KL-UCB-switch: optimal regret bounds for stochastic bandits from both a distribution-dependent and a distribution-free viewpoints

Aurélien Garivier, Hédi Hadiji, Pierre Ménard, Gilles Stoltz

1–41, May 2018

KL-UCB-switch: optimal regret bounds for stochastic bandits from both a distribution-dependent and a distribution-free viewpoints

Aurélien Garivier AURELIEN.GARIVIER@MATH.UNIV-TOULOUSE.FR
Institut de mathématiques de Toulouse, Université Paul Sabatier, Toulouse

Hédi Hadiji HEDI.HADIJI@MATH.U-PSUD.FR
Laboratoire de mathématiques d'Orsay, Université Paris Sud, Orsay

Pierre Ménard PIERRE.MENARD@MATH.UNIV-TOULOUSE.FR
Institut de mathématiques de Toulouse, Université Paul Sabatier, Toulouse

Gilles Stoltz GILLES.STOLTZ@MATH.U-PSUD.FR
Laboratoire de mathématiques d'Orsay, Université Paris Sud, Orsay

Abstract

In the context of K–armed stochastic bandits with distribution only assumed to be supported by $[0, 1]$, we introduce a new algorithm, KL-UCB-switch, and prove that is enjoys simultaneously a distribution-free regret bound of optimal order $\sqrt{KT}$ and a distribution-dependent regret bound of optimal order as well, that is, matching the $\kappa \ln T$ lower bound by Lai and Robbins (1985) and Burnetas and Katehakis (1996).

Keywords: K–armed stochastic bandits, distribution-dependent regret bounds, distribution-free regret bounds

1. Introduction and brief literature review

Great progress has been made, over the last decades, in the understanding of the stochastic K–armed bandit problem. In this simplistic and yet paradigmatic sequential decision model, an agent can at each step $t \in \mathbb{N}^*$ sample one out of K independent sources of randomness and receive the corresponding outcome as a reward. The most investigated challenge is to minimize the (pseudo-) regret, which is defined as the difference between the cumulated rewards obtained by the agent and by an oracle knowing in hindsight the distribution with largest expectation.

After Thompson's seminal paper (Thompson, 1933) and Gittins' Bayesian approach in the 1960s, Lai and his co-authors wrote in the 1980s a series of articles laying the foundations of a frequentist analysis of bandit strategies based on confidence regions. Lai and Robbins (1985) provided a general asymptotic lower bound, for parametric bandit models: for any reasonable strategy, the regret after T steps grows at least as $\kappa \ln(T)$, where κ is an informational complexity measure of the problem. In the 1990s, Agrawal (1995) and Burnetas and Katehakis (1996) analyzed the UCB algorithm, a simple procedure where at step t the arm with highest upper confidence bound is chosen. The same authors also extended the lower bound by Lai and Robbins to non-parametric models.

In the early 2000s, the much noticed contributions of Auer et al. (2002a) and Auer et al. (2002b) promoted three important ideas.

1. First, a bandit strategy should not address only specific statistical models, but general and non-parametric families of probability distributions, e.g., bounded distributions.

2. Second, the regret analysis should not only be asymptotic, but should provide finite-time bounds.

3. Third, a good bandit strategy should be competitive with respect to two concurrent notions of optimality: distribution-dependent optimality (it should reach the asymptotic lower bound of Lai and Robbins and have a regret not much larger than $\kappa \ln(T)$) and distribution-free optimality (the maximal regret over all considered probability distributions should be of the optimal order $\sqrt{KT}$).

These efforts were pursued by further works in those three directions. Maillard et al. (2011) and Garivier and Cappé (2011) simultaneously proved that the distribution-dependent lower bound could be reached with exactly the right multiplicative constant in simple settings (for example, for binary rewards) and provided finite-time bounds to do so. They were followed by similar results for other index policies like BayesUCB (Kaufmann et al., 2012) or Thompson sampling (Korda et al., 2013).

Initiated by Honda and Takemura for the IMED algorithm (see Honda and Takemura, 2015 and references to earlier works of the authors therein) and followed by Cappé et al. (2013) for the KL-UCB algorithm, the use of the empirical likelihood method for the construction of the upper confidence bounds was proved to be optimal as far as distribution-dependent bounds are concerned. The analysis for IMED was led for all (semi-)bounded distributions, while the analysis for KL-UCB was only sucessfully achieved in some classes of distributions (e.g., bounded distributions with finite supports). A contribution in passing of the present article is to also provide optimal distribution-dependent bounds for KL-UCB for families of bounded distributions.

On the other hand, classical UCB strategies were proved not to enjoy distribution-free optimal regret bounds. A modified strategy named MOSS was proposed by Audibert and Bubeck (2009) to address this issue: minimax (distribution-free) optimality was proved, but distribution-dependent optimality was then not considered. It took a few more years before Ménard and Garivier (2017) and Lattimore (2016) proved that, in simple parametric settings, a strategy can enjoy, at the same time, regret bounds that are optimal both from a distribution-dependent and a distribution-free viewpoints.

Main contributions. In this work, we generalize the latter bi-optimality result to the non-parametric class of distributions with bounded support, say, $[0, 1]$. Namely, we propose the KL-UCB-switch algorithm, a bandit strategy belonging to the family of upper-confidence-bounds strategies. We prove that it is simultaneously optimal from a distribution-free viewpoint (Theorem 1) and from a distribution-dependent viewpoint in the considered class of distributions (Theorem 2).

We go one step further by providing, as Honda and Takemura (2015) already achieved for IMED, a second-order term of the optimal order $-\ln(\ln(T))$ in the distribution-dependent bound (Theorem 3). This explains from a theoretical viewpoint why simulations consistently show strategies having a regret smaller than the main term of the lower bound of Lai and Robbins (1985). Note that, to the best of our knowledge, IMED is not proved to enjoy an optimal distribution-free regret bound; only a distribution-dependent regret analysis was provided for it.

Beyond these results, we took special care of the clarity and simplicity of all the proofs, and all our bounds are finite time, with closed-form expressions. In particular, we provide for the first time an elementary analysis of performance of the KL-UCB algorithm on the class of all distributions

over a bounded interval. The study of KL-UCB in Cappé et al. (2013) indeed remained somewhat intricate and limited to finitely supported distributions. Furthermore, our simplified analysis allowed us to derive similar optimality results for the anytime version of this new algorithm, with little if no additional effort (see Theorems 4 and 5).

Organization of the paper. Section 2 contains the presentation of the KL-UCB-switch algorithm, the precise statement of the aforementioned theorems, and corresponding results for an anytime version of the KL-UCB-switch algorithm. Section 3 discusses some numerical experiments comparing the performance of the KL-UCB-switch algorithm to competitors like IMED or KL-UCB. Section 5 contains the statements and the proofs of several results that were already known before, but for which we sometimes propose a simpler derivation. All technical results needed in this article are thus stated and proved from scratch (e.g., on the $\mathcal{K}_{\inf}$ quantity that is central to the analysis of IMED and KL-UCB, and on the analysis of the performance of MOSS), which makes our paper fully self-contained. These known results are used as building blocks in Section 4, where the main results of this article are proved, up to some more sophisticated bound whose analysis is detailed in Section 6. Technical arguments are deferred to the appendices.

2. Setting and statement of the main results

We consider the simplest case of a stochastic bandit problem, with finitely many arms indexed by $a \in \{1, \dots, K\}$. Each of these arms is associated with an unknown probability distribution ν_a over $[0,1]$. We call $\underline{\nu} = (\nu_1, \dots, \nu_K)$ a bandit problem over $[0,1]$. At each round $t \geqslant 1$, the player pulls the arm A_t and gets a real-valued reward Y_t drawn independently at random according to the distribution ν_{A_t}. This reward is the only piece of information available to the player.

A typical measure of the performance of a strategy is given by its *regret*. To recall its definition, we denote by $\mathrm{E}(\nu_a) = \mu_a$ the expected payoff of arm a and by Δ_a its gap to an optimal arm:

$$\mu^\star = \max_{a=1,\dots,K} \mu_a \qquad \text{and} \qquad \Delta_a = \mu^\star - \mu_a \,.$$

Arms a such that $\Delta_a > 0$ are called suboptimal arms. The expected regret of a strategy equals

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

The first equality above follows from the tower rule. To control the expected regret, it is thus sufficient to control the $\mathbb{E}\big[N_a(T)\big]$ quantities for suboptimal arms a.

Reminder of the existing lower bounds. The distribution-free lower bound of Auer et al. (2002b) states that for all strategies, for all $T \geqslant 1$ and all $K \geqslant 2$,

$$\sup_{\underline{\nu}} R_T \geqslant \frac{1}{20} \min\Big\{\sqrt{KT},\, T\Big\}\,, \tag{1}$$

where the supremum is taken over all bandit problems $\underline{\nu}$ over $[0,1]$.

We denote by $\mathcal{P}[0,1]$ the set of all distributions over $[0,1]$. The key quantity in stating distribution-dependent lower bounds is based on KL, the Kullback-Leibler divergence between two probability distributions. For $\nu_a \in \mathcal{P}[0,1]$ and $x \in [0,1]$,

$$\mathcal{K}_{\inf}(\nu_a, x) = \inf\Big\{\mathrm{KL}(\nu_a, \nu'_a) : \;\; \nu'_a \in \mathcal{P}[0,1] \;\text{ and }\; \mathrm{E}(\nu'_a) > x\Big\}\,,$$

where $\mathrm{E}(\nu'_a)$ denotes the expectation of the distribution ν'_a and where by convention, the infimum of the empty set equals $+\infty$. As essentially proved by Lai and Robbins (1985) and Burnetas and Katehakis (1996)—see also Garivier et al., 2018—, for any "reasonable" strategy, for any bandit problem $\underline{\nu}$ over $[0,1]$, for any suboptimal arm a,

$$\liminf_{T\to\infty} \frac{\mathbb{E}\big[N_a(T)\big]}{\ln T} \geqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a,\mu^\star)}\,. \tag{2}$$

By "reasonable" strategy, we mean a strategy that is uniformly fast convergent on $\mathcal{P}[0,1]$, that is, such that for all bandit problems $\underline{\nu}$ over $[0,1]$, for all suboptimal arms a,

$$\forall\,\alpha>0, \qquad \mathbb{E}\big[N_a(T)\big] = \mathrm{o}(T^\alpha)\,.$$

For uniformly super-fast convergent strategies, that is, strategies for which there exists a constant C such for all bandit problems $\underline{\nu}$ over $[0,1]$, for all suboptimal arms a,

$$\frac{\mathbb{E}\big[N_a(T)\big]}{\ln T} \leqslant \frac{C}{\Delta_a^2}\,,$$

the lower bound above can be strengthened into: for any bandit problem $\underline{\nu}$ over $[0,1]$, for any suboptimal arm a,

$$\mathbb{E}\big[N_a(T)\big] \geqslant \frac{\ln T}{\mathcal{K}_{\inf}(\nu_a,\mu^\star)} - \Omega\big(\ln(\ln T)\big)\,, \tag{3}$$

see Garivier et al. (2018, Section 4). This order of magnitude $-\ln(\ln T)$ for the second-order term in the regret bound is optimal, as follows from the upper bound exhibited by Honda and Takemura (2015, Theorem 5).

2.1. The KL-UCB-switch algorithm

Algorithm 1 Generic index policy

Inputs: index functions U_a
Initialization: Play each arm $a=1,\dots,K$ once and compute the $U_a(K)$
for $t=K+1,\dots,T$ **do**
  Pull an arm $A_t \in \arg\max_{a=1,\dots,K} U_a(t-1)$
  Get a reward Y_t drawn independently at random according to ν_{A_t}
end for

For any index policy as described above, we have $N_a(t)\geqslant 1$ for all arms a and $t\geqslant K$ and may thus define, respectively, the empirical distribution of the rewards associated with arm a up to round t included and their empirical mean:

$$\widehat{\nu}_a(t) = \frac{1}{N_a(t)}\sum_{s=1}^{t}\delta_{Y_s}\,\mathbb{1}_{\{A_s=a\}} \qquad \text{and} \qquad \widehat{\mu}_a(t) = \mathrm{E}\big[\widehat{\nu}_a(t)\big] = \frac{1}{N_a(t)}\sum_{s=1}^{t} Y_s\,\mathbb{1}_{\{A_s=a\}}\,,$$

where δ_y denotes the Dirac point-mass distribution at $y\in[0,1]$.

