A minimax and asymptotically optimal algorithm for stochastic bandits

Pierre Ménard, Aurélien Garivier

A minimax and asymptotically optimal algorithm for stochastic bandits

Pierre Ménard PIERRE.MENARD@MATH.UNIV-TOULOUSE.FR

Aurélien Garivier AURELIEN.GARIVIER@MATH.UNIV-TOULOUSE.FR

Institut de Mathématiques de Toulouse; UMR5219
Université de Toulouse; CNRS
UPS IMT, F-31062 Toulouse Cedex 9, France

Editors: Steve Hanneke and Lev Reyzin

Abstract

We propose the kl-UCB^{++} algorithm for regret minimization in stochastic bandit models with exponential families of distributions. We prove that it is simultaneously asymptotically optimal (in the sense of Lai and Robbins' lower bound) and minimax optimal. This is the first algorithm proved to enjoy these two properties at the same time. This work thus merges two different lines of research with simple and clear proofs.

Keywords: Stochastic multi-armed bandits, regret analysis, upper confidence bound (UCB), minimax optimality, asymptotic optimality.

1. Introduction

For regret minimization in stochastic bandit problems, two notions of time-optimality coexist. On the one hand, one may consider a fixed model: the famous lower bound by Lai and Robbins (1985) showed that the regret of any consistent strategy should grow at least as $C(\mu)\log(T)\big(1-o(1)\big)$ when the horizon T goes to infinity. Here, $C(\mu)$ is a constant depending solely on the model. A strategy with a regret upper-bounded by $C(\mu)\log(T)\big(1+o(1)\big)$ will be called in this paper *asymptotically-optimal*. Lai and Robbins provided a first example of such a strategy in their seminal work. Later, Garivier and Cappé (2011) and Maillard et al. (2011) provided finite-time analysis for variants of the UCB algorithm (see Agrawal (1995); Burnetas and Katehakis (1996); Auer et al. (2002a)) which imply asymptotic optimality. Since then, other algorithms like Bayes-UCB (Kaufmann et al., 2012) and Thompson Sampling (Korda et al., 2013) have also joined the family.

On the other hand, for a fixed horizon T one may assess the quality of a strategy by the greatest regret suffered in all possible bandit models. If the regret of a bandit strategy is upper-bounded by $C'\sqrt{KT}$ (the optimal rate: see Auer et al. (2002b) and Cesa-Bianchi and Lugosi (2006)) for some numeric constant C', this strategy is called *minimax-optimal*. The PolyINF and the MOSS strategies by Audibert and Bubeck (2009) were the first proved to be minimax-optimal.

Hitherto, as far as we know, no algorithm was proved to be *at the same time* asymptotically- and minimax-optimal. Two limited exceptions may be mentioned: the case of two Gaussian arms is treated in Garivier et al. (2016a); and the OC-UCB algorithm of Lattimore (2015) is proved to be minimax-optimal and almost problem-dependent optimal for Gaussian multi-armed bandit prob-

lems. Notably, the OC-UCB algorithm satisfies another worthwhile property of *finite-time instance near-optimality*, see Section 2 of Lattimore (2015) for a detailed discussion.

Contributions. In this work, we put forward the kl-UCB^{++} algorithm, a slightly modified version of kl-UCB^{+} algorithm discussed in Garivier et al. (2016a) as an empirical improvement of UCB, and analyzed in Kaufmann (2016). This bandit strategy is designed for some exponential distribution families, including for example Bernoulli and Gaussian laws. It borrows from the MOSS algorithm of Audibert and Bubeck (2009) the idea to divide the horizon by the number of arms in order to reach minimax optimality. We prove that it is at the same time asymptotically- and minimax-optimal. This work thus merges the progress which has been made in different directions towards the understanding of the optimism principle, finally reconciling the two notions of time-optimality.

Insofar, our contribution answers a very simple and natural question. The need for simultaneous minimax- and problem-dependent optimality could only be addressed in very limited settings by means that could not be generalized to the framework adopted in our paper. Indeed, for a given horizon T, the worst problem depends on T: it involves arms separated by a gap of order $\sqrt{K/T}$. Treating the T-dependent problems correctly for all T appears as a quite different task than catching the optimal, problem-dependent speed of convergence for every fixed bandit model. We show in this paper that the two goals can indeed be achieved simultaneously.

Combining the two notions of optimality requires a modified exploration rate. We stick as much as possible to existing algorithms and methods, introducing just what is necessary to obtain the desired results. Starting from that of kl-UCB (so as to have a tight asymptotic analysis), one has to completely cancel the exploration bonus of the arms that have been drawn roughly T/K times. The consequence is very slight and harmless in the case where the best arm is much better than the others, but essential in order to minimize the regret in the worst case where the best arm is barely distinguishable from the others. Indeed, when the best arm is separated by a gap of order $\sqrt{K/T}$ from the suboptimal arms, we can not afford to draw more than T/K times a suboptimal arm so as to get a regret of order $\sqrt{KT}$.

We present a general yet simple proof, combining the best elements of the above-cited sources which are simplified as much as possible and presented in a unified way. To this end, we develop new deviation inequalities, improving the analysis of the different terms contributing to the regret. This analysis is made in the framework which we believe is the best compromise between simplicity and generality (simple exponential families). This permits us to treat, among others, the Bernoulli and the Gaussian case at the same time. More fundamentally, this appears to us as the right, simple framework for the analysis, which emphasizes what is really required to have simple lower- and upper-bounds (the possibility to make adequate changes of measure, and Chernoff-type deviation bounds).

The paper is organized as follows. In Section 2, we introduce the setting and assumptions required for the main results, Theorems 1 and 2, which are presented in Section 3. We give the entire proofs of these results in Sections 4 and 5, with only a few technical lemmas proved in Appendix A. We conclude in Section 6 with some brief references to possible future prospects.

