Asymptotically Optimal Algorithms for Multiple Play Bandits with Partial Feedback

Alexander R. Luedtke, Emilie Kaufmann, Antoine Chambaz

▶ **To cite this version:**

Alexander R. Luedtke, Emilie Kaufmann, Antoine Chambaz. Asymptotically Optimal Algorithms for Multiple Play Bandits with Partial Feedback. 2016. hal-01338733v1

HAL Id: hal-01338733

https://hal.science/hal-01338733v1

Preprint submitted on 29 Jun 2016 (v1), last revised 3 Sep 2019 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Asymptotically Optimal Algorithms for Multiple Play Bandits with Partial Feedback

Alexander R. Luedtke[1,2], Emilie Kaufmann[3], and Antoine Chambaz[2,4,5]

[1]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
[2]Division of Biostatistics, University of California, Berkeley, Berkeley, CA, USA
[3]CNRS & CRIStAL, Université de Lille, Cité Scientifique, 59655 Villeneuve d'Ascq, France
[4]MAP5, UMR CNRS 8145, Université Paris Descartes, France
[5]Modal'X, Université Paris Ouest Nanterre, France

June 29, 2016

Abstract

We study a variant of the multi-armed bandit problem with multiple plays in which the user wishes to sample the m out of k arms with the highest expected rewards, but at any given time can only sample $\ell \leq m$ arms. When $\ell = m$, Thompson sampling was recently shown to be asymptotically efficient. We derive an asymptotic regret lower bound for any uniformly efficient algorithm in our new setting where ℓ may be less than m. We then establish the asymptotic optimality of Thompson sampling for Bernoulli rewards, where our proof technique differs from earlier methods even when $\ell = m$. We also prove the asymptotic optimality of an algorithm based on upper confidence bounds, KL-CUCB, for single-parameter exponential families and bounded, finitely supported rewards, a result which is new for all values of ℓ.

1 Introduction

In the classical multi-armed bandit problem, a user is repeatedly confronted with a set of k arms and must select one of the available arms to pull based on their knowledge from previous rounds of the game. Each arm presents the user with a reward drawn from some distribution, and the user's objective is to maximize the expected sum of their rewards over time or, equivalently, minimize the total regret (the expected reward of pulling the optimal arm at every time step minus the expected sum of the rewards corresponding to their selected

actions). To play the game well, the user must balance the need to gather new information about the reward distribution of each arm (exploration) with the need to take advantage of the information that they already have by pulling the arm for which they believe the reward will be the highest (exploitation).

While the general setup of the bandit problem is timeless, the problem first started receiving rigorous mathematical attention slightly under a century ago (Thompson, 1933). This early work focused on Bernoulli rewards, that are relevant in the simplest modeling of a sequential clinical trial, and presented a Bayesian algorithm now known as Thompson sampling. Since that time, many authors have contributed to a deeper understanding of the multi-armed bandit problem, both with Bernoulli and other reward distributions and either from a Bayesian (Gittins, 1979) or frequentist perspective (Robbins, 1952). Lai and Robbins (1985) established a lower bound on the (frequentist) regret of any algorithm which satisfies a general uniform efficiency condition. This lower bound provides a concise definition of asymptotic (regret) optimality for an algorithm: an algorithm is asymptotically optimal when it achieves this lower bound. Lai and Robbins also introduced what are known as upper confidence bound (UCB) procedures for deciding which arm to pull at a given time step. In short, these procedures compute a UCB for the expected reward of each arm at each time and pull the arm with the highest UCB. Many variants of UCB algorithms have been proposed since then (see the Introduction to Cappé et al., 2013a for a thorough review), with more explicit indices and/or finite-time regret guarantees. Among them the KL-UCB algorithm Cappé et al. (2013a) is proved to be asymptotically optimal for rewards that belong to a one-parameter exponential family and finitely-supported rewards. Meanwhile, there has been a recent interest for the theoretical understanding of the previously discussed Thompson sampling algorithm, whose first regret bound was obtained by Agrawal and Goyal (2011). Since then, Thompson Sampling has been proved to be asymptotically optimal for Bernoulli rewards (Kaufmann et al., 2012b; Agrawal and Goyal, 2012) and for reward distributions belonging to univariate exponential families (Korda et al., 2013).

There has recently been a surge of interest in the multi-armed bandit problem, due to its applications to (online) sequential content recommendation. In this context each arm models the feedback of a user to a specific item that can be displayed (e.g. an advertisement). In this framework, it might be relevant to display several items at a time, and some variants

of the classic bandit problems that have been proposed in the literature may be considered. In the *multi-armed bandit with multiple plays*, $m \geq 1$ out of k arms are sampled at each round and all the associated rewards are observed by the agent, who receives their sum. Anantharam et al. (1987) present a regret lower bound for this problem, together with a (non-explicit) matching strategy. More explicit strategies can be obtained when viewing this problem as a particular instance of a *combinatorial bandit problem with semi-bandit feedback*. Combinatorial bandits, originally introduced by Cesa-Bianchi and Lugosi (2012) in a non-stochastic setting, present the user with possibly structured subsets of arms at each round: once a subset is chosen, the agent receives the sum of their rewards. The semi-bandit feedback corresponds to the case when the user is able to see the reward of each of the sampled arms (Audibert et al., 2011). Two extensions of UCB procedures to this semi-bandit combinatorial setting have been proposed. Chen et al. (2013) introduce the CUCB algorithm, while Combes et al. (2015b) propose an adaption of the KL-UCB algorithm called efficient sampling for combinatorial bandits (ESCB). Both of these algorithms can be used in the multiple play setting, and enjoy rate-optimal but constant-suboptimal regret upper bounds for simple distributions (e.g. Bernoulli). Recently, Komiyama et al. (2015) prove the optimality of Thompson sampling for multiple play bandits with Bernoulli rewards in the case where the arm with the m^{th} largest mean is unique. An important consequence of the uniqueness of the m^{th} largest mean is that the optimal set of m arms is necessarily unique, which may not be plausible in practice.

In this paper, we consider a generalized version of multi-armed bandit problem with multiple plays that incorporates some *random partial feedback*. While we still assume that the user only wishes to sample from the m out of k arms with the largest mean reward, we allow the situation where they are forced to sample only $\ell < m$ arms at each time step, chosen uniformly at random among the subsets of size l of the m chosen arms. This setting may be relevant in the context of online advertising: a website will display ℓ advertisements on a given page, but these ℓ ads will be selected from a total of m relevant ads so that clients experience variety on their site. When $\ell = m$, we also analyze the corresponding combinatorial bandit problem with semi-bandit feedback.

Our contributions are the following:

- We generalize two existing algorithms for multiple-play bandits ($\ell = m$) to the case

where ℓ can be less than m. Among other contributions to this new problem, we generalize the lower bound in Anantharam et al. (1987) to this setting, under general assumptions on the reward distributions.

- We provide a novel technique for proving asymptotic optimality that leverages the asymptotic lower bound on the number of draws of any suboptimal arm. While this lower bound on suboptimal arm draws is typically used to prove an asymptotic lower bound on the regret of any reasonable algorithm, we use it as a key ingredient for our proof of an asymptotically optimal *upper bound* on the regret of KL-CUCB and Thompson sampling, i.e. to prove the asymptotic optimality of these two algorithms. The optimality result is new for the proposed KL-CUCB algorithm. The result is new for Thompson sampling when $\ell < m$, and when $\ell = m$ the proof technique is distinct from that of Komiyama et al. (2015).

- We do not require that the set of optimal arms be unique.

This article is organized as follows. Section 2 outlines our problem of interest. Section 3 provides an asymptotic lower bound on the number of suboptimal arm draws and on the regret. Section 4 presents the two sampling algorithms we consider in this paper and theorems establishing their asymptotic optimality: KL-CUCB (Section 4.1) and Thompson sampling (Section 4.2). Section 5 presents numerical experiments supporting our theoretical findings. Section 6 presents the proofs of our asymptotic optimality results for KL-CUCB and Thompson Sampling. Section 7 gives concluding remarks. Technical proofs are postponed to the appendix. Appendix A contains proofs establishing the asymptotic lower bound on the number of suboptimal arm draws and regret. Appendices B and C contain technical proofs for KL-CUCB and the Thompson sampling, respectively.

2 Bandit Problem

We now formally present our problem of interest. We closely follow the notation presented in Section 2 of Cappé et al. (2013a).

Consider the bandit problem with finitely many arms $a \in \{1, \ldots, k\}$, where each arm has real-valued reward distribution ν_a whose mean we denote by both μ_a and $E(\nu_a)$. We denote

the (possibly nonparametric) model of each ν_a by $\mathcal{D}$, and we use $\mathcal{V}$ to denote $(\nu_1, \ldots, \nu_k)$, where $\mathcal{V}$ has model $\mathcal{D}^k$. Throughout we assume without loss of generality that $\mu_1 \geq \ldots \geq \mu_k$. We also assume that arm m does not fall on the upper boundary of the parameter space, i.e. $\mu_m < \sup\{E(\nu) : \nu \in \mathcal{D}\}$. Let $\{(Y_1(t), \ldots, Y_k(t))\}_{t=1}^{\infty}$ denote an i.i.d. sample from the product distribution $\prod_{a=1}^{k} \nu_a$. At round t, the user will see ℓ action-reward pairs $(A, Y_A(t))$, where $1 \leq \ell < k$ is a fixed integer. We denote the set of actions seen at time t by $\mathcal{A}(t)$, and we emphasize that the user is aware that reward $Y_a(t)$ corresponds to the action $a \in \mathcal{A}(t)$. Following each time $t \geq 0$, the user defines a set $\widehat{\mathcal{S}}(t)$ of m arms, where $\ell \leq m < k$ is a fixed integer. The set $\widehat{\mathcal{S}}(t)$ may be based on the sigma-field $\mathcal{F}(t)$ generated by all action-reward pairs seen at times $1, \ldots, t$, and possibly also some exogenous stochastic mechanism as is typical in Thompson sampling. The user then gets to observe a subset $\mathcal{A}(t+1)$ uniformly selected among the collection of ℓ-cardinality subsets of $\widehat{\mathcal{S}}(t)$. He observes the rewards $Y_a(t+1)$ for each $a \in \mathcal{A}(t+1)$ and receives the reward $R_{t+1} = \sum_{a \in \mathcal{A}(t+1)} Y_a(t+1)$.

The agent aims at maximizing the expected sum of his rewards, $\mathbb{E}[\sum_{t=1}^{T} R_t]$, which an oracle strategy would achieve by always choosing the (possibly non-unique) optimal subset $\mathcal{S}^\star = \{1, \ldots, m\}$, as a subset with the m largest expected rewards. Letting

$$\mu^\star \equiv \frac{1}{m} \sum_{a^\star \in \mathcal{S}^\star} \mu_{a^\star},$$

the agent equivalently aims at finding a strategy minimizing the expected regret at time T, defined as

$$R(T) \equiv \mathbb{E}\left[\ell T \mu^* - \sum_{t=1}^{T} \sum_{a \in \mathcal{A}(t)} Y_a(t)\right].$$

Additionally to fixing $\mathcal{S}^\star$, we split the arms into three non-overlapping subsets:

$$\begin{aligned} \text{optimal arms away from the margin:} \quad & \mathcal{L} \equiv \{a : \mu_a > \mu_m\}, \\ \text{arms on the margin:} \quad & \mathcal{M} \equiv \{a : \mu_a = \mu_m\}, \\ \text{suboptimal arms away from the margin:} \quad & \mathcal{N} \equiv \{a : \mu_a < \mu_m\}. \end{aligned}$$

Note that $\mathcal{S}^\star \subseteq \mathcal{L} \cup \mathcal{M}$ is non-unique precisely when $\mathcal{L} \cup \mathcal{M}$ is of cardinality greater than m.

At any time T, the total number of pulls of arm a is given by $N_a(T) \equiv \sum_{t=1}^{T} \mathbb{1}\{a \in \mathcal{A}(t)\}$.

By the arm sampling procedure, for each arm a we have that

$$\mathbb{E}[N_a(T)] = \mathbb{E}\left[\sum_{t=0}^{T-1} \mathbb{P}\left\{a \in \mathcal{A}(t+1) | \mathcal{F}_t\right\}\right] \leq \frac{\ell T}{m}. \tag{1}$$

The arm sampling procedure also yields that

$$\begin{aligned}
\mathbb{P}\left(a^\star \in \mathcal{A}(t+1) | \mathcal{F}(t)\right) &= \mathbb{P}\left(a^\star \in \mathcal{A}(t+1), a^\star \in \widehat{\mathcal{S}}(t) \middle| \mathcal{F}(t)\right) \\
&= \mathbb{P}\left(a^\star \in \mathcal{A}(t+1) \middle| a^\star \in \widehat{\mathcal{S}}(t), \mathcal{F}(t)\right) \mathbb{P}\left(a^\star \in \widehat{\mathcal{S}}(t) \middle| \mathcal{F}(t)\right) \\
&= \frac{\ell}{m} \mathbb{P}\left(a^\star \in \widehat{\mathcal{S}}(t) \middle| \mathcal{F}(t)\right).
\end{aligned} \tag{2}$$

We also define the average gap between each arm a and $\mathcal{S}^\star$ as $\Delta_a \equiv \mu^\star - \mu_a = \frac{1}{m}\sum_{a^\star \in \mathcal{S}^\star}(\mu_{a^\star} - \mu_a)$. For arms $a > m$, $\Delta_a \geq 0$. The sum of the gap between optimal arms and $\mathcal{S}^\star$ is zero: $\sum_{a^\star \in \mathcal{S}^\star} \Delta_{a^\star} = 0$. Using this notation, the regret can be rewritten

$$\begin{aligned}
R(T) &= \mathbb{E}\left[\ell T \mu^* - \sum_{t=1}^{T} \sum_{a \in \mathcal{A}(t)} Y_a(t)\right] = \mathbb{E}\left[\ell T \mu^* - \sum_{a=1}^{k} \sum_{t=1}^{T} \mu_a \mathbb{1}\left\{a \in \mathcal{A}(t)\right\}\right] \\
&= \sum_{a=1}^{k} \Delta_a \, \mathbb{E}[N_a(T)].
\end{aligned} \tag{3}$$

It can be checked that the regret is indeed positive through the following argument. As $\mathbb{E}[N_a(T)] \leq \ell T/m$, $\sum_{a=1}^{k} N_a(T) = \ell T$, $R(T)$ is lower bounded by the solution to

$$\text{Minimize } \sum_{a=1}^{k} \Delta_a n_a \text{ subject to } \sum_{a=1}^{k} n_a = \ell T \text{ and } n_a \in \left[0, \frac{\ell T}{m}\right] \text{ for all } a.$$

As $\Delta_1 \leq \ldots \leq \Delta_k$, the above optimization resolves to $\frac{\ell T}{m}\sum_{a^\star \in \mathcal{S}^\star} \Delta_{a^\star} = 0$. Thus $R(T) \geq 0$.

For each arm a and natural number n, define $\tau_{a,n} = \min\{t \geq 1 : N_a(t) = n\}$ to be the (stopping) time at which the n^{th} draw of arm a occurs. Let $X_{a,n} \equiv Y_a(\tau_{a,n})$ denote the n^{th} draw from ν_a. One can show that $\{X_{a,n}\}_{n=1}^{\infty}$ is an i.i.d. sequence of draws from ν_a for each a, and that these sequences are independent for two arms $a \neq a'$[1]. We will denote the

[1]It is *a priori* possible that $\tau_{a,n} = \infty$ for all n large enough (though, as we show in Section 3, not possible for any reasonable algorithm). To deal with this case, let $X_{a,n} \equiv Y_a(\tau_{a,n})$ denote the n^{th} draws from ν_a for all $\tau_{a,n} < \infty$ and let $\{X_{a,n}\}_{n:\tau_{a,n}=\infty}$ denote an i.i.d. sequence independent of $\{X_{a,n}\}_{n:\tau_{a,n}<\infty}$.

empirical distribution function of observations drawn from arm a by any time T by

$$\hat{\nu}_a(T) \equiv \frac{1}{N_a(T)} \sum_{t=1}^{T} \delta_{Y_a(t)} \mathbb{1}\{a \in \mathcal{A}(t)\} = \frac{1}{N_a(T)} \sum_{n=1}^{N_a(T)} \delta_{X_{a,n}}.$$

We similarly define $\hat{\nu}_{a,n}$ to be the empirical distribution function of observations $X_{a,1}, \ldots, X_{a,n}$. Thus, $\hat{\nu}_a(t) = \hat{\nu}_{a,N_a(t)}$. We further define $\hat{\mu}_a(t)$ to be the empirical mean of observations drawn from arm a by time t and $\hat{\mu}_{a,N_a(t)} = \hat{\mu}_a(t)$.

