_{\boldsymbol{w}\in\mathcal{W}} m(\boldsymbol{w}, \nu_k)| \\
&\geq \Big||m(\boldsymbol{w}^*, \nu) - m(\boldsymbol{w}^*, \nu_k)| - |m(\boldsymbol{w}^*, \nu_k) - \max_{\boldsymbol{w}\in\mathcal{W}} m(\boldsymbol{w}, \nu_k)|\Big| \\
&\overset{(a)}{\geq} \varepsilon_2(C(\varepsilon)),
\end{aligned}$$

where for (a), we used (i) $m(\boldsymbol{w}^*, \nu) \to m(\boldsymbol{w}^*, \nu_k)$: from the continuity of $m(\boldsymbol{w}, \nu)$ with respect to ν for a fixed w (Lemma 9) with the convergence assumption of $(\nu_k)_{k\geq 1}$ and (ii) $|m(\boldsymbol{w}^*, \nu_k) - \max_{\boldsymbol{w}\in\mathcal{W}} m(\boldsymbol{w}, \nu_k)| \geq C(\varepsilon)$: from the optimality gap (10). Therefore, $\max_{\boldsymbol{w}\in\mathcal{W}} m(\boldsymbol{w}, \nu_k)$ does not converge to $\max_{\boldsymbol{w}\in\mathcal{W}} m(\boldsymbol{w}, \nu)$; hence contradiction. ■

F.6 Proof of Lemma 13

Proof Take any $(x^*_{a,x}), (y^*_{a,x}) \in \Phi(\nu)$ and any $\alpha \in [0, 1]$. We have

$$
\begin{aligned}
&m(\alpha(x^*_{a,x}) + (1-\alpha)(y^*_{a,x}), \nu) \\
&= \inf_{((\lambda_{a,x}),\zeta)\in \mathrm{Alt}(\nu)} \sum_{x\in\mathcal{X}} \zeta_x \sum_{a=1}^{K} (\alpha x^*_{a,x} + (1-\alpha) y^*_{a,x}) \mathrm{kl}(\mu_{a,x}, \lambda_{a,x}) \\
&\geq \alpha \inf_{((\lambda_{a,x}),\zeta)\in \mathrm{Alt}(\nu)} \sum_{x\in\mathcal{X}} \zeta_x \sum_{a=1}^{K} x^*_{a,x} \mathrm{kl}(\mu_{a,x}, \lambda_{a,x}) + (1-\alpha) \inf_{((\lambda_{a,x}),\zeta)\in \mathrm{Alt}(\nu)} \sum_{x\in\mathcal{X}} \zeta_x \sum_{a=1}^{K} y^*_{a,x} \mathrm{kl}(\mu_{a,x}, \lambda_{a,x}) \\
&= \max_{\boldsymbol{w}'\in\mathcal{W}} m(\boldsymbol{w}', \nu).
\end{aligned}
$$

Hence, $\alpha(x^*_{a,x}) + (1-\alpha)(y^*_{a,x}) \in \Phi(\nu)$. This concludes the proof. ■

Appendix G. Proofs for Sections 6.1–6.3 and the CTS Algorithm

G.1 Proof of Lemma 14

Our proof for the tracking lemma is inspired by that of D-tracking for linear bandits by Jedra and Proutiere (2020). Let us denote by C what we want to track. For a sequence that converges to C, in the following lemma, we show how to design a sampling rule so that $\frac{N_{a,x}(t)}{t}$ also converges to C.

Lemma 27 *(Tracking a set C) Let $(w(t))_{t\geq 1}$ be a sequence taking values in $\mathcal{W}$, such that there exists a compact, convex and non empty subset C in $\mathcal{W}$, there exists $\varepsilon > 0$ and $t_0(\varepsilon) \geq 1$ such that $\forall t \geq t_0(\varepsilon)$, it holds that*

$$\min_{\boldsymbol{w}'\in C} \max_{a\in[K], x\in\mathcal{X}} \left| w_{a,x}(t) - w'_{a,x} \right| \leq \varepsilon$$

Let $g : \mathbb{N} \to \mathbb{R}$ be a non-decreasing function that $g(0) = 0$, $g(t)/t \to 0$ as $t \to \infty$ and $\forall n, m \geq 1$,

$$\inf\left\{n \in \mathbb{N} : g(n) \geq m\right\} > \inf\left\{n \in \mathbb{N} : g(n) \geq m-1\right\} + K.$$

Define for every $t' \in \{0, \ldots, t-1\}$, $\varphi^g_{x,t'} = \{a : N_{a,x}(t') < g(N_x(t'))\}$ and a sampling rule as (5) Then for all $a \in [K]$ and $x \in \mathcal{X}$,

$$N_{a,x}(t) > g(N_x(t)) - 1,$$

and there exists $t_1(\varepsilon) \geq t_0(\varepsilon)$ *such that* $\forall t \geq t_1(\varepsilon)$*, Let* $D = |\mathcal{X}|$*. Then,*

$$\min_{\boldsymbol{w} \in C} \max_{a \in [K], x \in \mathcal{X}} \left| \frac{N_{a,x}(t)}{t} - \frac{1}{t} \sum_{s=1}^{t} \mathbb{1}[X_s = x] w_{a,x} \right| \leq 3(KD-1)\varepsilon.$$

The proof of Lemma 27 is inspired by the proof of Lemma 3 in Antos et al. (2008), Lemma 17 in Garivier and Kaufmann (2016), and Lemma 6 and Proposition 2 of Jedra and Proutiere (2020). We show the proof of Lemma 27 as follows.

Proof We separately show that

$$N_{a,x}(t) > g(N_x(t)) - 1$$

and

$$\min_{\boldsymbol{w} \in C} \max_{a \in [K], x \in \mathcal{X}} \left| \frac{N_{a,x}(t)}{t} - \frac{1}{t} \sum_{s=1}^{t} \mathbb{1}[X_s = x] w_{a,x} \right| \leq 3(KD-1)\varepsilon.$$

Proof of $N_{a,x}(t) > g(N_x(t)) - 1$**.** First, we justify that $N_{a,x}(t) > g(N_x(t)) - 1$.

For all $m \in \mathbb{N}$, let us define

$$\begin{aligned} k_m &= \inf\{n \in \mathbb{N} : g(n) \geq m\}, \\ \mathcal{I}_m &= \{k_m, \ldots, k_{m+1} - 1\}. \end{aligned}$$

From our assumptions on g, we have

$$\begin{aligned} &|\mathcal{I}_m| > K, \\ &m \leq g(n) < m+1 \quad \forall n \in \mathcal{I}_m. \end{aligned}$$

We consider the following statement for all $m \in \mathbb{N}$ and for all $x \in \mathcal{X}$:

$$\begin{aligned} &\text{for all } t \in \mathbb{N} \text{ such that } N_x(t) \in \mathcal{I}_m, \text{we have for all } a \in [K], N_{a,x}(t) \geq m; \\ &\text{for all } t \in \mathbb{N} \text{ such that } N_x(t) \geq k_m + K, \text{ we have } \varphi^g_{x,t} = \emptyset \text{ and } N_{a,x}(t) \geq m+1. \end{aligned} \tag{11}$$

If (11) holds for all m, then using that for all t and for all $a \in [K]$,

$$N_{a,x}(n) > g(N_x(t)) - 1$$

because from the definitions of $\mathcal{I}_m$ and k_m, for t such that

$$N_x(t) \in \mathcal{I}_m,$$

we have

$$N_{a,x}(t) \geq m > g(N_x(t)) - 1.$$

Here, we used $g(N_x(t)) \geq m$ and $g(N_x(t)) < m+1$ from the definition of k_m.

We prove (11) by induction with respect to $m \in \mathbb{N}$. First, we show the statement holds for $m = 0$. For all t such that $N_x(t) \in \mathcal{I}_0$, it holds that for all $a \in [K]$ and for all $x \in \mathcal{X}$,

$$\varphi^g_{x,t} = \{a : 0 \leq N_{a,x}(t) < g(N_x(t)) < 1\} = \{a : N_{a,x}(t) = 0\}.$$

Here, we used $\mathcal{I}_0 = \{k_0, \dots, k_1 - 1\}$ with $k_0 = \inf\{n \in \mathbb{N} : g(n) \geq 0\}$ and $k_1 = \inf\{n \in \mathbb{N} : g(n) \geq 1\}$. Therefore, for t such that $N_x(t) \geq K = k_0 + K$, we have $N_{a,x}(t) \geq 1$ and $\varphi^g_{x,t} = \emptyset$. Thus, the statement holds for $m = 0$.

Suppose that for $m = m' \geq 0$, the statement is true; that is,

$$\text{for all } t \in \mathbb{N} \text{ such that } N_x(t) \in \mathcal{I}_{m'}, \text{we have for all } a \in [K], N_{a,x}(t) \geq m';$$
$$\text{for all } t \in \mathbb{N} \text{ such that } N_x(t) \geq k_{m'} + K, \text{ we have } \varphi^g_{x,t} = \emptyset \text{ and } N_{a,x}(t) \geq m' + 1.$$

Then, we show the statement holds for $m = m' + 1$. From the inductive hypothesis and assumption $k_{m'+1} > k_{m'} + K$, since $k_{m'+1} - 1 \geq k_{m'} + K$, it holds that for all $a \in [K]$ and for all $x \in \mathcal{X}$,

$$N_{a,x}(k_{m'+1} - 1) \geq m' + 1.$$

From the definition of $\mathcal{I}_{m'+1}$, for t such that $N_x(t) \in \mathcal{I}_{m'+1} = \{k_{m'+1}, \dots, k_{m'+2} - 1\}$, $N_{a,x}(t) \geq N_{a,x}(k_{m'+1} - 1)$. Therefore,

$$N_{a,x}(t) \geq N_{a,x}(k_{m'+1} - 1) \geq m' + 1$$

Besides, for t such that $N_x(t) \in \mathcal{I}_{m'+1}$ and for all $x \in \mathcal{X}$,

$$m' + 1 \leq g(N_x(t)) < m' + 2.$$

This leads to

$$\varphi^g_{x,t} = \{a : m' + 1 \leq N_{a,x}(t) < g(N_x(t)) < m' + 2\} = \{a : N_{a,x}(t) = m' + 1\}.$$

Then, A_t is chosen among this set while it is non empty. Therefore, for t such that $N_x(t) \geq k_{m'+1} + K$, it holds that for all $a \in [K]$ and $x \in \mathcal{X}$, $\varphi^g_{x,t} = \emptyset$ and $N_{a,x}(t) \geq m' + 2$. Thus, the statement (11) holds when $m = m' + 1$.