2. Notation and Setting

Exponential families. We consider a simple stochastic bandit problem with K arms indexed by $a \in \{1, \dots, K\}$, with $K \geqslant 2$. Each arm is assumed to be a probability distribution of some

canonical one-dimensional exponential family ν_θ indexed by $\theta \in \Theta$. The probability law ν_θ is assumed to be absolutely continuous with respect to a dominating measure ρ on $\mathbb{R}$, with a density given by

$$\frac{\mathrm{d}\nu_\theta}{\mathrm{d}\rho}(x) = \exp(x\theta - b(\theta)), \quad \text{where } b(\theta) = \log \int_{\mathbb{R}} e^{x\theta} \mathrm{d}\rho(x) \text{ and } \Theta = \left\{\theta \in \mathbb{R} : \ b(\theta) < +\infty\right\}.$$

It is well-known that b is convex, twice differentiable on Θ, that $b'(\theta) = E(\nu_\theta)$ and $b''(\theta) = V(\nu_\theta) > 0$ are respectively the mean and the variance of the distribution ν_θ. The family can thus be parametrized by the mean $\mu = b'(\theta)$, for $\mu \in I = b'(\Theta) := (\bar{\mu}^-, \bar{\mu}^+)$. The Kullback-Leibler divergence between two distributions is $\mathrm{KL}(\nu_\theta, \nu_{\theta'}) = b(\theta') - b(\theta) - b'(\theta)(\theta' - \theta)$. This permits to define the following divergence on the set of arm expectations: for $\mu = E(\nu_\theta)$ and $\mu' = E(\nu_{\theta'})$, we write

$$\mathrm{kl}(\mu, \mu') := \mathrm{KL}(\nu_\theta, \nu_{\theta'}) \ .$$

For a minimax analysis, we need to restrict the set of means to bounded interval: we suppose that each arm ν_θ satisfies $\mu = b'(\theta) \in [\mu-, \mu+] \subset I$ for two fixed real numbers μ^+, μ^-. Our analysis requires a Pinsker-like inequality; we therefore assume that the variance is bounded in the exponential family: there exists $V > 0$ such that

$$\sup_{\mu \in I} b''\big(b'^{-1}(\mu)\big) = \sup_{\mu \in I} V\big(\nu_{b'^{-1}(\mu)}\big) \leqslant V < +\infty \ .$$

This implies that for all $\mu, \mu' \in I$,

$$\mathrm{kl}(\mu, \mu') \geqslant \frac{1}{2V}(\mu - \mu')^2 \, . \tag{1}$$

In the sequel, we denote by $\mathcal{F}$ the set of bandit problems ν satisfying these assumptions. By the usual Pinsker inequality, this setting includes in particular Bernoulli bandits with $V = 1/4$ and $\mathrm{kl}(\mu, \mu') = \mu \log(\mu/\mu') + (1-\mu) \log\big((1-\mu)/(1-\mu')\big)$ (by convention, $0 \log 0 = 0 \log 0/0 = 0$). This also includes (bounded) Gaussian bandits with known variance σ^2, with the choice $V = \sigma^2$ and $\mathrm{kl}(\mu, \mu') = (\mu - \mu')^2/(2\sigma^2)$.

Regret. The K arms are denoted $\nu_{\theta_1}, \dots \nu_{\theta_K}$, and the expectation of arm $a \in \{1, \dots, K\}$ is denoted by μ_a. At each round $1 \leqslant t \leqslant T$, the player pulls an arm A_t and receives an independent draw Y_t of the distribution $\nu_{\theta_{A_t}}$. This reward is the only piece of information available to the player. The best mean is $\mu^\star = \max_{a=1,\dots,K} \mu_a$. We denote by $N_a(T) = \sum_{t=1}^T \mathbb{I}_{\{A_t = a\}}$ the number of draws of arm a up to and including time T. In this work, the goal is to minimize the *expected regret*

$$R_T = T\mu^\star - \mathbb{E}\left[\sum_{t=1}^T Y_t\right] = \mathbb{E}\left[\sum_{t=1}^T \big(\mu^\star - \mu_{A_t}\big)\right] = \sum_{a=1}^K \big(\mu^\star - \mu_a\big)\, \mathbb{E}\big[N_a(T)\big] \, .$$

Lai and Robbins (1985) proved that if a strategy is uniformly efficient, that is if it is such that under any bandit model of a sufficiently rich family (such as an exponential family described above) $R_T = o(T^\alpha)$ holds for every $\alpha > 0$, then it needs to draw any suboptimal arm a at least as often as

$$\mathbb{E}\big[N_a(T)\big] \geqslant \frac{\log(T)}{\mathrm{kl}(\mu_a, \mu^\star)} \, \big(1 - o(1)\big) \ .$$

In light of the previous equality, this directly implies an asymptotic lower bound on $R_T/\log(T)$.

On the other side, a straightforward adaptation of the the proof of Theorem A.2 of Auer et al. (2002b) shows that there exists a constant C' depending only on the considered family $\mathcal{F}$ of distributions such that

$$\sup_{\nu\in\mathcal{F}} R_T \geqslant C' \min\left(\sqrt{KT}, T\right),$$

where the supremum is taken over all bandit problems ν in $\mathcal{F}$. Note that the notion of minimax-optimality is defined here up to a multiplicative constant, in contrast to the definition of (problem-dependent) asymptotic optimality. For a discussion on the minimax and asymptotic lower bounds, we refer to Garivier et al. (2016b) and references therein.

3. The kl-UCB^{++} Algorithm

We denote by $\widehat{\mu}_{a,n}$ the empirical mean of the first n rewards from arm a. The empirical mean of arm a after t rounds is

$$\widehat{\mu}_a(t) = \widehat{\mu}_{a,N_a(t)} = \frac{1}{N_a(t)} \sum_{s=1}^{t} Y_s \, \mathbb{I}_{\{A_s=a\}} .$$

Parameters: The horizon T and an exploration function $g : \mathbb{N} \mapsto \mathbb{R}^+$.
Initialization: Pull each arm of $\{1, .., K\}$ once.
For $t = K$ to $T-1$, **do**

1. Compute for each arm a the quantity

$$U_a(t) = \sup\left\{\mu \in I \ : \ \mathrm{kl}\big(\widehat{\mu}_a(t), \mu\big) \leqslant \frac{g\big(N_a(t)\big)}{N_a(t)}\right\}. \tag{2}$$

2. Play $A_{t+1} \in \arg\max_{a\in\{1,..,K\}} U_a(t)$.

The kl-UCB^{++} algorithm is a slight modification of algorithm kl-UCB$^+$ of Garivier and Cappé (2011) and of the kl-UCB-H$^+$ analyzed in Kaufmann (2016). It uses the exploration function g given by

$$g(n) = \log_+\left(\frac{T}{Kn}\left(\log_+^2\left(\frac{T}{Kn}\right) + 1\right)\right), \tag{3}$$

where $\log_+(x) := \max\big(\log(x), 0\big)$. The exploration function g borrows the general form with the extra exploration rate from the kl-UCB algorithm, the division by the number of draws from kl-UCB$^+$, and the division by the number of arm from MOSS.
The following results state that the kl-UCB^{++} algorithm is simultaneously minimax- and asymptotically-optimal.