The MOSS algorithm (see Audibert and Bubeck 2009) uses the index functions

$$U_a^{\mathrm{M}}(t) \overset{\text{def}}{=} \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)} \ln_+\left(\frac{T}{KN_a(t)}\right)}\,, \tag{4}$$

where $\ln_+$ denotes the nonnegative part of the natural logarithm, $\ln_+ = \max\{\ln, 0\}$.

We also consider a slight variation of the KL-UCB algorithm (see Cappé et al. 2013), which we call KL-UCB$^+$ and which relies on the index functions

$$U_a^{\mathrm{KL}}(t) \overset{\text{def}}{=} \sup\left\{\mu \in [0,1] \;\middle|\; \mathcal{K}_{\inf}\big(\widehat{\nu}_a(t), \mu\big) \leqslant \frac{1}{N_a(t)} \ln_+\left(\frac{T}{KN_a(t)}\right)\right\}. \tag{5}$$

We introduce a new algorithm KL-UCB-switch. The novelty here is that this algorithm switches from the KL-UCB-type index to the MOSS index once it has pulled an arm more than $f(T,K)$ times. In the sequel we will take $f(T,K) = \lfloor (T/K)^{1/5} \rfloor$. More precisely, we define the index functions

$$U_a(t) = \begin{cases} U_a^{\mathrm{KL}}(t) & \text{if } N_a(t) \leqslant f(T,K) \\ U_a^{\mathrm{M}}(t) & \text{if } N_a(t) > f(T,K) \end{cases}$$

2.2. Optimal distribution-dependent and distribution-free regret bounds (known horizon T)

We first consider a fixed and beforehand-known value of T. The proofs of the theorems below are provided in Section 4.

Theorem 1 (Distribution-free bound) *Given $T \geqslant 1$, the regret of the KL-UCB-switch algorithm, tuned with the knowledge of T and the switch function $f(T,K) = \lfloor (T/K)^{1/5} \rfloor$, is uniformly bounded over all bandit problems $\underline{\nu}$ over $[0,1]$ by*

$$R_T \leqslant (K-1) + 25\sqrt{KT}\,,$$

KL-UCB-switch thus enjoys a distribution-free regret bound of optimal order $\sqrt{KT}$, see (1). The MOSS strategy by Audibert and Bubeck (2009) already enjoyed this optimal regret bound.

Theorem 2 (Distribution-dependent bound) *Given $T \geqslant 1$, the KL-UCB-switch algorithm, tuned with the knowledge of tT and the switch function $f(T,K) = \lfloor (T/K)^{1/5} \rfloor$, ensures that for all bandit problems $\underline{\nu}$ over $[0,1]$, for all sub-optimal arms a,*

$$\mathbb{E}[N_a(T)] \leqslant \frac{\ln T}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} + \mathcal{O}_T\big((\ln T)^{2/3}\big)\,,$$

where a finite-time, closed-form expression of the $\mathcal{O}_T\big((\ln T)^{2/3}\big)$ term is the sum of the bounds (13) and (16) for the choice $\delta = (\ln T)^{-1/3}$.

By considering the exact same algorithm but by following a more sophisticated proof we may in fact get a stronger result.

Theorem 3 (Distribution-dependent bound with a second-order term) *We actually have*

$$\mathbb{E}[N_a(T)] \leqslant \frac{\ln T - \ln\ln T}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} + \mathcal{O}_T(1)\,,$$

where a finite-time, closed-form expression of the $\mathcal{O}_T(1)$ term is the sum of the bounds (13) and (39) for the choice $\delta = T^{-1/8}$.

KL-UCB-switch thus enjoys a distribution-dependent regret bounds of optimal orders, see (2) and (3). This optimal order was already reached by the IMED strategy by Honda and Takemura (2015) on the model $\mathcal{P}[0,1]$. The KL-UCB algorithm studied, e.g., by Cappé et al. (2013), only enjoyed optimal regret bounds for more limited models; for instance, for distributions over $[0,1]$ with finite support. In the analysis of KL-UCB-switch we actually provide in passing an analysis of KL-UCB for the model $\mathcal{P}[0,1]$ of all distributions over $[0,1]$.

2.3. Adaptation to the horizon T (an anytime version of KL-UCB-switch)

A standard doubling trick fails to provide a meta-strategy that would not require the knowledge of T and have optimal $\mathcal{O}\big(\sqrt{KT}\big)$ and $\big(1+\mathrm{o}(1)\big)(\ln T)/\mathcal{K}_{\inf}(\nu_a,\mu^\star)$ bounds. Indeed, there are first, two different rates, $\sqrt{T}$ and $\ln T$, to accommodate simultaneously and each would require different regime lengths, e.g., 2^r and 2^{2^r}, respectively, and second, any doubling trick on the distribution-dependent bound would result in an additional multiplicative constant in front of the $1/\mathcal{K}_{\inf}(\nu_a,\mu^\star)$ factor. This is why a dedicated anytime version of our algorithm is needed.

For technical reasons, it was useful in our proof to perform some additional exploration, which deteriorates the second-order terms in the regret bound. Indeed, we define the augmented exploration function

$$\varphi(x) = \ln_+\big(x(1+\ln_+^2 x)\big) \tag{6}$$

and the corresponding anytime index

$$U_a^{\mathrm{ANY}}(t) = \begin{cases} \sup\left\{ \mu \in [0,1] \;\middle|\; \mathcal{K}_{\inf}\big(\widehat{\nu}_a(t), \mu\big) \leqslant \dfrac{1}{N_a(t)}\,\varphi\left(\dfrac{t}{KN_a(t)}\right) \right\} & \text{if } N_a(t) \leqslant f(t,K) \\ \widehat{\mu}_a(t) + \sqrt{\dfrac{1}{2N_a(t)}\,\varphi\left(\dfrac{t}{KN_a(t)}\right)} & \text{if } N_a(t) > f(t,K) \end{cases}$$

Theorem 4 (Anytime distribution-free bound) *The regret of the anytime version of KL-UCB-switch algorithm above, tuned with the switch function $f(t,K) = \lfloor (t/K)^{1/5} \rfloor$, is uniformly bounded over all bandit problems $\underline{\nu}$ over $[0,1]$ as follows: for all $T \geqslant 1$,*

$$R_T \leqslant (K-1) + 46\sqrt{KT}$$

Theorem 5 (Anytime distribution-dependent bound) *The anytime version of KL-UCB-switch algorithm above, tuned with the switch function $f(t,K) = \lfloor (t/K)^{1/5} \rfloor$, ensures that for all bandit problems $\underline{\nu}$ over $[0,1]$, for all sub-optimal arms a,*

$$\limsup_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\ln T} \leqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a,\mu^\star)}$$

We provide the proofs of the two theorems in Appendix C. The distribution-free analysis is essentially the same as in the case of a known horizon, although the additional exploration required an adaptation of most of the calculations. Note also that the simulations detailed below suggest that all anytime variants of the KL-UCB algorithms (KL-UCB-switch included) behave better without the additional exploration required, i.e., with $\ln_+$ as the exploration function.

3. Numerical experiments

We provide here some numerical experiments comparing the different algorithms we refer to in this work. The KL-UCB-switch, KL-UCB, and MOSS algorithms are used in their anytime versions as described in Section 2.1 and Section 2.3. However, we stick to the natural exploration function $\ln_+\big(t/(KN_a(t))\big)$, i.e. without extra-exploration.

For KL-UCB-switch we actually consider a slightly delayed switch function, different from the one in our theoretical analysis : $f(t,K) = \lfloor t/K \rfloor^{8/9}$. While our choice $f(t,K) = \lfloor t/K \rfloor^{1/5}$ appeared most naturally in the proofs, many other choices were possible at the cost of higher constants in one of the two regret bounds.

Distribution-dependent bounds. We compare in Figure 1 the distribution-dependent behaviors of the algorithms. For the two scenarios with truncated exponential or Gaussian rewards we also consider the appropriate version of the kl-UCB algorithm for one-parameter exponential family (see Cappé et al., 2013), with the same exploration function as for the other algorithms; we call these algorithms kl-UCB-exp or kl-UCB-Gauss, respectively. The parameters of the middle and right scenarios were chosen in a way that, even with the truncation, the kl-UCB algorithms have a significantly better performance than the other algorithms. (This is the case because they are able to exploit the form of the underlying distributions.) Note that the kl-UCB-Gauss algorithm reduces to the MOSS algorithm with the constant $2\sigma^2$ instead of $1/2$.

As expected the regret of KL-UCB-switch is an interpolation between the one of MOSS and of KL-UCB.

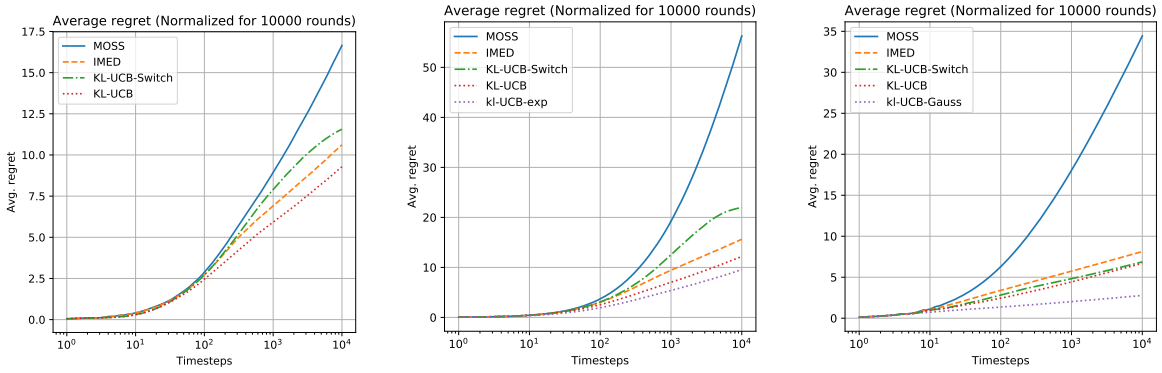

Figure 1: Regrets approximated over $10,000$ runs, shown on a log-scale; distributions of the arms consist of:
Left: Bernoulli distributions with parameters $(0.9,\ 0.8)$
Middle: Exponential distributions truncated on $[0,1]$, with parameters $(0.15,\ 0.12,\ 0.1,\ 0.05)$
Right: Gaussian distributions truncated on $[0,1]$, with means $(0.7,\ 0.5,\ 0.3,\ 0.2)$ and same standard deviation $\sigma = 0.1$

Distribution-free bounds. Here we also consider the UCB algorithm of Auer et al. (2002a) with the exploration function $\ln(t)$. We plot the behavior of the normalized regret, $R_T/\sqrt{KT}$, either as a function of T (Figure 2 left) or of K (Figure 2 right). This quantity should not increase without a bound as T or K increases. KL-UCB-switch and KL-UCB have a normalized regret that seems to not depend too much on T and K. (KL-UCB may perhaps satisfy a distribution-free bound of the optimal order, but we were unable to prove this fact.) The regret of IMED seems to suffer from a suboptimal dependence in K.