Proof of $\min_{\boldsymbol{w} \in C} \max_{a \in [K], x \in \mathcal{X}} \left| \frac{N_{a,x}(t)}{t} - \frac{1}{t} \sum_{s=1}^t \mathbb{1}[X_s = x] w_{a,x} \right| \leq 3(KD - 1)\varepsilon$**.** First, the condition

$$\min_{\boldsymbol{w}' \in C} \max_{a \in [K], x \in \mathcal{X}} \left| w_{a,x}(t) - w'_{a,x} \right| \leq \varepsilon$$

for $\forall t \geq t_0(\varepsilon)$ ensures that for large t,

$$\min_{\boldsymbol{w}' \in C} \max_{a \in [K], x \in \mathcal{X}} \left| \frac{1}{t} \sum_{s=1}^t \mathbb{1}[X_s = x] w_{a,x}(s) - \frac{1}{t} \sum_{s=1}^t \mathbb{1}[X_s = x] w'_{a,x} \right| \leq \varepsilon.$$

For all $t \geq 1$, we define

$$\overline{\eta}_{a,x}(t) = \frac{1}{t} \sum_{s=1}^t \mathbb{1}[X_s = x] w_{a,x}(s).$$

Next, since C is non-empty and compact, we can define

$$\widetilde{\boldsymbol{w}}(t) = \arg\min_{\boldsymbol{w}^\dagger \in C} \max_{a \in [K], x \in \mathcal{X}} \left| \overline{\eta}_{a,x}(t) - \frac{1}{t} \sum_{s=1}^t \mathbb{1}[X_s = x] w^\dagger_{a,x} \right|.$$

Here, by convexity of C, there exists $t_0'(\varepsilon) \geq t_0(\varepsilon)$ such that $\forall t \geq t_0'(\varepsilon)$, we can obtain the following inequalities:

$$\min_{\boldsymbol{w}^\dagger \in C} \max_{a\in[K],x\in\mathcal{X}} \left| \frac{N_{a,x}(t)}{t} - \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] w_{a,x}^\dagger \right| \leq \max_{a\in[K],x\in\mathcal{X}} \left| \frac{N_{a,x}(t)}{t} - \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] \widetilde{w}_{a,x}(t) \right| \tag{12}$$

and

$$\max_{a\in[K],x\in\mathcal{X}} \left| \overline{\eta}_{a,x}(t) - \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] \widetilde{w}_{a,x}(t) \right| \leq 2\varepsilon. \tag{13}$$

The first result can be directly obtained from the definition. We show the second result. To see that (13) holds, let us define for all $t \geq 1$,

$$\boldsymbol{v}(t) = \arg\min_{\boldsymbol{w}^\dagger \in C} \max_{a\in[K],x\in\mathcal{X}} |w_{a,x}(t) - w_{a,x}^\dagger|,$$

and observe that for all $a \in [K]$ and $x \in \mathcal{X}$, we have

$$\begin{aligned}
&\left| \overline{\eta}_{a,x}(t) - \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] v_{a,x}(s) \right| \\
&= \left| \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] w_{a,x}(s) - \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] v_{a,x}(s) \right| \\
&\leq \frac{1}{t}\sum_{s=1}^{t_0} \mathbb{1}[X_s = x] \left| w_{a,x}(s) - v_{a,x}(s) \right| + \frac{1}{t}\sum_{s=t_0+1}^{t} \mathbb{1}[X_s = x] \left| w_{a,x}(s) - v_{a,x}(s) \right| \\
&\leq \frac{t_0(\varepsilon)}{t} + \frac{t - t_0(\varepsilon)}{t}\varepsilon.
\end{aligned}$$

Note that $t_0(\varepsilon)$ is defined in the statement. Thus if $t \geq t_0' = \frac{t_0(\varepsilon)}{\varepsilon}$, then

$$\max_{a\in[K],x\in\mathcal{X}} \left| \overline{\eta}_{a,x}(t) - \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] v_{a,x}(s) \right| \leq 2\varepsilon.$$

Finally since the convexity of C leads to

$$\left(\frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = 1] \boldsymbol{v}_1(s), \ldots, \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = D] \boldsymbol{v}_D(s) \right)^\top \in C,$$

it follows that $\forall t \geq t_0'$

$$\max_{a\in[K],x\in\mathcal{X}} \left| \overline{\eta}_{a,x}(t) - \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] \widetilde{w}_{a,x}(s) \right| \leq \max_{a\in[K],x\in\mathcal{X}} \left| \overline{\eta}_{a,x}(t) - \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] v_{a,x}(s) \right| \leq 2\varepsilon.$$

Thus, we showed that (13) holds. By using (12) and (13), we consider bounding the term

$$\min_{\boldsymbol{w}\in C}\max_{a\in[K],x\in\mathcal{X}}\left|\frac{N_{a,x}(t)}{t}-\frac{1}{t}\sum_{s=1}^{t}\mathbb{1}[X_s=x]w_{a,x}\right|.$$

Let us define for $a\in[K]$ and for all $t\geq 1$,

$$E_{a,x,t}=N_{a,x}(t)-\sum_{s=1}^{t}\mathbb{1}[X_s=x]\widetilde{w}_{a,x}(t).$$

From (12), there exists $t_1\geq t_0'(\varepsilon)$ such that for all $t\geq t_1$,

$$\min_{\boldsymbol{w}\in C}\max_{a\in[K],x\in\mathcal{X}}\left|\frac{N_{a,x}(t)}{t}-\frac{1}{t}\sum_{s=1}^{t}\mathbb{1}[X_s=x]w_{a,x}\right|\leq\max_{a\in[K],x\in\mathcal{X}}\left|\frac{E_{a,x,t}}{t}\right|\ ,$$

Therefore, we consider bounding $\max_{a\in[K],x\in\mathcal{X}}\left|\frac{E_{a,x,t}}{t}\right|$. Since

$$\sum_{a=1}^{K}\sum_{x\in\mathcal{X}}E_{a,x,t}=\sum_{a=1}^{K}\sum_{x\in\mathcal{X}}N_{a,x}(t)-\sum_{a=1}^{K}\sum_{x\in\mathcal{X}}\sum_{s=1}^{t}\mathbb{1}[X_s=x]\widetilde{w}_{a,x}(t)=t-t=0$$

we have

$$\sup_{a,x}|E_{a,x,t}|\leq(KD-1)\sup_{a,x}\ E_{a,x,t}.$$

Then, for every $a\in[K]$ and $x\in\mathcal{X}$, we have $E_{a,x,t}\leq\sup_{a'\in[K]}\sup_{x'\in\mathcal{X}}E_{a',x',t}$ and

$$E_{a,x,t}=-\sum_{(a',x')\neq(a,x)}E_{a',x',t}\geq-\sum_{(a',x')\neq(a,x)}\sup_{a',x'}E_{a',x',t}=-(KD-1)\sup_{a',x'}E_{a',x',t}\ .$$

Next, we give an upper bound on $\sup_{a,x}E_{a,x,t}$, for t large enough. Let $t_0'\geq t_0$ such that

$$\forall t\geq t_0',\quad g(t)\leq 2t\epsilon\ \text{ and }\ 1/t\leq\epsilon\ .$$

We first show that for $t\geq t_0'$,

$$(A_{t+1}=a)\subseteq(E_{a,x,t}\leq 2t\epsilon)\tag{14}$$

To prove this, we write

$$(A_{t+1}=a)\subseteq\mathcal{E}_1\cup\mathcal{E}_2,$$

where

$$\mathcal{E}_1=\left(a\in\arg\min_{a\in[K]}\left(N_{a,x}(t)-t\sum_{s=1}^{t}\mathbb{1}[X_s=x]w_{a,x}(s)\right)\right)$$
$$\mathcal{E}_2=(N_{a,X_t}(t)\leq g(N_x(t)))$$

This inclusion is immediate by construction. Therefore, we show that

$$\mathcal{E}_1\cup\mathcal{E}_2\subseteq(E_{a,x,t}\leq 2t\epsilon)\,.$$

For the second case ($\mathcal{E}_2$), if $N_{a,x}(t) \leq g(N_x(t))$, we have

$$E_{a,x,t} \leq g(N_x(t)) - \sum_{s=1}^{t} \mathbb{1}[X_s = x] w_{a,x}(s) \leq g(N_x(t)) \leq g(t) \leq 2t\epsilon,$$

by definition of t_0'.

In the first case ($\mathcal{E}_1$), for $t \geq t_0$, we have

$$\begin{aligned}
E_{a,x,t} &= N_{a,x}(t) - \sum_{s=1}^{t} \mathbb{1}[X_s = x] \widetilde{w}_{a,x}(t) \\
&= N_{a,x}(t) - \sum_{s=1}^{t} \mathbb{1}[X_s = x] w_{a,x}(s) + \sum_{s=1}^{t} \mathbb{1}[X_s = x] w_{a,x}(s) - \sum_{s=1}^{t} \mathbb{1}[X_s = x] \widetilde{w}_{a,x}(t) \\
&\leq N_{a,x}(t) - \sum_{s=1}^{t} \mathbb{1}[X_s = x] w_{a,x}(s) + 2t\varepsilon \qquad \left(\text{since } \max_{a \in [K], x \in \mathcal{X}} \left| \overline{\eta}(t), \frac{1}{t} \sum_{s=1}^{t} \mathbb{1}[X_s = x] \widetilde{w}(t) \right| \leq 2\varepsilon \right) \\
&\leq \min_{a \in [K]} \left(N_{a,x}(t) - t \sum_{s=1}^{t} \mathbb{1}[X_s = x] w_{a,x}(s) \right) + 2t\varepsilon \qquad \text{(since } \mathcal{E}_1 \text{ holds)} \\
&\leq 2t\varepsilon.
\end{aligned}$$

where the last inequality holds because $\min_{a,x} E_{a,x,t} \leq 0$ holds from $\sum_{a=1}^{K} \sum_{x \in \mathcal{X}} E_{a,x,t} = 0$. This proves (14).