Theorem 1 (Minimax optimality) *For any family $\mathcal{F}$ satisfying the assumptions detailed in Section 2, and for any bandit model $\nu \in \mathcal{F}$, the expected regret of the kl-UCB^{++} algorithm is upper-bounded as*

$$R_T \leqslant 76\sqrt{VKT} + (\mu^+ - \mu^-)K . \tag{4}$$

Theorem 2 (Asymptotic optimality) *For any bandit model $\nu \in \mathcal{F}$, for any suboptimal arm a and any δ such that $\sqrt{22VK/T} \leqslant \delta \leqslant (\mu^\star - \mu_a)/3$,*

$$\mathbb{E}\big[N_a(T)\big] \leqslant \frac{\log(T)}{\mathrm{kl}(\mu_a + \delta, \mu^\star - \delta)} + O\left(\frac{\mathrm{loglog}(T)}{\delta^2}\right) \tag{5}$$

which implies the asymptotic optimality (see the end of the proof in Section 5 for an explicit bound).

Theorems 1 and 2 are proved in Sections 4 and 5 respectively. The main differences between the two proofs are discussed at the beginning of Section 5. Note that the two regret bounds of Theorems 1 and 2 also apply to all $[0,1]$-valued bandit models, with the value $V = 1/4$, as the deviations of $[0,1]$-valued random variables are dominated by those of a Bernoulli distribution with the same mean (this is discussed for example in Cappé et al. (2013)). However, the kl-UCB^{++} algorithm is not asymptotically optimal then: the regret bound in $\log(T)/\mathrm{kl}(\mu_a, \mu^*)$ is not optimal in that case. Asymptotic optimality would require tight distribution-dependent, non-parametric upper confidence bounds (for example based on the empirical-likelihood method, as in the above cited paper). This is out of the scope of this work (and would require a lot more space).

4. Proof of Theorem 1

This proof merges merges ideas presented in Bubeck and Liu (2013) for the analysis of the MOSS algorithm and from the analysis of kl-UCB in Cappé et al. (2013) (see also Kaufmann (2016)). It is divided into the following steps:

Decomposition of the regret. Let a^* be the index of an optimal arm. Since by definition of the strategy $U_{a^*}(t) \leqslant U_{A_{t+1}}(t)$ for all $t \geqslant K-1$, the regret can be decomposed as follows:

$$R_T \leqslant K(\mu^+ - \mu^-) + \underbrace{\sum_{t=K}^{T-1} \mathbb{E}\big[\mu^\star - U_{a^*}(t)\big]}_{A} + \underbrace{\sum_{t=K}^{T-1} \mathbb{E}\big[U_{A_{t+1}}(t) - \mu_{A_{t+1}}\big]}_{B} . \tag{6}$$

We define $\delta_0 = \sqrt{22VK/T}$; since the bound (4) is otherwise trivial, we assume in the sequel that $\delta_0 \leqslant 1$. For the first term A, as in the proof of MOSS algorithm, we carefully upper bound the probability that appears inside the integral thanks to a 'peeling trick'. The second term B is easier to handle since we can reduce the index to UCB-like-index thanks to the Pinsker inequality (1) and proceed as in Bubeck and Liu (2013).

Step 1: Upper-bounding A. Term A is concerned with the optimal arm a^* only. Two words of intuition: since $U_{a^*}(t)$ is meant to be an upper confidence bound for $\mu^\star$, this term should not be too large, at least as long as the the confidence level controlled by function g is large enough – but when the confidence level is low, the number of draws is large and deviations are unlikely.

Upper-bounding term A boils down to controlling the probability that $\mu^\star$ is under-estimated at time t. Indeed,

$$\begin{aligned}\mathbb{E}\big[\mu^\star - U_{a^*}(t)\big] \leqslant \mathbb{E}\Big[\big(\mu^\star - U_{a^*}(t)\big)_+\Big] &\leqslant \int_0^{+\infty} \mathbb{P}\big(u < \mu^\star - U_{a^*}(t)\big)\mathrm{d}u \\ &\leqslant \delta_0 + \int_{\delta_0}^{+\infty} \mathbb{P}\big(U_{a^*}(t) \leqslant \mu^\star - u\big)\mathrm{d}u\,, \end{aligned} \tag{7}$$

and we need to upper bound the left-deviations of the mean of arm a^*. On the event $\{U_{a^*}(t) \leqslant \mu^\star - u\}$, we have that $\widehat{\mu}_{a^*}(t) \leqslant U_{a^*}(t) \leqslant \mu^\star - u < \mu^\star$, and by definition of $U_{a^*}(t)$ it holds that

$$\mathrm{kl}\big(\widehat{\mu}_{a^*}(t), \mu^\star\big) \geqslant \frac{g\big(N_{a^*}(t)\big)}{N_{a^*}(t)}\,.$$

Consequently,

$$\begin{aligned}\mathbb{P}\big(U_{a^*}(t) \leqslant \mu^\star - u\big) &\leqslant \mathbb{P}\Big(\widehat{\mu}_{a^*}(t) \leqslant \mu^\star - u \text{ and } \mathrm{kl}\big(\widehat{\mu}_{a^*}(t), \mu^\star\big) \geqslant g\big(N_{a^*}(t)\big)/N_{a^*}(t)\Big) \\ &\leqslant \mathbb{P}\big(\exists 1 \leqslant n \leqslant T,\ \widehat{\mu}_{a^*,n} \leqslant \mu^\star - u \text{ and } \mathrm{kl}(\widehat{\mu}_{a^*,n}, \mu^\star) \geqslant g(n)/n\big)\,. \qquad (8)\end{aligned}$$

For small values of n, the dominant term is given by $\mathrm{kl}(\widehat{\mu}_{a^*,n}, \mu^\star) \geqslant g(n)/n$, whereas for large n the event $\widehat{\mu}_{a^*,n} \leqslant \mu^\star - u$ is quite unlikely. This is why we split the probability in two terms, proceeding as follows. Let f be the function defined, for $u \geqslant \delta_0$, by

$$f(u) = \frac{2V}{u^2}\log\!\left(\frac{Tu^2}{2VK}\right)\,.$$

Our choice of δ_0 implies that $f(u)K/T \leqslant \exp(-3/2)$, and thus

$$f(u) < \frac{T}{K} \qquad \text{and} \qquad \log\!\left(\frac{T}{Kf(u)}\right) \geqslant 3/2\,. \qquad (9)$$