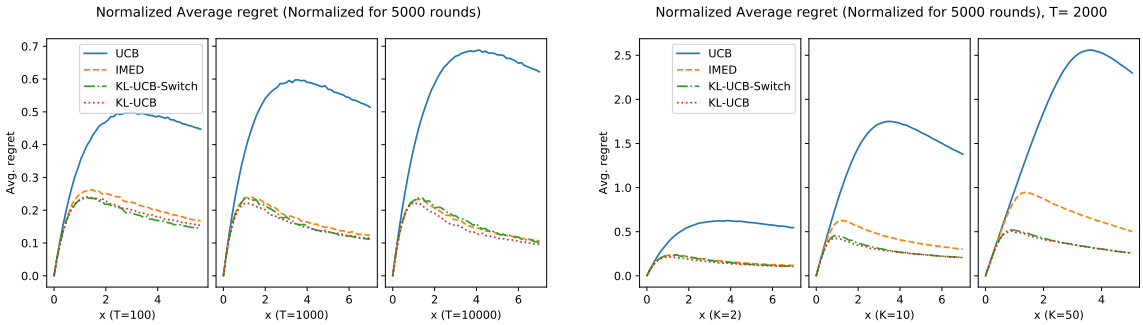

Figure 2: Expected regret $R_T/\sqrt{KT}$, approximated over $5,000$ runs
Left: as a function of x, for a Bernoulli bandit problem with parameters $\big(0.8,\ 0.8 - x\sqrt{K/T}\big)$ and for time horizons $T \in \{100,\ 1000,\ 10000\}$
Right: as a function of x, for a Bernoulli bandit problem with parameters $(0.8,\ 0.8 - x\sqrt{K/T},\ \ldots,\ 0.8 - x\sqrt{K/T})$ and K arms, where $K \in \{2,\ 10,\ 50\}$

4. Proofs of our main results: the first two theorems of Section 2.2

The proof of Theorem 1 is quite standard: it is similar the proof of MOSS and involves no particular difficulty. Some difficulties had to be overcome for the proof of Theorem 2.

Proof of Theorem 1 The first step is standard, see Bubeck and Liu (2013); we use $U_{A_t}(t) \geqslant U_{a^\star}(t)$ to decompose the regret as

$$R_T = \sum_{t=1}^{T} \mathbb{E}\big[\mu^\star - \mu_{A_t}\big] \leqslant (K-1) + \sum_{t=K+1}^{T} \mathbb{E}\big[\mu^\star - U_{a^\star}(t)\big] + \sum_{t=K+1}^{T} \mathbb{E}\big[U_{A_t}(t) - \mu_{A_t}\big] \qquad (7)$$

Each term in the second sum in (7) is bounded in a crude way: by the application (28) of Pinsker's inequality, $U_a(t) \leqslant U_a^{\mathrm{M}}(t)$ so that

$$\mathbb{E}\big[U_{A_t}(t) - \mu_{A_t}\big] \leqslant \sqrt{\frac{K}{T}} + \mathbb{E}\left[\left(U_{A_t}(t) - \mu_{A_t} - \sqrt{\frac{K}{T}}\right)^+\right] \leqslant \sqrt{\frac{K}{T}} + \mathbb{E}\left[\left(U_{A_t}^{\mathrm{M}}(t) - \mu_{A_t} - \sqrt{\frac{K}{T}}\right)^+\right]$$
$$\leqslant \sqrt{\frac{K}{T}} + \sum_{a=1}^{K}\sum_{n=1}^{T} \mathbb{E}\left[\left(U_{a,n}^{\mathrm{M}} - \mu_a - \sqrt{\frac{K}{T}}\right)^+\right]$$

where for the final inequality we used optional skipping (see Section 5.1). The first sum in (7) is dealt with by substituting the value $U_{a^\star}^{\mathrm{KL}}(t)$ or $U_{a^\star}^{\mathrm{M}}(t)$ of $U_{a^\star}(t)$ depending on $N_{a^\star}(t) \leqslant f(T,K)$ or

$N_{a^\star}(t) > f(T,K)$:

$$\sum_{t=K+1}^{T} \mathbb{E}\big[\mu^\star - U_{a^\star}(t)\big] \leqslant \sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{KL}}_{a^\star}(t)\big)^+ \mathbb{1}_{\{N_{a^\star}(t)\leqslant f(T,K)\}}\Big] + \sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{M}}_{a^\star}(t)\big)^+ \underbrace{\mathbb{1}_{\{N_{a^\star}(t)> f(T,K)\}}}_{\leqslant 1}\Big]$$

Collecting the inequalities above into (7), we see that the regret of KL-UCB-switch is less than the claimed $(K-1)+25\sqrt{KT}$ bound,

$$R_T \leqslant (K-1) + \sum_{t=K+1}^{T} \underbrace{\mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{KL}}_{a^\star}(t)\big)^+ \mathbb{1}_{\{N_{a^\star}(t)\leqslant f(T,K)\}}\Big]}_{\text{we show below that } \leqslant 8\sqrt{K/T} \text{ for each } t} + \underbrace{\sqrt{KT} + \sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{M}}_{a^\star}(t)\big)^+\Big] + \sum_{a=1}^{K}\sum_{n=1}^{T} \mathbb{E}\Big[\big(U^{\mathrm{M}}_{a,n} - \mu_a - \sqrt{K/T}\big)^+\Big]}_{\leqslant 17\sqrt{KT} \quad \text{by (25)}}$$

Indeed, by optional skipping (see Section 5.1),

$$\mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{KL}}_{a^\star}(t)\big)^+ \mathbb{1}_{\{N_{a^\star}(t)\leqslant f(T,K)\}}\Big] \leqslant \sum_{n=1}^{f(T,K)} \mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{KL}}_{a^\star,n}\big)^+\Big]$$

where by Fubini-Tonelli, for each $1 \leqslant n \leqslant f(T,K) < T/K$,

$$\mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{KL}}_{a^\star,n}\big)^+\Big] = \int_0^{\mu^\star} \mathbb{P}\big[\mu^\star - U^{\mathrm{KL}}_{a^\star,n} > u\big]\,\mathrm{d}u \leqslant \int_0^{+\infty} \mathrm{e}(2n+1)\frac{Kn}{T}\mathrm{e}^{-2nu^2}\,\mathrm{d}u$$

as for all $u \in (0,\mu^\star)$, by using successively (31) and Proposition 12,

$$\mathbb{P}\big[\mu^\star - U^{\mathrm{KL}}_{a^\star,n} > u\big] \leqslant \mathbb{P}\left[\mathcal{K}_{\inf}(\widehat{\nu}_{a^\star,n},\mu^\star) > \frac{1}{n}\ln\Big(\frac{T}{Kn}\Big) + 2u^2\right] \leqslant \mathrm{e}(2n+1)\frac{Kn}{T}\mathrm{e}^{-2nu^2}$$

The proof of the desired $8\sqrt{K/T}$ bound is concluded by straightforward calculations,

$$\sum_{n=1}^{f(T,K)} \int_0^{+\infty} \mathrm{e}(2n+1)\frac{Kn}{T}\mathrm{e}^{-2nu^2}\,\mathrm{d}u = \sum_{n=1}^{f(T,K)} \mathrm{e}(2n+1)\frac{Kn}{T}\,\frac{1}{\sqrt{2n}}\sqrt{\frac{\pi}{2}} = \frac{\mathrm{e}\sqrt{\pi}}{2}\frac{K}{T}\sum_{n=1}^{f(T,K)}(2n+1)\sqrt{n}$$
$$\leqslant \frac{\mathrm{e}\sqrt{\pi}}{2}\frac{K}{T}\,f(T,K)\left(2f(T,K)^{3/2} + f(T,K)^{1/2}\right) \leqslant \frac{\mathrm{e}\sqrt{\pi}}{2}\frac{K}{T}\,3f(T,K)^{5/2} \leqslant 8\sqrt{K/T}$$

since by definition of $f(T,K)$, we have $f(T,K)^{5/2} \leqslant (T/K)^{1/2}$. ■

The proof of the first distribution-dependent bound (Theorem 2) relies entirely on elementary applications of concentration inequalities, after some careful cutting of events.

Proof of Theorem 2 Given $\delta > 0$ sufficiently small (to be determined by the analysis), we decompose $\mathbb{E}\big[N_a(T)\big]$ as

$$\mathbb{E}\big[N_a(T)\big] = 1 + \sum_{t=K}^{T-1} \mathbb{P}\big[U_a(t) < \mu^\star - \delta \text{ and } A_{t+1} = a\big] + \sum_{t=K}^{T-1} \mathbb{P}\big[U_a(t) \geqslant \mu^\star - \delta \text{ and } A_{t+1} = a\big] \tag{8}$$

Control of the first sum in (8). When $A_{t+1} = a$, we have $U_{a^\star}(t) \leqslant U_a(t)$ by definition of the index policy and this is the only piece of information that traditional proofs, as the one of, e.g., Cappé et al. (2013), use: they are left to bound $\sum \mathbb{P}\big[U_{a^\star}(t) < \mu^\star - \delta\big]$. We proceed slightly more carefully by introducing a possible cutting at $(\mu^\star + \mu_a)/2$ and by distinguishing whether $U_{a^\star}(t)$ is smaller or larger than this value; in the latter case, $U_a(t)$ is also larger than it. In addition we set a threshold $n_0 \geqslant 1$ (to be determined by the analysis) and distinguish whether $N_a(t) \geqslant n_0$ or $N_a(t) \leqslant n_0 - 1$. We thus get the decomposition

$$\begin{aligned}\big\{U_a(t) < \mu^\star - \delta \text{ and } A_{t+1} = a\big\} \subseteq \quad & \big\{U_{a^\star}(t) < (\mu^\star + \mu_a)/2\big\} \\ & \cup \big\{U_a(t) \geqslant (\mu^\star + \mu_a)/2 \text{ and } A_{t+1} = a \text{ and } N_a(t) \geqslant n_0\big\} \\ & \cup \big\{U_{a^\star}(t) < \mu^\star - \delta \text{ and } A_{t+1} = a \text{ and } N_a(t) \leqslant n_0 - 1\big\}\end{aligned}$$

For $u \in (0,1)$, we introduce the event

$$\mathcal{E}_\star(u) = \Big\{\exists\, \tau \in \{K, \dots, T-1\} : U_{a^\star}(\tau) < u\Big\}$$

We now rewrite the second event in the set decomposition above. To that end, we note that by (28) and by definition of the MOSS index, we have, when $N_a(t) \geqslant n_0$,

$$U_a(t) \leqslant U_a^{\mathrm{M}}(t) = \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)} \ln_+\!\left(\frac{T}{KN_a(t)}\right)} \leqslant \widehat{\mu}_a(t) + \underbrace{\sqrt{\frac{1}{2n_0} \ln_+\!\left(\frac{T}{Kn_0}\right)}}_{\leqslant \Delta_a/4} \tag{9}$$

where the inequality in the root of the right-most term comes from the choice

$$n_0 = \left\lceil \frac{8}{\Delta_a^2} \ln\!\left(\frac{T}{K}\right) \right\rceil \tag{10}$$

In particular, we get the inclusion

$$\big\{U_a(t) \geqslant (\mu^\star + \mu_a)/2\big\} = \big\{U_a(t) \geqslant \mu_a + \Delta_a/2\big\} \subseteq \big\{\widehat{\mu}_a(t) \geqslant \mu_a + \Delta_a/4\big\}$$

Collecting all elements together and substituting the definition of $\mathcal{E}_\star$, we established the cruder decomposition

$$\begin{aligned}\big\{U_a(t) < \mu^\star - \delta \text{ and } A_{t+1} = a\big\} \subseteq \quad & \mathcal{E}_\star\big((\mu^\star + \mu_a)/2\big) \\ & \cup \big\{\widehat{\mu}_a(t) \geqslant \mu_a + \Delta_a/4 \text{ and } A_{t+1} = a \text{ and } N_a(t) \geqslant n_0\big\} \\ & \cup \Big(\mathcal{E}_\star(\mu^\star - \delta) \cap \big\{A_{t+1} = a \text{ and } N_a(t) \leqslant n_0 - 1\big\}\Big)\end{aligned}$$