We let $\mathrm{KL}(\nu, \nu')$ denote the KL-divergence between distributions ν and ν'. If ν and ν' are uniquely parameterized by their respective means μ and μ' as in a single parameter exponential family (e.g. Bernoulli distributions), then we abuse notation and let $\mathrm{KL}(\mu, \mu') \equiv \mathrm{KL}(\nu, \nu')$. For real μ and a distribution $\nu \in \mathcal{D}$, we define

$$\mathcal{K}_{\inf}(\nu, \mu) \equiv \inf \left\{ \mathrm{KL}(\nu, \nu') : \nu' \in \mathcal{D} \text{ and } \mu < E(\nu') \text{ and } \nu \ll \nu' \right\}, \tag{4}$$

with the convention that $\mathcal{K}_{\inf}(\nu, \mu) = \infty$ if there does not exist a $\nu \ll \nu'$ with $\mu < E(\nu')$.

3 Regret Lower Bound

We first give in Lemma 1 asymptotic lower bounds on the number of draws of suboptimal arms, either in high-probability or in expectation, in the spirit of those obtained by Lai and Robbins (1985); Anantharam et al. (1987). Compared to these works, the lower bounds obtained here hold under our more general assumptions on the arms distributions, which is reminiscent of the work of Burnetas and Katehakis (1996). While (6) could also easily be obtained using the recent change-of-distribution tools introduced by Garivier et al. (2016), note that we need to go back to Lai and Robbins' technique to prove the high-probability result (5), which will be crucial in the sequel. Indeed, we will use it to prove optimal regret of our algorithms: in essence we need to ensure that we have enough information about arms in $\mathcal{M} \cup \mathcal{N}$ to ensure that we pull the optimal arms in $\mathcal{L}$ sufficiently often. We first define a uniformly efficient algorithm. For any $\widetilde{\mathcal{V}} \equiv (\tilde{\nu}_1, \ldots, \tilde{\nu}_k) \in \mathcal{D}^k$, denote the m^{th} largest $E(\tilde{\nu}_a)$ by $\tilde{\mu}_{(m)}$. When emphasis is needed, we subscript $\mathbb{E}$ by $\widetilde{\mathcal{V}}$ to emphasize that the reward distribution is $\widetilde{\mathcal{V}}$. Nonetheless, this expectation also integrates out any additional

randomness in the algorithm. The algorithm is called uniformly efficient if, for all $\widetilde{\mathcal{V}} \in \mathcal{D}^k$ and $\alpha > 0$,

1. $\frac{\ell T}{m} - \mathbb{E}_{\widetilde{\mathcal{V}}}\left[N_{a^\star}(T)\right] = o\left(T^\alpha\right)$ for all $a^\star$ such that $E(\tilde{\nu}_{a^\star}) > \tilde{\mu}_{(m)}$;

2. $\ell T - \sum_{a^\star : E(\tilde{\nu}_{a^\star}) \geq \tilde{\mu}_{(m)}} \mathbb{E}_{\widetilde{\mathcal{V}}}\left[N_{a^\star}(T)\right] = o\left(T^\alpha\right)$[2].

An equivalent definition of uniform efficiency is that $R(T) = o(T^\alpha)$ under all bandit models in $\mathcal{D}$, though directly stating the above conditions is useful in our setting.

Lemma 1 (Lower bound on suboptimal arm draws)**.** *If the algorithm is uniformly efficient, then, for any arm a with $\mu_a \leq \mu_m$ and any $\delta \in (0,1)$ and $\epsilon > 0$,*

$$\lim_T \mathbb{P}\left\{ N_a(T) < (1-\delta)\frac{\log T}{\mathcal{K}_{\inf}(\nu_a, \mu_m) + \epsilon} \right\} = 0. \tag{5}$$

One can take $\epsilon = 0$ if $\mu_a < \mu_m$. Furthermore, for any suboptimal arm a with $\mu_a < \mu_m$,

$$\liminf_T \frac{\mathbb{E}[N_a(T)]}{\log T} \geq \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu_m)}. \tag{6}$$

We defer the proof of this result to Appendix A. We now present a corollary to this result which provides a regret lower bound, as well as sufficient conditions for an algorithm to asymptotically match it. As already noted by Komiyama et al. (2015) in the Bernoulli case, an algorithm achieving the asymptotic lower bound (6) on the expected number of draws of arms in $\mathcal{N}$ does not necessarily achieve optimal regret, unlike in classic bandit problems. Thus, we emphasize that condition (8) alone is not sufficient to prove asymptotic optimality, and we will later prove that our two algorithms satisfy (8) and (9). These two conditions can be easily obtained from the regret decomposition (A.3) given in Appendix A.

Theorem 2 (Regret lower bound)**.** *If an algorithm is uniformly efficient, then*

$$\liminf_T \frac{R(T)}{\log T} \geq \sum_{a \in \mathcal{N}} \frac{\mu_m - \mu_a}{\mathcal{K}_{\inf}(\nu_a, \mu_m)}. \tag{7}$$

[2]If the set of optimal arms is unique under $\widetilde{\mathcal{V}}$ then, in light of the fact that $\max_a \mathbb{E}[N_a(T)] \leq \ell T/m$, Condition 2 implies that $\frac{\ell T}{m} - \mathbb{E}_{\widetilde{\mathcal{V}}}\left[N_{a^\star}(T)\right] = o(T^\alpha)$ for all $\alpha > 0$ and optimal arms $a^\star$, including arms $a^\star \in \mathcal{M}$.

Moreover, any algorithm satisfying

$$\text{for arms } a \in \mathcal{N}: \quad \mathbb{E}[N_a(T)] = \frac{\log T}{\mathcal{K}_{\inf}(\nu_a, \mu_m)} + o(\log T), \tag{8}$$

$$\text{for arms } a^\star \in \mathcal{L}: \quad \mathbb{E}[N_{a^\star}(t)] = \frac{\ell T}{m} - o(\log T), \tag{9}$$

is asymptotically optimal, in the sense that it satisfies

$$\limsup_T \frac{R(T)}{\log T} \leq \sum_{a \in \mathcal{N}} \frac{\mu_m - \mu_a}{\mathcal{K}_{\inf}(\nu_a, \mu_m)}.$$

4 Algorithms

4.1 KL-CUCB

At time t, UCB algorithms leverage high probability upper bound $U_a(t)$ on μ_a for each a. The methods used to build these confidence bounds vary, as does the way the algorithm uses these confidence bounds. In our setting, we derive these bounds using the same technique as for KL-UCB in Cappé et al. (2013a), and then defines the subset $\widehat{\mathcal{S}}(t)$ as the subset of arms with m largest upper confidence bounds, just like the CUCB algorithm of Chen et al. (2013). The proposed algorithm is referred to as KL-CUCB.

The definition of the upper bound $U_a(t)$ is closely related to that of $\mathcal{K}_{\inf}$ given in (4). Let $\Pi_{\mathcal{D}}$ be an operator mapping from empirical distribution functions $\hat{\nu}_a(t)$ to an element of the model $\mathcal{D}$. Furthermore, let $f : \mathbb{N} \to \mathbb{R}$ be a non-decreasing function, where this function is usually chosen so that $f(t) \approx \log t$. The UCB is then defined as

$$U_a(t) \equiv \sup \left\{ E(\nu) : \nu \in \mathcal{D} \text{ and } \mathrm{KL}\left(\Pi_{\mathcal{D}}\left(\hat{\nu}_a(t)\right), \nu\right) \leq \frac{f(t)}{N_a(t)} \right\}. \tag{10}$$

As we will see, the closed form expression for $U_a(t)$ can be made slightly more explicit for exponential family models, though the expression still has the same general flavor. If a number μ satisfies $\mu \geq U_a(t)$, then this implies that, for every $\nu \in \mathcal{D}$ for which $E(\nu) > \mu$, $\mathrm{KL}\left(\Pi_{\mathcal{D}}(\hat{\nu}_a(t)), \nu\right) > \frac{f(t)}{N_a(t)}$. Consequently, $\mathcal{K}_{\inf}(\Pi_{\mathcal{D}}(\hat{\nu}_a(t)), \mu) \geq \frac{f(t)}{N_a(t)}$.

We now describe two settings in which the algorithm that we have described achieves the optimal asymptotic regret bound. These two settings and the presentation thereof follows

Algorithm KL-CUCB

Parameters A non-decreasing function $f : \mathbb{N} \to \mathbb{R}$ and an operator $\Pi_{\mathcal{D}}$ mapping from empirical distribution functions $\hat{\nu}_a(t)$ to an element of the model $\mathcal{D}$

Initialization Pull each arm of $\{1, \ldots, k\}$ ℓ times[3][4]

for $t = k, k+1, \ldots$ **do**

Let $U_a(t)$ be defined as in (10).

Let $\widehat{\mathcal{S}}(t)$ contain the m arms a for which $U_a(t)$ is maximal (ties broken arbitrarily)

Draw $\mathcal{A}(t+1)$ uniformly from the set of ℓ-cardinality subsets of $\widehat{\mathcal{S}}(t)$

Draw the corresponding rewards $Y_a(t+1)$ independently from ν_a, $a \in \mathcal{A}(t+1)$

Cappé et al. (2013a). The first family of distributions we consider for $\mathcal{D}$ is a canonical one-dimensional exponential family $\mathcal{E}$. For some dominating measure ρ, open set $H \subseteq \mathbb{R}$, and twice-differentiable strictly convex function $b : H \to \mathbb{R}$, $\mathcal{E}$ is a set of distributions ν_η such that

$$\frac{d\nu_\eta}{d\rho}(x) = \exp\left[x\eta - b(\eta)\right].$$

We assume that the open set H is the natural parameter space, i.e. the set of all $\eta \in \mathbb{R}$ such that $\int \exp(x\eta) d\rho(x) < \infty$. We define the corresponding (open) set of expectations by $I \equiv \{E(\nu_\eta) : \eta \in H\} \equiv (\mu_-, \mu_+)$ and its closure by $\bar{I} = [\mu_-, \mu_+]$. We have omitted the dependence of $\mathcal{E}$ on ρ and b in the notation. It is easily verified that $\mathcal{K}_{\inf}(\nu_a, \mu_m) = \mathrm{KL}(\nu_a, \nu_m)$

For the moment suppose that $\hat{\nu}_a(t)$ is such that $\hat{\mu}_a(t) \in I$. In this case we let $\Pi_{\mathcal{D}}$ denote the maximum likelihood operator so that $\Pi_{\mathcal{D}}\left(\hat{\nu}_a(t)\right)$ returns the unique distribution in $\mathcal{D}$ indexed by the η satisfying $b'(\eta) = \hat{\mu}_a(t)$. Thus, in this setting where $\hat{\mu}_a(t) \in I$, the UCB $U_a(t)$ then takes the form of the expression in (10).

More generally, we must deal with the case that $\hat{\mu}_a(t)$ equals μ_+ or μ_-. For $\mu \in I$, define by convention $\mathrm{KL}(\mu_-, \mu) = \lim_{\mu' \to \mu_-} \mathrm{KL}(\mu_-, \mu)$, $\mathrm{KL}(\mu_+, \mu) = \lim_{\mu' \to \mu_+} \mathrm{KL}(\mu', \mu)$, and analogously for $\mathrm{KL}(\mu, \mu_-)$ and $\mathrm{KL}(\mu, \mu_+)$. Finally, define $\mathrm{KL}(\mu_-, \mu_-)$ and $\mathrm{KL}(\mu_+, \mu_+)$ to be

[3] It is always possible to do this using a total of k draws of cardinality ℓ subsets of $\{1, \ldots, k\}$. For notational simplicity, consider the special case where $k = 4$ and $\ell = 3$. Then this is accomplished by drawing $\{1, 2, 3\}$, $\{2, 3, 4\}$, $\{3, 4, 1\}$, and $\{4, 1, 2\}$. The generalization to other values of k, ℓ is obvious.

[4] For $\ell < m$, the sampling strategy for the first ℓ draws of each arm does not technically conform to the sampling procedure described in Section 2, as we do not define a subset $\widehat{\mathcal{S}}(t)$ at each time point. Nonetheless, this has no implications on the asymptotic performance of our algorithm. Furthermore, one can confirm that the lower bounds presented in Section 3 still applies in this setting.

zero. This then gives the following general expression for $U_a(t)$ that we use to replace (10) in the KL-CUCB Algorithm:

$$U_a(t) \equiv \sup \left\{ \mu \in \bar{I} : \mathrm{KL}\left(\hat{\mu}_a(t), \mu\right) \leq \frac{f(t)}{N_a(t)} \right\}. \tag{11}$$

Note that this definition of $U_a(t)$ does not explicitly include a mapping $\Pi_{\mathcal{D}}$ mapping from an empirical distribution function to an element of the model $\mathcal{D}$. Thus we have avoided any problems that could arise in defining such a mapping when $\hat{\mu}_a(t)$ falls on the boundary of $\bar{I}$.

The KL-CUCB variant that we have presented achieves the asymptotic regret bound in the setting where $\mathcal{D} = \mathcal{E}$.

Theorem 3 (Optimality for single parameter exponential families)**.** *Suppose that $\mathcal{D} = \mathcal{E}$. Further let $f(t) = \log t + 3 \log\log t$ for $t \geq 3$ and $f(1) = f(2) = f(3)$. This variant of KL-CUCB satisfies (8) and (9). Thus KL-CUCB achieves the asymptotic regret lower bound (2) for uniformly efficient algorithms.*

Another interesting family of distributions for $\mathcal{D}$ is a set $\mathcal{B}$ of distributions on $[0, 1]$ with finite support. If the support of $\mathcal{D}$ is instead bounded in some $[-M, M]$, then the observations can be rescaled to $[0, 1]$ when selecting which arm to pull using the linear transformation $x \mapsto (x + M)/(2M)$.

If $\mathcal{D}$ is equal to $\mathcal{B}$, then Cappé et al. (2013a) observe that (10) rewrites as

$$U_a(t) = \sup \left\{ E(\nu) : \mathrm{Support}[\nu] \subseteq \mathrm{Support}\left[\hat{\nu}_a(t)\right] \cup \{1\} \text{ and } \mathrm{KL}\left(\hat{\nu}_a(t), \nu\right) \leq \frac{f(t)}{N_a(t)} \right\},$$

where, for a measure ν', we use $\mathrm{Support}[\nu']$ to denote the support of ν'. They furthermore observe that this expression admits an explicit solution via the method of Lagrange multipliers.

Theorem 4 (Optimality for finitely supported distributions)**.** *Suppose that $\mathcal{D} = \mathcal{B}$. Let $\Pi_{\mathcal{D}}$ denote the identity map and $f(t) = \log t + \log\log t$ for $t \geq 2$ and $f(1) = f(2)$. Suppose that $\mu_1 < 1$, $\mu_k > 0$, and $\mu_m > \mu_{m+1}$. The variant of KL-CUCB satisfies (8) and (9). Thus KL-CUCB achieves the asymptotic regret lower bound (2) for uniformly efficient algorithms.*

In both theorems, the little-oh notation hides the problem-dependent but T-independent quantities. In the proofs of Theorems 3 and 4 we refer to equations in Cappé et al. (2013b)

where the reader can find explicit finite-sample, problem-dependent expressions for the $o(\log T)$ term in (8) for the settings of Theorems 3 and 4. The argument used to establish (9) in these settings is asymptotic in nature so does not appear to easily yield finite sample constants.

4.2 Thompson Sampling

Algorithm Thompson Sampling

Parameters For each arm $a = 1, \ldots, k$, let $\Pi_a(0)$ be a prior distribution on μ_a
for $t = 0, 1, \ldots$ **do**
For each arm $a = 1, \ldots, k$, draw $\theta_a(t) \sim \Pi_a(t)$
Let $\widehat{\mathcal{S}}(t)$ contain the m arms a for which $\theta_a(t)$ is maximal (ties broken arbitrarily)
Draw $\mathcal{A}(t+1)$ uniformly from the set of ℓ-cardinality subsets of $\widehat{\mathcal{S}}(t)$
Draw the corresponding rewards $Y_a(t+1)$ independently from ν_a, $a \in \mathcal{A}(t+1)$
For each $a \in \mathcal{A}(t+1)$, obtain a new posterior $\Pi_a(t+1)$ by updating $\Pi_a(t)$ with the observation $Y_a(t+1)$
For each $a \notin \mathcal{A}(t+1)$, let $\Pi_a(t+1) = \Pi_a(t)$

Thompson sampling uses Bayesian ideas to account for the uncertainty in the estimated reward distributions. That is, one first posits a (typically non-informative) prior over the means of the reward distributions, and then at each time updates the posterior and takes a random draw of the k means from the posterior and pulls the arm whose posterior draw is the largest. This correspond to drawing a subset of m arms at random according to its posterior probability of being the optimal subset $\mathcal{S}^\star$, which generalizes the idea initially proposed by Thompson (1933). In the above algorithm we focus on independent priors so that the only posteriors updated at time $t+1$ are those of arms in $\mathcal{A}(t+1)$.