In particular, for $n \leqslant f(u)$ it holds that

$$g(n) = \log\!\left(\frac{T}{Kn}\left(1 + \log^2\!\left(\frac{T}{Kn}\right)\right)\right)\,.$$

It appears that $f(u)$ is the right place where to split the probability of Equation (8): defining $\mathrm{kl}_+(p,q) := \mathrm{kl}(p,q)\mathbb{I}_{\{p\leqslant q\}}$, we write

$$\begin{aligned}&\mathbb{P}\big(\exists 1 \leqslant n \leqslant T,\ \widehat{\mu}_{a^*,n} \leqslant \mu^\star - u \text{ and } \mathrm{kl}(\widehat{\mu}_{a^*,n}, \mu^\star) \geqslant g(n)/n\big) \leqslant \\ &\qquad \underbrace{\mathbb{P}\big(\exists 1 \leqslant n \leqslant f(u),\ \mathrm{kl}_+(\widehat{\mu}_{a^*,n}, \mu^\star) \geqslant g(n)/n\big)}_{A_1} + \underbrace{\mathbb{P}\big(\exists f(u) \leqslant n \leqslant T,\ \widehat{\mu}_{a^*,n} \leqslant \mu^\star - u\big)}_{A_2}\,. \qquad (10)\end{aligned}$$

Controlling terms A_1 and A_2 is a matter of deviation inequalities.

Step 1.1: Upper-bounding A_1. The term A_1, which involves self-normalized devation probabilities, can be upper-bounded thanks to a 'peeling trick' as in the proof of Theorem 5 from Audibert and Bubeck (2009). We assume that $f(u) \geqslant 1$, for otherwise $A_1 = 0$. We use the grid $f(u)/\beta^{\ell+1} \leqslant n \leqslant f(u)/\beta^\ell$, where the real $\beta > 1$ will be chosen later. We write

$$A_1 \leqslant \sum_{\ell=0}^{+\infty} \underbrace{\mathbb{P}\left(\exists \frac{f(u)}{\beta^{\ell+1}} \leqslant n \leqslant \frac{f(u)}{\beta^\ell},\ \mathrm{kl}_+(\widehat{\mu}_{a^*,n}, \mu^\star) \geqslant \gamma_\ell\right)}_{A_1^\ell}\,, \qquad (11)$$

where

$$\gamma_\ell = \frac{\log\left(\frac{T\beta^\ell}{Kf(u)}\left(1+\log^2\left(\frac{T}{Kf(u)}\right)\right)\right)}{f(u)/\beta^\ell}\,.$$

Thanks to Doob's maximal inequality (see Lemma 4 in Appendix A),

$$A_1^\ell \leqslant \exp\left(-\frac{f(u)}{\beta^{\ell+1}}\,\gamma_\ell\right) = e^{-\ell\log(\beta)/\beta - C/\beta}\,,$$

where

$$C := \log\left(\frac{T}{Kf(u)}\left(1+\log^2\left(\frac{T}{Kf(u)}\right)\right)\right). \tag{12}$$

Plugging this last inequality into (11), together with the numerical inequality of Lemma 3 (see Appendix A), we get

$$\begin{aligned} A_1 \leqslant \sum_{\ell=0}^{+\infty} e^{-\ell\log(\beta)/\beta - C/\beta} &= \frac{1}{1-e^{-\log(\beta)/\beta}}e^{-C/\beta} \\ &\leqslant \frac{e}{e^{\log(\beta)/\beta}-1}e^{-C/\beta} \leqslant 2e\max\big(\beta,\beta/(\beta-1)\big)e^{-C/\beta}\,. \end{aligned}$$

But thanks to Equation (9),

$$C = \log\left(\frac{T}{Kf(u)}\left(1+\log^2\left(\frac{T}{Kf(u)}\right)\right)\right) \geqslant \log\left(\frac{T}{Kf(u)}\right) \geqslant \frac{3}{2}\,.$$

It is now time to choose $\beta := C/(C-1)$, so that $\beta \leqslant 2C$ and $\beta/(\beta-1) = C$. Together with the definition of f, this choice yields

$$A_1 \leqslant 4e^2 Ce^{-C} = 4e^2\frac{\log\left(\frac{T}{Kf(u)}\left(1+\log^2\left(\frac{T}{Kf(u)}\right)\right)\right)}{1+\log^2\left(\frac{T}{Kf(u)}\right)}\frac{Kf(u)}{T}\,, \tag{13}$$

and therefore

$$A_1 \leqslant 4e^2\frac{Kf(u)}{T} = \frac{16e^2VK}{Tu^2}\log\left(\sqrt{\frac{T}{2VK}}u\right) \tag{14}$$

as, for all $x \geqslant 1$,

$$\frac{\log\Big(x\big(1+\log^2(x)\big)\Big)}{1+\log^2(x)} \leqslant 1\,.$$

Step 1.2: Upper-bounding A_2. The term A_2 is more simple to handle, as it does not involve self-normalized deviations. Thanks to the maximal inequality (recalled in Equation (33) of Appendix A) and thanks to the Pinsker-like inequality (1),

$$A_2 \leqslant e^{-u^2 f(u)/2V} = \frac{2VK}{Tu^2}\,. \tag{15}$$

Putting Equations (7) to (15) together, we obtain that

$$\mathbb{E}\big[\mu^\star - U_{a^*}(t)\big] \leqslant \delta_0 + \int_{\delta_0}^{+\infty} \frac{16e^2VK}{Tu^2} \log\left(\sqrt{\frac{T}{2VK}}u\right) + \frac{2VK}{Tu^2}\mathrm{d}u\,. \tag{16}$$

It remains only to conclude with some calculus:

$$\begin{aligned}\int_{\delta_0}^{+\infty} \frac{16e^2VK}{Tu^2} \log\left(\sqrt{\frac{T}{2VK}}u\right) \mathrm{d}u &= \left[-\frac{16e^2VK}{Tu} \log\left(e\sqrt{\frac{T}{2VK}}u\right)\right]_{\delta_0}^{+\infty} \\ &= \frac{16e^2\sqrt{V}}{\sqrt{22}} \log\left(e\sqrt{11}\right)\sqrt{\frac{K}{T}}\,.\end{aligned}$$