Taking probabilities, resorting to a union bound, summing over t, and using the deterministic control

$$\sum_{t=K}^{T-1} \mathbb{1}_{\{A_{t+1}=a \text{ and } N_a(t)\leqslant n_0-1\}} \leqslant n_0$$

we get (and this is where it is so handy that the $\mathcal{E}_\star$ do not depend on a particular t):

$$\begin{aligned}\sum_{t=K}^{T-1} \mathbb{P}\big[U_a(t) < \mu^\star - \delta \text{ and } A_{t+1}=a\big] \leqslant \sum_{t=K}^{T-1} \mathbb{P}\left[\widehat{\mu}_a(t) \geqslant \mu_a + \frac{\Delta_a}{4} \text{ and } A_{t+1}=a \text{ and } N_a(t) \geqslant n_0\right] \\ + T\,\mathbb{P}\Big(\mathcal{E}_\star\big((\mu^\star+\mu_a)/2\big)\Big) + n_0\,\mathbb{P}\big(\mathcal{E}_\star(\mu^\star-\delta)\big) \qquad (11)\end{aligned}$$

By optional skipping (see Section 5.1), the first sum above is bounded by

$$\sum_{t=K}^{T-1} \mathbb{P}\left[\widehat{\mu}_a(t) \geqslant \mu_a + \frac{\Delta_a}{4} \text{ and } A_{t+1}=a \text{ and } N_a(t) \geqslant n_0\right] \leqslant \sum_{n=n_0}^{T} \mathbb{P}\left[\widehat{\mu}_{a,n} \geqslant \mu_a + \frac{\Delta_a}{4}\right]$$

and we continue the upper bounding by applying Hoeffding's inequality (Proposition 7, actually not using the maximal form):

$$\sum_{n=n_0}^{T} \mathbb{P}\left[\widehat{\mu}_{a,n} \geqslant \mu_a + \frac{\Delta_a}{4}\right] \leqslant \sum_{n=n_0}^{T} \mathrm{e}^{-n\Delta_a^2/8} = \frac{\mathrm{e}^{-n_0\Delta_a^2/8}}{1-\mathrm{e}^{-\Delta_a^2/8}} \leqslant \frac{K/T}{1-\mathrm{e}^{-\Delta_a^2/8}} \qquad (12)$$

where we substituted the value (10) of n_0. The two other terms in (11) are bounded using the lemma right below, respectively with $x=\Delta_a/2$ and $x=\delta$, and we get the final upper bound

$$\begin{aligned}\sum_{t=K}^{T-1} \mathbb{P}\big[U_a(t) < \mu^\star - \delta \text{ and } A_{t+1}=a\big] \leqslant \left(\frac{320\,\mathrm{e}}{(1-\mathrm{e}^{-2})^3}\frac{K}{\Delta_a^6} + T\mathrm{e}^{-T\Delta_a^2/(2K)}\right) \\ + \left\lceil \frac{8}{\Delta_a^2}\ln\left(\frac{T}{K}\right)\right\rceil \left(\frac{5\mathrm{e}}{(1-\mathrm{e}^{-2})^3}\frac{K}{T\delta^6} + \mathrm{e}^{-2\delta^2 T/K}\right) + \frac{K/T}{1-\mathrm{e}^{-\Delta_a^2/8}}\end{aligned} \qquad (13)$$

The bound above is a $\mathcal{O}_T(1)$ for the choices $\delta=(\ln T)^{-1/3}$ and $\delta=T^{-1/8}$ respectively considered in Theorems 2 and 3.

Lemma 6 *For all $x \in (0,\mu^\star)$,*

$$\mathbb{P}\Big(\mathcal{E}_\star\big(\mu^\star-x\big)\Big) = \mathbb{P}\Big[\exists\,\tau\in\{K,\dots,T-1\} : U_{a^\star}(\tau) < \mu^\star - x\Big] \leqslant \frac{5\mathrm{e}}{(1-\mathrm{e}^{-2})^3}\frac{K}{Tx^6} + \mathrm{e}^{-2x^2T/K}$$

Proof The $\ln_+$ in the definition of $U_{a^\star}(\tau)$ vanishes when $N_{a^\star}(\tau) \geqslant T/K$. Therefore, by distinguishing the cases $N_{a^\star}(\tau) < T/K$ and $N_{a^\star}(\tau) \geqslant T/K$, by Pinsker's inequality (28), by optional skipping (see Section 5.1) and by the definition of the index as a given supremum, we successively

get

$$\begin{aligned}
&\mathbb{P}\Big[\exists\, \tau \in \{K,\dots,T-1\} : U_{a^\star}(\tau) < \mu^\star - x\Big] \\
=&\mathbb{P}\Big[\exists\, \tau \in \{K,\dots,T-1\} : U_{a^\star}(\tau) < \mu^\star - x \text{ and } N_{a^\star}(\tau) < T/K\Big] \\
&+ \mathbb{P}\Big[\exists\, \tau \in \{K,\dots,T-1\} : U_{a^\star}(\tau) < \mu^\star - x \text{ and } N_{a^\star}(\tau) \geqslant T/K\Big] \\
=&\mathbb{P}\Big[\exists\, \tau \in \{K,\dots,T-1\} : U^{\mathrm{KL}}_{a^\star}(\tau) < \mu^\star - x \text{ and } N_{a^\star}(\tau) < T/K\Big] \\
&+ \mathbb{P}\Big[\exists\, \tau \in \{K,\dots,T-1\} : \widehat{\mu}_{a^\star}(\tau) < \mu^\star - x \text{ and } N_{a^\star}(\tau) \geqslant T/K\Big] \\
\leqslant&\mathbb{P}\Big[\exists\, m \in \big\{1,\dots,\lfloor T/K\rfloor\big\} : U^{\mathrm{KL}}_{a^\star,m} < \mu^\star - x\Big] + \mathbb{P}\Big[\exists\, m \in \big\{\lceil T/K\rceil,\dots,T\big\} : \widehat{\mu}_{a^\star,m} < \mu^\star - x\Big] \\
\leqslant&\mathbb{P}\left[\exists\, m \in \big\{1,\dots,\lfloor T/K\rfloor\big\} : \mathcal{K}_{\inf}\big(\widehat{\nu}_{a^\star,m}, \mu^\star - x\big) > \frac{1}{m}\ln\left(\frac{T}{Km}\right)\right] \\
&+ \mathbb{P}\Big[\exists\, m \in \big\{\lceil T/K\rceil,\dots,T\big\} : \widehat{\mu}_{a^\star,m} < \mu^\star - x\Big]
\end{aligned}$$

where by a union bound, by the deviation inequality (35) stated as a consequence of Proposition 12, and by some elementary calculations detailed below,

$$\begin{aligned}
&\mathbb{P}\left[\exists\, m \in \big\{1,\dots,\lfloor T/K\rfloor\big\} : \mathcal{K}_{\inf}\big(\widehat{\nu}_{a^\star,m}, \mu^\star - x\big) > \frac{1}{m}\ln\left(\frac{T}{Km}\right)\right] \\
\leqslant& \sum_{m=1}^{\lfloor T/K\rfloor} \mathrm{e}(2m+1)\,\frac{Km}{T}\mathrm{e}^{-2mx^2} \leqslant \frac{\mathrm{e}K}{T}\sum_{m=1}^{+\infty} m(2m+1)\,\mathrm{e}^{-2mx^2} \leqslant \frac{5e}{(1-\mathrm{e}^{-2})^3}\frac{K}{Tx^6}
\end{aligned} \tag{14}$$

while by Hoeffding's maximal inequality (Proposition 7)

$$\mathbb{P}\Big[\exists\, m \in \big\{\lceil T/K\rceil,\dots,T\big\} : \widehat{\mu}_{a^\star,m} < \mu^\star - x\Big] \leqslant \mathrm{e}^{-2\,\lceil T/K\rceil\, x^2} \leqslant \mathrm{e}^{-2x^2T/K}$$

More precisely, the elementary calculations leading to the final inequality in (14) are based on differentiating the defining series for the exponential distribution: for all $\theta > 0$,

$$\begin{aligned}
\sum_{m=0}^{+\infty} \mathrm{e}^{-m\theta} &= \frac{1}{1-\mathrm{e}^{-\theta}}\,, \\
-\sum_{m=1}^{+\infty} m\,\mathrm{e}^{-m\theta} &= \frac{-\mathrm{e}^{-\theta}}{(1-\mathrm{e}^{-\theta})^2} \geqslant \frac{-1}{(1-\mathrm{e}^{-\theta})^2} \geqslant \frac{-1}{(1-\mathrm{e}^{-\theta})^3}\,, \\
\sum_{m=1}^{+\infty} m^2\,\mathrm{e}^{-m\theta} &= \frac{\mathrm{e}^{-\theta}(1+\mathrm{e}^{-\theta})}{(1-\mathrm{e}^{-\theta})^3} \leqslant \frac{2}{(1-\mathrm{e}^{-\theta})^3}
\end{aligned}$$

Hence

$$\sum_{m=1}^{+\infty} m(2m+1)\,\mathrm{e}^{-m\,2x^2} \leqslant \frac{5}{(1-\mathrm{e}^{-2x^2})^3}$$

Now, since $\theta \in (0,+\infty) \mapsto \theta/(1-\mathrm{e}^{-\theta})$ is increasing and since $2x^2 \leqslant 2$, we have

$$\frac{2x^2}{1-\mathrm{e}^{-2x^2}} \leqslant \frac{2}{1-\mathrm{e}^{-2}} \qquad \text{thus} \qquad \frac{1}{(1-\mathrm{e}^{-2x^2})^3} \leqslant \frac{1}{x^6(1-\mathrm{e}^{-2})^3}$$

which concludes the proof of the final inequality in (14), thus the proof of this lemma, and finally, the treatment of the first sum in (8). ■

Control of the second sum in (8). By what are routine manipulations now, namely, distinguishing whether $N_a(t)$ is larger or smaller than $f(T,K)$ and by optional skipping (see Section 5.1), we have

$$\sum_{t=K}^{T-1} \mathbb{P}\big[U_a(t) \geqslant \mu^\star - \delta \text{ and } A_{t+1} = a\big] \leqslant \sum_{n=f(T,K)+1}^{T} \mathbb{P}\big[U_{a,n}^{\mathrm{M}} \geqslant \mu^\star - \delta\big] + \sum_{n=1}^{f(T,K)} \mathbb{P}\big[U_{a,n}^{\mathrm{KL}} \geqslant \mu^\star - \delta\big] \tag{15}$$

Let us denote

$$\gamma_\star = \frac{1}{\sqrt{1-\mu^\star}}\left(16\mathrm{e}^{-2} + \ln^2\left(\frac{1}{1-\mu^\star}\right)\right)$$

as in the statement of Proposition 13. For T large enough to satisfy (17) and for δ small enough so that $\delta \leqslant \Delta_a/2$ and $\delta^2 \leqslant \gamma_\star(1-\mu^\star)^2/2$, we further upper bound below (15) by

$$\frac{K\,f(T,K)/T}{1-\mathrm{e}^{\Delta_a^2/8}} + \frac{1}{1-\mathrm{e}^{-\delta^2/(2\gamma_\star(1-\mu^\star)^2)}} + \frac{\ln(T/K)}{\mathcal{K}_{\inf}(\nu_a,\mu^\star) - 2\delta/(1-\mu^\star)} \tag{16}$$

The obtained bound indeed equals $(\ln T)/\mathcal{K}_{\inf}(\nu_a,\mu^\star) + \mathcal{O}_T\big((\ln T)^{2/3}\big)$ for the considered choice $\delta = (\ln T)^{-1/3}$, as can be seen by noting that the first term in (16) can be bounded by a constant, while the second and third terms can be dealt with by resorting, respectively, to $1-\mathrm{e}^{-u} = u + \mathrm{o}(u)$ and $1/(1-u) = 1+u+o(u)$ as $u \to 0$, where u is proportional, respectively, to δ^2 and δ.