Here, $E_{a,x,t}$ satisfies $E_{a,x,t+1} = E_{a,x,t} + \mathbb{1}[A_{t+1} = a, X_{t+1} = x] - \mathbb{1}[X_s = x] \widetilde{w}_{a,x}(t+1)$, therefore, if $t \geq t_0'$,

$$\begin{aligned}
E_{a,x,t+1} &\leq E_{a,x,t} + \mathbb{1}[A_{t+1} = a, X_{t+1} = x] - \mathbb{1}[X_s = x] \widetilde{w}_{a,x}(t+1) \\
&\leq E_{a,x,t} + \mathbb{1}[E_{a,x,t} \leq 2t\epsilon] - \mathbb{1}[X_s = x] \widetilde{w}_{a,x}(t+1).
\end{aligned}$$

We now prove by induction that for every $t \geq t_0'$, we have

$$E_{a,x,t} \leq \max(E_{a,x,t_0'}, 2t\epsilon + 1).$$

For $t = t_0'$, this statement clearly holds. Let $t \geq t_0'$ such that the statement holds. If $E_{a,x,t} \leq 2t\epsilon$, we have

$$\begin{aligned}
E_{a,x,t+1} &\leq 2t\epsilon + 1 - \widetilde{w}_{a,x}(t+1) \leq 2t\epsilon + 1 \leq \max(E_{a,x,t_0'}, 2t\epsilon + 1) \\
&\leq \max(E_{a,x,t_0'}, 2(t+1)\epsilon + 1).
\end{aligned}$$

If $E_{a,x,t} > 2t\epsilon$, the indicator is zero and

$$E_{a,x,t+1} \leq \max(E_{a,x,t_0'}, 2t\epsilon + 1) - \widetilde{w}_{a,x}(t+1) \leq \max(E_{a,x,t_0'}, 2(t+1)\epsilon + 1),$$

which concludes the induction.

For all $t \geq t_0'$, using that $E_{a,x,t_0'} \leq t_0'$ and $1/t \leq \epsilon$, it follows that

$$\max_{a \in [K], x \in \mathcal{X}} \left| \frac{E_{a,x,t}}{t} \right| \leq (KD - 1) \max \left(2\epsilon + \frac{1}{t}, \frac{t_0'}{t} \right) \leq (KD - 1) \max \left(3\epsilon, \frac{t_0}{t} \right).$$

Hence, as mentioned above, from (12), there exists $t_1 \geq t'_0(\varepsilon)$ such that for all $t \geq t_1$,

$$\min_{\boldsymbol{w}\in C} \max_{a\in[K],x\in\mathcal{X}} \left| \frac{N_{a,x}(t)}{t} - \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] w_{a,x} \right| \leq \max_{a\in[K],x\in\mathcal{X}} \left| \frac{E_{a,x,t}}{t} \right| \leq 3(KD-1)\epsilon,$$

which concludes the proof. ■

Then, we can prove Lemma 14 as follows.

Proof Let $g(n) = (\sqrt{n} - K/2)_+$. Let $\varepsilon' = \frac{\varepsilon}{3KD-1} > 0$ and $C = \Phi(\nu)$. First, by Lemma 13, and Lemma 12, there exists $\xi(\varepsilon') > 0$ such that for all $\nu' = ((\mu'_{a,x}), (\zeta'_x))$ such that

$$|\mu_{a,x} - \mu'_{a,x}| < \xi(\varepsilon')$$

and

$$|\zeta_x - \zeta'_x| \leq \xi(\varepsilon'),$$

we have

$$\max_{\boldsymbol{w}\in\Phi(\nu')} \min_{\boldsymbol{w}'\in\Phi(\nu)} \max_{a\in[K],x\in\mathcal{X}} \left| \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] w_{a,x}(s) - \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] w'_{a,x} \right| \leq \varepsilon'/2.$$

From the law of large numbers, there exists $t_0(\varepsilon') \geq 0$ such that for all $t \geq t_0(\varepsilon')$, we have $|\mu_{a,x} - \hat{\mu}_{a,x}(t)| \leq \xi(\varepsilon')$ and $|\zeta_x - \hat{\zeta}_x(t)| \leq \xi(\varepsilon')$. Here, the $\hat{\nu}(t)$ in the plug-in estimate $\Phi(\hat{\nu}(t))$ is $\hat{\nu}(t) = ((\hat{\mu}_{a,x}(t)), (\hat{\zeta}_x(t)))$. The condition (4) states that

$$\lim_{t\to\infty} \min_{\boldsymbol{w}'\in\Phi(\nu)} \max_{a\in[K],x\in\mathcal{X}} \left| w_{a,x}(t) - w'_{a,x} \right| = 0$$

almost surely. This guarantees that there exist $t_1 \geq 1$ such that for all $t \geq t_1$, we have

$$\min_{\boldsymbol{w}'\in\Phi(\nu)} \max_{a\in[K],x\in\mathcal{X}} \left| \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] w_{a,x}(s) - \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] w'_{a,x} \right| < \varepsilon'/2.$$

Now for all $t \geq \max(t_0(\varepsilon'), t_1)$, we have

$$\begin{aligned}
&\min_{\boldsymbol{w}'\in\Phi(\nu)} \max_{a\in[K],x\in\mathcal{X}} \left| \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] w_{a,x}(s) - \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] w'_{a,x} \right| \\
&\leq \min_{\boldsymbol{w}'\in\Phi(\nu)} \max_{a\in[K],x\in\mathcal{X}} \left| \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] w_{a,x}(s) - \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] w'_{a,x} \right| \\
&\quad + \max_{\boldsymbol{w}\in\Phi(\hat{\nu}(t))} \min_{\boldsymbol{w}'\in\Phi(\nu)} \max_{a\in[K],x\in\mathcal{X}} \left| \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] w_{a,x}(s) - \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] w'_{a,x} \right| < \varepsilon'.
\end{aligned}$$

Thus, we have shown that

$$\min_{\boldsymbol{w}'\in\Phi(\nu)} \max_{a\in[K],x\in\mathcal{X}} \left| \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] w_{a,x}(s) - \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] w'_{a,x} \right| \to 0$$

almost surely.

Next, we recall that by Lemmas 8 and 13, $\Phi(\nu)$ is non empty, compact and convex. Thus, applying the (strong) law of large numbers and Lemma 27 yields immediately that with

$$\mathbb{P}\left(\min_{\boldsymbol{w}^* \in \Phi(\nu)} \left\{ \lim_{t\to\infty} \frac{N_{a,x}(t)}{t} = \zeta_x w^*_{a,x} \right\}\right) = 1$$

Here, we used

$$\begin{aligned}
&\min_{\boldsymbol{w}^* \in \Phi(\nu)} \left| \frac{N_{a,x}(t)}{t} - p(x) w^*_{a,x}(t) \right| \\
&= \min_{\boldsymbol{w}^* \in \Phi(\nu)} \left| \frac{N_{a,x}(t)}{t} - \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] w^*_{a,x}(t) + \frac{1}{t}\sum_{s=0}^{t} \mathbb{1}[X_s = x] w^*_{a,x}(t) - p(x)\widetilde{w}_{a,x}(t) \right| \\
&\leq \min_{\boldsymbol{w}^* \in \Phi(\nu)} \left| \frac{N_{a,x}(t)}{t} - \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] w^*_{a,x}(t) \right| + \left| \left\{ \frac{1}{t}\sum_{s=0}^{t} \mathbb{1}[X_s = x] - p(x) \right\} w^*_{a,x}(t) \right| \\
&\leq \min_{\boldsymbol{w}^* \in \Phi(\nu)} \left| \frac{N_{a,x}(t)}{t} - \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] w^*_{a,x}(t) \right| + \left| \frac{1}{t}\sum_{s=0}^{t} \mathbb{1}[X_s = x] - p(x) \right|,
\end{aligned}$$

and for $t \geq t_0(\varepsilon')$

$$\begin{aligned}
&\min_{\boldsymbol{w}^* \in C} \max_{a\in[K], x\in\mathcal{X}} \left| \frac{N_{a,x}(t)}{t} - \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] w^*_{a,x} \right| \leq \varepsilon' \\
&\|\boldsymbol{\zeta} - \hat{\boldsymbol{\zeta}}_t\| \leq \xi(\varepsilon').
\end{aligned}$$

■

G.2 Proof of Theorem 15

We proceed similarly to Garivier and Kaufmann (2016). Introducing, for $a, b \in [K]$, $T_{a,b} := \inf\{t \in \mathbb{N} : Z_{a,b}(t) > \beta(t,\delta)\}$, we have

$$\begin{aligned}
\mathbb{P}_\nu(\tau_\delta < \infty, \hat{a}_{\tau_\delta} \neq a^*) &\leq \mathbb{P}_\nu\left(\exists a \in [K] \setminus \{a^*\}, \exists t \in \mathbb{N} : Z_{a,a^*}(t) > \beta(t,\delta)\right) \\
&\leq \sum_{a \in [K]\setminus\{a^*\}} \mathbb{P}_\nu(T_{a,a^*} < \infty)\,.
\end{aligned}$$

We show that if $\beta(t,\delta) = \log(2t(K-1)/\delta)$ and $\mu_a < \mu_b$, then $\mathbb{P}_\nu(T_{a,b} < \infty) \leq \frac{\delta}{K-1}$. For such a pair of arms, observe that on the event $\left\{T_{a,b} = t\right\}$ time t is the first moment when $Z_{a,b}(t)$ exceeds the threshold $\beta(t,\delta)$, which implies by definition that

$$1 \leq e^{-\beta(t,\delta)} \frac{\max_{\xi_a \geq \xi_b} p_{\xi_a}\left(\underline{R}_{a,x}(t), \underline{X}(t)\right) p_{\xi_b}\left(\underline{R}_{b,x}(t), \underline{X}(t)\right)}{\max_{\xi_a \leq \xi_b} p_{\xi_a}\left(\underline{R}_{a,x}(t), \underline{X}(t)\right) p_{\xi_b}\left(\underline{R}_{b,x}(t), \underline{X}(t)\right)}\,.$$

It thus holds that

$$\mathbb{P}_\nu(T_{a,b}<\infty)=\sum_{t=1}^{\infty}\mathbb{P}_\nu(T_{a,b}=t)=\sum_{t=1}^{\infty}\mathbb{E}_\nu\Big[\mathbb{1}[T_{a,b}=t]\Big]$$
$$\leq\sum_{t=1}^{\infty}\exp\big(-\beta(t,\delta)\big)\mathbb{E}_\nu\left[\mathbb{1}\left[T_{a,b}=t\right]\frac{\max_{\xi_a\geq\xi_b}p_{\xi_a}\big(\underline{R}_{a,x}(t),\underline{X}(t)\big)p_{\xi_b}\big(\underline{R}_{b,x}(t),\underline{X}(t)\big)}{\max_{\xi_a\leq\xi_b}p_{\xi_a}\big(\underline{R}_{a,x}(t),\underline{X}(t)\big)p_{\xi_b}\big(\underline{R}_{b,x}(t),\underline{X}(t)\big)}\right]$$
$$\leq\sum_{t=1}^{\infty}\exp\big(-\beta(t,\delta)\big)\mathbb{E}_\nu\left[\mathbb{1}\left[T_{a,b}=t\right]\frac{\max_{\xi_a\geq\xi_b}p_{\xi_a}\big(\underline{R}_{a,x}(t),\underline{X}(t)\big)p_{\xi_b}\big(\underline{R}_{b,x}(t),\underline{X}(t)\big)}{p_{\mu_a}\big(\underline{R}_{a,x}(t),\underline{X}(t)\big)p_{\mu_b}\big(\underline{R}_{b,x}(t),\underline{X}(t)\big)}\right].$$