Similarly,

$$\int_{\delta_0}^{+\infty} \frac{2VK}{Tu^2}\mathrm{d}u = 2\sqrt{\frac{V}{22}}\sqrt{\frac{K}{T}}\,,$$

and replacing δ_0 by its value we obtain from Equation (16) the following relation:

$$\mathbb{E}\big[\mu^\star - U_{a^*}(t)\big] \leqslant \sqrt{V}\left(\sqrt{22} + \frac{16e^2}{\sqrt{22}} \log\left(e\sqrt{11}\right) + \frac{2}{\sqrt{22}}\right)\sqrt{\frac{K}{T}}\,.$$

Summing over t from K to $T-1$, this yields:

$$A \leqslant \sqrt{V}\left(\sqrt{22} + \frac{16e^2}{\sqrt{22}} \log\left(e\sqrt{11}\right) + \frac{2}{\sqrt{22}}\right)\sqrt{KT}\,. \tag{17}$$

Step 2: Upper-bounding B. Term B is of different nature, since typically $U_{A_{t+1}}(t) > \mu_{A_{t+1}}$. However, as for the term A, we first reduce the problem to the upper-bounding of a probability:

$$\begin{aligned}B &\leqslant \sum_{t=K}^{T-1} \delta_0 + \int_{\delta_0}^{+\infty} \mathbb{P}\big(U_{A_{t+1}}(t) - \mu_{A_{t+1}} \geqslant u\big)\mathrm{d}u \\ &\leqslant T\delta_0 + \int_{\delta_0}^{+\infty} \sum_{t=K}^{T-1} \mathbb{P}\big(U_{A_{t+1}}(t) - \mu_{A_{t+1}} \geqslant u\big)\mathrm{d}u\,. \end{aligned} \tag{18}$$

The event $\big\{U_{A_{t+1}}(t) - \mu_{A_{t+1}} \geqslant u\big\}$ is typical if $N_{A_{t+1}}(t)$ is small, and corresponds to a deviation of the sample mean otherwise. In order to handle this correctly, we first get rid of the randomness of $N_{A_{t+1}}(t)$ by the pessimistic trajectorial upper bound from Bubeck and Liu (2013)

$$\sum_{t=K}^{T-1} \mathbb{I}_{\left\{U_{A_{t+1}}(t) - \mu_{A_{t+1}} \geqslant u\right\}} \leqslant \sum_{n=1}^{T}\sum_{a=1}^{K} \mathbb{I}_{\left\{U_{a,n} - \mu_a \geqslant u\right\}}\,.$$

In addition, we simplify the upper bound thanks to our assumption (1) that some Pinsker type inequality is available:

$$U_{a,n} := \sup\left\{\mu \in I \;:\; \mathrm{kl}\big(\widehat{\mu}_{a,n}, \mu\big) \leqslant \frac{g(n)}{n}\right\} \leqslant B_{a,n} := \widehat{\mu}_{a,n} + \sqrt{2V\frac{g(n)}{n}}\,. \tag{19}$$

Hence, B can be upper-bounded as

$$B \leqslant T\delta_0 + \sum_{a=1}^{K} \int_{\delta_0}^{+\infty} \sum_{n=1}^{T} \mathbb{P}(B_{a,n} - \mu_a \geqslant u) \mathrm{d}u \,. \tag{20}$$

Then, we need only to upper bound $\sum_{n=1}^{T} \mathbb{P}(B_{a,n} - \mu_a \geqslant u)$ for each arm $a \in \{1, \dots, K\}$. We cut the sum at the critical sample size $n(u)$ where the event $\{B_{a,n} - \mu_a > u\}$ becomes atypical: for $u \geqslant \delta_0$, let $n(u)$ be the integer such that

$$n(u) = \left\lceil \frac{8V}{u^2} \log\!\left(\frac{Tu^2}{8VK}\right) \right\rceil .$$

For $n \geqslant n(u)$ it holds that

$$\sqrt{2V \frac{g(n)}{n}} \leqslant \frac{u}{\sqrt{2}} \,. \tag{21}$$

Indeed, as $\log(1 + x^2) \leqslant x$ for all $x \geqslant 0$, we have

$$2V \frac{g(n)}{n} \leqslant \frac{4V}{n} \log_+\!\left(\frac{T}{Kn}\right) ,$$

also observe that $h(x) := \log\big(x/\log(x)\big)/\log(x)$ is such that $h(x) \leqslant 1$ for $x \geqslant 11/4$, and thus for $n \geqslant n(u)$ and $u \geqslant \delta_0$

$$2V \frac{g(n)}{n} \leqslant \frac{4V}{n(u)} \log_+\!\left(\frac{T}{Kn(u)}\right) \leqslant \frac{u^2}{2} h\left(\frac{Tu^2}{8VK}\right) \leqslant \frac{u^2}{2} \,.$$

Therefore, cutting the sum in (20) at $n(u)$, we obtain:

$$\begin{aligned}
\sum_{n=1}^{T} \mathbb{P}(B_{a,n} - \mu_a \geqslant u) &\leqslant n(u) - 1 + \sum_{n=n(u)}^{T} \mathbb{P}\big(\widehat{\mu}_{a,n} - \mu_a \geqslant u - \sqrt{2Vg(n)/n}\big) \\
&\leqslant n(u) - 1 + \sum_{n=n(u)}^{T} \mathbb{P}\Big(\widehat{\mu}_{a,n} - \mu_a \geqslant u\big(1 - 1/\sqrt{2}\big)\Big) \\
&\leqslant \frac{8V}{u^2} \log\!\left(\frac{Tu^2}{8VK}\right) + \sum_{n=n(u)}^{T} \mathbb{P}(\widehat{\mu}_{a,n} - \mu_a \geqslant cu) \,,
\end{aligned} \tag{22}$$

where $c := 1 - 1/\sqrt{2}$. It remains to integrate Inequality (22) from $u = \delta_0$ to infinity. The first summand involves the same integral as we have already met in the upper bound of term A_1:

$$\int_{\delta_0}^{+\infty} \frac{8V}{u^2} \log\!\left(\frac{Tu^2}{8VK}\right) \mathrm{d}u = 16\sqrt{\frac{V}{22}} \log\!\left(e\sqrt{\frac{11}{4}}\right) \sqrt{\frac{T}{K}} \,.$$