We prove the optimality of Thompson sampling for Bernoulli rewards, for the particular choice of a uniform prior distribution on the mean of each arm. Note that the algorithm is easy to implement in that case, since $\Pi_a(t)$ is a beta distribution with parameters $N_a(t)\hat{\mu}_a(t)+1$ and $N_a(t)(1-\hat{\mu}_a(t))+1$. Our proof makes use of the techniques used to prove the optimality of Thompson sampling in the standard bandit setting for Bernoulli rewards in Agrawal and Goyal (2012). We note that Komiyama et al. (2015) also made use of some of the techniques in Agrawal and Goyal (2012) to prove the optimality of Thompson sampling for Bernoulli rewards in the multiple play bandit setting.

Theorem 5 (Optimality for Bernoulli rewards)**.** *If the reward distributions are Bernoulli and* $\Pi_a(0)$ *is a standard uniform distribution for each* a*, then Thompson sampling satisfies (8) and (9). Thus Thompson sampling achieves the asymptotic regret lower bound (2) for uniformly efficient algorithms.*

For any $\epsilon > 0$, the proof shows that Thompson sampling satisfies

$$\mathbb{E}[N_a(T)] \leq (1+\epsilon)^2 \frac{f(T)}{\mathrm{KL}(\mu_a, \mu_m)} + o(\log T).$$

The proof gives an explicit bound on the $o(\log T)$ term that depends on both the problem and the choice of ϵ. The derivation of (9) in the setting of this theorem is asymptotic in nature and does not yield a finite sample bound.

5 Numerical Experiments

We now run two simulations to evaluate our theoretical results in practice, both with Bernoulli reward distributions and a horizon of $T = 100\,000$. For Simulation 1, the $k = 5$ arms have mean rewards $(0.5, 0.45, 0.45, 0.4, 0.3)$, $m = 2$, and we consider $\ell = 1, 2$. Note that $\mu_{m+1} \in \mathcal{M} = \{2, 3\}$ in this simulation. For Simulation 2, the $k = 5$ arms have mean rewards $(0.7, 0.6, 0.5, 0.3, 0.2)$, $m = 3$, and we consider $\ell = 1, 3$. Here, $\mathcal{M} = \{3\}$. All simulations are run using 500 Monte Carlo repetitions.

For $c \in \mathbb{R}$, we define the KL-CUCB c algorithm as the instance of KL-CUCB using the function $f(t) = \log t + c \log\log t$. Note that the use of both KL-CUCB 3 and KL-CUCB 1 are theoretically justified by the results of Theorems 3 and 4, as Bernoulli distributions satisfy the conditions of both theorems. We compare Thompson sampling and KL-CUCB to the ESCB algorithm of Combes et al. (2015b). ESCB is a different generalization of the KL-UCB algorithm from Cappé et al. (2013a) that computes an upper confidence bound for the sum of the arm means for each of the $\binom{k}{m}$ candidate sets $\mathcal{S}$, defined by the optimal value to

$$\sup_{(\mu_1,\ldots,\mu_k)\in[0,1]^k} \sum_{a\in\mathcal{S}} \mu_k \text{ subject to } \sum_{a\in\mathcal{S}} N_a(t)\,\mathrm{KL}\left(\hat{\mu}_a(t), \mu_a\right) \leq f(t) \tag{12}$$

and takes $\widehat{\mathcal{S}}(t)$ to be equal to the $\mathcal{S}$ with the maximal index. While this algorithm is

introduced for the case where $\ell = m$, it can also be applied to choose $\widehat{\mathcal{S}}(t)$ when $\ell < m$. Just like KL-CUCB, ESCB uses confidence bounds whose level rely on a function f such that $f(t) \approx \log t$. Because the optimization problem solved to compute the indices (11) and (12) are different, the f functions used by KL-CUCB and ESCB are not directly comparable. Nonetheless, a side-by-side comparison of the two algorithms seems to indicate that $f(t) = \log t + cm \log\log t$ for ESCB is comparable to $f(t) = \log t + c \log\log t$ for KL-CUCB. Combes et al. prove an $O(\log T)$ regret bound (with a sub-optimal constant) for the version of ESCB corresponding to the constant $c = 4$, that we refer to as ESCB $4m$.

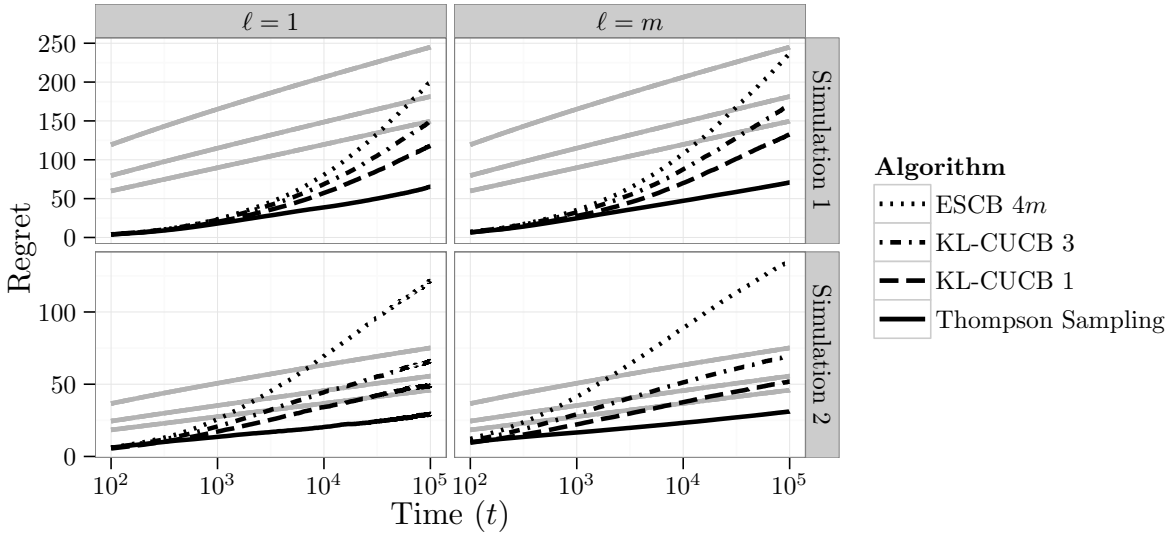

Figure 1: Regret of the four algorithms with theoretical guarantees. The bottom of the three gray lines represents the asymptotic regret bound given in Theorem 2. The middle and upper of the three Grey lines represent the asymptotic regret bound with $\log t$ replaced by $\log t + \log\log t$ and $\log t + 3\log\log t$, respectively. We expect the middle line to be an approximate upper bound on the regret of the KL-CUCB 1 algorithm, and the upper line to be an approximate upper bound on the KL-CUCB 3 algorithm. Asymptotically all three Grey lines agree in first order, though clearly they can be quite different in finite samples.

Figure 1 displays the regret of the four algorithms with theoretical guarantees. All but ESCB $4m$ have been proven to be asymptotically optimal, and thus are guaranteed to achieve the theoretical lower bound asymptotically. In our finite sample simulation, Thompson sampling performs better than this theoretical guarantee may suggest. Indeed, Thompson sampling outperforms the KL-CUCB algorithms in all four settings, while KL-

CUCB 1 outperforms KL-CUCB 3 and KL-CUCB 3 outperforms ESCB $4m$. To give the reader intuition on the relative performance of KL-CUCB variants, note that in the proofs of Theorems 3 and 4 we prove that the number of pulls on each suboptimal arm a is upper bounded by $f(T)/\mathcal{K}_{\inf}(\nu_a, \mu_m) + o(\log T)$, with an explicit finite sample constant for the $o(\log T)$ term. While $f(T) = \log T + o(\log T)$ for KL-CUCB 1 and KL-CUCB 3, for finite T the quantities $\log T$ and $\log T + c \log\log T$, $c = 1, 3$, are quite different. At $T = 10^5$, $\log T + \log\log T$ is 20% larger than $\log T$, and $\log T + 3\log\log T$ is 60% larger. This difference does not decay quickly with sample size: at $T = 10^{15}$, these two quantities are still respectively 10% and 30% larger than $\log T$. This makes clear the practical benefit to choosing $f(t)$ as close to $\log t$ as is theoretically justifiable: for Bernoullis, the choice of $f(t)$ in Theorem 4 yields much better results than the choice of $f(t)$ in Theorem 3.

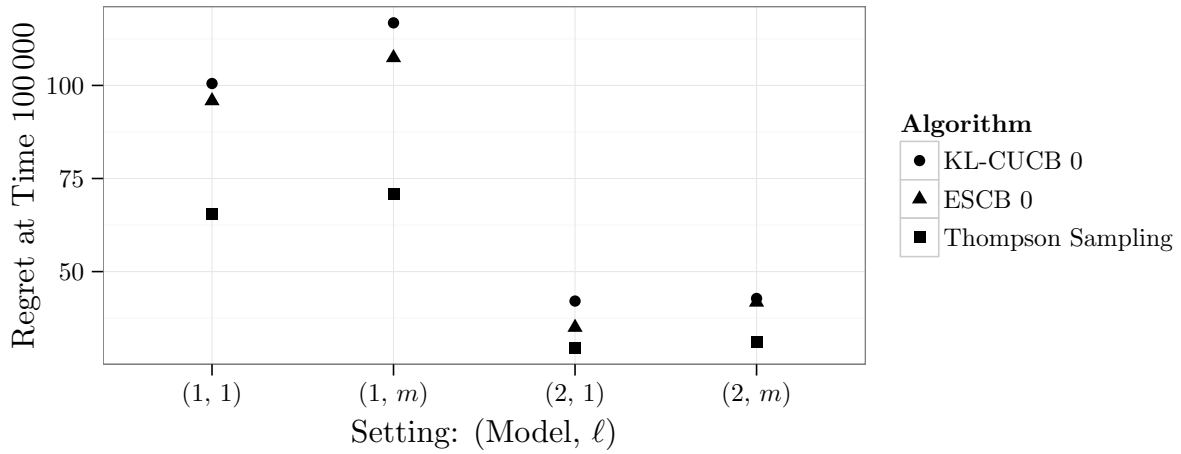

Figure 2: Regret at time 100 000 for KL-CUCB 0, ESCB 0, and Thompson sampling. Across all settings, ESCB 0 slightly outperforms KL-CUCB, with Thompson sampling significantly outperforming both algorithms. There are currently no theoretical guarantees for KL-CUCB 0 or ESCB 0.

Figure 2 compares the performance of KL-CUCB 0 and ESCB 0 on the same simulation settings considered earlier in this section. Though not theoretically justified, this choice of $f(t) = \log t$ has been used quite a lot in practice. The ordering of the three algorithms is the same across all models and values of ℓ: Thompson Sampling performs best while ESCB 0 slightly outperforms KL-CUCB 0. This should however be mitigated by the gap of numerical complexity between the two algorithms, especially when m and k are large and m/k is not

close to 0 or 1: while KL-CUCB only requires running k univariate root-finding procedures regardless of m, the current proposed ESCB algorithm requires running $\binom{k}{m}$ univariate root-finding procedures. For $k = 100$ and $m = 10$, this is a difference of running 100 root-finding procedures versus more than 10^{13} of them.

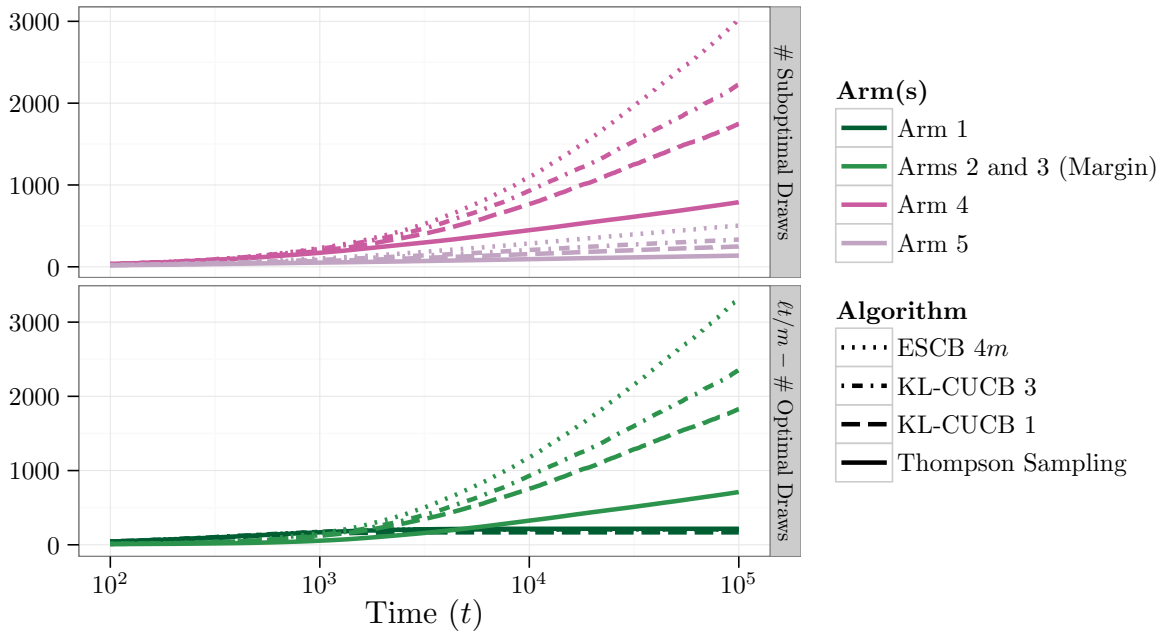

Figure 3: Number of suboptimal arm draws (top) and $\ell t/m$ minus the number of optimal arm draws (bottom) in Simulation 1 with $m = \ell = 2$. The former plot agrees with our theoretical results for KL-CUCB and Thompson sampling that suboptimal arms closer to the margin will receive more draws than those far from the margin. The latter agrees with our result that most suboptimal arm draws occur in place of draws of arms on the margin (here arms 2 and 3), rather than arms away from the margin (here arm 1).

Figure 3 displays the number of optimal and suboptimal arm draws in Simulation 1 with $m = 2$ and $\ell = 2$. Observe that all of the algorithms sample the optimal arm that is not on the margin (arm 1) approximately $\ell t/m$ times for each time t, and also sample the sum of the arms on the margin (arms 2 and 3) approximately $\ell t/m$ times, though indeed the margin arms are drawn far less frequently than arm 1. This agrees with the theoretical guarantee for KL-CUCB and Thompson sampling that we have established in this paper, namely that both satisfy (8) and (9).

6 Proofs of Optimality of KL-CUCB and Thompson Sampling Schemes

We now outline our proofs of optimality for the KL-CUCB and Thompson sampling schemes. We break this section into two subsections. Section 6.1 establishes that the arms in $\mathcal{N}$, i.e. the suboptimal arms, are not pulled often (satisfy Equation 8). Due to the differences in proof methods, we consider the KL-CUCB and Thompson sampling schemes separately in this subsection. Section 6.2 establishes that the arms in $\mathcal{L}$, i.e. the optimal arms away from the margin, are pulled often (satisfy Equation 9). We give the outline of the proofs for the KL-CUCB and Thompson sampling schemes simultaneously, though provide the detailed arguments separately in Appendices B and C, respectively. We note that the order of presentation of the two subsections is important: the arguments used in Subsection 6.2 rely on the validity of (8), which is established in Section 6.1.