We expand the expectation $\mathbb{E}_\nu\left[\mathbb{1}\left[T_{a,b}=t\right]\frac{\max_{\xi_a\geq\xi_b}p_{\xi_a}\left(\underline{R}_{a,x}(t),\underline{X}(t)\right)p_{\xi_b}\left(\underline{R}_{b,x}(t),\underline{X}(t)\right)}{p_{\mu_a}\left(\underline{R}_{a,x}(t),\underline{X}(t)\right)p_{\mu_b}\left(\underline{R}_{b,x}(t),\underline{X}(t)\right)}\right]$ as follows:

$$\begin{aligned}
&\mathbb{E}_\nu\left[\mathbb{1}\left[T_{a,b}=t\right]\frac{\max_{\xi_a\geq\xi_b}p_{\xi_a}\big(\underline{R}_{a,x}(t),\underline{X}(t)\big)p_{\xi_b}\big(\underline{R}_{b,x}(t),\underline{X}(t)\big)}{p_{\mu_a}\big(\underline{R}_{a,x}(t),\underline{X}(t)\big)p_{\mu_b}\big(\underline{R}_{b,x}(t),\underline{X}(t)\big)}\right]\\
&=\sum_{\underline{r}_t\in\{0,1\}^t}\sum_{\underline{a}_t\in[K]^t}\sum_{\underline{X}(t)\in\mathcal{X}^t}\mathbb{1}\left[T_{a,b}=t\right](\underline{r}_t,\underline{a}_t,\underline{x}_t)\max_{\xi_a\geq\xi_b}p_{\xi_a}\big(\underline{R}_{a,x}(t)=\underline{r}_t,\underline{X}(t)=\underline{x}_t\big)p_{\xi_b}\big(\underline{R}_{b,x}(t)=\underline{r}_t,\underline{X}(t)=\underline{x}_t\big)\\
&\quad\cdot\prod_{c\in[K]\setminus\{a,b\}}\left[\prod_{s=1}^{t}p_{\mu_c}\big(R_{s,c}=r_s\mid X_s=x_s\big)p(A_1=c\mid X_1=x_1)\prod_{s=2}^{t}p(A_s=c\mid X_s=x_s,\Omega_{s-1})\right]\frac{1}{p(\underline{X}(t)=\underline{x}_t)}\\
&=\sum_{\underline{r}_t\in\{0,1\}^t}\sum_{\underline{a}_t\in[K]^t}\sum_{\underline{x}_t\in\mathcal{X}^t}\mathbb{1}\left[T_{a,b}=t\right](\underline{r}_t,\underline{a}_t,\underline{x}_t)\max_{\xi_a\geq\xi_b}p_{\xi_a}\big(\underline{R}_{a,x}(t)=\underline{r}_t\mid\underline{X}(t)=\underline{x}_t\big)p_{\xi_b}\big(\underline{R}_{b,x}(t)=\underline{r}_t\mid\underline{X}(t)=\underline{x}_t\big)\\
&\quad\cdot\prod_{c\in[K]\setminus\{a,b\}}\left[\prod_{s=1}^{t}p_{\mu_c}\big(R_{s,c}=r_s\mid X_s=x_s\big)p(A_1=c\mid X_1=x_1)\prod_{s=2}^{t}p(A_s=c\mid X_s=x_s,\Omega_{s-1})\right]p(\underline{X}(t)=\underline{x}_t),
\end{aligned}\tag{15}$$

where $\underline{r}_t$ denotes the sequence $\{r_s\}_{s=1}^t$, $\underline{a}_t$ denotes the sequence $\{a_s\}_{s=1}^t$, $\underline{x}_t$ denotes the sequence $\{x_s\}_{s=1}^t$, $p_{\mu_b}\big(R_{t,c}=r_t\mid X_t=x_t\big)$ denotes the conditional density of r_t given x_t Note that $\mathbb{1}\left[T_{a,b}=t\right]$ is a random variable depending on $(\underline{R}_t,\underline{A}_t,\underline{X}(t))$, therefore, we denote it as $\mathbb{1}\left[T_{a,b}=t\right](\underline{r}_t,\underline{a}_t,\underline{x}_t)$. For a vector x, let us introduce the Krichevsky-Trofimov distribution

$$\mathrm{kt}(x)=\int_0^1\frac{1}{\pi\sqrt{u(1-u)}}p_u(x),$$

as defined in Lemma 11 of Garivier and Kaufmann (2016). Then, following the same procedure as Garivier and Kaufmann (2016), we bound (15) by

$$\begin{aligned}
&\sum_{t=1}^{\infty}2t\exp\big(-\beta(t,\delta)\big)\sum_{\underline{r}_t\in\{0,1\}^t}\sum_{\underline{a}_t\in[K]^t}\sum_{\underline{x}_t\in\mathcal{X}^t}\mathbb{1}\left[T_{a,b}=t\right](\underline{r}_t,\underline{a}_t,\underline{x}_t)\mathrm{kt}\big(\underline{R}_{a,x}(t)\big)\mathrm{kt}\big(\underline{R}_{b,x}(t)\big)\\
&\times\prod_{c\in[K]\setminus\{a,b\}}\left[\prod_{s=1}^{t}p_{\mu'_c}\big(R_{s,c}=r_s\mid X_s=x_s\big)p(A_1=c\mid X_1=x_1)\prod_{s=2}^{t}p(A_s=c\mid X_s=x_s,\Omega_{s-1})\right]p(\underline{X}(t)=\underline{x}_t)\\
&=\sum_{t=1}^{\infty}2t\exp\big(-\beta(t,\delta)\big)\sum_{\underline{r}_t\in\{0,1\}^t}\sum_{\underline{a}_t\in[K]^t}\sum_{\underline{x}_t\in\mathcal{X}^t}\mathbb{1}\left[T_{a,b}=t\right](\underline{r}_t,\underline{a}_t,\underline{x}_t)I(\underline{r}_t,\underline{a}_t,\underline{x}_t)p(\underline{X}(t)=\underline{x}_t),
\end{aligned}$$

where the partially integrated likelihood

$$I(\underline{r}_t, \underline{a}_t, \underline{x}_t) = \mathrm{kt}\big(\underline{R}_{a,x}(t)\big)\mathrm{kt}\big(\underline{R}_{b,x}(t)\big)$$
$$\times \prod_{c\in[K]\setminus\{a,b\}} \left[\prod_{s=1}^{t} p_{\mu'_c}\big(R_{s,c} = r_s \mid X_s = x_s\big) p(A_1 = c \mid X_1 = x_1) \prod_{s=2}^{t} p(A_s = c \mid X_s = x_s, \Omega_{s-1})\right]$$

is the density of an alternative probability measure $\widetilde{\mathbb{P}}$, under which μ_a and μ_b are drawn from a Beta$(1/2, 1/2)$ distribution at the beginning of the sampling process. This is bounded as

$$\leq \sum_{t=1}^{\infty} 2t \exp\big(-\beta(t,\delta)\big) \sum_{\underline{x}_t \in \mathcal{X}^t} \widetilde{\mathbb{P}}\left(T_{a,b} = t\right) p(\underline{X}(t) = \underline{x}_t)$$
$$\leq \frac{\delta}{K-1} \sum_{t=1}^{\infty} \widetilde{\mathbb{P}}\left(T_{a,b} = t\right) = \frac{\delta}{K-1} \widetilde{\mathbb{P}}\left(T_{a,b} < \infty\right) \leq \frac{\delta}{K-1},$$

Thus, for any $\mu_a < \mu_b$, then $\mathbb{P}_\nu(T_{a,b} < \infty) \leq \frac{\delta}{K-1}$. Therefore,

$$\mathbb{P}_\nu(\tau_\delta < \infty, \hat{a}_{\tau_\delta} \neq a^*) \leq \sum_{a\in[K]\setminus\{a^*\}} \mathbb{P}_\nu(T_{a,a^*} < \infty) \leq K - 1\frac{\delta}{K-1} = \delta.$$

Appendix H. Proofs for Section 6.4

H.1 Proof of Lemma 18

Proof In a Bernoulli bandit model, let $\mathcal{E}$ be an event such that

$$\mathcal{E} = \left\{\forall a \in [K], \forall x \in \mathcal{X},\ \min_{\boldsymbol{w}^* \in \Phi(\nu)} \left|\lim_{t\to\infty} \frac{N_{a,x}(t)}{t} - \zeta_x w^*_{a,x}\right| = 0,\ \hat{\mu}_{a,x}(t) \overset{t\to\infty}{\to} \mu_{a,x},\ \frac{N_x(t)}{t} \overset{t\to\infty}{\to} \zeta_x\right\}.$$

When considering a bandit model that belongs to a canonical one-parameter exponential family, suppose that the true parameter ζ_x is given; that is, $\hat{\zeta}_x(t) = \zeta_x$. From the assumption on the sampling strategy (see Lemma 14) and the law of large numbers, $\mathcal{E}$ is of probability 1. On $\mathcal{E}$, there exists t_0 such that for all $t \geq t_0$, $\hat{\mu}_1(t) > \max_{a\neq 1} \hat{\mu}_a(t)$ and

$$Z(t) = \min_{a\neq 1} Z_{1,a}(t)$$
$$= t \min_{a\neq 1} \sum_{x\in\mathcal{X}} \left\{ \frac{N_{1,x}(t)}{t} \left\{\hat{\mu}_{1,x}(t) \log \frac{\hat{\mu}_{1,x}(t)}{1-\hat{\mu}_{1,x}(t)} + \log(1-\hat{\mu}_{1,x}(t))\right\}\right.$$
$$+ \frac{N_{a,x}(t)}{t} \left\{\hat{\mu}_{a,x}(t) \log \frac{\hat{\mu}_{a,x}(t)}{1-\hat{\mu}_{a,x}(t)} + \log(1-\hat{\mu}_{a,x}(t))\right\}$$
$$- \frac{N_{1,x}(t)}{t} \left\{\hat{\mu}_{1,x}(t) \log \frac{\widetilde{\xi}_{1,x}(t)}{1-\widetilde{\xi}_{1,x}(t)} + \log(1-\widetilde{\xi}_{1,x}(t))\right\}$$
$$\left. - \frac{N_{a,x}(t)}{t} \left\{\hat{\mu}_{a,x}(t) \log \frac{\widetilde{\xi}_{a,x}(t)}{1-\widetilde{\xi}_{a,x}(t)} + \log(1-\widetilde{\xi}_{a,x}(t))\right\}\right\}.$$