We turn to the proof of (16). We deal with the first sum in the right-hand side of (15) as around (9): provided that T is large enough, so that

$$\frac{\ln\Big(T/\big(K\,f(T,K)\big)\Big)}{f(T,K)} \leqslant \frac{\Delta_a^2}{8}\,, \tag{17}$$

we have, for $f(T,K)+1 \leqslant n < T/K$,

$$U_{a,n}^{\mathrm{M}} \leqslant U_{a,\,f(T,K)}^{\mathrm{M}} = \widehat{\mu}_{a,n} + \sqrt{\frac{1}{2f(T,K)}\ln\left(\frac{T}{K\,f(T,K)}\right)} \leqslant \widehat{\mu}_{a,n} + \frac{\Delta_a}{4} \tag{18}$$

Note that the $U_{a,n}^{\mathrm{M}} \leqslant \widehat{\mu}_{a,n} + \Delta_a/4$ bound is valid even when $n \geqslant T/K$, as the exploration then vanishes. Therefore, as in (12), as soon as $\delta \leqslant \Delta_a/2$,

$$\begin{aligned}
&\sum_{n=f(T,K)+1}^{T} \mathbb{P}\big[U_{a,n}^{\mathrm{M}} \geqslant \mu^\star - \delta\big] \leqslant \sum_{n=f(T,K)+1}^{T} \mathbb{P}\left[U_{a,n}^{\mathrm{M}} \geqslant \mu^\star - \frac{\Delta_a}{2}\right] \\
&\leqslant \sum_{n=f(T,K)+1}^{T} \mathbb{P}\left[\widehat{\mu}_{a,n} + \frac{\Delta_a}{4} \geqslant \mu^\star - \frac{\Delta_a}{2}\right] = \sum_{n=f(T,K)+1}^{T} \mathbb{P}\left[\widehat{\mu}_{a,n} \geqslant \mu_a + \frac{\Delta_a}{4}\right] \\
&\leqslant \sum_{n=f(T,K)+1}^{T} \mathrm{e}^{-n\Delta_a^2/8} \leqslant \frac{\mathrm{e}^{-f(T,K)\Delta_a^2/8}}{1-\mathrm{e}^{-\Delta_a^2/8}} \leqslant \frac{K\,f(T,K)/T}{1-\mathrm{e}^{-\Delta_a^2/8}}
\end{aligned} \tag{19}$$

where we used again condition (17) to get the last inequality.

It only remains to deal with the second sum in the right-hand side of (15). Let

$$n_1 = \left\lceil \frac{\ln(T/K)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star) - 2\delta/(1-\mu^\star)} \right\rceil$$

For $n_1 \leqslant n \leqslant f(T,K) < T/K$, by definition of the index as a supremum and by left-continuity of $\mathcal{K}_{\inf}$ (see the comments after Lemma 10),

$$\begin{aligned}
\big\{U_{a,n}^{\mathrm{KL}} \geqslant \mu^\star - \delta\big\} &\subseteq \left\{\mathcal{K}_{\inf}\big(\widehat{\nu}_{a,n}, \mu^\star - \delta\big) \leqslant \frac{1}{n}\ln\left(\frac{T}{Kn}\right)\right\} \\
&\subseteq \Big\{\mathcal{K}_{\inf}\big(\widehat{\nu}_{a,n}, \mu^\star - \delta\big) \leqslant \mathcal{K}_{\inf}(\nu_a, \mu^\star) - 2\delta/(1-\mu^\star)\Big\} \\
&\subseteq \Big\{\mathcal{K}_{\inf}\big(\widehat{\nu}_{a,n}, \mu^\star\big) \leqslant \mathcal{K}_{\inf}(\nu_a, \mu^\star) - \delta/(1-\mu^\star))\Big\}
\end{aligned} \tag{20}$$

where the second inclusion only uses $n \geqslant n_1$ and the definition of n_1, and the last inclusion holds by the regularity inequality (29). Therefore we may resort to the concentration inequality on $\mathcal{K}_{\inf}$, stated as Proposition 13: we get, for all $n_1 \leqslant n \leqslant f(T,K)$,

$$\mathbb{P}\big[U_{a,n}^{\mathrm{KL}} \geqslant \mu^\star - \delta\big] \leqslant \max\left\{\mathrm{e}^{-n/4},\ \exp\left(-\frac{n\delta^2}{2\gamma_\star(1-\mu^\star)^2}\right)\right\}$$

and whether the first or the second argument of the maximum is the largest is independent of n. Therefore, a summation over $n \geqslant n_1$ keeping the latter remark in mind leads to

$$\sum_{n=n_1}^{f(T,K)} \mathbb{P}\big[U_{a,n}^{\mathrm{KL}} \geqslant \mu^\star - \delta\big] \leqslant \max\left\{\frac{1}{1-\mathrm{e}^{-1/4}},\ \frac{1}{1-\mathrm{e}^{-\delta^2/(2\gamma_\star(1-\mu^\star)^2)}}\right\} \tag{21}$$

Now if $\delta^2 \leqslant \gamma_\star(1-\mu^\star)^2/2$ we may keep only the second term in the maximum. For such δ, by bounding by 1 the first $n_1 - 1$ probabilities in the sum in (15), we finally obtain

$$\sum_{n=1}^{f(T,K)} \mathbb{P}\big[U_{a,n}^{\mathrm{KL}} \geqslant \mu^\star - \delta\big] \leqslant \frac{\ln(T/K)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star) - 2\delta/(1-\mu^\star)} + \frac{1}{1-\mathrm{e}^{-\delta^2/(2\gamma_\star(1-\mu^\star)^2)}} \tag{22}$$

which concludes the proof of (16). ■

5. Results (more or less) extracted from the literature

We gather in this section results that are all known and published elsewhere (or almost). For the sake of self-completeness we provide a proof of each of them (sometimes this proof is shorter or simpler than the known proofs, and we then comment on this fact).

5.1. Optional skipping

The trick detailed here is standard in the bandit literature, see, e.g., its application in Auer et al. (2002a).

We detail how to reindex various quantities like $U_a(t)$, $\widehat{\mu}_a(t)$, etc., that are indexed by the global time t, into versions indexed by the local number of times $N_a(t) = n$ the specific arm considered has been pulled. The corresponding quantities will be denoted by $U_{a,n}$, $\widehat{\mu}_{a,n}$, etc.

The reindexation is possible as soon as the considered algorithm pulls each arm infinitely often; it is the case for all algorithms considered in this article (exploration never stops even if it becomes rare after a certain time).

We denote by $\mathcal{F}_0 = \{\emptyset, \Omega\}$ the trivial σ–algebra and by $\mathcal{F}_t$ the σ–algebra generated by $A_1, Y_1, \ldots, A_t, Y_t$, when $t \geqslant 1$. We fix an arm a. For each $n \geqslant 1$, we denote by

$$\tau_{a,n} = \min\bigl\{t \geqslant 1 : \ \ N_a(t) = n\bigr\}$$

the round at which arm a was pulled for the n–th time. Doob's optional skipping (see, e.g., Chow and Teicher, 1988, Section 5.3 for a reference) ensures that the random variables $X_{a,n} = Y_{\tau_{a,n}}$ are independent and identically distributed according to ν_a.

We can then define, for instance, for $n \geqslant 1$,

$$\widehat{\mu}_{a,n} = \frac{1}{n}\sum_{k=1}^{n} X_{a,k}$$

and have the equality $\widehat{\mu}_a(t) = \widehat{\mu}_{a,N_a(t)}$ for $t \geqslant K$. Here is an example of how to use this rewriting. Recall that $N_a(t) \geqslant 1$ for $t \geqslant K$ and $N_a(t) \leqslant t - K + 1$ as each arm was pulled once in the first rounds. Given a subset $\mathcal{E} \subseteq [0,1]$, we get the inclusion

$$\bigl\{\widehat{\mu}_a(t) \in \mathcal{E}\bigr\} = \bigcup_{n=1}^{t-K+1} \bigl\{\widehat{\mu}_a(t) \in \mathcal{E} \text{ and } N_a(t) = n\bigr\} = \bigcup_{n=1}^{t-K+1} \bigl\{\widehat{\mu}_{a,n} \in \mathcal{E} \text{ and } N_a(t) = n\bigr\}$$

so that, by a union bound,

$$\mathbb{P}\bigl[\widehat{\mu}_a(t) \in \mathcal{E}\bigr] \leqslant \sum_{n=1}^{t-K+1} \mathbb{P}\bigl[\widehat{\mu}_{a,n} \in \mathcal{E} \text{ and } N_a(t) = n\bigr] \leqslant \sum_{n=1}^{t-K+1} \mathbb{P}\bigl[\widehat{\mu}_{a,n} \in \mathcal{E}\bigr]\,.$$

The last sum above only deals with independent and identically distributed random variables; we took care of all dependency issues that are so present in bandit problems. The price to pay, however, is that we bounded one probability by a sum of probabilities.

5.2. Maximal Hoeffding's inequality

This standard result from Hoeffding (1963) was already used in the proof of the regret bound of MOSS (Audibert and Bubeck, 2009).

Proposition 7 *Let $X_1, \dots, X_n$ be a sequence of i.i.d. random variables bounded in $[0,1]$ and let $\widehat{\mu}_n$ denote their empirical mean. Then for all $u > 0$ and for all $N \geqslant 1$:*

$$\mathbb{P}\left[\max_{n \geqslant N} \left(\widehat{\mu}_n - \mu\right) \geqslant u\right] \leqslant \mathrm{e}^{-2Nu^2} \tag{23}$$

Corollary 8 *Under the same assumptions, for all $\varepsilon > 0$,*

$$\mathbb{E}\left[\left(\max_{n \geqslant N} \left(\mu - \widehat{\mu}_n - \varepsilon\right)\right)^+\right] \leqslant \sqrt{\frac{\pi}{8}}\sqrt{\frac{1}{N}}\,\mathrm{e}^{-2N\varepsilon^2} \tag{24}$$

Proof By Fubini-Tonelli, an integration of the maximal concentration inequality yields

$$\begin{aligned}\mathbb{E}\left[\left(\max_{n \geqslant N} \left(\mu - \widehat{\mu}_n - \varepsilon\right)\right)^+\right] &= \int_0^{+\infty} \mathbb{P}\left[\max_{n \geqslant N} \left(\widehat{\mu}_n - \mu - \varepsilon\right) \geqslant u\right] \mathrm{d}u \\ &\leqslant \int_0^{+\infty} \mathrm{e}^{-2N(u+\varepsilon)^2}\,\mathrm{d}u \leqslant \mathrm{e}^{-2N\varepsilon^2} \int_0^{+\infty} \mathrm{e}^{-2Nu^2}\,\mathrm{d}u = \sqrt{\frac{\pi}{8}}\sqrt{\frac{1}{N}}\,\mathrm{e}^{-2N\varepsilon^2}\end{aligned}$$

■

5.3. Analysis of the MOSS algorithm

This analysis was already performed in the literature, both for a known horizon T (see Audibert and Bubeck 2009) and for an anytime version (see Degenne and Perchet 2016). We provide slightly shorter and more focused proofs of these results based on Proposition 8 in Appendix B; the main difference to the mentioned proofs lies in elegance. Typically, the peeling trick was used on the probabilities of deviations (see Proposition 7) and had to be performed separately and differently for each deviation u; then, these probabilities were integrated to obtain a control on the needed expectations. In contrast, we perform the peeling trick directly on the expectations at hand, and we do so by applying it only once, at fixed times depending solely on T, which makes the proof more readable. Put differently, we do not claim any improvement on the results themselves, just a clarification of their proof.