6.1 Suboptimal arms not pulled often

KL-CUCB

Preliminary: a general analysis of KL-CUCB. We start by giving a general analysis of KL-CUCB in our setting, and then use it to prove Theorems 3 and 4. The arguments in this section generalize those given in Cappé et al. (2013a,b) for the case where $m = \ell = 1$. Let $\mu^\dagger$ be some real number that we will choose to be either equal to μ_m or slightly smaller than μ_m. For all $t \geq k$,

$$\begin{aligned}\{a \in \mathcal{A}(t+1)\} &= \left\{a \in \mathcal{A}(t+1), \mu^\dagger \geq \min_{a^\star \in \mathcal{S}^\star} U_{a^\star}(t)\right\} \cup \left\{a \in \mathcal{A}(t+1), \mu^\dagger < \min_{a^\star \in \mathcal{S}^\star} U_{a^\star}(t)\right\} \\ &\subseteq \left[\cup_{a^\star \in \mathcal{S}^\star} \left\{\mu^\dagger \geq U_{a^\star}(t)\right\}\right] \cup \left\{a \in \mathcal{A}(t+1), \mu^\dagger < U_a(t)\right\}. \end{aligned} \tag{13}$$

In the second line, we use that by definition of the algorithm $\min_{i \in \hat{\mathcal{S}}(t)} U_i(t) \geq \min_{a^* \in \mathcal{S}^*} U_{a^\star}(t)$, in particular for any $a \in \mathcal{A}(t+1)$, $U_a(t) \geq \min_{a^* \in \mathcal{S}^\star} U_{a^*}(t)$. For each $\gamma > 0$, we now introduce the set $\mathcal{C}_{\mu^\dagger,\gamma}$. In the setting of Theorem 3,

$$\mathcal{C}_{\mu^\dagger,\gamma} \equiv \left\{\nu' : \text{Support}[\nu'] \subseteq \bar{I}\right\} \cap \left\{\nu' : \exists\, \mu \in (\mu^\dagger, \mu_+] \text{ with } \mathrm{KL}(E(\nu'), \mu) \leq \gamma\right\},$$

where above $\mathrm{KL}(E(\nu'), \mu)$ is the KL-divergence in the canonical exponential family $\mathcal{E}$. In the setting of Theorem 4,

$$\mathcal{C}_{\mu^\dagger,\gamma} \equiv \{\nu' : \text{Support}[\nu'] \subseteq [0,1]\} \cap \left\{\nu' : \exists\, \nu \in \mathcal{B} \text{ with } \mu^\dagger < E(\nu) \text{ and } \mathrm{KL}(\Pi_{\mathcal{D}}(\nu'), \nu) \leq \gamma\right\}.$$

The set $\mathcal{C}_{\mu^\dagger,\gamma}$ is defined in both settings so that $\mu^\dagger < U_a(t)$ if and only if $\hat{\nu}_a(t) \in \mathcal{C}_{\mu^\dagger, f(t)/N_a(t)}$. Recalling that $\mathbb{E}[N_a(T)] = \sum_{t=0}^{T-1} \mathbb{P}\{a \in \mathcal{A}(t+1)\}$, a union bound gives

$$\mathbb{E}[N_a(T)] \leq \ell + \sum_{a^\star \in \mathcal{S}^\star} \sum_{t=k}^{T-1} \mathbb{P}\left\{\mu^\dagger \geq U_{a^\star}(t)\right\} + \sum_{t=k}^{T-1} \mathbb{P}\left\{a \in \mathcal{A}(t+1), \hat{\nu}_{a,N_a(t)} \in \mathcal{C}_{\mu^\dagger, f(t)/N_a(t)}\right\}.$$

In analogue to Equation 8 in Cappé et al. (2013a), the above rightmost term satisfies

$$\begin{aligned}
&\sum_{t=k}^{T-1} \mathbb{P}\left\{a \in \mathcal{A}(t+1), \hat{\nu}_{a,N_a(t)} \in \mathcal{C}_{\mu^\dagger, f(t)/N_a(t)}\right\} \\
\leq &\sum_{t=k}^{T-1} \mathbb{P}\left\{a \in \mathcal{A}(t+1), \hat{\nu}_{a,N_a(t)} \in \mathcal{C}_{\mu^\dagger, f(T)/N_a(t)}\right\} \\
= &\sum_{t=k}^{T-1} \sum_{n=\ell+1}^{T-k+1} \mathbb{P}\left\{\hat{\nu}_{a,n-1} \in \mathcal{C}_{\mu^\dagger, f(T)/(n-1)}, \tau_{a,n} = t+1\right\} \qquad (14) \\
\leq &\sum_{n=\ell}^{T-k} \mathbb{P}\left\{\hat{\nu}_{a,n} \in \mathcal{C}_{\mu^\dagger, f(T)/n}\right\},
\end{aligned}$$

where the final inequality holds because, for each n, $\tau_{a,n} = t+1$ for at most one t in $\{k, \ldots, T-1\}$. We will upper bound the terms with $n = \ell, \ldots, b_a^m(T)$ in the sum on the right by 1, where

$$b_a^m(T) \equiv \left\lceil \frac{f(T)}{\mathcal{K}_{\inf}(\nu_a, \mu_m)} \right\rceil \leq \frac{f(T)}{\mathcal{K}_{\inf}(\nu_a, \mu_m)} + 1.$$

This gives the bound

$$\sum_{n=\ell}^{T-k} \mathbb{P}\left\{\hat{\nu}_{a,n} \in \mathcal{C}_{\mu^\dagger, f(T)/n}\right\} \leq \frac{f(T)}{\mathcal{K}_{\inf}(\nu_a, \mu_m)} - \ell + 2 + \sum_{n=b_a^m(T)+1}^{\infty} \mathbb{P}\left\{\hat{\nu}_{a,n} \in \mathcal{C}_{\mu^\dagger, f(T)/n}\right\}.$$

Hence,

$$\mathbb{E}[N_a(T)] \leq \frac{f(T)}{\mathcal{K}_{\inf}(\nu_a, \mu_m)} + \underbrace{\sum_{n=b_a^m(T)+1}^{\infty} \mathbb{P}\left\{\hat{\nu}_{a,n} \in \mathcal{C}_{\mu^\dagger, f(T)/n}\right\}}_{\text{Term 1}} + \sum_{a^\star \in \mathcal{S}^\star} \underbrace{\sum_{t=k}^{T-1} \mathbb{P}\left\{\mu^\dagger \geq U_{a^\star}(t)\right\}}_{\text{Term } 2a^\star} + 2. \tag{15}$$

Up until this point we have not committed to any particular choice of $\mu^\dagger$ or f. We now give proofs of (8) in the settings of Theorems 3 and 4. For each proof we make a particular choice of $\mu^\dagger$ and use the choice of f from the theorem statement.

Proof of (8) in the settings of Theorems 3 and 4. In the setting of Theorem 3 let $\mu^\dagger = \mu_m$ and in the setting of Theorem 4 let $\mu^\dagger = \left[1 - \log(T)^{-1/5}\right] \mu_m$. Lemma A.1 shows that Term 1 is $o(\log T)$ for both settings and includes references on where to find an explicit finite sample upper bound. Fix $a^\star \in \mathcal{S}^\star$. Noting that $\mu^\dagger \leq \mu_{a^\star}$, Term $2a^\star$ is $o(\log T)$ in both settings by Lemma A.2, with an exact finite sample upper bound given in the proof thereof. Thus $\sum_{a^\star \in \mathcal{S}^\star} \text{Term } 2a^\star = o(\log T)$ in both settings. □

Thompson Sampling

This proof is inspired by the analysis of Thompson sampling proposed by Agrawal and Goyal (2012). Fix a suboptimal arm $a \in \mathcal{N}$. Let $\mu^\dagger$ and $\theta^\dagger$ be numbers (to be specified later) satisfying $\mu^\dagger < \theta^\dagger$. Observe that

$$\begin{aligned}\{a \in \mathcal{A}(t+1)\} &= \left\{a \in \mathcal{A}(t+1), \theta_a(t) \leq \theta^\dagger\right\} \cup \left\{a \in \mathcal{A}(t+1), \theta_a(t) > \theta^\dagger\right\} \\ &\subseteq \left[\cup_{a^\star \in \mathcal{S}^\star} \left\{a \in \mathcal{A}(t+1), \theta_a(t) \leq \theta^\dagger, a^\star \notin \widehat{\mathcal{S}}(t)\right\}\right] \cup \left\{a \in \mathcal{A}(t+1), \theta_a(t) > \theta^\dagger\right\}.\end{aligned}$$

Furthermore,

$$\begin{aligned}&\left\{a \in \mathcal{A}(t+1), \theta_a(t) > \theta^\dagger\right\} \\ &\subseteq \left\{a \in \mathcal{A}(t+1), \theta_a(t) > \theta^\dagger, \hat{\mu}_a(t) \leq \mu^\dagger\right\} \cup \left\{a \in \mathcal{A}(t+1), \hat{\mu}_a(t) > \mu^\dagger\right\}.\end{aligned}$$

Recalling that $\mathbb{E}[N_a(T)] = \sum_{t=0}^{T-1} \mathbb{P}\left\{a \in \mathcal{A}(t+1)\right\}$,

$$\begin{aligned}\mathbb{E}[N_a(T)] \leq & \sum_{a^\star \in \mathcal{S}^\star} \underbrace{\sum_{t=0}^{T-1} \mathbb{P}\left\{a \in \mathcal{A}(t+1), \theta_a(t) \leq \theta^\dagger, a^\star \notin \widehat{\mathcal{S}}(t)\right\}}_{\text{Term I}a^\star} \\ & + \underbrace{\sum_{t=0}^{T-1} \mathbb{P}\left\{a \in \mathcal{A}(t+1), \theta_a(t) > \theta^\dagger, \hat{\mu}_a(t) \leq \mu^\dagger\right\}}_{\text{Term II}} \\ & + \underbrace{\sum_{t=0}^{T-1} \mathbb{P}\left\{a \in \mathcal{A}(t+1), \hat{\mu}_a(t) > \mu^\dagger\right\}}_{\text{Term III}}. \end{aligned} \tag{16}$$

The above decomposition does not depend on the algorithm. Bounding Terms Ia^*, $a^\star \in \mathcal{S}^\star$, and Term II will rely on arguments that are specific to Thompson Sampling. Fix $a^\star \in \mathcal{S}^\star$ and let $p^{\theta^\dagger}_{a^\star}(t) \equiv \mathbb{P}(\theta_{a^\star}(t) > \theta^\dagger \big| \mathcal{F}(t))$. Note that $p^{\theta^\dagger}_{a^\star}(t) \neq p^{\theta^\dagger}_{a^\star}(t+1)$ implies $a^\star \in \mathcal{A}(t+1)$. Thus $p^{\theta^\dagger}_{a^\star}(t)$ is equal to $p^{\theta^\dagger}_{a^\star,n} \equiv p^{\theta^\dagger}_{a^\star}(\tau_{a^\star,n})$ for all t such that $N_{a^\star}(t) = n$. We now state Lemma 6, that generalizes Lemma 1 in Agrawal and Goyal (2012) beyond the case $\ell = m = 1$.

Lemma 6. *If a is some arm, $a^\star \in \mathcal{S}^\star$ and $\theta^\dagger < 1$, then, for all $t \geq 0$,*

$$\mathbb{P}\left(a \in \mathcal{A}(t+1), \theta_a(t) \leq \theta^\dagger, a^\star \notin \widehat{\mathcal{S}}(t) \Big| \mathcal{F}(t)\right) \leq \frac{1 - p^{\theta^\dagger}_{a^\star}(t)}{p^{\theta^\dagger}_{a^\star}(t)} \mathbb{P}\left(a^\star \in \widehat{\mathcal{S}}(t) \Big| \mathcal{F}(t)\right).$$

This lemma is not specific to Thompson sampling (beyond the condition that $\theta^\dagger < 1$). Rather, it will apply to any randomized algorithm which, conditional on $\mathcal{F}(t)$, independently draws $\{\theta_a(t) : a = 1, \dots k\}$ and lets $\widehat{\mathcal{S}}(t)$ contain the m arms a for which $\theta_a(t)$ is maximal.

Observe that the upper bound in the above lemma does not rely on a. We also have the following lemma, whose proof can be found in Appendix C with that of Lemma 6.

Lemma 7. *If $a^\star \in \mathcal{S}^\star$ and $\theta^\dagger < 1$, then, for all $t \geq 0$,*

$$\mathbb{E}\left[\sum_{t=0}^{T-1} \frac{1 - p^{\theta^\dagger}_{a^\star}(t)}{p^{\theta^\dagger}_{a^\star}(t)} \mathbb{P}\left(a^\star \in \widehat{\mathcal{S}}(t) \Big| \mathcal{F}(t)\right)\right] \leq \frac{m}{\ell} \mathbb{E}\left[\sum_{n=0}^{T-1} \frac{1 - p^{\theta^\dagger}_{a^\star,n}}{p^{\theta^\dagger}_{a^\star,n}}\right].$$

Combining the two preceding lemmas yield the inequality

$$\text{Term I}a^\star \leq \frac{m}{\ell}\,\mathbb{E}\left[\sum_{n=0}^{T-1}\frac{1-p^{\theta^\dagger}_{a^\star,n}}{p^{\theta^\dagger}_{a^\star,n}}\right]. \tag{17}$$

Note crucially that we have upper bounded the sum over time on the left-hand side by a sum over the number of pulls of arm $a^\star$ on the right-hand side. Transferring from a sum over time to a sum over counts is difficult in stochastic algorithms such as Thompson sampling. Indeed, there appears to be a steep price to pay for this transfer: the right-hand side inverse weights by a conditional probability, which may be small for certain realizations of the data. Lemma 2 in Agrawal and Goyal (2012), that we restate below using our modified notation, establishes that this inverse weighting does not cause a problem in the special case of Thompson sampling with Bernoulli rewards and independent beta priors. If $\theta^\dagger < \mu_m$, then the proceeding lemma implies that, for each $a^\star \in \mathcal{S}^\star$, Term I$a^\star$ is $O(1)$, i.e. is $o(\log T)$ with much to spare. Obviously, this implies that $\sum_{a^\star\in\mathcal{S}^\star}$ Term I$a^\star = o(\log T)$ as well.

Lemma 8 (Lemma 2 from Agrawal and Goyal, 2012)**.** *If $a^\star \in \mathcal{S}^\star$ and $\theta^\dagger < \mu_{a^\star}$, then, with $\Delta \equiv \mu_{a^\star} - \theta^\dagger$,*

$$\mathbb{E}\left[\frac{1-p^{\theta^\dagger}_{a^\star,n}}{p^{\theta^\dagger}_{a^\star,n}}\right] = \begin{cases} \frac{3}{\Delta}, & \text{for } n < \frac{8}{\Delta} \\ \Theta\left(e^{-\Delta^2 n/2} + \frac{1}{(n+1)\Delta^2}e^{-\mathrm{KL}(\theta^\dagger,\mu_{a^\star})n} + \frac{1}{\exp(\Delta^2 n/4)-1}\right), & \text{for } n \geq \frac{8}{\Delta}. \end{cases}$$

Above $\Theta(\cdot)$ is used to represent big-Theta notation.

We now turn to Term II. The following result mimics Lemma 4 in Agrawal and Goyal (2012), and is a consequence of the close link between beta and binomial distributions and the Chernoff-Hoeffding bound. We provide a proof of this result in Appendix C.

Lemma 9. *If a is some arm and $\mu_a < \mu^\dagger < \theta^\dagger$, then*

$$\text{Term II} \equiv \sum_{t=0}^{T-1}\mathbb{P}\left\{a \in \mathcal{A}(t+1), \theta_a(t) > \theta^\dagger, \hat{\mu}_a(t) \leq \mu^\dagger\right\} \leq \frac{\log T}{\mathrm{KL}(\mu^\dagger, \theta^\dagger)}.$$

We now turn to Term III. Note that

$$\begin{aligned}\text{Term III} &= \mathbb{E}\left[\sum_{t=0}^{T-1} \mathbb{1}\left\{a \in \mathcal{A}(t+1), \hat{\mu}_{a,N_a(t)} > \mu^\dagger\right\}\right] \\ &= \mathbb{E}\left[\sum_{t=0}^{T-1}\sum_{n=0}^{T-1} \mathbb{1}\left\{\tau_{a,n+1} = t+1, \hat{\mu}_{a,n} > \mu^\dagger\right\}\right] \\ &\leq \sum_{n=0}^{T-1} \mathbb{P}\left\{\hat{\mu}_{a,n} > \mu^\dagger\right\}, \end{aligned} \tag{18}$$

where the latter inequality holds because $\tau_{a,n+1} = t+1$ for at most one t in $\{0,\dots,T-1\}$. The following lemma controls the right-hand side of the above.

Lemma 10. *Fix an arm a. If $\mu^\dagger > \mu_a$, then*

$$\sum_{n=0}^{T-1} \mathbb{P}\left\{\hat{\mu}_{a,n} > \mu^\dagger\right\} \leq 1 + \frac{1}{\mathrm{KL}(\mu^\dagger, \mu_a)}.$$

The proof is omitted, but is an immediate consequence of the Chernoff-Hoeffding bound and the additional bounding from the proof of Lemma 3 in Agrawal and Goyal (2012). Thus we have shown that Term III is $o(\log T)$, with much to spare as well.

The proof of (8) in the setting of Theorem 5 is now straightforward.