By continuity of m, there exists an open neighborhood $\mathcal{N}(\nu,\varepsilon)$ of $\Phi(\nu)\times\{\boldsymbol{\mu}\}\times\{\boldsymbol{\zeta}\}$ such that for all $(\boldsymbol{w}',\boldsymbol{\mu}',\boldsymbol{\zeta}')\in\mathcal{N}(\nu,\varepsilon)$, it holds that

$$m(\boldsymbol{w}',\nu')\geq(1-\varepsilon)m(\boldsymbol{w}',\nu),$$

where where $\nu'=(\boldsymbol{\mu}',\boldsymbol{\zeta}')$, and $\boldsymbol{w}^\star$ is some element in $\Phi(\nu)$. Recall that the function m is defined in Section 5.2 Now, observe that under the event $\mathcal{E}$, there exists $t_1\geq t_0$ such that for all $t\geq t_1$ it holds that $((\hat{\mu}_{a,x}(t)),(\hat{\zeta}_x(t)))\in\mathcal{N}(\nu,\varepsilon)$, thus for all $t\geq t_0$, it follows that

$$m((N_a(t)/t)_{a\in[K]},\hat{\nu}_t)\geq\frac{1}{1+\epsilon}m(\boldsymbol{w}^*,\nu),$$

where $\hat{\nu}_t=(\hat{\boldsymbol{\mu}}_t,\hat{\boldsymbol{\zeta}}_t)$. Therefore, on $\mathcal{E}$, for all $t\geq t_1$,

$$Z(t)=tm((N_a(t)/t)_{a\in[K]},\hat{\boldsymbol{\mu}}_t,\hat{\boldsymbol{\zeta}}_t)\geq\frac{t}{1+\epsilon}m(\boldsymbol{w}^*,\nu)=\frac{t}{(1+\epsilon)T^\star(\nu)}.$$

Consequently,

$$\begin{aligned}\tau_\delta &= \inf\{t\in\mathbb{N}:Z(t)\geq\beta(t,\delta)\}\\ &\leq t_1\vee\inf\{t\in\mathbb{N}:t(1+\epsilon)^{-1}T^\star(\nu)^{-1}\geq\log(r(t)/\delta)\}\\ &\leq t_1\vee\inf\{t\in\mathbb{N}:t(1+\epsilon)^{-1}T^\star(\nu)^{-1}\geq\log(Ct^\alpha/\delta)\},\end{aligned}$$

for some positive constant C. Using the technical Lemma 18 in Garivier and Kaufmann (2016), it follows that on $\mathcal{E}$, as $\alpha\in[1,e/2]$,

$$\tau_\delta\leq t_1\vee\alpha(1+\epsilon)T^\star(\nu)\left[\log\left(\frac{Ce((1+\epsilon)T^\star(\boldsymbol{\mu}))^\alpha}{\delta}\right)+\log\log\left(\frac{C((1+\epsilon)T^\star(\nu))^\alpha}{\delta}\right)\right].$$

Thus τ_δ is finite on $\mathcal{E}$ for every $\delta\in(0,1)$, and

$$\limsup_{\delta\to0}\frac{\tau_\delta}{\log(1/\delta)}\leq(1+\epsilon)\,\alpha\,T^\star(\nu)\,.$$

Letting ϵ go to zero concludes the proof.

■

H.2 Proof of Theorem 19

This proof also mainly follows Garivier and Kaufmann (2016). We use the following proposition from Garivier and Kaufmann (2016).

Proposition 28 (Lemma 18 of Garivier and Kaufmann (2016)) *For every $\alpha\in[1,e/2]$, for any two constants $c_1,c_2>0$,*

$$x=\frac{\alpha}{c_1}\left[\log\left(\frac{c_2e}{c_1^\alpha}\right)+\log\log\left(\frac{c_2}{c_1^\alpha}\right)\right]$$

is such that $c_1x\geq\log(c_2x^\alpha)$.

To ease the notation, we assume that the bandit model ν is such that $\mu_1 > \mu_2 \geq \cdots \geq \mu_K$. Let $\epsilon > 0$. From Lemma 12, there exists $\Upsilon = \Upsilon(\epsilon) \leq (\mu_1 - \mu_2)/4$ such that

$$\mathcal{I}_{\mu,\epsilon} := \prod_{x\in\mathcal{X}} \Big([\mu_{1,x} - \Upsilon, \mu_{1,x} + \Upsilon] \times [\mu_{2,x} - \Upsilon, \mu_{2,x} + \Upsilon] \times \cdots \times [\mu_{K,x} - \Upsilon, \mu_{K,x} + \Upsilon] \Big),$$
$$\mathcal{I}_{\zeta,\epsilon} := \prod_{x\in\mathcal{X}} [\zeta_x - \Upsilon, \zeta_x + \Upsilon]$$

satisfy that for all $\nu' \in \mathcal{I}_{\mu,\epsilon} \times \mathcal{I}_{\zeta,\epsilon}$, for $\boldsymbol{w}' \in \Phi(\nu')$,

$$\min_{\boldsymbol{w}\in\Phi(\nu)} \max_{a\in[K], x\in\mathcal{X}} \left| \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] w_{a,x}(s) - \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] w'_{a,x} \right| \leq \varepsilon.$$

In particular, whenever $(\hat{\mu}_{a,x}(t), \hat{\zeta}_x)_{x\in\mathcal{X}} \in \mathcal{I}_{\mu,\epsilon} \times \mathcal{I}_{\zeta,\epsilon}$, the empirical best arm is $\hat{a}_t = 1$.

Let $T \in \mathbb{N}$ and define $h(T) := T^{1/4}$ and the event

$$\mathcal{E}_T(\epsilon) = \bigcap_{t=h(T)}^{T} \left((\hat{\mu}_{a,x}(t), \hat{\zeta}_x)_{x\in\mathcal{X}} \in \mathcal{I}_{\mu,\epsilon} \times \mathcal{I}_{\zeta,\epsilon} \right).$$

The following proposition is a consequence of the proposed CTS algorithm, which ensures that each arm is drawn at least of order $\sqrt{t}$ times at round t.

Lemma 29 *There exist two constants B, C (that depend on ν and ϵ) such that*

$$\mathbb{P}_\nu(\mathcal{E}_T^c) \leq BT\exp(-CT^{1/8}).$$

Using these ingredients, we prove Theorem 19.

Proof On the event $\mathcal{E}_T$, it holds for $t \geq h(T)$ that $\hat{a}_t = 1$ and the Chernoff stopping statistic rewrites

$$\begin{aligned}
&\max_{a\in[K]} \min_{b\neq 1} Z_{1,b}(t) = \min_{a\neq 1} Z_{1,a}(t) \\
&= t \min_{a\neq 1} \sum_{x\in\mathcal{X}} \left\{ \frac{N_{1,x}(t)}{t} \left\{ \hat{\mu}_{1,x}(t) \log \frac{\hat{\mu}_{1,x}(t)}{1-\hat{\mu}_{1,x}(t)} + \log(1-\hat{\mu}_{1,x}(t)) \right\} \right. \\
&\qquad + \frac{N_{a,x}(t)}{t} \left\{ \hat{\mu}_{a,x}(t) \log \frac{\hat{\mu}_{a,x}(t)}{1-\hat{\mu}_{a,x}(t)} + \log(1-\hat{\mu}_{a,x}(t)) \right\} \\
&\qquad - \frac{N_{1,x}(t)}{t} \left\{ \hat{\mu}_{1,x}(t) \log \frac{\widetilde{\xi}_{1,x}(t)}{1-\widetilde{\xi}_{1,x}(t)} + \log(1-\widetilde{\xi}_{1,x}(t)) \right\} \\
&\qquad \left. - \frac{N_{a,x}(t)}{t} \left\{ \hat{\mu}_{a,x}(t) \log \frac{\widetilde{\xi}_{a,x}(t)}{1-\widetilde{\xi}_{a,x}(t)} + \log(1-\widetilde{\xi}_{a,x}(t)) \right\} \right\} \\
&= t g\left((\hat{\mu}_{a,x}(t))_{x\in\mathcal{X}}, (\hat{\zeta}_x(t))_{x\in\mathcal{X}}, \left(\frac{N_{a,x}(t)}{t} \right)_{a\in[K], x\in\mathcal{X}} \right),
\end{aligned}$$

where we introduce the function

$$
\begin{aligned}
&g\left((\hat{\mu}_{a,x}(t))_{x\in\mathcal{X}}, (\hat{\zeta}_x(t))_{x\in\mathcal{X}}, \left(\frac{N_{a,x}(t)}{t}\right)_{a\in[K],x\in\mathcal{X}}\right) \\
&= \min_{a\neq 1} \sum_{x\in\mathcal{X}} \left\{ \frac{N_{1,x}(t)}{t}\left\{\hat{\mu}_{1,x}(t)\log\frac{\hat{\mu}_{1,x}(t)}{1-\hat{\mu}_{1,x}(t)} + \log(1-\hat{\mu}_{1,x}(t))\right\}\right. \\
&\qquad + \frac{N_{a,x}(t)}{t}\left\{\hat{\mu}_{a,x}(t)\log\frac{\hat{\mu}_{a,x}(t)}{1-\hat{\mu}_{a,x}(t)} + \log(1-\hat{\mu}_{a,x}(t))\right\} \\
&\qquad - \frac{N_{1,x}(t)}{t}\left\{\hat{\mu}_{1,x}(t)\log\frac{\widetilde{\xi}_{1,x}(t)}{1-\widetilde{\xi}_{1,x}(t)} + \log(1-\widetilde{\xi}_{1,x}(t))\right\} \\
&\qquad \left. - \frac{N_{a,x}(t)}{t}\left\{\hat{\mu}_{a,x}(t)\log\frac{\widetilde{\xi}_{a,x}(t)}{1-\widetilde{\xi}_{a,x}(t)} + \log(1-\widetilde{\xi}_{a,x}(t))\right\}\right\}.
\end{aligned}
$$