We first recall the distribution-free bound on the regret of MOSS, when T is known. We also extract an intermediary result from its proof, which will be used in the analysis of our algorithm. We denote by A_t^{M} the arm played by the index strategy maximizing, at each step $t+1$ with $t \geqslant K$, the quantity

$$U_a^{\mathrm{M}}(t) \stackrel{\text{def}}{=} \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)} \ln_+\left(\frac{T}{KN_a(t)}\right)}$$

Proposition 9 *For all bandit problems $\underline{\nu}$ over $[0,1]$, the regret of MOSS satisfies*

$$R_T \leqslant (K-1) + 17\sqrt{KT}$$

More precisely, we have the inequalities

$$R_T - (K-1) \leqslant \sqrt{KT} + \underbrace{\sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star}^{\mathrm{M}}(t)\big)^+\Big] + \sum_{a=1}^{K}\sum_{n=1}^{T} \mathbb{E}\Big[\big(U_{a,n}^{\mathrm{M}} - \mu_a - \sqrt{K/T}\big)^+\Big]}_{\leqslant 16\sqrt{KT}} \tag{25}$$

Our proof in Appendix B reveals that designing an adaptive version of MOSS comes at no effort; indeed, MOSS-anytime relies on the indexes, for $t \geqslant K$,

$$U_a^{\text{M-A}}(t) \stackrel{\text{def}}{=} \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)} \ln_+\left(\frac{t}{KN_a(t)}\right)} \tag{26}$$

and satisfies a regret bound of $(K-1) + 29\sqrt{KT}$.

5.4. Regularity and deviation/concentration results on $\mathcal{K}_{\inf}$

Many results of this section rely on Pinsker's inequality. One of its most basic consequences is in terms of a lower bound on $\mathcal{K}_{\inf}$. Indeed, since we are considering distributions over $[0,1]$, the data-processing inequality for Kullback-Leibler divergences ensures (see, e.g., Garivier et al., 2018, Lemma 1) that for all $\nu \in \mathcal{P}[0,1]$ and all $\mu \in \big(\mathrm{E}(\nu), 1\big)$,

$$\mathcal{K}_{\inf}(\nu,\mu) \geqslant \inf_{\nu' : \mathrm{E}(\nu') > \mu} \mathrm{KL}\Big(\mathrm{Ber}\big(\mathrm{E}(\nu)\big), \mathrm{Ber}\big(\mathrm{E}(\nu')\big)\Big) = \mathrm{KL}\Big(\mathrm{Ber}\big(\mathrm{E}(\nu)\big), \mathrm{Ber}\big(\mu\big)\Big)\,,$$

where $\mathrm{Ber}(p)$ denotes the Bernoulli distribution with parameter p. Therefore, by Pinsker's inequality for Bernoulli distributions,

$$\mathcal{K}_{\inf}(\nu,\mu) \geqslant 2\big(\mathrm{E}(\nu) - \mu\big)^2\,, \qquad \text{thus} \qquad U_a^{\mathrm{KL}}(t) \leqslant U_a^{\mathrm{M}}(t) \tag{27}$$

for all arms a and all rounds $t \geqslant K$. In particular, for KL-UCB-switch,

$$U_a^{\mathrm{KL}}(t) \leqslant U_a(t) \leqslant U_a^{\mathrm{M}}(t) \tag{28}$$

Another consequence of Pinsker's inequality is given by the inequality (30) below, while the inequality (29) appears as Lemma 7 in Honda and Takemura (2015), see also Garivier et al. (2018, Lemma 3) for a later but much simpler proof of (29). These two inequalities are proved in details in Section D; the proposed proofs are slightly simpler or lead to sharper bounds than in the mentioned references.

Lemma 10 (regularity of $\mathcal{K}_{\inf}$) *For all $\nu \in \mathcal{P}[0,1]$ and all $\mu \in (0,1)$,*

$$\forall \varepsilon \in (0, \mu)\,, \qquad \mathcal{K}_{\inf}(\nu,\mu) \leqslant \mathcal{K}_{\inf}(\nu, \mu-\varepsilon) + \frac{\varepsilon}{1-\mu}\,, \tag{29}$$

and

$$\forall \varepsilon \in \big[0, \mu - \mathrm{E}(\nu)\big]\,, \qquad \mathcal{K}_{\inf}(\nu,\mu) \geqslant \mathcal{K}_{\inf}(\nu, \mu-\varepsilon) + 2\varepsilon^2\,. \tag{30}$$

A consequence of Lemma 10 is the left-continuity of $\mathcal{K}_{\inf}$: for all $\nu \in \mathcal{P}[0,1]$ and all $\mu \in (0,1)$, we have $\mathcal{K}_{\inf}(\nu, \mu-\varepsilon) \nearrow \mathcal{K}_{\inf}(\nu,\mu)$ as $\varepsilon \searrow 0$. Therefore, by a sandwich argument, $\mathcal{K}_{\inf}\big(\nu, \mathrm{E}(\nu)\big) = 0$ whenever $\mathrm{E}(\nu) \in (0,1)$.

A consequence of (30) is the following. For all $B > 0$, all $\tilde{\mu} \in (0,1)$, all $\varepsilon \in [0, \tilde{\mu})$, and all distributions ν over $[0,1]$ with $\mathrm{E}(\nu) < \tilde{\mu} - \varepsilon$,

$$\Big\{\sup\big\{\mu \in [0,1] \;\big|\; \mathcal{K}_{\inf}(\nu,\mu) \leqslant B\big\} < \tilde{\mu}-\varepsilon\Big\} \subseteq \Big\{\mathcal{K}_{\inf}\big(\nu, \tilde{\mu}-\varepsilon\big) > B\Big\} \subseteq \Big\{\mathcal{K}_{\inf}\big(\nu,\tilde{\mu}\big) > B+2\varepsilon^2\Big\}, \tag{31}$$

and these inclusions still hold even when $\mathrm{E}(\nu) \geqslant \tilde{\mu} - \varepsilon$, as in this case, the left-most set is empty.

The variational formula appears in Honda and Takemura (2015) as Theorem 2 (and Lemma 6) and is an essential tool for deriving concentration results for the $\mathcal{K}_{\inf}$. We re-derive it in an elegant and direct way in Section D.

Lemma 11 (variational formula for $\mathcal{K}_{\inf}$) *For all $\nu \in \mathcal{P}[0,1]$ and all $0 < \mu < 1$,*

$$\mathcal{K}_{\inf}(\nu,\mu) = \max_{0 \leqslant \lambda \leqslant 1} \mathbb{E}\left[\ln\left(1 - \lambda \frac{X-\mu}{1-\mu}\right)\right] \qquad \text{where } X \sim \nu \tag{32}$$

Moreover, if we denote by $\lambda^\star$ the value at which the above maximum is reached, then

$$\mathbb{E}\left[\frac{1}{1-\lambda^\star(X-\mu)/(1-\mu)}\right] \leqslant 1 \tag{33}$$

The following deviation inequality on $\mathcal{K}_{\inf}$ was provided by Cappé et al. (2013, Lemma 6) in all cases where the variational formula (32) holds. For the sake of completeness, we recall its proof in Section D.

Proposition 12 (deviation result on $\mathcal{K}_{\inf}$) *Let $\widehat{\nu}_n$ denote the empirical distribution associated with a sequence of n i.i.d. random variables with distribution ν over $[0,1]$ with $\mathrm{E}(\nu) \in (0,1)$. Then, for all $u \geqslant 0$,*

$$\mathbb{P}\Big[\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu)\big) \geqslant u\Big] \leqslant \mathrm{e}(2n+1)\,\mathrm{e}^{-nu} \tag{34}$$

A consequence of the proposition above and of Lemma 10 is the following one: for all $u > 0$ and all $\varepsilon \in \big[0, \mathrm{E}(\nu)\big)$,

$$\mathbb{P}\Big[\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu) - \varepsilon\big) \geqslant u\Big] \leqslant \mathrm{e}(2n+1)\,\mathrm{e}^{-n(u+2\varepsilon^2)} \tag{35}$$

Indeed, when ε is such that $\mathrm{E}(\nu) - \varepsilon < \widehat{\mu}_n$, where $\widehat{\mu}_n$ denotes the average of the considered i.i.d. random variables, then $\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu) - \varepsilon\big) = 0$ by definition, while otherwise, by (30), since $\varepsilon \in \big[0, \mathrm{E}(\nu) - \widehat{\mu}_n\big]$, we have

$$\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu) - \varepsilon\big) + 2\varepsilon^2 \leqslant \mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu)\big)\,.$$

Therefore, the inclusion

$$\Big\{\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu) - \varepsilon\big) \geqslant u\Big\} \subseteq \Big\{\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu)\big) \geqslant u + 2\varepsilon^2\Big\}$$

is valid for all $u > 0$ and (35) follows from Proposition 12.

The next proposition is similar in spirit to Honda and Takemura (2015, Proposition 11) but is better suited to our needs. We prove it in Appendix A.

Proposition 13 (concentration result on $\mathcal{K}_{\inf}$) *With the same notation and assumptions as in the previous proposition, consider a real number $\mu^\star \in \big(\mathrm{E}(\nu), 1\big)$ and define*

$$\gamma_\star = \frac{1}{\sqrt{1-\mu^\star}}\left(16\mathrm{e}^{-2} + \ln^2\left(\frac{1}{1-\mu^\star}\right)\right) \tag{36}$$

Then for all $x < \mathcal{K}_{\inf}(\nu, \mu^\star)$,

$$\mathbb{P}\big[\mathcal{K}_{\inf}(\widehat{\nu}_n, \mu^\star) \leqslant x\big] \leqslant \begin{cases} \exp(-n\gamma_\star/8) \leqslant \exp(-n/4) & \text{if } x \leqslant \mathcal{K}_{\inf}(\nu,\mu^\star) - \gamma_\star/2 \\ \exp\Big(-n\big(\mathcal{K}_{\inf}(\nu,\mu^\star) - x\big)^2/(2\gamma_\star)\Big) & \text{if } x > \mathcal{K}_{\inf}(\nu,\mu^\star) - \gamma_\star/2 \end{cases}$$

6. Proof of the more advanced bound of Theorem 3

The proof of the sharper bound of Theorem 3 relies on the following lemma, which was (almost) stated in Honda and Takemura (2015, Lemma 18): our assumptions and result are slightly different (they are tailored to our needs), which is why we provide below a proof of this lemma.

By convention, the infimum over an empty set equals $+\infty$. In what follows, $\wedge$ denotes the minimum of two numbers; the considered stopping time τ is thus always bounded by T. We recall that Lambert's function W is defined, for $x > 0$, as the unique solution $W(x)$ of the equation $w\,\mathrm{e}^w = x$, with unknown $w > 0$. We recall (see, e.g., Hoorfar and Hassani, 2008, Corollary 2.4) that it is increasing and that

$$\forall\, x > \mathrm{e}, \qquad \ln x - \ln\ln x \leqslant W(x) \leqslant \ln x - \ln\ln x + \ln\big(1 + \mathrm{e}^{-1}\big) \tag{37}$$

and in particular, $W(x) = \ln x - \ln\ln x + \mathcal{O}(1)$ as $x \to +\infty$.

Lemma 14 *Let (Z_i) be a sequence of i.i.d. variables with a positive expectation $\mathbb{E}[Z_1] > 0$ and such that $Z_i \leqslant \alpha$ for some $\alpha > 0$. For an integer $T \geqslant 1$, consider the stopping time*

$$\tau \overset{\text{def}}{=} \inf\left\{ n \geqslant 1 \;\middle|\; \sum_{i=1}^n Z_i > \ln\left(\frac{T}{Kn}\right) \right\} \wedge T$$

Then, for all $T \geqslant K\mathrm{e}^\alpha$,

$$\mathbb{E}[\tau] \leqslant \frac{W(\alpha T/K) + \alpha + \ln 2}{\mathbb{E}[Z_1]}$$

where W is Lambert's function.