Proof of (8) in the setting of Theorem 5. Fix $a \notin \mathcal{L} \cup \mathcal{M}$. Fix $\mu^\dagger < \theta^\dagger$ and $\theta^\dagger$ (to be specified shortly) so that $\mu_a < \mu^\dagger < \theta^\dagger < \mu_m$ and $\epsilon \in]0, 1]$ a constant. Plugging our results on each Term I$a^\star$ and on Terms II and III into (16) then yields that

$$\mathbb{E}\left[N_a(T)\right] \leq \frac{\log T}{\mathrm{KL}(\mu^\dagger, \theta^\dagger)} + 1 + \frac{1}{\mathrm{KL}(\mu^\dagger, \mu_a)} + O(1).$$

The argument concludes by selecting $\mu^\dagger$ so that $\mathrm{KL}(\mu^\dagger, \mu_m) = \frac{\mathrm{KL}(\mu_a, \mu_m)}{1+\epsilon}$ and $\theta^\dagger$ so that $\mathrm{KL}(\mu^\dagger, \theta^\dagger) = \frac{\mathrm{KL}(\mu^\dagger, \mu_m)}{1+\epsilon}$, since this gives $\mathrm{KL}(\mu^\dagger, \theta^\dagger) = \frac{\mathrm{KL}(\mu_a, \mu_m)}{(1+\epsilon)^2}$. Hence,

$$\mathbb{E}[N_a(T)] \leq (1+\epsilon)^2 \frac{f(T)}{\mathrm{KL}(\mu_a, \mu_m)} + o(\log T).$$

Dividing both sides by $\log T$, and then taking $T \to \infty$ followed by $\epsilon \to 0$ gives the result. □

6.2 Optimal arms away from margin pulled $\ell T/m - o(\log T)$ times

We now show that the optimal arms away from the margin ($a^\star \in \mathcal{L}$) are pulled often. We start by giving a general analysis that does not rely on the algorithm under consideration, and then we specialize the discussion to KL-CUCB and Thompson sampling, respectively. For the remainder of this section, we fix an optimal arm $a^\star$ with $\mu_{a^\star} > \mu_m$. Observe that, for $t \geq k$ (KL-CUCB) or $t \geq 0$ (Thompson sampling),

$$\left\{a^\star \notin \widehat{\mathcal{S}}(t)\right\} = \cup_{a:\mu_a \leq \mu_m} \left\{a \in \widehat{\mathcal{S}}(t), a^\star \notin \widehat{\mathcal{S}}(t)\right\}.$$

Recalling (2), we see that, for Thompson sampling,

$$\begin{aligned}
\frac{\ell T}{m} - \mathbb{E}[N_{a^\star}(T)] &= \frac{\ell T}{m} - \mathbb{E}\left[\sum_{t=0}^{T-1} \mathbb{P}\left\{a^\star \in \mathcal{A}(t+1) | \mathcal{F}(t)\right\}\right] \\
&= \frac{\ell}{m}\left(T - \mathbb{E}\left[\sum_{t=0}^{T-1} \mathbb{P}\left\{a^\star \in \widehat{\mathcal{S}}(t)\middle|\mathcal{F}(t)\right\}\right]\right) \\
&= \frac{\ell}{m}\sum_{t=0}^{T-1} \mathbb{P}\{a^\star \notin \widehat{\mathcal{S}}(t)\} \\
&\leq \frac{\ell}{m}\sum_{a:\mu_a \leq \mu_m}\sum_{t=0}^{T-1} \mathbb{P}\left\{a \in \widehat{\mathcal{S}}(t), a^\star \notin \widehat{\mathcal{S}}(t)\right\}, \qquad (19)
\end{aligned}$$

where the final inequality holds by the preceding display. We have a similar identity for KL-CUCB, though the identity is slightly different due to the initiation of each of the k arms. Specifically,

$$\frac{\ell(T-k)}{m} + \ell - \mathbb{E}[N_{a^\star}(T)] \leq \frac{\ell}{m}\sum_{a:\mu_a \leq \mu_m}\sum_{t=k}^{T-1} \mathbb{P}\left\{a \in \widehat{\mathcal{S}}(t), a^\star \notin \widehat{\mathcal{S}}(t)\right\}. \qquad (20)$$

Similar to (2), for each a such that $\mu_a \leq \mu_m$ and $t \geq k$ (KL-CUCB) or $t \geq 0$ (Thompson sampling),

$$\mathbb{P}\left\{a \in \mathcal{A}(t+1), a^\star \notin \widehat{\mathcal{S}}(t)\middle|\mathcal{F}(t)\right\} = \frac{\ell}{m}\mathbb{P}\left\{a \in \widehat{\mathcal{S}}(t), a^\star \notin \widehat{\mathcal{S}}(t)\middle|\mathcal{F}(t)\right\},$$

and thus

$$\sum_{t=0}^{T-1} \mathbb{P}\left\{a \in \widehat{\mathcal{S}}(t), a^\star \notin \widehat{\mathcal{S}}(t)\right\} = \frac{m}{\ell} \sum_{t=0}^{T-1} \mathbb{P}\left\{a \in \mathcal{A}(t+1), a^\star \notin \widehat{\mathcal{S}}(t)\right\}.$$

For each a with $\mu_a \leq \mu_m$, let

$$M_a^{a^\star}(T) \equiv \sum_{t=0}^{T-1} \mathbb{1}\{a \in \mathcal{A}(t+1), a^\star \notin \widehat{\mathcal{S}}(t)\}.$$

The bounds (20) and (19) yield the key observation that we use in this section:

$$\begin{aligned} \text{for KL-CUCB:} \quad & \frac{\ell(T-k)}{m} + \ell - \mathbb{E}[N_{a^\star}(T)] \leq \sum_{a:\mu_a \leq \mu_m} \mathbb{E}[M_a^{a^\star}(T)]; \\ \text{for Thompson sampling:} \quad & \frac{\ell T}{m} - \mathbb{E}[N_{a^\star}(T)] \leq \sum_{a:\mu_a \leq \mu_m} \mathbb{E}[M_a^{a^\star}(T)]. \end{aligned} \tag{21}$$

The remainder of the analysis involves controlling $\mathbb{E}[M_a^{a^\star}(T)]$ for arms $a \in \mathcal{M} \cup \mathcal{N}$.

Let G be some integer in $[0, +\infty($ and $\delta \in (0,1)$ be a constant to be specified shortly. For convenience, we let $T^{(g)} \equiv \lfloor T^{(1-\delta)^g} \rfloor$ for $g \in \mathbb{N}$. Our analysis relies on the following bound (for which we provide the arguments below):

$$\begin{aligned} \sum_{a:\mu_a \leq \mu_m} \mathbb{E}[M_a^{a^\star}(T)] &= \sum_{a:\mu_a \leq \mu_m} \mathbb{E}[M_a^{a^\star}(T^{(G)})] + \sum_{g=1}^{G} \sum_{a:\mu_a \leq \mu_m} \mathbb{E}[M_a^{a^\star}(T^{(g-1)}) - M_a^{a^\star}(T^{(g)})] \\ &\leq (1-\delta)^G \sum_{a \in \mathcal{N}} \frac{\log T}{\mathcal{K}_{\inf}(\nu_a, \mu_m)} + o(\log T) + \underbrace{\sum_{a \in \mathcal{M}} \mathbb{E}[M_a^{a^\star}(T^{(G)})]}_{\text{Term A}} \\ &\quad + \underbrace{\sum_{g=1}^{G} \sum_{a:\mu_a \leq \mu_m} \mathbb{E}[M_a^{a^\star}(T^{(g-1)}) - M_a^{a^\star}(T^{(g)})]}_{\text{Term B}}. \end{aligned} \tag{22}$$

The equality is a telescoping series and the inequality holds using that (i) $\mathbb{E}[M_a^{a^\star}(T^{(G)})] \leq \mathbb{E}[N_a(T^{(G)})]$ and (ii) the fact that the algorithm achieves the asymptotically optimal number of suboptimal arm draws: indeed it has been proved in Section 6.1 that (8) holds in the settings of Theorems 3, 4, and 5.

We now present the key ingredients to bound Term A and B. Each lemma stated below

holds for both KL-CUCB in the settings of Theorems 3 and 4 and for Thompson sampling in the setting of Theorem 5. Though these lemmas hold for both algorithms, the methods of proof for KL-CUCB and for Thompson sampling are quite different. Thus we give the proofs of the lemmas in the settings of Theorems 3 and 4 in Appendix B and the proofs in the setting of Theorem 5 in Appendix C.

Lemma 11. *Fix $G \geq 0$. In the settings of Theorem 3, 4, and 5,*

$$\text{Term A} \leq (1-\delta)^G \sum_{a \in \mathcal{M}} \frac{\log T}{\mathcal{K}_{\inf}(\nu_a, \mu_{a^\star})} + o(\log T).$$

The proof of Lemma 11 borrows a lot from the proof of (8) for each algorithm. Note that an exact finite sample upper bound on the $o(\log T)$ term in the settings of Theorems 3 and 4 can be found via the use of Lemmas A.1 and A.2 in the appendix.

Controlling Term B relies on a careful choice of $\delta > 0$, which is specified in Lemma 12 below. The proof of this lemma is highly original: indeed we first prove that the considered algorithm is uniformly efficient, which allows to exploit the lower bound (5) given in Theorem 2. Its proof is provided in the appendix for both KL-CUCB and Thompson Sampling, and we sketch it below.

Lemma 12. *Let $c \in (0,1)$ and δ chosen such*

$$\delta = c \left[1 - \left(\max_{a \in \mathcal{N}} \frac{\mathcal{K}_{\inf}(\nu_a, \mu_m)}{\mathcal{K}_{\inf}(\nu_a, \mu_{a^\star})} \right)^{1/2} \right], \tag{23}$$

and $\delta = c$ if $\mathcal{N} = \emptyset$. Then in the setting of Theorems 3, 4, and 5, Term B is $o(\log T)$.

Sketch of proof of Lemma 12. We first show that the algorithms are uniformly efficient in the sense defined in Section 3.

Lemma 13. *KL-CUCB is uniformly efficient in the settings of Theorems 3 and 4 and Thompson sampling is uniformly efficient in the setting of Theorem 5.*

Proof. Fix an arbitrary reward distribution $\mathcal{V} \in \mathcal{D}^k$. By the already proven (8) in the

settings of Theorems 3, 4, and 5 and by Lemma 11, both of which hold for $\mathcal{V}$,

$$\begin{aligned}\frac{\ell T}{m} - \mathbb{E}_{\mathcal{V}}[N_{a^\star}(T)] &\leq \sum_{a:\mu_a \leq \mu_m} \mathbb{E}_{\mathcal{V}}[M_a^{a^\star}(T)] + O(1) \\ &\leq \sum_{a \in \mathcal{N}} \mathbb{E}_{\mathcal{V}}[N_a(T)] + \sum_{a \in \mathcal{M}} \mathbb{E}_{\mathcal{V}}[M_a^{a^\star}(T)] + O(1) = O(\log T)\end{aligned}$$

for any $a^\star \in \mathcal{L}$, where the $O(1)$ term is equal to zero for Thompson sampling and, by (21), is $k\ell/m - \ell$ for KL-CUCB. Clearly $\log T = o(T^\alpha)$ for any $\alpha > 0$. □

Fix $g \in \mathbb{N}$ and an arm $a \in \mathcal{N}$. By the uniform efficiency of the algorithm established in Lemma 13, we will be able to apply (5) from Lemma 1 to show that $N_a(T^{(g)}) \geq (1-\delta)\frac{\log T^{(g)}}{\mathcal{K}_{\inf}(\nu_a, \mu_m)}$ with probability approaching 1. For now suppose this holds almost surely (in the proofs we deal with the fact that this happens with probability approaching rather than exactly 1). Our objective will be to show that this lower bound on $N_a(T^{(g)})$ suffices to ensure that $M_a^{a^\star}(T^{(g-1)}) - M_a^{a^\star}(T^{(g)})$ is $o(\log T)$, in words that arm a is pulled while arm $a^\star \notin \widehat{\mathcal{S}}(t)$ at most $o(\log T)$ times from time $t = T^{(g)}, \ldots, T^{(g-1)}$.

We will see that $\frac{\log T^{(g-1)}}{\mathcal{K}_{\inf}(\nu_a, \mu_{a^\star})}$ pulls of arm a by time $T^{(g)}$ suffices to ensure this in both settings. Using that $(1-\delta)\log T^{(g)} \approx (1-\delta)^2 \log T^{(g-1)}$, it will follow that we can control the sum in Term B for each $a \in \mathcal{N}$ provided we choose $\delta \in (0,1)$ so that

$$(1-\delta)^2 \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu_m)} > \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu_{a^\star})} \text{ for all } a \in \mathcal{N}. \tag{24}$$

It is easy to check to for any $c \in (0,1)$, δ as defined in Lemma 12 satisfies this inequality. Note that $\mathcal{K}_{\inf}(\nu_a, \mu_{a^\star}) \geq \mathcal{K}_{\inf}(\nu_a, \mu_m)$, and thus $\delta \in (0,1)$. So far we have only considered suboptimal arms, but the fact that, for any $a \in \mathcal{M}$, Lemma 1 ensures that $N_a(T^{(g)}) > \log T^{(g)}/\epsilon$ with probability approaching 1 for *any* $\epsilon > 0$ shows that $a \in \mathcal{N}$ is indeed the harder case. Indeed, this is what we see in our proofs controlling Term B for the two algorithms.

□

We now conclude the analysis. Combining Equations (21) and (22) with the bounds on Term A and B obtained in Lemma 11 and Lemma 12 yield, for any finite G and for the

particular choice of $\delta \in (0,1)$ given in (23)

$$\limsup_T \frac{T/m - \mathbb{E}[N_{a^\star}(T)]}{\log T} \leq (1-\delta)^G \left[\sum_{a \in \mathcal{N}} \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu_m)} + \sum_{a \in \mathcal{M}} \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu_{a^\star})} \right]. \tag{25}$$

Taking G to infinity yields the result.

7 Conclusion

We have established the asymptotic efficiency of KL-CUCB and Thompson sampling for a generalization of the multiple play bandit problem that allows for random partial feedback on the chosen subset of arms. In our simulation, it appeared that Thompson sampling outperforms KL-CUCB. We provided some hypotheses following our simulation as to why this may be the case, but we have not at this point proven the regret superiority of Thompson sampling, where this regret superiority would occur in the $o(\log T)$ second-order term. Despite the strong performance of Thompson sampling for Bernoulli rewards, we have been able to prove stronger results about KL-CUCB in this work, dealing with more general distributions. Understanding for which distributions one of these algorithms is preferable to the other is an interesting area for future work.

In this work we considered a simple example of random partial feedback. Indeed, the sampling scheme is restricted so that the user can only observe ℓ arms, chosen at random out of the m arms selected. Other types of random partial feedback have recently been studied to account for user behavior in recommender's system (Kveton et al., 2015; Combes et al., 2015a; Lagrée et al., 2016), and we hope to extend our results to these models. Further, we will investigate whether our new proof technique can be used to study KL-CUCB and Thompson Sampling in more general combinatorial bandit models with semi-bandit feedback. We have shown that KL-CUCB attains comparable, if not better, performance than the recently proposed ESCB algorithm Combes et al. (2015b) in the particular case of multiple-plays bandits, even with a possibly random feedback. It would be interesting to investigate the possible gap of performance between these two algorithms in more general settings, as KL-CUCB is much easier to implement.

Acknowledgements

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR), under grant ANR-13-BS01-0005 (project SPADRO). Alex Luedtke gratefully acknowledges the support of a Berkeley Fellowship.

References

S Agrawal and N Goyal. Analysis of thompson sampling for the multi-armed bandit problem. *arXiv Prepr. arXiv1111.1797*, 2011.

S Agrawal and N Goyal. Further optimal regret bounds for thompson sampling. *arXiv Prepr. arXiv1209.3353*, 2012.

V Anantharam, P Varaiya, and J Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part I: IID rewards. *Autom. Control. IEEE Trans.*, 32(11):968–976, 1987.

J-Y Audibert, S Bubeck, and G Lugosi. Minimax policies for combinatorial prediction games. *arXiv Prepr. arXiv1105.4871*, 2011.

A N Burnetas and M Katehakis. Optimal adaptive policies for sequential allocation problems. *Adv. Appl. Math.*, 17(2):122–142, 1996.

O Cappé, A Garivier, O A Maillard, R Munos, and G Stoltz. Kullback–leibler upper confidence bounds for optimal sequential allocation. *Ann. Statist.*, 41(3):1516–1541, 2013a.

O Cappé, A Garivier, O A Maillard, R Munos, and G Stoltz. Kullback–leibler upper confidence bounds / supplemental argticle. *Ann. Stat.*, 41(3):1516–1541, 2013b.