For the remaining summand, Inequality (33) yields

$$\sum_{n=n(u)}^{T} \mathbb{P}(\widehat{\mu}_{a,n} - \mu_a \geqslant cu) \leqslant \sum_{n=n(u)}^{T} e^{-\frac{u^2c^2n}{2V}} \leqslant \frac{1}{e^{\frac{u^2c^2}{2V}} - 1} \,.$$

Thus, as $e^x - 1 \geqslant x$ for all $x \geqslant 0$,

$$\int_{\delta_0}^{+\infty} \frac{1}{e^{\frac{u^2c^2}{2V}} - 1} \mathrm{d}u \leqslant \int_{\delta_0}^{+\infty} \frac{2V}{u^2c^2} \mathrm{d}u = \frac{2}{c^2}\sqrt{\frac{V}{22}}\sqrt{\frac{T}{K}}\,,$$

Putting everything together starting from Inequality (22), we have proved that

$$\int_{\delta_0}^{+\infty} \sum_{n=1}^{T} \mathbb{P}(B_{a,n} - \mu_a \geqslant u)\mathrm{d}u \leqslant \sqrt{\frac{V}{22}} \left(16 \log\left(e\sqrt{\frac{11}{4}} \right) + \frac{2}{c^2} \right) \sqrt{\frac{T}{K}}\,.$$

By Equation (20), replacing δ_0 by its value finally yields

$$B \leqslant \sqrt{V} \left(\sqrt{22} + \frac{16}{\sqrt{22}} \log\left(e\sqrt{\frac{11}{4}} \right) + \frac{2}{\sqrt{22}c^2} \right) \sqrt{KT}\,. \tag{23}$$

Conclusion of the proof. It just remains to plug Inequalities (17) and (23) into Equation (6):

$$\begin{aligned} A + B &\leqslant \sqrt{V} \left(2\sqrt{22} + \frac{16e^2}{\sqrt{22}} \log\left(e\sqrt{11}\right) + \frac{2}{\sqrt{22}} + \frac{16}{\sqrt{22}} \log\left(e\sqrt{\frac{11}{4}} \right) + \frac{2}{\sqrt{22}c^2} \right) \sqrt{KT} \\ &\leqslant 76\sqrt{VKT}\,, \end{aligned}$$

which concludes the proof.

5. Proof of Theorem 2

The analysis of asymptotic optimality shares many elements with the minimax analysis, with some differences however. The decomposition of the regret into two terms A and B is similar, but localized on a fixed sub-optimal arm $a \in \{1, \dots, K\}$: we analyze the number of draws of a and not directly the regret (and we do not need to integrate the deviations at the end). We proceed roughly as in the proof of Theorem 1 for term A, which involves the deviations of an optimal arm. For term B, which stands for the behavior of the sub-optimal arm a, a different (but classical) argument is used, as one cannot simply use the Pinsker-like Inequality (1) if one wants to obtain the correct constant (and thus asymptotic optimality).

Decomposition of $\mathbb{E}\big[N_a(T)\big]$. If arm a is pulled at time $t+1$, then by definition of the strategy $U_{a^*}(t) \leqslant U_a(t)$ for any index a^* of an optimal arm. Thus, for any fixed δ to be chosen later,

$$\begin{aligned} \big\{A_{t+1} = a\big\} &\subseteq \big\{\mu^* - \delta \geqslant U_a(t)\big\} \cup \big\{\mu^* - \delta < U_a(t) \text{ and } A_{t+1} = a\big\} \\ &\subseteq \big\{\mu^* - \delta \geqslant U_{a^*}(t)\big\} \cup \big\{\mu^* - \delta < U_a(t) \text{ and } A_{t+1} = a\big\}\,. \end{aligned}$$

As a consequence,

$$\mathbb{E}\big[N_a(T)\big] \leqslant 1 + \underbrace{\sum_{t=K}^{T-1} \mathbb{P}\big(U_{a^*}(t) \leqslant \mu^* - \delta\big)}_{\text{A}} + \underbrace{\sum_{t=K}^{T-1} \mathbb{P}\big(\mu^* - \delta < U_a(t) \text{ and } A_{t+1} = a\big)}_{\text{B}}\,, \tag{24}$$

and it remains to bound each of these terms.

Step 1: Upper-bounding A. As in the proof of Theorem 1, we write

$$\mathbb{P}\big(U_{a^*}(t) \leqslant \mu^* - \delta\big) \leqslant \underbrace{\mathbb{P}\big(\exists 1 \leqslant n \leqslant f(\delta),\ \ \mathrm{kl}_+(\widehat{\mu}_{a^*,n}, \mu^\star) \geqslant g(n)/n\big)}_{A_1} + \underbrace{\mathbb{P}\big(\exists f(\delta) \leqslant n \leqslant T,\ \ \widehat{\mu}_{a^*,n} \leqslant \mu^\star - \delta\big)}_{A_2}, \tag{25}$$

where we use the same function

$$f(\delta) = \frac{2V}{\delta^2}\log\left(\frac{T\delta^2}{2KV}\right).$$

Thanks to the Inequality (13) that we saw in the proof of Theorem 1, we obtain that

$$A_1 \leqslant 4e^2 \frac{\log\left(\frac{T}{Kf(\delta)}\left(1+\log^2\left(\frac{T}{Kf(\delta)}\right)\right)\right)}{\log\left(\frac{T}{Kf(\delta)}\right)} \frac{f(\delta)}{\log\left(\frac{T}{Kf(\delta)}\right)}\frac{K}{T} \leqslant \frac{16e^2}{\delta^2}\frac{2VK}{T}.$$

Here, we used that for all $x \geqslant e^{3/2}$, since the condition $\delta^2 \geqslant 22VK/T$ implies that $f(\delta)K/T \leqslant e^{-3/2}$,

$$\frac{\log\left(x\big(1+\log^2(x)\big)\right)}{\log(x)} \leqslant 2 \qquad \text{and} \qquad \frac{\log(x)}{\log\big(x/\log(x)\big)} \leqslant 2,$$

and that

$$\frac{f(\delta)}{\log\left(\frac{T}{Kf(\delta)}\right)} = \frac{2V}{\delta^2}\frac{\log\left(\frac{T\delta^2}{2VK}\right)}{\log\left(\frac{T\delta^2}{2VK}\frac{1}{\log\left(T\delta^2/(2VK)\right)}\right)}.$$