Proof We consider the martingale $(M_n)_{n\geqslant 0}$ defined by

$$M_n = \sum_{i=1}^n \big(Z_i - \mathbb{E}[Z_1]\big)$$

As τ is a finite stopping time, Doob's optional stopping theorem indicates that $\mathbb{E}[M_\tau] = \mathbb{E}[M_0] = 0$, that is,

$$\mathbb{E}[\tau]\,\mathbb{E}[Z_1] = \mathbb{E}\left[\sum_{i=1}^\tau Z_i\right]$$

That first step of the proof was similar to the one of Honda and Takemura (2015, Lemma 18). The idea is now to upper bound the right-hand side of the above equality, which we do by resorting to the very definition of τ. An adaptation is needed with respect to the original argument as the value $\ln\big(T/(Kn)\big)$ of the barrier varies with n.

We proceed as follows. Since $Z_1 \leqslant \alpha$ and $T \geqslant K\mathrm{e}^{\alpha}$ by assumption, we necessarily have $\tau \geqslant 2$; using again the boundedness by α, we have, by definition of τ,

$$\sum_{i=1}^{\tau-1} Z_i + Z_\tau \leqslant \ln\left(\frac{T}{K(\tau-1)}\right) + \alpha = \ln\left(\frac{T}{K\tau}\right) + \ln\left(\frac{\tau}{\tau-1}\right) + \alpha \leqslant \ln\left(\frac{T}{K\tau}\right) + \ln 2 + \alpha$$

In addition, when $\tau < T/K$,

$$\ln\left(\frac{T}{K\tau}\right) < \sum_{i=1}^{\tau} Z_i \leqslant \tau\alpha \qquad \text{thus} \qquad 0 < \frac{T}{K\tau}\,\ln\left(\frac{T}{K\tau}\right) \leqslant \frac{T\alpha}{K}$$

Applying the increasing function W to all sides of the latter inequality, we get, when $\tau < T/K$,

$$\ln\left(\frac{T}{K\tau}\right) \leqslant W\left(\frac{T\alpha}{K}\right)$$

This inequality also holds when $\tau \geqslant T/K$ as the left-hand side then is non-positive, while the right-hand side is positive. Putting all elements together, we successively proved

$$\mathbb{E}[\tau]\,\mathbb{E}[Z_1] = \mathbb{E}\left[\sum_{i=1}^{\tau} Z_i\right] \leqslant \mathbb{E}\left[\ln\left(\frac{T}{K\tau}\right)\right] + \ln 2 + \alpha \leqslant W\left(\frac{T\alpha}{K}\right) + \ln 2 + \alpha$$

which concludes the proof. ∎

Proof of Theorem 3 All inequalities of the proof of Theorem 2 hold in the present case as well, given that we are studying exactly the same algorithm. The regret is decomposed as in the mentioned proof, and inequality (13) holds as a first part of the final regret bound. Now, the second part consists of (15), which we bound as (19) plus the bound

$$\sum_{n=1}^{f(T,K)} \mathbb{P}\big[U_{a,n}^{\mathrm{KL}} \geqslant \mu^\star - \delta\big] \leqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a,\mu^\star) - \delta/(1-\mu^\star)}\left(W\left(\frac{\ln\big(1/(1-\mu^\star)\big)}{K}\,T\right) + \ln\big(2/(1-\mu^\star)\big)\right) + 5 + \frac{1}{1-\mathrm{e}^{-\mathcal{K}_{\inf}(\nu,\mu^\star)^2/(8\gamma_\star)}} \tag{38}$$

where again,

$$\gamma_\star = \frac{1}{\sqrt{1-\mu^\star}}\left(16\mathrm{e}^{-2} + \ln^2\left(\frac{1}{1-\mu^\star}\right)\right)$$

To do so, we use the conditions (17) and $T > K/(1-\mu^\star)$ on T, and the conditions $\delta \leqslant \Delta_a/2$ and $\delta \leqslant \mathcal{K}_{\inf}(\nu_a,\mu^\star)/\big(2(1-\mu^\star)\big)$ on δ; all in all, we get

$$\sum_{t=K}^{T-1} \mathbb{P}\big[U_a(t) \geqslant \mu^\star - \delta \text{ and } A_{t+1} = a\big] \leqslant \frac{K\,f(T,K)/T}{1-\mathrm{e}^{\Delta_a^2/8}} + 5 + \frac{1}{1-\mathrm{e}^{-\mathcal{K}_{\inf}(\nu,\mu^\star)^2/(8\gamma_\star)}} + \frac{1}{\mathcal{K}_{\inf}(\nu_a,\mu^\star) - \delta/(1-\mu^\star)}\left(W\left(\frac{\ln\big(1/(1-\mu^\star)\big)}{K}\,T\right) + \ln\big(2/(1-\mu^\star)\big)\right) \tag{39}$$

Since $1/(1-u) = 1 + \mathcal{O}(u)$ as $u \to 0$, for the choice $\delta = T^{-1/8}$ contemplated in Theorem 3, the bound above equals

$$\frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star) - \delta/(1-\mu^\star)} W\left(\frac{\ln\big(1/(1-\mu^\star)\big)}{K}\, T\right) + \mathcal{O}_T(1)$$
$$= \frac{W(T)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}\big(1 + \mathcal{O}_T(T^{-1/8})\big) + \mathcal{O}_T(1) = \frac{\ln T - \ln\ln T}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} + \mathcal{O}_T(1)\,,$$

where the final equality follows from the asymptotic expansion (37).

The difference with the proof of Theorem 2 lies in a sharper bound of the quantity (38), given by the last two terms in the above inequality (39). We follow exactly the same method as in the analysis of the IMED policy of Honda and Takemura (2015, Theorem 5): their idea was to deal with the deviations in a more careful way and relate the sum (38) to the behaviour of a biased random walk.

We start by following the same steps as in the proof of Proposition 13 in Appendix A and link the deviations in $\mathcal{K}_{\inf}$ divergence to the ones of a random walk. The variational formulation (Lemma 11) for $\mathcal{K}_{\inf}$ entails the existence of $\lambda_{a,\delta} \in [0,1]$ such that

$$\mathcal{K}_{\inf}(\nu_a, \mu^\star - \delta) = \mathbb{E}\left[\ln\left(1 - \lambda_{a,\delta}\frac{X_a - (\mu^\star - \delta)}{1 - (\mu^\star - \delta)}\right)\right] \qquad \text{where} \qquad X_a \sim \nu_a$$

Note that $\mathcal{K}_{\inf}(\nu_a, \mu^\star - \delta) > 0$ by (27) given that $\delta \leqslant \Delta_a/2$. We consider i.i.d. copies $X_{a,1}, \dots, X_{a,n}$ of X and form the random variables

$$Z_{a,i} = \ln\left(1 - \lambda_{a,\delta}\frac{X_{a,i} - (\mu^\star - \delta)}{1 - (\mu^\star - \delta)}\right)$$

where, since $X_{a,i} \geqslant 0$ and $\lambda_{a,\delta} \in [0,1]$, we have

$$Z_{a,i} = \ln\left(1 - \lambda_{a,\delta}\frac{X_{a,i} - (\mu^\star - \delta)}{1 - (\mu^\star - \delta)}\right) \leqslant \ln\left(1 + \lambda_{a,\delta}\frac{\mu^\star - \delta}{1 - (\mu^\star - \delta)}\right)$$
$$\leqslant \ln\left(1 + \frac{\mu^\star - \delta}{1 - (\mu^\star - \delta)}\right) = \ln\left(\frac{1}{1 - (\mu^\star - \delta)}\right) \leqslant \ln\left(\frac{1}{1-\mu^\star}\right) \stackrel{\text{def}}{=} \alpha$$

By the variational formulation again, applied this time to $\mathcal{K}_{\inf}(\widehat{\nu}_{a,n}, \mu^\star - \delta)$,

$$\mathcal{K}_{\inf}\big(\widehat{\nu}_{a,n}, \mu^\star - \delta\big) \geqslant \frac{1}{n}\sum_{i=1}^{n} Z_{a,i}$$

which entails, for each $n \geqslant 1$,

$$\big\{U_{a,n}^{\mathrm{KL}} \geqslant \mu^\star - \delta\big\} \subseteq \left\{\mathcal{K}_{\inf}\big(\widehat{\nu}_{a,n}, \mu^\star - \delta\big) \leqslant \frac{1}{n}\ln\left(\frac{T}{Kn}\right)\right\} \subseteq \left\{\sum_{i=1}^{n} Z_{a,i} \leqslant \ln\left(\frac{T}{Kn}\right)\right\} \qquad (40)$$

where the first inclusion holds for the same reasons (including left-continuity of $\mathcal{K}_{\inf}$) as in (20). Therefore, the quantity of interest (38) is bounded by

$$\sum_{n=1}^{f(T,K)} \mathbb{P}\big[U_{a,n}^{\mathrm{KL}} \geqslant \mu^\star - \delta\big] \leqslant \sum_{n=1}^{f(T,K)} \mathbb{P}\left[\sum_{i=1}^{n} Z_{a,i} \leqslant \ln\left(\frac{T}{Kn}\right)\right] = \mathbb{E}\left[\sum_{n=1}^{f(T,K)} \mathbb{1}_{\left\{\sum_{i=1}^{n} Z_{a,i} \leqslant \ln(T/(Kn))\right\}}\right]$$
$$\leqslant \mathbb{E}\left[\sum_{n=1}^{T} \mathbb{1}_{\left\{\sum_{i=1}^{n} Z_{a,i} \leqslant \ln(T/(Kn))\right\}}\right]$$

This latter sum can be reinterpreted as the expected number of times a random walk with positive bias stays under a decreasing logarithmic barrier. We exploit this interpretation to our advantage by decomposing this sum into the expected hitting time of the barrier and a sum of deviation probabilities for the walk.

Let us therefore define the first hitting time τ of the barrier

$$\tau = \inf\left\{ n \geqslant 1 \;\middle|\; \sum_{i=1}^{n} Z_{a,i} > \ln\left(\frac{T}{Kn}\right)\right\} \wedge T \tag{41}$$

which is is a stopping time with respect to the filtration generated by the family $(Z_{a,i})_{1\leqslant i\leqslant n}$. By distinguishing according to whether or not the condition in the defining infimum of τ is met for a $1 \leqslant n \leqslant T$ or not, i.e., whether or not the barrier is hit for $1 \leqslant n \leqslant T$, we get

$$\mathbb{E}\left[\sum_{n=1}^{T} \mathbb{1}_{\left\{\sum_{i=1}^{n} Z_{a,i} \leqslant \ln(T/(Kn))\right\}}\right] \leqslant \mathbb{E}[\tau] + \mathbb{E}\left[\sum_{n=\tau+1}^{T} \mathbb{1}_{\left\{\sum_{i=1}^{n} Z_{a,i} \leqslant \ln(T/(Kn))\right\}}\right] \tag{42}$$

where the sum from $\tau+1$ to T is void thus null when $\tau = T$ (this is the case, in particular, when the barrier is hit for no $n \leqslant T$). Lemma 14 applies, as, among others, $Z_{a,i} \leqslant \alpha = \ln\big(1/(1-\mu^\star)\big)$ as shown above and $T > K/(1-\mu^\star)$; it yields

$$\mathbb{E}[\tau] \leqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star - \delta)}\left(W\left(\frac{\ln\big(1/(1-\mu^\star)\big)}{K}\,T\right) + \ln\big(2/(1-\mu^\star)\big)\right)$$

We apply the regularity inequality on $\mathcal{K}_{\inf}$, see also (45) below, to get the claimed bound on the first part of (42) We now bound its second part. We may assume that $\tau < T$ so that

$$\ln\left(\frac{T}{K\tau}\right) < \sum_{i=1}^{\tau} Z_{a,i}$$

For $n \geqslant \tau$, we then have

$$\sum_{i=1}^{n} Z_{a,i} \leqslant \ln\left(\frac{T}{Kn}\right) \qquad \text{implies} \qquad \sum_{i=1}^{n} Z_{a,i} \leqslant \ln\left(\frac{T}{K\tau}\right) \leqslant \sum_{i=1}^{\tau} Z_{a,i} \tag{43}$$