N Cesa-Bianchi and G Lugosi. Combinatorial Bandits. *J. Comput. Syst. Sci.*, 78:1404–1422, 2012.

W Chen, Y Wang, and Y Yuan. Combinatorial multi-armed bandit: General framework and applications. In *Proc. 30th Int. Conf. Mach. Learn.*, pages 151–159, 2013.

R Combes, S Magureanu, A Proutière, and C Laroche. Learning to Rank: Regret Lower Bounds and Efficient Algorithms. In *Proc. 2015 ACM SIGMETRICS Int. Conf. Meas. Model. Comput. Syst.*, pages 231–244, 2015a.

R Combes, M S T M Shahi, A Proutiere, and M Lelarge. Combinatorial Bandits Revisited. In *Adv. Neural Inf. Process. Syst.*, pages 2107–2115, 2015b.

A Garivier, P Ménard, and G Stoltz. Explore First, Exploit Next: The True Shape of Regret in Bandit Problems. *arXiv Prepr. arXiv1602.07182*, 2016.

J C Gittins. Bandit processes and dynamic allocation indices. *J. R. Stat. Soc. Ser. B*, 41 (2):148–177, 1979.

E Kaufmann, O Cappé, and A Garivier. On Bayesian upper confidence bounds for bandit problems. In *Int. Conf. Artif. Intell. Stat.*, pages 592–600, 2012a.

E Kaufmann, N Korda, and R Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learn. Theory*, pages 199–213. Springer, 2012b.

J Komiyama, J Honda, and H Nakagawa. Optimal regret analysis of thompson sampling in stochastic multi-armed bandit problem with multiple plays. *arXiv Prepr. arXiv1506.00779*, 2015.

N Korda, E Kaufmann, and R Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Adv. Neural Inf. Process. Syst.*, pages 1448–1456, 2013.

B Kveton, C Szepesvári, Z Wen, and A Ashkan. Cascading Bandits: Learning to Rank in the Cascade Model. In *Proc. 32nd Int. Conf. Mach. Learn.*, pages 767–776, 2015.

P Lagrée, C Vernade, and O Cappé. Multiple-Play Bandits in the Postition-Based Model. *Prepr. arXiv1606.02448*, 2016.

T L Lai and H Robbins. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.*, 6(1):4–22, 1985.

H Robbins. Some aspects of the sequential design of experiments. *Bull. Am. Math. Soc.*, 58(5):527–535, 1952.

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

Appendix

We begin with an outline of the results proven in this appendix and how they are related to one another. Lemma 1 gives a lower bound on the number of draws of each suboptimal arm for a uniformly efficient algorithm. Deduced from Lemma 1, Theorem 2 gives an asymptotic regret lower bound (2) for a uniformly efficient algorithm. The asymptotic lower bound is achieved whenever the expected number of draws of each suboptimal arm satisfies the asymptotic condition (8) and the expected number of draws of each optimal arm away from the margin satisfies the asymptotic condition (9). Theorems 3 and 4 state that the variants of KL-CUCB are uniformly efficient and achieve (2) for rewards sampled either from a single parameter exponential family or from bounded and finitely supported distributions. Theorem 5 states that Thompson sampling is uniformly efficient and achieves (2) for Bernoulli distributed rewards.

The first step of the proof of Theorems 3, 4, and 5 consists in showing that KL-CUCB and Thompson sampling achieve the asymptotically optimal expected number of suboptimal arm draws, i.e. that (8) holds in their contexts. For KL-CUCB, this is a consequence of a preliminary analysis given in Lemmas A.1 and A.2. For Thompson sampling, this is a consequence of another preliminary analysis given in Lemmas 6 through 9. The proof of Lemma 9 relies on a link between the beta and binomial distributions given in Lemma A.3.

The second step of the proof of Theorems 3, 4, and 5 consists in showing that KL-CUCB and Thompson sampling are uniformly efficient in their respective contexts. This is a consequence of yet another preliminary analysis, (8), and Lemma 11.

The third step of the proof of Theorems 3, 4, and 5 consists in showing that KL-CUCB and Thompson sampling achieve the asymptotically optimal expected number of optimal draws away from the margin, i.e. that (9) holds in their contexts. This is a consequence of the preliminary analysis undertaken in step two and of Lemmas 11 and 12. The proofs of Lemmas 11 and 12 hinge on Lemmas 6 through 9. The proof of Lemma 12 also relies on Lemma A.3.

The fourth and final step of the proof of Theorems 3, 4, and 5 boils down to applying Theorem 2.

A Proof of Lower Bound on Suboptimal Arm Draws

Proof of Lemma 1. Fix some arm a with $\mu_a \leq \mu_m$, natural number T, and $\delta \in (0,1)$. Let $\mathcal{V}'$ be some distribution that is equal to $\mathcal{V}$ except in the a^{th} component, where its a^{th} component $\nu'_a \in \mathcal{D}$ is such that $\mu'_a \equiv E(\nu'_a) > \mu_m$ and $\nu_a \ll \nu'_a$. Such a distribution $\mathcal{V}'$ is guaranteed to exist by our assumption that $\mu_m < \sup\{E(\nu) : \nu \in \mathcal{D}\}$. Observe that $\mu'_a > \mu_a$ implies that $\text{KL}(\nu_a, \nu'_a) > 0$. Define the log-likelihood ratio random variable $L_a(T) \equiv L_{a,N_a(T)} \equiv \sum_{n=1}^{N_a(T)} \log \frac{d\nu_a}{d\nu'_a}(X_{a,n})$. Let $b_a(T) \equiv (1-\delta)\frac{\log T}{\text{KL}(\nu_a,\nu'_a)}$ and $c(T) \equiv (1-\delta/2)\log T$. We have that

$$\begin{aligned}
&\mathbb{P}_{\mathcal{V}}\left\{N_a(T) < b_a(T)\right\} \\
&\leq \mathbb{P}_{\mathcal{V}}\left\{N_a(T) < b_a(T), L_a(T) \leq c(T)\right\} + \mathbb{P}_{\mathcal{V}}\left\{N_a(T) < b_a(T), L_a(T) > c(T)\right\} \\
&\leq e^{c(T)}\,\mathbb{P}_{\mathcal{V}'}\left\{N_a(T) < b_a(T)\right\} + \mathbb{P}_{\mathcal{V}}\left\{N_a(T) < b_a(T), L_a(T) > c(T)\right\}, \qquad \text{(A.1)}
\end{aligned}$$

where the final inequality holds because, for any event $D \subseteq \{N_a(T) = b, L_a(T) \leq c(T)\}$, a change of measure shows that $\mathbb{P}_{\mathcal{V}}\{D\} = e^{L_{a,b}}\,\mathbb{P}_{\mathcal{V}'}\{D\} \leq e^{c(T)}\,\mathbb{P}_{\mathcal{V}'}\{D\}$ (see Equation 2.6 in Lai and Robbins, 1985). Using the uniform efficiency of the algorithm and the fact that arm a under the reward distribution involving ν'_a is either (i) part of a unique set of optimal arms or (ii) has mean larger than the m^{th} largest arm mean[5], Markov's inequality yields that,

$$\mathbb{P}_{\mathcal{V}'}\left\{N_a(T) < b_a(T)\right\} = \mathbb{P}_{\mathcal{V}'}\left\{\frac{\ell T}{m} - N_a(T) > \frac{\ell T}{m} - b_a(T)\right\} = o\left(T^{\delta/2-1}\right).$$

Thus, by the choice of $c(T)$, the first term in (A.1) converges to zero as $T \to \infty$. For the second term, observe that

$$\{N_a(T) < b_a(T), L_a(T) > c(T)\} \subseteq \left\{\max_{n \leq b_a(T)} \frac{L_{a,n}}{b_a(T)} > \frac{c(T)}{b_a(T)}\right\}$$

[5]For a general model $\mathcal{D}$, (i) does not imply (ii) and (ii) does not imply (i). For example, if $\{E(\nu) : \nu \in \mathcal{D}\} = \{0,1,2\}$, $k = 3$, $m = 2$, $a = 3$, and $(\mu_1, \mu_2, \mu_3) = (2,1,0)$, then $\mathcal{V}'$ must have mean vector $(2,1,2)$ so that (i) holds and (ii) does not. In the same scenario but with $(\mu_1, \mu_2, \mu_3) = (1,1,0)$, then $\mathcal{V}'$ must have mean vector $(1,1,2)$ so that (ii) holds and (i) does not.

$$= \left\{ \max_{n \le b_a(T)} \frac{L_{a,n}}{b_a(T)} > \frac{1-\delta/2}{1-\delta} \mathrm{KL}(\nu_a, \nu_a') > \mathrm{KL}(\nu_a, \nu_a') \right\}.$$

By the strong law of large numbers, $b_a(T)^{-1} L_{a,\lfloor b_a(T) \rfloor} \to \mathrm{KL}(\nu_a, \nu_a')$ almost surely under ν_a. Further, $\max_{n \le b_a(T)} b_a(T)^{-1} L_{a,n} \to \mathrm{KL}(\nu_a, \nu_a')$ almost surely as $T \to \infty$. It follows that the second term in (A.1) converges to zero as $T \to \infty$ so that

$$\mathbb{P}_{\mathcal{V}} \left\{ N_a(T) < (1-\delta) \frac{\log(T)}{\mathrm{KL}(\nu_a, \nu_a')} \right\} \to 0. \tag{A.2}$$

For convenience, we let $\mathcal{K} \equiv \mathcal{K}_{\inf}(\nu_a, \mu_m)$ in what follows. By the definition of the infimum, for every $\epsilon > 0$ there exists some ν_a' such that $\mathcal{K} + \epsilon > \mathrm{KL}(\nu_a, \nu_a')$. This proves (5). If $a \in \mathcal{N}$ so that $\mathcal{K} > 0$, then take $\epsilon = \left[(1-\delta)^{-1/2} - 1 \right] \mathcal{K}$ and write

$$\mathbb{P}_{\mathcal{V}} \left\{ N_a(T) < (1-\delta)^{3/2} \frac{\log(T)}{\mathcal{K}_{\inf}(\nu_a, \mu_m)} \right\} \to 0.$$

Applying the above to $\delta' = 1 - (1-\delta)^{2/3}$ (such that $(1-\delta')^{3/2} = (1-\delta)$) yield the result for $a \in \mathcal{N}$. For $a \in \mathcal{N}$, it also follows that for all $\delta \in (0,1)$ one has

$$\mathbb{E}[N_a(T)] \ge (1-\delta) \frac{\log T}{\mathcal{K}_{\inf}(\nu_a, \mu_m)} \mathbb{P}_{\mathcal{V}} \left\{ N_a(T) \ge (1-\delta) \frac{\log T}{\mathcal{K}_{\inf}(\nu_a, \mu_m)} \right\} \underset{T \to \infty}{\sim} (1-\delta) \frac{\log T}{\mathcal{K}_{\inf}(\nu_a, \mu_m)},$$

which yields (6), letting δ go to zero. □

Proof of Theorem 2. The proof follows from a new regret decomposition.

$$\begin{aligned}
R(T) &= \sum_{a^\star \in \mathcal{L}} \Delta_{a^\star} \mathbb{E}[N_{a^\star}(T)] + \sum_{a^\star \in \mathcal{M}} \Delta_m \mathbb{E}[N_{a^\star}(T)] + \sum_{a \in \mathcal{N}} \Delta_a \mathbb{E}[N_a(T)] \\
&= \sum_{a^\star \in \mathcal{L}} \Delta_{a^\star} \mathbb{E}[N_{a^\star}(T)] + \Delta_m \Big(\ell T - \sum_{a^\star \in \mathcal{L}} \mathbb{E}[N_{a^\star}(T)] - \sum_{a \in \mathcal{N}} \mathbb{E}[N_a(T)] \Big) + \sum_{a \in \mathcal{N}} \Delta_a \mathbb{E}[N_a(T)] \\
&= \sum_{a^\star \in \mathcal{L}} (\Delta_{a^\star} - \Delta_m) \mathbb{E}[N_{a^\star}(T)] + \Delta_m \ell T + \sum_{a \in \mathcal{N}} (\Delta_a - \Delta_m) \mathbb{E}[N_a(T)] \\
&= \sum_{a^\star \in \mathcal{L}} (\mu_{a^\star} - \mu_m) \left(\frac{\ell T}{m} - \mathbb{E}[N_{a^\star}(T)] \right) + \frac{\ell T}{m} \sum_{a^\star \in \mathcal{L}} (\Delta_{a^*} - \Delta_m) + \Delta_m \ell T \\
&\quad + \sum_{a \in \mathcal{N}} (\mu_m - \mu_a) \mathbb{E}[N_a(T)].
\end{aligned}$$

Now using that $\Delta_m = \Delta_{a^\star}$ for all $a^\star \in \mathcal{S}^\star \backslash \mathcal{L} \subseteq \mathcal{M}$ and that $\sum_{a^\star \in \mathcal{S}^\star} \Delta_{a^\star} = 0$, one can write

$$\sum_{a^\star \in \mathcal{L}} (\Delta_{a^\star} - \Delta_m) = \sum_{a^\star \in \mathcal{S}^\star} (\Delta_{a^\star} - \Delta_m) = -m\Delta_m$$

and it follows that

$$R(T) = \sum_{a^\star \in \mathcal{L}} (\mu_{a^\star} - \mu_m) \left(\frac{\ell T}{m} - \mathbb{E}[N_{a^\star}(T)] \right) + \sum_{a \in \mathcal{N}} (\mu_m - \mu_a)\, \mathbb{E}[N_a(T)]. \tag{A.3}$$

As the first term in this decomposition is positive, the lower bound easily follows from the lower bound on the expected number of draws of an arm a in $\mathcal{N}$ obtained in (6):

$$\liminf_{T\to\infty} \frac{R(T)}{\log(T)} \geq \sum_{a \in \mathcal{N}} (\mu_m - \mu_a) \liminf_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\log(T)} \geq \sum_{a \in \mathcal{N}} \frac{\mu_m - \mu_a}{\mathcal{K}_{\inf}(\nu_a, \mu_m)}.$$

Still using (A.3), a strategy satisfying (8) and (9) is asymptotically optimal. □

B Supplementary Proofs for KL-CUCB

Lemma A.1. *Fix arms a and $a^\star$ with $\mu_a < \mu_{a^\star}$. In the setting of Theorem 3 with $\mu^\dagger = \mu_{a^\star}$ or in the setting of Theorem 4 with $\mu^\dagger = \left[1 - \log(T)^{-1/5}\right] \mu_{a^\star}$, it holds that*

$$\sum_{n=b(T)+1}^{\infty} \mathbb{P}\left\{ \hat{\nu}_{a,n} \in \mathcal{C}_{\mu^\dagger, f(T)/n} \right\} = o(\log T),$$

where $b(T)$ is any number satisfying

$$b(T) \geq \left\lceil \frac{f(T)}{\mathcal{K}_{\inf}(\mu_a, \mu_{a^\star})} \right\rceil.$$

An explicit finite sample bound on the $o(\log T)$ term can be found in Cappé et al. (2013b).

Proof. In the setting of Theorem 3, Equation 25 in Cappé et al. (2013b) gives the result for $\mu^\dagger = \mu_{a^\star}$. We refer the readers to that equation for the explicit finite sample bound that we are summarizing with little-oh notation.

In the setting of Theorem 4, Equation 33 combined with the unnumbered equation

preceding Equation 36 in Section B.4 of Cappé et al. (2013b) gives the result for $\mu^\dagger = \left[1 - \log(T)^{-1/5}\right] \mu_{a^\star}$. An explicit finite sample upper bound on this quantity can be found in Section B.4 of Cappé et al. (2013b). □

Lemma A.2. *Fix an arm $a^\star \in \mathcal{S}^\star$. In the setting of Theorem 3 with $\mu^\dagger \leq \mu_{a^\star}$ or in the setting of Theorem 4 with $\mu^\dagger \leq \left[1 - \log(T)^{-1/5}\right] \mu_{a^\star}$, it holds that*

$$\sum_{t=k}^{T-1} \mathbb{P}\left\{\mu^\dagger \geq U_{a^\star}(t)\right\} = o(\log T).$$

Explicit finite sample constants can be found in the proof.