Thanks to the maximal inequality recalled in Appendix A as Equation (33), it holds that

$$A_2 \leqslant e^{-\delta^2 f(\delta)/(2V)} = \frac{2VK}{T\delta^2}. \tag{26}$$

Putting Equations (25) to (26) together yields:

$$A \leqslant (16e^2+1)\frac{2VK}{\delta^2}. \tag{27}$$

Step 2: Upper-bounding B. Thanks to the definition of $U_a(t)$ it holds that

$$\big\{\mu^* - \delta < U_a(t) \text{ and } A_{t+1} = a\big\} \subseteq \Big\{\mathrm{kl}\big(\widehat{\mu}_a(t), \mu^* - \delta\big) \leqslant g\big(N_a(t)\big)/N_a(t) \text{ and } A_{t+1} = a\Big\}$$

Together with the following classical argument for regret analysis in bandit models, this yields:

$$\begin{aligned} B &\leqslant \sum_{t=K}^{T-1} \mathbb{P}\big(\mathrm{kl}\big(\widehat{\mu}_a(t), \mu^* - \delta\big) \leqslant g\big(N_a(t)\big)/N_a(t) \text{ and } A_{t+1} = a\big) \\ &\leqslant \sum_{n=1}^{T} \mathbb{P}\big(\mathrm{kl}(\widehat{\mu}_{a,n}, \mu^* - \delta) \leqslant g(n)/n\big) \\ &\leqslant \sum_{n=1}^{T} \mathbb{P}\left(\mathrm{kl}(\widehat{\mu}_{a,n}, \mu^* - \delta) \leqslant \log\Big(T/K\big(1+\log^2(T/K)\big)\Big)/n\right), \end{aligned} \tag{28}$$

as it holds $g(n) \leqslant g(1)$. Now, let $n(\delta)$ be the integer defined as

$$n(\delta) = \left\lceil \frac{\log\Big(T/K\big(1+\log^2(T/K)\big)\Big)}{\mathrm{kl}(\mu_a + \delta, \mu^* - \delta)} \right\rceil .$$

Then, for $n \geqslant n(\delta)$,

$$\log\Big(T/K\big(1+\log^2(T/K)\big)\Big)/n \leqslant \mathrm{kl}(\mu_a + \delta, \mu^* - \delta) \ .$$

We cut the sum in (28) at $n(\delta)$, so that

$$\begin{aligned} B &\leqslant n(\delta) - 1 + \sum_{n=n(\delta)}^{T} \mathbb{P}\big(\mathrm{kl}(\widehat{\mu}_{a,n}, \mu^* - \delta) \leqslant \mathrm{kl}(\mu_a + \delta, \mu^* - \delta)\big) \\ &\leqslant \frac{\log\Big(T/K\big(1+\log^2(T/K)\big)\Big)}{\mathrm{kl}(\mu_a + \delta, \mu^* - \delta)} + \sum_{n=n(\delta)}^{T} \mathbb{P}\big(\mathrm{kl}(\widehat{\mu}_{a,n}, \mu^* - \delta) \leqslant \mathrm{kl}(\mu_a + \delta, \mu^* - \delta)\big) \,. \end{aligned} \tag{29}$$

Recall that by assumption $\delta < (\mu^\star - \mu_a)/3$, using the inclusion

$$\big\{\mathrm{kl}(\widehat{\mu}_{a,n}, \mu^* - \delta) \leqslant \mathrm{kl}(\mu_a + \delta, \mu^* - \delta)\big\} \subseteq \{\widehat{\mu}_{a,n} \geqslant \mu_a + \delta\} \,,$$

together with Inequality (33), we obtain that

$$\begin{aligned} \sum_{n=n(\delta)}^{T} \mathbb{P}\big(\mathrm{kl}(\widehat{\mu}_{a,n}, \mu^* - \delta) \leqslant \mathrm{kl}(\mu_a + \delta, \mu^* - \delta)\big) \ &\leqslant \sum_{n=n(\delta)}^{T} \mathbb{P}\big(\widehat{\mu}_{a,n} \geqslant \mu_a + \delta\big) \\ &\leqslant \sum_{n=1}^{\infty} e^{-n\delta^2/(2V)} = \frac{1}{e^{\delta^2/(2V)} - 1} \leqslant \frac{2V}{\delta^2} \,, \end{aligned}$$

and Equation (29) yields

$$B \leqslant \frac{\log(T)}{\mathrm{kl}(\mu_a + \delta, \mu^* - \delta)} + \frac{\log\Big(1/K\big(1+\log^2(T/K)\big)\Big)}{\mathrm{kl}(\mu_a + \delta, \mu^* - \delta)} + \frac{2V}{\delta^2} \,. \tag{30}$$

Conclusion of the proof. It just remains to plug Inequalities (27) and (30) into Equation (24):

$$\mathbb{E}\big[N_a(T)\big] \leqslant \frac{\log(T)}{\mathrm{kl}(\mu_a + \delta, \mu^* - \delta)} + \frac{\log\Big(1/K\big(1+\log^2(T/K)\big)\Big)}{\mathrm{kl}(\mu_a + \delta, \mu^* - \delta)} + (16e^2 + 2)\frac{2VK}{\delta^2} + 1 \,,$$

and we obtain Equation (5). Choosing δ of order $1/\log\log(T)^{1/2}$ yields the asymptotic optimality.

6. Conclusion and Perspectives

We have proved that the kl-UCB^{++} algorithm is both minimax- and asymptotically-optimal for the exponential distribution families described in Section 2. So far, this algorithm requires the horizon T as a parameter: to keep the proofs clear and simple, we have deferred to future work the analysis of an anytime variant. We believe, though, that obtaining such an extension should be possible by using the tools developed in Degenne and Perchet (2016). In addition, we have focused in this paper on asymptotic optimality without trying to derive explicit finite-time bounds: we believe that this would have impaired the clarity and simplicity of the reasoning. But it is certainly a challenging and important objective to design a general strategy that would, in addition to minimax- and asymptotic optimality, would also reach the important notion of *finite-time instance near optimality* of Lattimore (2015).