From Lemma 27, there exists a constant for T_ϵ such that the following inequality holds on $\mathcal{E}_T$:

$$
\forall t \geq \sqrt{T},\ \min_{\boldsymbol{w}\in\Phi(\nu)} \max_{a\in[K],x\in\mathcal{X}} \left| \frac{N_{a,x}(t)}{t} - \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] w_{a,x} \right| \leq 3(KD-1)\varepsilon.
$$

Then, we introduce

$$
H^*_\epsilon(\nu) = \inf_{\substack{\mu'_{a,x}:|\mu'_{a,x}-\mu_{a,x}|\leq\Upsilon(\epsilon) \\ \zeta'_x:|\zeta'_x-\zeta_x|\leq\Upsilon(\epsilon) \\ w'_{a,x}:|w'_{a,x}-w^*_{a,x}|\leq 3(KD-1)\epsilon}} g(\boldsymbol{\mu}', \boldsymbol{w}')\ ,
$$

where

$$
\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}\in\Phi(\nu)} \max_{a\in[K],x\in\mathcal{X}} \left| \frac{N_{a,x}(t)}{t} - \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}[X_s = x] w_{a,x} \right| \leq 3(KD-1)\varepsilon.
$$

Here, on the event $\mathcal{E}_T$ it holds that for every $t \geq \sqrt{T}$,

$$
\left(\max_{a\in[K]} \min_{b\neq a}\ Z_{a,b}(t) \geq t H^*_\epsilon(\nu) \right)\ .
$$

Let us define$T \geq T_\epsilon$. Then, on the event $\mathcal{E}_T$,

$$
\begin{aligned}
\min(\tau_\delta, T) &\leq \sqrt{T} + \sum_{t=\sqrt{T}}^{T} \mathbb{1}\left[\tau_\delta > t\right] \leq \sqrt{T} + \sum_{t=\sqrt{T}}^{T} \mathbb{1}\left[\max_{a\in[K]} \min_{b\neq a}\ Z_{a,b}(t) \leq \beta(t,\delta)\right] \\
&\leq \sqrt{T} + \sum_{t=\sqrt{T}}^{T} \mathbb{1}\left[t H^*_\epsilon(\nu) \leq \beta(T,\delta)\right] \leq \sqrt{T} + \frac{\beta(T,\delta)}{H^*_\epsilon(\nu)}\ .
\end{aligned}
$$

Introducing

$$
T_0(\delta) = \inf\left\{ T \in \mathbb{N} : \sqrt{T} + \frac{\beta(T,\delta)}{H^*_\epsilon(\nu)} \leq T \right\},
$$

for every $T \geq \max(T_0(\delta), T_\epsilon)$, we have $\mathcal{E}_T \subseteq (\tau_\delta \leq T)$, therefore

$$\mathbb{P}_\nu\left(\tau_\delta > T\right) \leq \mathbb{P}(\mathcal{E}_T^c) \leq BT\exp(-CT^{1/8})$$

and

$$\mathbb{E}_\nu[\tau_\delta] \leq T_0(\delta) + T_\epsilon + \sum_{T=1}^{\infty} BT\exp(-CT^{1/8})\ .$$

We now provide an upper bound on $T_0(\delta)$. Let us define $\eta > 0$ and the constant

$$C(\eta) = \inf\{T \in \mathbb{N} : T - \sqrt{T} \geq T/(1+\eta)\}.$$

Then, we have

$$\begin{aligned} T_0(\delta) &\leq C(\eta) + \inf\left\{T \in \mathbb{N} : \frac{1}{H_\epsilon^*(\nu)}\log\left(\frac{r(T)}{\delta}\right) \leq \frac{T}{1+\eta}\right\} \\ &\leq C(\eta) + \inf\left\{T \in \mathbb{N} : \frac{H_\epsilon^*(\nu)}{1+\eta}T \geq \log\left(\frac{Dt^{1+\alpha}}{\delta}\right)\right\}, \end{aligned}$$

where the constant D is such that $r(T) \leq DT^\alpha$. By using Proposition 28, we obtain, for $\alpha \in [1, e/2]$,

$$T_0(\delta) \leq C(\eta) + \frac{\alpha(1+\eta)}{H_\epsilon^*(\nu)}\left[\log\left(\frac{De(1+\eta)^\alpha}{\delta(H_\epsilon^*(\nu))^\alpha}\right) + \log\log\left(\frac{D(1+\eta)^\alpha}{\delta(H_\epsilon^*(\nu))^\alpha}\right)\right].$$

The last upper bound yields, for every $\eta > 0$ and $\epsilon > 0$,

$$\limsup_{\delta\to 0}\frac{\mathbb{E}_\nu[\tau_\delta]}{\log(1/\delta)} \leq \frac{\alpha(1+\eta)}{H_\epsilon^*(\nu)}.$$

As η and ϵ go to zero, by continuity of g and by definition of w^*,

$$\lim_{\epsilon\to 0} H_\epsilon^*(\nu) = T^\star(\nu)^{-1}.$$

This yields

$$\liminf_{\delta\to 0}\frac{\mathbb{E}_\nu[\tau_\delta]}{\log(1/\delta)} \leq \alpha T^\star(\nu)\ .$$

■

Appendix I. Proofs for Section 6.5

I.1 Proof of Theorem 21

Proof. Let $U = \phi(X)$ and fix any $u \in [M]$. By the law of total probability and the definition of $\mu_{a,u}$,

$$\mathbb{P}(R_{t,a} = 1 \mid U_t = u) = \mathbb{E}\left[\mathbb{P}(R_{t,a} = 1 \mid X_t) \mid U_t = u\right] = \mathbb{E}\left[\mu_a(X_t) \mid U_t = u\right] = \mu_{a,u}.$$

Therefore, conditional on $U_t = u$, the reward of arm a is Bernoulli with mean $\mu_{a,u}$. Moreover, letting $\zeta_u = \mathbb{P}(U_t = u)$,

$$\mu_a = \mathbb{E}[\mu_a(X_t)] = \sum_{u=1}^{M} \mathbb{P}(U_t = u)\ \mathbb{E}\left[\mu_a(X_t) \mid U_t = u\right] = \sum_{u=1}^{M} \zeta_u\, \mu_{a,u}.$$

Hence the best arm defined by maximizing μ_a coincides with the best arm in the aggregated finite-context instance $\nu_\phi = ((\mu_{a,u}), (\zeta_u))$. Finally, Theorem 15 applies to ν_ϕ and guarantees that CTS run on the finite contexts (U_t) is δ-PAC. Since the recommendation is made according to the same marginalized means μ_a, this implies δ-PAC for the original instance.

I.2 Proof of Lemma 22

Proof. Let $\phi : \mathcal{X} \to [M]$ and $\phi' : \mathcal{X} \to [M']$ be aggregations such that ϕ' refines ϕ, i.e., there exists a map $\pi : [M'] \to [M]$ with $\phi = \pi \circ \phi'$. For $u \in [M]$, denote $S_u := \pi^{-1}(u)$.

Write $\zeta_u := \mathbb{P}(\phi(X) = u)$ and $\zeta'_v := \mathbb{P}(\phi'(X) = v)$ for $v \in [M']$. Then $\zeta_u = \sum_{v \in S_u} \zeta'_v$. For any arm $a \in [K]$, recall that

$$\mu_{a,u} := \mathbb{E}[\mu_a(X) \mid \phi(X) = u] \qquad \text{and} \qquad \mu'_{a,v} := \mathbb{E}[\mu_a(X) \mid \phi'(X) = v].$$

By the tower property, for every $u \in [M]$,

$$\mu_{a,u} = \sum_{v \in S_u} \frac{\zeta'_v}{\zeta_u} \mu'_{a,v}.$$

Fix any $\boldsymbol{w} = (w_{a,u})_{a \in [K],\, u \in [M]} \in \mathcal{W}$ for the induced instance ν_ϕ and define $\boldsymbol{w}' = (w'_{a,v})_{a \in [K],\, v \in [M']} \in \mathcal{W}$ for $\nu_{\phi'}$ by $w'_{a,v} := w_{a,\pi(v)}$. (In other words, $\boldsymbol{w}'$ ignores the refinement.)

We show that $m(\boldsymbol{w}', \nu_{\phi'}) \geq m(\boldsymbol{w}, \nu_\phi)$, where $m(\cdot, \cdot)$ is defined in Section 5.2. Fix any $a \neq 1$ and consider an arbitrary feasible pair $(\lambda'_{1,v}, \lambda'_{a,v})_{v \in [M']}$ satisfying the constraint $\sum_{v \in [M']} \zeta'_v \lambda'_{a,v} = \sum_{v \in [M']} \zeta'_v \lambda'_{1,v}$. For each $u \in [M]$ and $b \in \{1, a\}$, define the corresponding coarse parameters by

$$\lambda_{b,u} := \sum_{v \in S_u} \frac{\zeta'_v}{\zeta_u} \lambda'_{b,v}.$$

Then $\sum_{u \in [M]} \zeta_u \lambda_{a,u} = \sum_{u \in [M]} \zeta_u \lambda_{1,u}$, so $(\lambda_{1,u}, \lambda_{a,u})_{u \in [M]}$ is feasible for the inner minimization that defines $m(\boldsymbol{w}, \nu_\phi)$ for arm a.

Moreover, the KL divergence is jointly convex, hence for every $u \in [M]$ and $b \in \{1, a\}$,

$$\sum_{v \in S_u} \frac{\zeta'_v}{\zeta_u} \mathrm{kl}(\mu'_{b,v}, \lambda'_{b,v}) \geq \mathrm{kl}\Big(\sum_{v \in S_u} \frac{\zeta'_v}{\zeta_u} \mu'_{b,v},\ \sum_{v \in S_u} \frac{\zeta'_v}{\zeta_u} \lambda'_{b,v} \Big) = \mathrm{kl}(\mu_{b,u}, \lambda_{b,u}).$$

Multiplying by $\zeta_u w_{b,u}$ and summing over u yields

$$\sum_{v \in [M']} \zeta'_v w'_{b,v} \mathrm{kl}(\mu'_{b,v}, \lambda'_{b,v}) = \sum_{u \in [M]} \sum_{v \in S_u} \zeta'_v w_{b,u} \mathrm{kl}(\mu'_{b,v}, \lambda'_{b,v}) \geq \sum_{u \in [M]} \zeta_u w_{b,u} \mathrm{kl}(\mu_{b,u}, \lambda_{b,u}).$$

Adding the inequalities for $b = 1$ and $b = a$ shows that, for this fixed a and this feasible λ', the objective value in the inner minimization defining $m(\boldsymbol{w}', \nu_{\phi'})$ is at least the corresponding objective value for ν_ϕ with the aggregated parameters λ. Since the inequality holds for every feasible λ', taking the minimum over λ' yields that the inner minimum for $\nu_{\phi'}$ (for arm a) is at least that for ν_ϕ (for the same arm a). Finally, taking the minimum over $a \neq 1$ gives $m(\boldsymbol{w}', \nu_{\phi'}) \geq m(\boldsymbol{w}, \nu_\phi)$.