Proof. In the setting of Theorem 3, it holds that $\{\mu^\dagger \geq U_{a^\star}(t)\} \subseteq \{\mu_{a^\star} \geq U_{a^\star}(t)\}$. Hence,

$$\sum_{t=k}^{T-1} \mathbb{P}\left\{\mu^\dagger \geq U_{a^\star}(t)\right\} \leq \sum_{t=k}^{T-1} \mathbb{P}\{\mu_{a^\star} \geq U_{a^\star}(t)\}.$$

Furthermore,

$$\{\mu_{a^\star} \geq U_{a^\star}(t)\} \subseteq \bigcup_{n=\ell}^{t-k+\ell} \left\{\mu_{a^\star} \geq \hat{\mu}_{a^\star,n}, \mathrm{KL}(\hat{\mu}_{a^\star,n}, \mu_{a^\star}) \geq \frac{f(t)}{n}\right\}.$$

Using the above, Equations 17 and 18 in Cappé et al. (2013b) show that $\sum_{t=k}^{T-1} \mathbb{P}\{\mu_{a^\star} \geq U_{a^\star}(t)\}$ is upper bounded by $3 + 4e \log\log T = o(\log T)$ provided $T \geq 3$. We note that the union above is over $n = \ell, \ldots, t-k+\ell$ rather than $n = 1, \ldots, t-k+1$ as was used in Equation 17 of Cappé et al. (2013b), but that the bound in their Equation 17 still holds in our setting.

In the setting of Theorem 4, it holds that $\{\mu^\dagger \geq U_{a^\star}(t)\} \subseteq \{\left[1 - \log(T)^{-1/5}\right] \mu_{a^\star} \geq U_{a^\star}(t)\}$. Hence,

$$\sum_{t=k}^{T-1} \mathbb{P}\left\{\mu^\dagger \geq U_{a^\star}(t)\right\} \leq \sum_{t=k}^{T-1} \mathbb{P}\left\{\left[1 - \log(T)^{-1/5}\right] \mu_{a^\star} \geq U_{a^\star}(t)\right\}. \tag{A.4}$$

Let $\epsilon \equiv \log(T)^{-1/5} \mu_{a^\star} > 0$. Arguments given in Section B.2 of Cappé et al. (2013b) show

that

$$\begin{aligned}\{\mu_{a^\star} - \epsilon \geq U_{a^\star}(t)\} &\subseteq \left\{\mathcal{K}_{\inf}(\hat{\nu}_{a^\star}(t), \mu_{a^\star} - \epsilon) \geq \frac{f(t)}{N_{a^\star}(t)}\right\} \\ &\subseteq \left\{\mathcal{K}_{\inf}(\hat{\nu}_{a^\star}(t), \mu_{a^\star}) \geq \frac{f(t)}{N_{a^\star}(t)} + \frac{\epsilon^2}{2}\right\} \\ &\subseteq \cup_{n=\ell}^{t-k+\ell} \left\{\mathcal{K}_{\inf}(\hat{\nu}_{a^\star,n}, \mu_{a^\star}) \geq \frac{f(t)}{n} + \frac{\epsilon^2}{2}\right\}.\end{aligned}$$

The remainder of the proof is now the same as in Cappé et al. (2013b). In particular, their Equation 26 combined with the bounds given after their Equation 35 shows that the right-hand side of (A.4) is upper bounded by $36\mu_{a^\star}^{-4}(2 + \log\log T)(\log T)^{4/5} = o(\log T)$. Again we note that their Equation 26 holds despite the union above being over $n = \ell, \ldots, t-k+\ell$ rather than $n = 1, \ldots, t-k+1$ as in Cappé et al. (2013b). □

Proof of Lemma 11 for KL-CUCB in the settings of Theorems 3 and 4. Fix $a \in \mathcal{M}$. For ease of notation, we analyze $\mathbb{E}[M_a^{a^\star}(T)]$ rather than $\mathbb{E}[M_a^{a^\star}(T^{(G)})]$, but for fixed $G < \infty$ there is no loss of generality in doing so. In the setting of Theorem 3 we let $\mu^\dagger = \mu_{a^\star}$, and in the setting of Theorem 4 we let $\mu^\dagger = \left[1 - \log(T)^{-1/5}\right]\mu_{a^\star}$. The same arguments as for (13) show that

$$\begin{aligned}&\left\{a \in \mathcal{A}(t+1), a^\star \notin \widehat{\mathcal{S}}(t)\right\} \\ &\subseteq \left\{a \in \mathcal{A}(t+1), a^\star \notin \widehat{\mathcal{S}}(t), \mu^\dagger \geq U_a(t)\right\} \cup \left\{a \in \mathcal{A}(t+1), \mu^\dagger < U_a(t), a^\star \notin \widehat{\mathcal{S}}(t)\right\} \\ &\subseteq \left\{\mu^\dagger \geq U_{a^\star}(t)\right\} \cup \left\{a \in \mathcal{A}(t+1), \mu^\dagger < U_a(t)\right\}. \qquad \text{(A.5)}\end{aligned}$$

Let

$$b_a^{a^\star}(T) \equiv \left\lceil \frac{f(T)}{\mathcal{K}_{\inf}(\nu_a, \mu_{a^\star})} \right\rceil.$$

Similarly to (15), we have that

$$\mathbb{E}[M_a^{a^\star}(T)] \leq \frac{f(T)}{\mathcal{K}_{\inf}(\nu_a, \mu_{a^\star})} + \sum_{n=b_a^{a^\star}(T)+1}^{\infty} \mathbb{P}\left\{\hat{\nu}_{a,n} \in \mathcal{C}_{\mu^\dagger, f(T)/n}\right\} + \sum_{t=k}^{T-1} \mathbb{P}\left\{\mu^\dagger \geq U_{a^\star}(t)\right\} + 2.$$

By Lemmas A.1 and A.2, $\mathbb{E}[M_a^{a^\star}(T)] \leq \frac{\log T}{\mathcal{K}_{\inf}(\nu_a, \mu_{a^\star})} + o(\log T)$. Replacing T by $T^{(G)}$ (for T

large enough so that $T^{(G)} > 1$) gives $\mathbb{E}[M_a^{a^\star}(T^{(G)})] \leq (1-\delta)^G \frac{\log T}{\mathcal{K}_{\inf}(\nu_a, \mu_{a^\star})} + o(\log T)$. $\square$

Proof of Lemma 12 for KL-CUCB in the settings of Theorems 3 and 4. Fix $g \in \mathbb{N}$, a with $\mu_a \leq \mu_m$, and $T^{(g)}$ such that $T^{(g)} > 1$. In the setting of Theorem 3 let $\mu^\dagger = \mu_{a^\star}$, and in the setting of Theorem 4 let $\mu^\dagger = [1 - \log(T)^{-1/5}]\mu_{a^\star}$. By (A.5) and the fact that $\{\mu^\dagger < U_a(t)\} = \left\{\hat{\nu}_{a,N_a(t)} \in \mathcal{C}_{\mu^\dagger, f(t)/N_a(t)}\right\}$,

$$\begin{aligned}
&\mathbb{E}[M_a^{a^\star}(T^{(g-1)}) - M_a^{a^\star}(T^{(g)})] \\
&\leq \sum_{t=T^{(g)}}^{T^{(g-1)}-1} \mathbb{P}\left\{\mu^\dagger \geq U_{a^\star}(t)\right\} + \sum_{t=T^{(g)}}^{T^{(g-1)}-1} \mathbb{P}\left\{a \in \mathcal{A}(t+1), \hat{\nu}_{a,N_a(t)} \in \mathcal{C}_{\mu^\dagger, f(t)/N_a(t)}\right\}.
\end{aligned}$$

The first term in the right hand side is upper bounded by the same sum from $t = k$ to $T-1$, and is thus $o(\log T)$ by Lemma A.2. For the second term, let $b'_a(T,g) \equiv \lceil (1-\delta)\frac{f(T^{(g)})}{\mathcal{K}_{\inf}(\nu_a,\mu_m)} \rceil$ if $a \in \mathcal{N}$ and let $b'_a(T,g) \equiv \lceil \frac{f(T^{(g)})}{(1-\delta)\mathcal{K}_{\inf}(\nu_a,\mu_{a^\star})} \rceil$ if $a \in \mathcal{M}$. Similar arguments to those used to derive (14) in Section 6.1 show that, for T large enough so that $T^{(g)} \geq k$,

$$\begin{aligned}
&\sum_{t=T^{(g)}}^{T^{(g-1)}-1} \mathbb{P}\left\{a \in \mathcal{A}(t+1), \hat{\nu}_{a,N_a(t)} \in \mathcal{C}_{\mu^\dagger, f(t)/N_a(t)}\right\} \\
&\qquad \leq \sum_{n=\ell}^{T^{(g-1)}-k} \sum_{t=T^{(g)}}^{T^{(g-1)}-1} \mathbb{P}\left\{\hat{\nu}_{a,n} \in \mathcal{C}_{\mu^\dagger, f(T^{(g-1)})/n}, \tau_{a,n+1} = t+1\right\}.
\end{aligned}$$

We split the sum over n into a sum S_1 from $n = \ell$ to $b'_a(T,g)$ and a sum S_2 from $n = b'_a(T,g)+1$ to $T^{(g-1)} - k$. For the latter sum, the fact that, for each n, $\tau_{a,n+1} = t+1$ for at most one t in a given interval, yields that

$$S_2 \leq \sum_{n=b'_a(T,g)+1}^{T^{(g-1)}-k} \mathbb{P}\left\{\hat{\nu}_{a,n} \in \mathcal{C}_{\mu^\dagger, f(T^{(g-1)})/n}\right\}.$$

If $a \in \mathcal{N}$, then δ satisfying (24) yields that $b'_a(T,g) > \frac{f(T^{(g-1)})}{\mathcal{K}_{\inf}(\nu_a,\mu_{a^\star})}$, and so the above sum is $o(\log T)$ by Lemma A.1. If $a \in \mathcal{M}$, then $b'_a(T,g) = \lceil \frac{f(T^{(g-1)})}{\mathcal{K}_{\inf}(\nu_a,\mu_{a^\star})} \rceil$, and so again the above sum is $o(\log T)$.

We now bound S_1. Note that if $N_a(T^{(g)} - 1) > b'_a(T,g)$, then, for every $n \leq b'_a(T,g)$,

$\tau_{a,n+1} < T^{(g)}$ and $S_1 = 0$ (the sum over t is void). Therefore,

$$\begin{aligned} S_1 &\leq \sum_{n=\ell}^{b'_a(T,g)} \mathbb{P}\left\{N_a(T^{(g)}-1) \leq b'_a(T,g)\right\} \\ &= [b'_a(T,g) - \ell + 1]\,\mathbb{P}\left\{N_a(T^{(g)}-1) \leq b'_a(T,g)\right\} \leq b'_a(T,g)\,\mathbb{P}\left\{N_a(T^{(g)}-1) \leq b'_a(T,g)\right\}. \end{aligned}$$

From Lemma 13 KL-CUCB is uniformly efficient. Thus by (5), for any $a \in \mathcal{M}$ one has

$$\lim_{T\to\infty} \mathbb{P}\left(N_a(T^{(g)}-1) \leq \frac{2\log(T^{(g)})}{(1-\delta)\,\mathcal{K}_{\inf}(\nu_a, \mu_{a^\star})}\right) = 0,$$

where we use the fact that $\mathcal{K}_{\inf}(\nu_a, \mu_m) = 0$ and choose $\epsilon = (1-\delta)\,\mathcal{K}_{\inf}(\nu_a, \mu_{a^\star})/2 > 0$. This yields that $\mathbb{P}\left\{N_a\left(T^{(g)}-1\right) < b'_a(T,g)\right\} \to 0$ as $T \to \infty$ and $S_1 = o(\log T)$.

If $a \in \mathcal{N}$, then Lemma 13 and (5) from Lemma 1 yield that

$$\mathbb{P}\left\{N_a\left(T^{(g)}-1\right) < (1-\delta)\frac{\log T^{(g)}}{\mathcal{K}_{\inf}(\nu_a, \mu_m)}\right\} \to 0 \text{ as } T \to 0.$$

The fact that $\lim_T f(T)/\log T = 1$ shows that $b'_a(T,g) = (1-\delta)\frac{\log T^{(g)}}{\mathcal{K}_{\inf}(\nu_a,\mu_m)} + o(\log T)$. Plugging this into (5) from Lemma 1 (which holds for *every* δ between 0 and 1) yields that $\mathbb{P}\left\{N_a\left(T^{(g)}-1\right) < b'_a(T,g)\right\} \to 0$ as $T \to \infty$. It follows that $S_1 = o(\log T)$.

We have then shown that $\mathbb{E}[M_a^{a^\star}(T^{(g-1)}) - M_a^{a^\star}(T^{(g)})] = o(\log T)$ for each a with $\mu_a \leq \mu_m$ and each $g \leq G$. As Term B is a sum of finitely many such terms, Term B is $o(\log T)$. □

C Supplementary Proofs for Thompson Sampling

We begin with a lemma.

Lemma A.3. *For any fixed real number L, arm a, $\mu_a < \mu^\dagger < \theta^\dagger$, and $t \geq 1$,*

$$I\left\{\hat{\mu}_a(t) \leq \mu^\dagger, N_a(t) \geq L\right\}\mathbb{P}\left\{\theta_a(t) > \theta^\dagger \middle| \mathcal{F}_t\right\} \leq e^{-(L+1)\,\mathrm{KL}(\mu^\dagger, \theta^\dagger)}.$$

Proof. From Fact 3 in Agrawal and Goyal (2012) (also used in Agrawal and Goyal, 2011;

Kaufmann et al., 2012a,b),

$$\mathbb{P}\left(\theta_a(t) > \theta^\dagger \middle| \mathcal{F}(t)\right) = \mathbb{P}\left(\sum_{n=1}^{N_a(T)+1} Z_n \leq \sum_{n=1}^{N_a(T)} \mathbb{1}\{X_{a,n} = 0\} \middle| \mathcal{F}(t)\right),$$

where $\{Z_n\}$ is an i.i.d. sequence (independent of all other quantities under consideration) of Bernoulli random variables with mean $\theta^\dagger$. Upper bounding the right-hand side yields

$$\mathbb{P}\left(\theta_a(t) > \theta^\dagger \middle| \mathcal{F}(t)\right) \leq \mathbb{P}\left(\frac{1}{N_a(T)+1} \sum_{n=1}^{N_a(T)+1} Z_n \leq \hat{\mu}_a(T) \middle| \mathcal{F}(t)\right).$$

Using that $\mu^\dagger < \theta^\dagger$, the Chernoff-Hoeffding bound gives that $\mathbb{P}\left(\theta_a(t) > \theta^\dagger \middle| \mathcal{F}(t)\right)$ is no larger than $e^{-[N_a(t)+1]\,\mathrm{KL}(\hat{\mu}_a(t),\theta^\dagger)}$. Multiplying the left-hand side by $I\left\{\hat{\mu}_a(t) \leq \mu^\dagger, N_a(t) \geq L\right\}$, this yields the upper bound $e^{-(L+1)\,\mathrm{KL}(\mu^\dagger,\theta^\dagger)}$. □

Proof of Lemma 6. Let $\theta_{a^\star,(m+1)}(t)$ denote the $(m+1)^{\text{th}}$ largest element in $\{\theta_{\tilde{a}}(t) : \tilde{a} \neq a^\star\}$. We first define the event $B \equiv \left\{\theta_a(t) \leq \theta^\dagger, \theta_a(t) > \theta_{a^\star,(m+1)}(t)\right\}$. Observe that

$$\begin{aligned}\mathbb{P}\left\{a \in \widehat{\mathcal{S}}(t), \theta_a(t) \leq \theta^\dagger, a^\star \notin \widehat{\mathcal{S}}(t) \middle| \mathcal{F}(t)\right\} &= \mathbb{P}\left(\left\{a \in \widehat{\mathcal{S}}(t), a^\star \notin \widehat{\mathcal{S}}(t)\right\} \cap B \middle| \mathcal{F}(t)\right) \\ &\leq \mathbb{P}\left(\left\{\theta_{a^\star}(t) \leq \theta^\dagger\right\} \cap B \middle| \mathcal{F}(t)\right). \qquad \text{(A.6)}\end{aligned}$$

The event $\{\theta_{a^\star}(t) > \theta^\dagger\}$ is independent of the event B conditional on $\mathcal{F}(t)$, and so the fact that $\left\{\theta_{a^\star}(t) > \theta^\dagger\right\} \cap B \subseteq \left\{a^\star \in \widehat{\mathcal{S}}(t)\right\}$ yields

$$\mathbb{P}(B|\mathcal{F}(t)) \leq \frac{\mathbb{P}\left(a^\star \in \widehat{\mathcal{S}}(t) \middle| \mathcal{F}(t)\right)}{\mathbb{P}\left(\theta_{a^\star}(t) > \theta^\dagger | \mathcal{F}(t)\right)}.$$