From a more technical point of view, it may be possible to suppress the extra $\log^2$ exploration term in the definition of the confidence bonus g in Equation (3). This is carried out in Garivier et al. (2016a) using some particularities of the Gaussian distributions; using an improved Chernoff bound such as Talagrand (1995) may allow considering more general cases. Finally, we defer the consideration of general bounded probability distributions (with non-parametric upper-confidence bounds) to future work.

Acknowledgments

This work was partially supported by the CIMI (Centre International de Mathématiques et d'Informatique) Excellence program. The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR), under grants ANR-13-BS01-0005 (project SPADRO) and ANR-13-CORD-0020 (project ALICIA).

Appendix A. Some Technical Lemmas

Lemma 3 *For all $\beta > 1$ we have*

$$\frac{1}{e^{\log(\beta)/\beta} - 1} \leqslant 2 \max\big(\beta, \beta/(\beta-1)\big)\,. \tag{31}$$

Proof Inequality (31) is equivalent to

$$e^{\log(\beta)/\beta} - 1 \geqslant \frac{1}{2\beta} \min(1, \beta - 1)\,.$$

If $\beta \geqslant 2$, then

$$e^{\log(\beta)/\beta} - 1 \geqslant e^{\log(2)/\beta} - 1 \geqslant \frac{\log(2)}{\beta} \geqslant \frac{1}{2\beta}\,.$$

Otherwise, if $1 < \beta < 2$, as the function $\beta \mapsto \log(\beta)/(\beta-1)$ is non-increasing one gets

$$\frac{\beta}{\beta-1}\Big(e^{\log(\beta)/\beta} - 1\Big) \geqslant \frac{\log(\beta)}{\beta-1} \geqslant \log(2) \geqslant 1/2\,.$$

■

Lemma 4 *(Maximal Inequality) Let N and M be two real numbers in $\mathbb{R}^+ \times \overline{\mathbb{R}^+}$, let γ be a real number in $\mathbb{R}^{+*}$, and let $\widehat{\mu}_n$ be the empirical mean of n random variables i.i.d. according to the distribution $\nu_{b'^{-1}(\mu)}$. Then*

$$\mathbb{P}\big(\exists N \leqslant n \leqslant M,\ \ \mathrm{kl}_+(\widehat{\mu}_n, \mu) \geqslant \gamma\big) \leqslant e^{-N\gamma}\,. \tag{32}$$

Proof If $\gamma > \mathrm{kl}(\bar{\mu}^-, \mu)$ or $\widehat{\mu}_n \geqslant \mu$ the Inequality (32) is trivial. Else, there exist two real numbers $z < \mu$ and $\lambda < 0$ such that

$$\gamma = \mathrm{kl}(z, \mu) = \lambda z - \varphi_\mu(\lambda)\,,$$

where φ_μ denotes the the log-moment generating function of $\nu_{b'^{-1}(\mu)}$. Since on the event $\big\{\exists N \leqslant n \leqslant M,\ \ \mathrm{kl}_+(\widehat{\mu}_n, \mu) \geqslant \gamma\big\}$ one has at the same time

$$\widehat{\mu}_n \leqslant \mu\,, \quad \lambda\widehat{\mu}_n - \varphi_\mu(\lambda) \geqslant \lambda z - \varphi_\mu(\lambda) = \gamma \quad \text{and} \quad \lambda n\widehat{\mu}_n - n\varphi_\mu(\lambda) \geqslant N\gamma\,,$$

we can write that

$$\begin{aligned}\mathbb{P}\big(\exists N \leqslant n \leqslant M,\ \ \mathrm{kl}_+(\widehat{\mu}_n, \mu) \geqslant \gamma\big) &\leqslant \mathbb{P}\big(\exists N \leqslant n \leqslant M,\ \ \lambda n\widehat{\mu}_n - n\varphi_\mu(\lambda) \geqslant N\gamma\big)\\ &\leqslant \exp(-N\gamma)\,,\end{aligned}$$

by Doob's maximal inequality for the exponential martingale $\exp\big(\lambda n\widehat{\mu}_n - n\varphi_\mu(\lambda)\big)$. ■

As a simple consequence of this Lemma 4 and Inequality (1), it holds that:

$$\text{for every } x \leqslant \mu, \qquad \mathbb{P}(\exists N \leqslant n \leqslant M,\ \ \widehat{\mu}_n \leqslant x) \leqslant e^{-N(x-\mu)^2/(2V)}\,, \tag{33}$$

$$\text{for every } x \geqslant \mu, \qquad \mathbb{P}(\exists N \leqslant n \leqslant M,\ \ \widehat{\mu}_n \geqslant x) \leqslant e^{-N(x-\mu)^2/(2V)}\,. \tag{34}$$

References

Rajeev Agrawal. Sample mean based index policies with o(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995. ISSN 00018678. URL http://www.jstor.org/stable/1427934.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, pages 217–226, 2009.

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

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multi-armed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b.

Sébastien Bubeck and Che-Yu Liu. Prior-free and prior-dependent regret bounds for thompson sampling. In *Advances in Neural Information Processing Systems*, pages 638–646, 2013.

Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

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

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Rémy Degenne and Vianney Perchet. Anytime optimal algorithms in stochastic multi-armed bandits. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 1587–1595. JMLR.org, 2016. URL http://dl.acm.org/citation.cfm?id=3045390.3045558.

Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *COLT*, pages 359–376, 2011.

Aurélien Garivier, Tor Lattimore, and Emilie Kaufmann. On explore-then-commit strategies. In *Advances in Neural Information Processing Systems*, pages 784–792, 2016a.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *arXiv preprint arXiv:1602.07182*, 2016b.

Emilie Kaufmann. On bayesian index policies for sequential resource allocation. *arXiv preprint arXiv:1601.01190*, 2016.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On bayesian upper confidence bounds for bandit problems. In *AISTATS*, pages 592–600, 2012.

Nathaniel Korda, Emilie Kaufmann, and Remi Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems*, pages 1448–1456, 2013.

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

Tor Lattimore. Optimally confident ucb: Improved regret for finite-armed bandits. *arXiv preprint arXiv:1507.07880*, 2015.

O-A. Maillard, R. Munos, and G. Stoltz. A finite-time analysis of multi-armed bandits problems with Kullback-Leibler divergences. In *Proceedings of the 23rd Annual Conference on Learning Theory*, Budapest, Hungary, 2011.

Michel Talagrand. The missing factor in hoeffding's inequalities. In *Annales de l'IHP Probabilités et statistiques*, volume 31, pages 689–702, 1995.