Since for every $\boldsymbol{w} \in \mathcal{W}$ we have constructed $\boldsymbol{w}' \in \mathcal{W}$ satisfying $m(\boldsymbol{w}', \nu_{\phi'}) \geq m(\boldsymbol{w}, \nu_\phi)$, we obtain $\max_{\boldsymbol{w}' \in \mathcal{W}} m(\boldsymbol{w}', \nu_{\phi'}) \geq \max_{\boldsymbol{w} \in \mathcal{W}} m(\boldsymbol{w}, \nu_\phi)$. Recalling that $T^\star(\nu)^{-1} = \max_{\boldsymbol{w} \in \mathcal{W}} m(\boldsymbol{w}, \nu)$, we conclude that $T^\star(\nu_{\phi'}) \leq T^\star(\nu_\phi)$, as claimed.

I.3 Proof of Proposition 23.

Let Π be a finite measurable partition of $\mathcal{X}$, and define the piecewise-constant mean functions

$$\mu_a^\Pi(x) := \mathbb{E}\big[\mu_a(X) \mid X \in B\big], \qquad x \in B \in \Pi.$$

Denote by $\mathcal{V}^\Pi$ the instance with the same context distribution ζ and Bernoulli rewards with conditional means $(\mu_a^\Pi)_{a \in [K]}$. By construction, $\mathcal{V}^\Pi$ is equivalent to the aggregated finite-context instance induced by Π (since, within each bin B, the reward distribution does not depend on the exact context $x \in B$), and therefore $T^\star(\mathcal{V}^\Pi) = T^\star(\widetilde{\mathcal{V}})$, where $\widetilde{\mathcal{V}}$ is the induced instance in Theorem 21.

Let $\omega_a(r) := \sup\{|\mu_a(x) - \mu_a(y)| : x, y \in \mathcal{X}, |x - y| \leq r\}$ be the modulus of continuity of μ_a on the compact set $\mathcal{X}$. Since each μ_a is uniformly continuous, $\omega_a(r) \to 0$ as $r \downarrow 0$. For any $x \in B \in \Pi$, we have

$$|\mu_a^\Pi(x) - \mu_a(x)| = \Big|\mathbb{E}\big[\mu_a(X) - \mu_a(x) \mid X \in B\big]\Big| \leq \sup_{y \in B} |\mu_a(y) - \mu_a(x)| \leq \omega_a\big(\operatorname{diam}(B)\big).$$

Consequently, letting

$$\delta(\Pi) := \max_{a \in [K]} \sup_{x \in \mathcal{X}} |\mu_a^\Pi(x) - \mu_a(x)|,$$

we obtain $\delta(\Pi) \leq \max_{a \in [K]} \omega_a\big(\max_{B \in \Pi} \operatorname{diam}(B)\big)$. For the nested sequence $(\Pi_m)_{m \geq 1}$ in Proposition 23, define $\delta_m := \delta(\Pi_m)$; by assumption, $\delta_m \to 0$.

Let $a^\star = a^\star(\mathcal{V})$ and denote the marginal means by $\mu_a := \int_{\mathcal{X}} \mu_a(x)\, \zeta(dx)$ and $\mu_a^{(m)} := \int_{\mathcal{X}} \mu_a^{\Pi_m}(x)\, \zeta(dx)$. Then $|\mu_a^{(m)} - \mu_a| \leq \delta_m$ for all $a \in [K]$. Since $a^\star(\mathcal{V})$ is unique, letting $\Delta := \mu_{a^\star} - \max_{b \neq a^\star} \mu_b > 0$, we have $a^\star(\mathcal{V}^{\Pi_m}) = a^\star$ for all sufficiently large m such that $\delta_m \leq \Delta/4$. For such m, the alternative set depends on $\mathcal{V}$ only through $a^\star$ (see the definition of $\mathrm{Alt}(\cdot)$ in Theorem 2), hence $\mathrm{Alt}(\mathcal{V}^{\Pi_m}) = \mathrm{Alt}(\mathcal{V})$.

Define the information rate $I(\cdot) := 1/T^\star(\cdot)$ as in Theorem 2:

$$I(\mathcal{U}) := \sup_{\boldsymbol{w} \in \mathcal{W}} \inf_{(\boldsymbol{q}, \zeta) \in \mathrm{Alt}(\mathcal{U})} \sum_{a=1}^{K} \int_{\mathcal{X}} w_{a,x}\, \mathrm{KL}(p_{a,x}, q_{a,x})\, \zeta(dx).$$

Under our restriction that both the true means and the alternative means lie in $[\eta, 1-\eta]$, we have $\mathrm{KL}(p_{a,x}, q_{a,x}) = d(\mu_a(x), \nu_a(x))$ where $d(\cdot, \cdot)$ is the Bernoulli KL divergence. Moreover, $d(p, q)$ is continuously differentiable on the compact set $[\eta, 1-\eta]^2$, and thus

$$L_\eta := \sup_{p, q \in [\eta, 1-\eta]} \left|\tfrac{\partial}{\partial p} d(p, q)\right| < \infty.$$

It follows that $|d(p,q) - d(p',q)| \le L_\eta |p - p'|$ for all $p, p', q \in [\eta, 1-\eta]$.

Fix $\boldsymbol{w} \in \mathcal{W}$ and $(\boldsymbol{q}, \zeta) \in \mathrm{Alt}(\mathcal{V})$. For all sufficiently large m (so that $\mathrm{Alt}(\mathcal{V}^{\Pi_m}) = \mathrm{Alt}(\mathcal{V})$), using $|\mu_a^{\Pi_m}(x) - \mu_a(x)| \le \delta_m$ and $\sum_{a=1}^K w_{a,x} = 1$, we obtain

$$\left| \sum_{a=1}^K \int w_{a,x} \, \mathrm{KL}(p_{a,x}^{\Pi_m}, q_{a,x}) \, \zeta(dx) - \sum_{a=1}^K \int w_{a,x} \, \mathrm{KL}(p_{a,x}, q_{a,x}) \, \zeta(dx) \right| \le K L_\eta \delta_m,$$

where $p_{a,x}^{\Pi_m} = \mathrm{Bern}(\mu_a^{\Pi_m}(x))$ and $p_{a,x} = \mathrm{Bern}(\mu_a(x))$. Taking the infimum over $(\boldsymbol{q}, \zeta) \in \mathrm{Alt}(\mathcal{V})$ and then the supremum over $\boldsymbol{w} \in \mathcal{W}$ yields

$$|I(\mathcal{V}^{\Pi_m}) - I(\mathcal{V})| \le K L_\eta \delta_m \xrightarrow[m \to \infty]{} 0.$$

Hence $T^\star(\mathcal{V}^{\Pi_m}) = 1/I(\mathcal{V}^{\Pi_m}) \to 1/I(\mathcal{V}) = T^\star(\mathcal{V})$, and recalling $T^\star(\mathcal{V}^{\Pi_m}) = T^\star(\widetilde{\mathcal{V}}^{(m)})$ proves the first claim of Proposition 23.

Finally, fix $\varepsilon > 0$ and choose m_ε such that $T^\star(\widetilde{\mathcal{V}}^{(m_\varepsilon)}) \le T^\star(\mathcal{V}) + \varepsilon$. By Theorem 19 applied to the fixed finite-context instance $\widetilde{\mathcal{V}}^{(m_\varepsilon)}$, we have

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\mathcal{V}}[\tau_\delta]}{\log(1/\delta)} = \limsup_{\delta \to 0} \frac{\mathbb{E}_{\widetilde{\mathcal{V}}^{(m_\varepsilon)}}[\tau_\delta]}{\log(1/\delta)} \le T^\star(\widetilde{\mathcal{V}}^{(m_\varepsilon)}) \le T^\star(\mathcal{V}) + \varepsilon,$$

where we used that the interaction induced by CTS depends on the original contexts only through their bin indices, so its distribution under $\mathcal{V}$ coincides with that under $\widetilde{\mathcal{V}}^{(m_\varepsilon)}$.

Appendix J. Details of Experiments

J.1 Calculation of an Optimal Weight

To update the allocation $w(t)$, we need to solve the minimax optimization problem defined in (3). Unlike Garivier and Kaufmann (2016), we do not have an analytical solution for this problem. Therefore, we solve this problem numerically, using sequential quadratic programming. In our experiments, we use the sequential least squares programming (SLSQP) algorithm implemented in the `optimize.minimize` method of scipy, which is a Python library. Note that Garivier and Kaufmann (2016) used only the bisection method for the numerical optimization with the help of the analytical solution of the inner optimization in $L_{a,b}(\cdot)$. Unlike Garivier and Kaufmann (2016), in our case, errors of optimization affect the results more.

J.2 Environment of Experiments

All experiments were conducted on a MacBook Pro with a 2.8 GHz quad-core Intel Core i7. We use Python. The version of Python is 3.7.5, and the version of SciPy is 1.4.1. To reduce the computational load, $\boldsymbol{w}$ is updated once every 10 trials. This is an asymptotically negligible heuristic.

J.3 Experimental Settings and Additional Results with Bernoulli bandit models

In all experiments with Bernoulli bandit models, we assume that there exist two contexts $X_t \in \{1, 2\}$ and each context is drawn with probability 0.5.

Table 2: Additional Bernoulli simulations (known-allocation baseline). For each instance, we report the mean and median stopping time (with interquartile range, IQR). In all settings below, the empirical error count was 0 out of the $n_{\text{runs}} = 10$ trials ($\delta = 0.05$).