We note that $\mathbb{P}\left(\theta_{a^\star}(t) > \theta^\dagger \middle| \mathcal{F}(t)\right)$ is positive (a beta distribution with at least one success is larger than $\theta^\dagger < 1$ with positive probability). Finally, since $a \in \mathcal{A}(t+1)$ implies $a \in \widehat{\mathcal{S}}(t)$, (A.6) yields

$$\begin{aligned}\mathbb{P}\left(a \in \mathcal{A}(t+1), \theta_a(t) \leq \theta^\dagger, a^\star \notin \widehat{\mathcal{S}}(t) \middle| \mathcal{F}(t)\right) &\leq \mathbb{P}\left(\left\{\theta_{a^\star}(t) \leq \theta^\dagger\right\} \cap B \middle| \mathcal{F}(t)\right) \\ &= \mathbb{P}\left(\theta_{a^\star}(t) \leq \theta^\dagger \middle| \mathcal{F}(t)\right) \mathbb{P}\left(B | \mathcal{F}(t)\right)\end{aligned}$$

$$\leq \mathbb{P}\left(\theta_{a^\star}(t) \leq \theta^\dagger \middle| \mathcal{F}(t)\right) \frac{\mathbb{P}\left(a^\star \in \widehat{\mathcal{S}}(t) \middle| \mathcal{F}(t)\right)}{\mathbb{P}\left(\theta_{a^\star}(t) > \theta^\dagger \middle| \mathcal{F}(t)\right)}.$$

□

Proof of Lemma 7. Using (2), one can write

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{T-1} \frac{1-p_{a^\star}^{\theta^\dagger}(t)}{p_{a^\star}^{\theta^\dagger}(t)} \mathbb{P}\left(a^\star \in \widehat{\mathcal{S}}(t) \middle| \mathcal{F}(t)\right)\right] \\
&= \frac{m}{\ell}\mathbb{E}\left[\sum_{t=0}^{T-1} \frac{1-p_{a^\star}^{\theta^\dagger}(t)}{p_{a^\star}^{\theta^\dagger}(t)} \mathbb{P}\left(a^\star \in \mathcal{A}(t+1) | \mathcal{F}(t)\right)\right] \\
&= \frac{m}{\ell}\mathbb{E}\left[\sum_{t=0}^{T-1} \frac{1-p_{a^\star,N_{a^\star}(t)}}{p_{a^\star,N_{a^\star}(t)}} \mathbb{1}\left\{a^\star \in \mathcal{A}(t+1)\right\}\right] \\
&= \frac{m}{\ell}\mathbb{E}\left[\sum_{t=0}^{T-1}\sum_{n=0}^{T-1} \frac{1-p_{a^\star,n}^{\theta^\dagger}}{p_{a^\star,n}^{\theta^\dagger}} \mathbb{1}\left\{\tau_{a^\star,n+1} = t+1\right\}\right] \\
&\leq \frac{m}{\ell}\mathbb{E}\left[\sum_{n=0}^{T-1} \frac{1-p_{a^\star,n}^{\theta^\dagger}}{p_{a^\star,n}^{\theta^\dagger}}\right],
\end{aligned}$$

where the latter inequality holds because $\tau_{a^\star,n+1} = t+1$ for at most one t in $\{0,\ldots,T-1\}$. □

Proof of Lemma 9. Let $L^\dagger(T) \equiv \frac{\log T}{\mathrm{KL}(\mu^\dagger,\theta^\dagger)}$. We have that

$$\begin{aligned}
&\sum_{t=0}^{T-1} \mathbb{P}\left\{a \in \mathcal{A}(t+1), \theta_a(t) > \theta^\dagger, \hat{\mu}_a(t) \leq \mu^\dagger\right\} \\
&\quad= \mathbb{E}\left[\sum_{t=0}^{T-1} \mathbb{1}\left\{N_a(t) < L^\dagger(T) - 1, \hat{\mu}_a(t) \leq \mu^\dagger\right\} \mathbb{P}\left(a \in \mathcal{A}(t+1), \theta_a(t) > \theta^\dagger \middle| \mathcal{F}(t)\right)\right] \\
&\qquad+ \mathbb{E}\left[\sum_{t=0}^{T-1} \mathbb{1}\left\{N_a(t) \geq L^\dagger(T) - 1, \hat{\mu}_a(t) \leq \mu^\dagger\right\} \mathbb{P}\left(a \in \mathcal{A}(t+1), \theta_a(t) > \theta^\dagger \middle| \mathcal{F}(t)\right)\right] \\
&\quad\leq \mathbb{E}\left[\sum_{t=0}^{T-1} \mathbb{1}\left\{N_a(t) < L^\dagger(T) - 1\right\} \mathbb{P}\left(a \in \mathcal{A}(t+1) \middle| \mathcal{F}(t)\right)\right] \\
&\qquad+ \mathbb{E}\left[\sum_{t=0}^{T-1} \mathbb{1}\left\{N_a(t) \geq L^\dagger(T) - 1, \hat{\mu}_a(t) \leq \mu^\dagger\right\} \mathbb{P}\left(\theta_a(t) > \theta^\dagger \middle| \mathcal{F}(t)\right)\right].
\end{aligned} \tag{A.7}$$

The first term in the right hand side equals $\mathbb{E}\left[\sum_{t=0}^{T-1} \mathbb{1}\left\{N_a(t) < L^\dagger(T) - 1, a \in \mathcal{A}(t+1)\right\}\right]$. Hence it is no larger than $L^\dagger(T) - 1$ (the sum has at most $L^\dagger(T) - 1$ nonzero terms). For

the second term, Lemma A.3 yields

$$\mathbb{1}\left\{\hat{\mu}_a(t) < \mu^\dagger, N_a(T) \geq L^\dagger(T) - 1\right\} \mathbb{P}\left(\theta_a(t) > \theta^\dagger \middle| \mathcal{F}(t)\right) \leq e^{-L^\dagger(T)\,\mathrm{KL}(\mu^\dagger, \theta^\dagger)} = T^{-1}.$$

It follows that the second term on the right of (A.7) is upper bounded by $\sum_{t=0}^{T-1} T^{-1} = 1$. This completes the proof. □

Proof of Lemma 11 for Thompson sampling in the setting of Theorem 5. Fix $a \in \mathcal{M}$ and $\epsilon \in [0, 1($. For ease of notation, we analyze $\mathbb{E}[M_a^{a^\star}(T)]$ rather than $\mathbb{E}[M_a^{a^\star}(T^{(G)})]$, but for fixed $G < \infty$ there is no loss of generality in doing so. Let $\mu^\dagger$ and $\theta^\dagger$ satisfy $\mu_a < \mu^\dagger < \theta^\dagger < \mu_{a^\star}$ (exact quantities to be specified at the end of the proof). Note that

$$\begin{aligned}
&\left\{a \in \mathcal{A}(t+1), a^\star \notin \widehat{\mathcal{S}}(t)\right\} \\
&\subseteq \left\{a \in \mathcal{A}(t+1), \theta_a(t) \leq \theta^\dagger, a^\star \notin \widehat{\mathcal{S}}(t)\right\} \cup \left\{a \in \mathcal{A}(t+1), \theta_a(t) > \theta^\dagger\right\}.
\end{aligned}$$

Recalling that $\mathbb{E}[M_a^{a^\star}(T)]$ is equal to $\sum_{t=0}^{T-1} \mathbb{P}\left\{a \in \mathcal{A}(t+1), a^\star \notin \widehat{\mathcal{S}}(t)\right\}$, the above yields

$$\begin{aligned}
\mathbb{E}[M_a^{a^\star}(T)] \leq &\sum_{t=0}^{T-1} \mathbb{P}\left\{a \in \mathcal{A}(t+1), \theta_a(t) \leq \theta^\dagger, a^\star \notin \widehat{\mathcal{S}}(t)\right\} \\
&+ \sum_{t=0}^{T-1} \mathbb{P}\left\{a \in \mathcal{A}(t+1), \hat{\mu}_a(t) > \mu^\dagger\right\} \\
&+ \sum_{t=0}^{T-1} \mathbb{P}\left\{a \in \mathcal{A}(t+1), \theta_a(t) > \theta^\dagger, \hat{\mu}_a(t) \leq \mu^\dagger\right\}. \qquad \text{(A.8)}
\end{aligned}$$

Note that the right-hand side of the above is almost identical to (16). Recall that $p_{a^\star,n}^{\theta^\dagger}$ denotes the probability that a posterior draw of the mean of arm $a^\star$ is greater than $\theta^\dagger$ after n draws of arm $a^\star$. Note that all of the results used to control the three terms on the right-hand side of (16) hold for any a with $\mu_a \leq \mu_m$ provided $\mu_a < \mu^\dagger < \theta^\dagger < \mu_{a^\star}$. In particular, we are referring to Lemma 6, (17), Lemma 8, (18), Lemma 10, and Lemma 9. Hence, $\mathbb{E}[M_a^{a^\star}(T)] \leq \frac{\log T}{\mathrm{KL}(\mu^\dagger, \theta^\dagger)} + o(\log T)$. Selecting $\mu^\dagger$ and $\theta^\dagger$ as in the proof of (8) from Theorem 5, with the only difference being that we replace μ_m by $\mu_{a^\star}$, yields $\mathbb{E}[M_a^{a^\star}(T)] \leq (1+\epsilon)^2 \frac{\log T}{\mathrm{KL}(\mu_a, \mu_{a^\star})} + o(\log T)$. As ϵ was arbitrary, dividing both sides by $\log T$ and taking $T \to \infty$ followed by $\epsilon \to 0$ yields that $\mathbb{E}[M_a^{a^\star}(T)] \leq \frac{\log T}{\mathrm{KL}(\mu_a, \mu_{a^\star})} + o(\log T)$. Replacing T by $T^{(G)}$ (for T large enough so that

$T^{(G)} > 1)$ gives $\mathbb{E}[M_a^{a^\star}(T^{(G)})] \le (1-\delta)^G \frac{\log T}{\mathcal{K}_{\inf}(\nu_a, \mu_{a^\star})} + o(\log T)$. $\square$

Proof of Lemma 12 for Thompson sampling in the setting of Theorem 5. Fix $g \in \mathbb{N}$, an arm a with $\mu_a \le \mu_m$, and $T^{(g)}$ such that $T^{(g)} > 1$. Let $\mu^\dagger$ and $\theta^\dagger$ satisfy $\mu_a < \mu^\dagger < \theta^\dagger < \mu_{a^\star}$ and $\mathrm{KL}(\mu^\dagger, \theta^\dagger) \ge (1-\delta)\,\mathrm{KL}(\mu_a, \mu_{a^\star})$. By the same arguments used for (A.8),

$$\begin{aligned}
\mathbb{E}[M_a^{a^\star}(T^{(g-1)}) - M_a^{a^\star}(T^{(g)})] &= \sum_{t=T^{(g)}}^{T^{(g-1)}-1} \mathbb{P}\left\{a \in \mathcal{A}(t+1), a^\star \notin \widehat{\mathcal{S}}(t)\right\} \\
&\le \sum_{t=T^{(g)}}^{T^{(g-1)}-1} \mathbb{P}\left\{a \in \mathcal{A}(t+1), \theta_a(t) \le \theta^\dagger, a^\star \notin \widehat{\mathcal{S}}(t)\right\} \\
&\quad + \sum_{t=T^{(g)}}^{T^{(g-1)}-1} \mathbb{P}\left\{a \in \mathcal{A}(t+1), \hat{\mu}_a(t) > \mu^\dagger\right\} \\
&\quad + \sum_{t=T^{(g)}}^{T^{(g-1)}-1} \mathbb{P}\left\{a \in \mathcal{A}(t+1), \theta_a(t) > \theta^\dagger, \hat{\mu}_a(t) \le \mu^\dagger\right\}. \qquad \text{(A.9)}
\end{aligned}$$

The first two sums are trivially upper bounded by the sums from $t = 0$ to $T-1$, and thus are $o(\log T)$ by Lemma 6, (17), Lemma 8, (18), and Lemma 10. If $a \in \mathcal{N}$, then let $b_a(T,g) \equiv (1-\delta)\frac{\log T^{(g)}}{\mathrm{KL}(\mu_a, \mu_m)}$, and if $a \in \mathcal{M}$ then let $b_a(T,g) \equiv \frac{\log T^{(g)}}{(1-\delta)\,\mathrm{KL}(\mu_a, \mu_{a^\star})}$. We have that

$$\begin{aligned}
&\sum_{t=T^{(g)}}^{T^{(g-1)}-1} \mathbb{P}\left\{a \in \mathcal{A}(t+1), \theta_a(t) > \theta^\dagger, \hat{\mu}_a(t) \le \mu^\dagger\right\} \\
&= \mathbb{E}\left[\sum_{t=T^{(g)}}^{T^{(g-1)}-1} \mathbb{1}\left\{\hat{\mu}_a(t) \le \mu^\dagger, N_a(t) \ge b_a(T,g)\right\} \mathbb{P}\left\{a \in \mathcal{A}(t+1), \theta_a(t) > \theta^\dagger \middle| \mathcal{F}(t)\right\}\right] \\
&\quad + \mathbb{E}\left[\sum_{t=T^{(g)}}^{T^{(g-1)}-1} \mathbb{1}\left\{\hat{\mu}_a(t) \le \mu^\dagger, N_a(t) < b_a(T,g)\right\} \mathbb{P}\left\{a \in \mathcal{A}(t+1), \theta_a(t) > \theta^\dagger \middle| \mathcal{F}(t)\right\}\right]. \qquad \text{(A.10)}
\end{aligned}$$

If $a \in \mathcal{N}$, then Lemma A.3 and $\mathrm{KL}(\mu^\dagger, \theta^\dagger) \ge (1-\delta)\,\mathrm{KL}(\mu_a, \mu_{a^\star})$ yield that the first term on the right is upper bounded by

$$\sum_{t=T^{(g)}}^{T^{(g-1)}-1} \exp\left[-(1-\delta)^2 \frac{\log T^{(g-1)}}{\mathrm{KL}(\mu_a, \mu_m)} \mathrm{KL}(\mu_a, \mu_{a^\star})\right]$$

$$\leq T^{(g-1)} \exp\left[-(1-\delta)^2 \frac{\log T^{(g-1)}}{\mathrm{KL}(\mu_a, \mu_m)} \mathrm{KL}(\mu_a, \mu_{a^\star})\right] \leq 1,$$

where the second inequality holds because δ satisfies (24). If $a \in \mathcal{M}$, then we instead have that this term is no larger than

$$\sum_{t=T^{(g)}}^{T^{(g-1)}-1} \exp\left[-\frac{\log T^{(g-1)}}{\mathrm{KL}(\mu_a, \mu_{a^\star})} \mathrm{KL}(\mu_a, \mu_{a^\star})\right] \leq 1.$$

For the second term in (A.10), note that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=T^{(g)}}^{T^{(g-1)}-1} \mathbb{1}\left\{\hat{\mu}_a(t) \leq \mu^\dagger, N_a(t) < b_a(T,g)\right\} \mathbb{P}\left\{a \in \mathcal{A}(t+1), \theta_a(t) > \theta^\dagger \middle| \mathcal{F}(t)\right\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=T^{(g)}}^{T^{(g-1)}-1} \mathbb{1}\left\{N_a(t) < b_a(T,g)\right\} \mathbb{P}\left\{a \in \mathcal{A}(t+1) | \mathcal{F}(t)\right\}\right] \\
&= \mathbb{E}\left[\sum_{t=T^{(g)}}^{T^{(g-1)}-1} \mathbb{1}\left\{N_a(t) < b_a(T,g), a \in \mathcal{A}(t+1)\right\}\right] \\
&= \mathbb{E}\left[\mathbb{1}\left\{N_a\left(T^{(g)}\right) < b_a(T,g)\right\} \sum_{t=T^{(g)}}^{T^{(g-1)}-1} \mathbb{1}\left\{N_a(t) < b_a(T,g), a \in \mathcal{A}(t+1)\right\}\right] \\
&\leq b_a(T,g)\, \mathbb{P}\left\{N_a\left(T^{(g)}\right) < b_a(T,g)\right\},
\end{aligned}$$

where the final inequality uses that the sum inside the expectation is at most $b_a(T,g)$. By the uniform efficiency of the algorithm established in Lemma 13 and (5) from Lemma 1, the probability in the final inequality is $o(1)$, and thus the above is $o(b_a(T,g)) = o(\log T)$. Thus (A.10) is $o(\log T)$.

Plugging this into (A.9) yields that $\mathbb{E}[M_a^{a^\star}(T^{(g-1)}) - M_a^{a^\star}(T^{(g)})] = o(\log T)$ for each a with $\mu_a \leq \mu_m$ and each $g \leq G$. As Term B is a sum of finitely many such terms, Term B is $o(\log T)$. □