Instance	K	$\lvert\mathcal{X}\rvert$	TS (known allocation)	CTS (known allocation)	Error
I1 (Fig. 3)	4	2	21301.0 (20302; 17967–23640)	9292.2 (9700; 8411–10198)	0/10
I2 (heterogeneous)	4	2	471.9 (544; 333–589)	329.7 (338; 242–399)	0/10

We conduct three additional experiments with different settings from the one in Section 7. For the Bernoulli bandit model, we consider a situation where the marginalized mean rewards are $\{\mu_1, \mu_2, \mu_3, \mu_4\} = \{0.5, 0.45, 0.43, 0.4\}$, which is the same as one of the scenarios used in Garivier and Kaufmann (2016). Suppose that for each context, the conditional mean rewards are given as $\{\mu_{1,1}, \mu_{2,1}, \mu_{3,1}, \mu_{4,1}\} = \{0.5, 0.01, 0.4, 0.01\}$ and $\{\mu_{1,2}, \mu_{2,2}, \mu_{3,2}, \mu_{4,2}\} = \{0.5, 0.89, 0.46, 0.79\}$. We show the evolutions of the GLRT statistic in Figure 4. As well as the result shown in Section 7, the CTS algorithm achieves a smaller sample complexity than TS. However, the variance is larger than the case discussed in Section 7. We believe that this is due to the gaps between the mean rewards are smaller than in the previous case and to the errors of the estimation/optimization affect the results more.

Next, we consider another scenario:

$$\{\mu_1, \mu_2, \mu_3, \mu_4\} = \{0.3, 0.21, 0.2, 0.19\}$$
$$\{\mu_1, \mu_2, \mu_3, \mu_4\} = \{0.5, 0.45, 0.43, 0.4\},$$

which are the same as Garivier and Kaufmann (2016) and our previous experiments. For each setting, we use the same conditional mean rewards as $\{\mu_{1,1}, \mu_{2,1}, \mu_{3,1}, \mu_{4,1}\} = \{0.5, 0.2, 0.2, 0.1\}$. The counterparts and $\{\mu_{1,2}, \mu_{2,2}, \mu_{3,2}, \mu_{4,2}\}$ for

$$\{\mu_1, \mu_2, \mu_3, \mu_4\} = \{0.3, 0.21, 0.2, 0.19\}$$

and

$$\{\mu_1, \mu_2, \mu_3, \mu_4\} = \{0.5, 0.45, 0.43, 0.4\}$$

are $\{\mu_{1,2}, \mu_{2,2}, \mu_{3,2}, \mu_{4,2}\} = \{0.1, 0.22, 0.2, 0.28\}$ and $\{\mu_{1,2}, \mu_{2,2}, \mu_{3,2}, \mu_{4,2}\} = \{0.5, 0.7, 0.66, 0.7\}$, respectively. Compared to these cases, the previous experiments take more extreme values of the conditional mean rewards. Therefore, in the current setting, we expect the difference between the results of track-and-stop and contextual track-and-stop to be less than in the previous ones. We show the value of the GLRT statistic in Figure 5. As we expect, improvement is limited in this case.

Additional Bernoulli instances. To complement Figure 3, Table 2 reports stopping times on two Bernoulli instances with $K = 4$ arms and $|\mathcal{X}| = 2$ contexts. We report the mean stopping time, the median, and the interquartile range (IQR) over 10 independent runs at $\delta = 0.05$. For this table we used fixed sampling proportions $w^\star$ (computed from the true instance) for both TS and CTS to isolate the gain from exploiting contextual information.

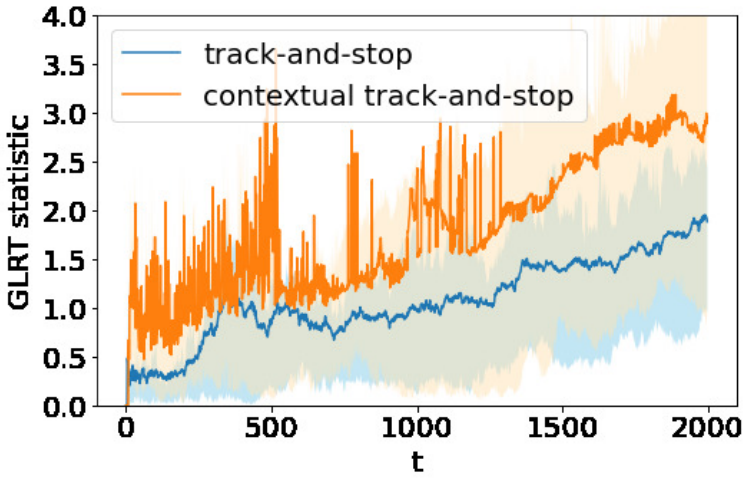

Figure 4: This graph illustrates the maximum GLRT statistic $\max_{a\in[K]} \min_{b\in[K]\setminus\{a\}} Z_{a,b}(t)$. The solid line represents the averaged value over 20 trials, and the light-colored area shows the values between the first and third quartiles.

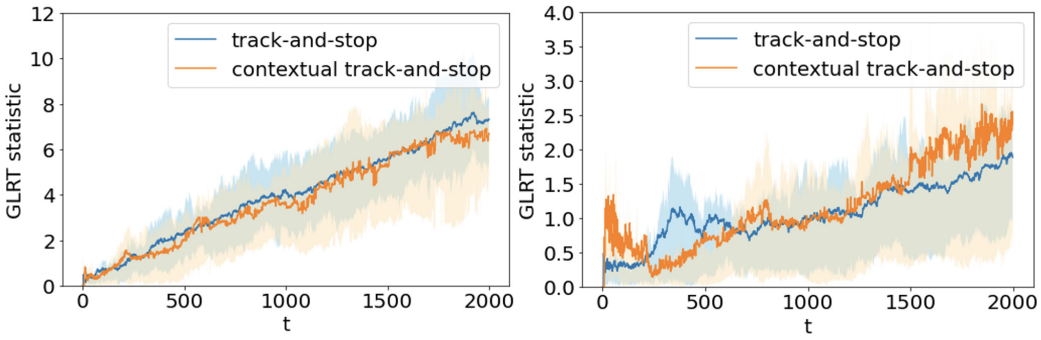

Figure 5: This graph illustrates the maximum GLRT statistic $\max_{a\in[K]} \min_{b\in[K]\setminus\{a\}} Z_{a,b}(t)$. The left figure shows when $\{\mu_1, \mu_2, \mu_3, \mu_4\} = \{0.3, 0.21, 0.2, 0.19\}$ is given. The right figure shows the results when $\{\mu_1, \mu_2, \mu_3, \mu_4\} = \{0.5, 0.45, 0.43, 0.4\}$ is given. The solid line represents the averaged value over 20 trials, and the light-colored area shows the values between the first and third quartiles.

We also ran a light-weight plug-in version that recomputes the sampling proportions infrequently (every 1000 rounds for TS and every 2000 rounds for CTS). On instance I1 this yielded mean stopping times 3334 (TS) and 2693 (CTS) over 10 runs, suggesting that the contextual gain persists when the sampling proportions are estimated online.

J.4 Additional Results for Plug-in α-elimination with Gaussian Rewards

A plug-in variant with unknown correlations and variances. The analysis of Section 3 assumes that the correlations and variances needed to form the control-variate estimator are known. In practice, these nuisance parameters can be estimated online from the

Table 3: Gaussian plug-in α-elimination. For each setting, we report the mean and median stopping time (with interquartile range, IQR). Empirical error rates over $n_{\text{runs}} = 300$ trials ($\delta = 0.05$) were 0% (known-nuisance baseline) and 3.3%, 2.7%, 3.3% (plug-in) for (G1, G2, G3), respectively.

Setting	Known-nuisance baseline mean (median; IQR)	Plug-in mean (median; IQR)
G1	39 (37; 25–52)	24 (12; 6–34)
G2	14 (13; 8–17)	10 (7; 6–11)
G3	64 (60; 44–83)	34 (17; 7–55)

observed pairs (X_t, R_{t,A_t}), for example by least-squares regression of $R_{t,a}$ on X_t using only the data collected when arm a is pulled. To assess the resulting behavior, we implemented a plug-in variant that (i) estimates the control-variate coefficient for each arm by least squares, (ii) estimates the residual variance $\sigma_a'^2$ by the sample variance of the residuals, and (iii) uses the estimated residual standard deviations to form $\hat{\alpha}_t$ and a D-tracking style sampling rule targeting the proportion $\hat{\alpha}_t$.

Gaussian settings (G1–G3). We consider a scalar context $X_t \sim \mathcal{N}(0,1)$ and two arms with rewards

$$R_{t,a} = \mu_a + \rho_a X_t + \sqrt{1-\rho_a^2}\,\varepsilon_{t,a}, \qquad \varepsilon_{t,a} \sim \mathcal{N}(0,1) \text{ i.i.d.}$$

so that $\mathrm{Var}(R_{t,a}) = 1$, the optimal control-variate coefficient equals $\beta_a^\star = \rho_a$, and the residual variance after control variates is $(\sigma_a')^2 = 1 - \rho_a^2$. We set $(\mu_1, \mu_2) = (1, 0)$ and use G1: $(\rho_1, \rho_2) = (0.9, 0.1)$, G2: $(0.9, 0.9)$, G3: $(0.5, -0.5)$. In the plug-in implementation, we avoid the ill-posed regression regime with $N_{t,a} \le 2$ (0 degrees of freedom for estimating $(\sigma_a')^2$ with an intercept+slope regression) by using a conservative baseline variance estimate and by starting to apply the stopping rule only once each arm has at least three samples.

Note that the plug-in variant can stop earlier than the known-nuisance baseline in these simulations. This is not a contradiction: the "known-nuisance baseline" column refers to the same α-elimination procedure with nuisance parameters known (not an information-theoretic lower bound), and the theoretical exploration rate $\beta(t, \delta)$ is conservative. Moreover, the plug-in widths rely on estimated nuisance quantities and do not explicitly account for the additional uncertainty from estimating the regression coefficient. Hence, the plug-in can behave more aggressively and stop earlier, at the cost of a slightly higher empirical error (which is still close to the target level δ in Table 3).

J.5 Discussion and Future Work: Towards δ-PAC Plug-in Variants

The plug-in variants above are practically motivated heuristics, and we do not claim that they satisfy the δ-PAC guarantee. A natural direction to endow such plug-in methods with guarantees is to replace point estimates of nuisance quantities by *anytime-valid confidence sequences* and to inflate the elimination thresholds to hold uniformly over these confidence sets. For example, in the Gaussian setting one can run online regression to estimate the

nuisance parameters and combine (i) a time-uniform confidence ellipsoid for regression coefficients (via self-normalized martingale bounds) with (ii) an anytime confidence sequence for the noise variance (e.g., empirical-Bernstein or t-type bounds for unknown variance). Elimination can then be performed using confidence widths computed for the worst case over the resulting confidence sets, yielding a principled (albeit conservative) δ-PAC plug-in procedure. Extending this idea beyond parametric regression—e.g., to kernel regression / kernel density plug-in approaches—would require time-uniform confidence bands for the estimated conditional mean (and possibly the context density), which is an interesting direction for future work.