Hence, in this case,

$$\sum_{i=1}^{n} Z_{a,i} \leqslant \ln\left(\frac{T}{Kn}\right) \qquad \text{implies} \qquad \sum_{i=\tau+1}^{n} Z_{a,i} \leqslant 0\,.$$

This, together with a breakdown according to the values of τ and the independence between $\{\tau = k\}$ and $X_{k+1}, \dots, X_T$, yields

$$\begin{aligned}
&\mathbb{E}\left[\sum_{n=\tau+1}^{T} \mathbb{1}_{\left\{\sum_{i=1}^{n} Z_{a,i} \leqslant \ln(T/(Kn))\right\}}\right] \\
&\leqslant \mathbb{E}\left[\sum_{n=\tau+1}^{T} \mathbb{1}_{\left\{\sum_{i=\tau+1}^{n} Z_{a,i} \leqslant 0\right\}}\right] = \sum_{k=1}^{T} \mathbb{E}\left[\mathbb{1}_{\{\tau=k\}} \sum_{n=k+1}^{T} \mathbb{1}_{\left\{\sum_{i=k+1}^{n} Z_{a,i} \leqslant 0\right\}}\right] \\
&= \sum_{k=1}^{T} \sum_{n=k+1}^{T} \mathbb{P}[\tau = k]\, \mathbb{P}\left[\sum_{i=k+1}^{n} Z_{a,i} \leqslant 0\right] \\
&= \sum_{k=1}^{T} \mathbb{P}[\tau = k] \left(\underbrace{\sum_{n=k+1}^{T} \mathbb{P}\left[\sum_{i=k+1}^{n} Z_{a,i} \leqslant 0\right]}_{\text{we show below } \leqslant \beta}\right) \leqslant \beta \overset{\text{def}}{=} 5 + \frac{1}{1 - \mathrm{e}^{-\mathcal{K}_{\inf}(\nu,\mu^\star)^2/(8\gamma_\star)}}
\end{aligned} \tag{44}$$

Indeed, by the concentration results on $\mathcal{K}_{\inf}$ (Proposition 13), denoting

$$\gamma_{\star,\delta} = \frac{1}{\sqrt{1 - (\mu^\star - \delta)}} \left(16\mathrm{e}^{-2} + \ln^2\left(\frac{1}{1 - (\mu^\star - \delta)}\right)\right) \leqslant \gamma_\star\,,$$

we get

$$\begin{aligned}
\mathbb{P}\left[\sum_{i=k+1}^{n} Z_{a,i} \leqslant 0\right] &\leqslant \max\left\{\mathrm{e}^{-(n-k)/4},\ \exp\left(-\frac{n-k}{2\gamma_{\star,\delta}}\Big(\mathcal{K}_{\inf}(\nu_a, \mu^\star - \delta)\Big)^2\right)\right\} \\
&\leqslant \mathrm{e}^{-(n-k)/4} + \exp\left(-\frac{n-k}{2\gamma_\star}\Big(\mathcal{K}_{\inf}(\nu_a, \mu^\star - \delta)\Big)^2\right) \\
&\leqslant \mathrm{e}^{-(n-k)/4} + \mathrm{e}^{-(n-k)\mathcal{K}_{\inf}(\nu,\mu^\star)^2/(8\gamma_\star)}
\end{aligned}$$

where the third inequality follows from the first regularity inequality of Lemma 10 and from our stated condition $\delta \leqslant \mathcal{K}_{\inf}(\nu_a, \mu^\star)/\big(2(1-\mu^\star)\big)$:

$$\mathcal{K}_{\inf}(\nu, \mu^\star - \delta) \geqslant \mathcal{K}_{\inf}(\nu, \mu^\star) - \frac{\delta}{1 - \mu^\star} \geqslant \frac{\mathcal{K}_{\inf}(\nu, \mu^\star)}{2} \tag{45}$$

We finally get, after summation over $n = k+1, \dots, T$,

$$\sum_{n=k+1}^{T} \mathbb{P}\left[\sum_{i=k+1}^{n} Z_{a,i} \leqslant 0\right] \leqslant \underbrace{\frac{1}{1 - \mathrm{e}^{-1/4}}}_{\leqslant 5} + \frac{1}{1 - \mathrm{e}^{-\mathcal{K}_{\inf}(\nu,\mu^\star)^2/(8\gamma_\star)}}\,,$$

which is the inequality claimed in (44). ∎

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).

References

R. Agrawal. Sample mean based index policies with $o(\ln n)$ regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995.

J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Annual Conference on Learning Theory*, COLT'09, pages 217–226, 2009.

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

P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities. A nonasymptotic theory of independence*. Oxford University Press, 2013.

S. Bubeck and C.-Y. Liu. Prior-free and prior-dependent regret bounds for Thompson sampling. arXiv:1304.5758, 2013.

A.N. Burnetas and M.N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 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. *The Annals of Statistics*, 41(3):1516–1541, 2013.

Y. Chow and H. Teicher. *Probability Theory*. Springer, 1988.

R. Degenne and V. Perchet. Anytime optimal algorithms in stochastic multi-armed bandits. In *Proceedings of the 2016 International Conference on Machine Learning*, ICML'16, pages 1587–1595, 2016.

A. Garivier and O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th Annual Conference on Learning Theory*, COLT'11, 2011.

A. Garivier, P. Ménard, and G. Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 2018. To appear; meanwhile, see arXiv preprint arXiv:1602.07182.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

J. Honda and A. Takemura. Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards. *Journal of Machine Learning Research*, 16:3721–3756, 2015.

A. Hoorfar and M. Hassani. Inequalities on the Lambert W function and hyperpower function. *Journal of Inequalities in Pure and Applied Mathematics*, 9(2):Article 51, 2008.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On bayesian upper confidence bounds for bandit problems. In *Proceedings of the 2012 International Conference on Artificial Intelligence and Statistics*, AIStats'12, pages 592–600, 2012.

N. Korda, E. Kaufmann, and R. Munos. Thompson sampling for 1–dimensional exponential family bandits. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pages 1448–1456, 2013.

T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

T. Lattimore. Regret analysis of the anytime optimally confident UCB algorithm. arXiv:1603.08661, 2016.

O.-A. Maillard, R. Munos, and G. Stoltz. Finite-time analysis of multi-armed bandits problems with Kullback-Leibler divergences. In *Proceedings of the 24th annual Conference on Learning Theory*, COLT'11, 2011.

P. Ménard and A. Garivier. A minimax and asymptotically optimal algorithm for stochastic bandits. In *Procedings of the 2017 Algorithmic Learning Theory Conference*, ALT'17, 2017.

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

Appendix A. Proof of Proposition 13

The proof of Proposition 13 relies on the following lemma via the variational formula (32). This lemma is a concentration result for random variables that are essentially bounded from one side only. Its holds also for possibly negative u (there is no lower bound on the u that can be considered).

Lemma 15 *Let $Z_1, \dots, Z_n$ be i.i.d. random variables such that there exist a, $b \geqslant 0$ with*

$$Z_1 \leqslant a \quad \text{a.s.} \quad \text{and} \quad \mathbb{E}\big[\mathrm{e}^{-Z_1}\big] \leqslant b$$

Define furthermore $\gamma = \sqrt{\mathrm{e}^a}\big(16\,\mathrm{e}^{-2}b + a^2\big)$. Then Z_1 in integrable and for all $u < \mathbb{E}[Z_1]$,

$$\mathbb{P}\left[\sum_{i=1}^n Z_i \leqslant nu\right] \leqslant \begin{cases} \exp(-n\gamma/8) & \text{if } u \leqslant \mathbb{E}[Z_1] - \gamma/2 \\ \exp\Big(-n\big(\mathbb{E}[Z_1] - u\big)^2/(2\gamma)\Big) & \text{if } u > \mathbb{E}[Z_1] - \gamma/2 \end{cases}$$

Indeed, denoting by $\lambda^\star \in [0,1]$ a real number achieving the maximum in the variational formula (32) for $\mathcal{K}_{\inf}(\nu, \mu^\star)$, we introduce the random variable

$$Z = \ln\left(1 - \lambda^\star \frac{X - \mu^\star}{1 - \mu^\star}\right) \qquad \text{where} \qquad X \sim \nu$$

and i.i.d. copies $Z_1, \dots, Z_n$ of Z. Then, $\mathcal{K}_{\inf}(\nu, \mu^\star) = \mathbb{E}\big[Z\big]$ and by the variational formula (32) again,

$$\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mu^\star\big) \geqslant \frac{1}{n}\sum_{i=1}^n Z_i\,, \qquad \text{therefore,} \qquad \mathbb{P}\big[\mathcal{K}_{\inf}(\widehat{\nu}_n, \mu^\star) \leqslant x\big] \leqslant \mathbb{P}\left[\sum_{i=1}^n Z_i \leqslant nx\right]$$

for all real numbers x. Now,

$$X \geqslant 0 \qquad \text{thus} \qquad Z \leqslant \ln\left(1 + \lambda^\star \frac{\mu^\star}{1 - \mu^\star}\right) \leqslant \ln\left(\frac{1}{1 - \mu^\star}\right) \overset{\text{def}}{=} a$$

and on the other hand,

$$\mathbb{E}\big[\mathrm{e}^{-Z}\big] = \mathbb{E}\left[\frac{1}{1 - \lambda^\star(X - \mu^\star)/(1 - \mu^\star)}\right] \overset{\text{def}}{=} b$$

where $b \leqslant 1$ follows from (33). This proves Lemma 15, except for the inequality $\mathrm{e}^{-n\gamma_\star/8} \leqslant \mathrm{e}^{-n/4}$ claimed therein. The latter is a consequence of $\gamma_\star \geqslant 2$, as $\gamma_\star$ is an increasing function of $\mu^\star > 0$,

$$\gamma_\star = \frac{1}{\sqrt{1 - \mu^\star}}\left(16\mathrm{e}^{-2} + \ln^2\left(\frac{1}{1 - \mu^\star}\right)\right) > 16\mathrm{e}^{-2} > 2\,.$$

Proof of Lemma 15

For the sake of completeness, we provide a proof of Lemma 15 which is a direct application of the Crámer-Chernoff method.

Proof We will make repeated uses of the fact that e^{-Z_1} is integrable (by the assumption on b), and that so is e^{Z_1}, as e^{Z_1} takes bounded values in $(0, \mathrm{e}^a]$. In particular, Z_1 is integrable, as by Jensen's inequality,

$$\mathbb{E}\big[|Z_1|\big] \leqslant \ln \mathbb{E}\Big[\mathrm{e}^{|Z_1|}\Big] \leqslant \ln\Big(\mathbb{E}\big[\mathrm{e}^{-Z_1}\big] + \mathbb{E}\big[\mathrm{e}^{Z_1}\big]\Big) < +\infty$$

We will show below that the log-moment generation function Λ of Z_1 is well-defined at least on the interval $[-1, 1]$,

$$\Lambda : x \in [-1, 1] \longmapsto \ln \mathbb{E}\big[\mathrm{e}^{xZ_1}\big]$$

and twice differentiable at least on $(-1, 1)$, with $\Lambda'(0) = \mathbb{E}[Z_1]$ and $\Lambda''(x) \leqslant \gamma$ for $x \in [-1/2,\, 0]$. By a Taylor expansion with a Cauchy remainder, we then have

$$\forall\, x \in [-1/2,\, 0], \qquad \Lambda(x) \leqslant \Lambda(0) + x\,\Lambda'(0) + \frac{x^2}{2} \sup_{y \in (-1/2,\, 0)} \Lambda''(y) \leqslant x\,\mathbb{E}[Z_1] + \frac{\gamma}{2} x^2$$

Therefore, by the Crámer-Chernoff method, for all $x \in [-1/2,\, 0]$, the probability of interest is bounded by

$$\begin{aligned}\mathbb{P}\left[\sum_{i=1}^n Z_i \leqslant nu\right] = \mathbb{P}\left[\prod_{i=1}^n \mathrm{e}^{xZ_i} \geqslant \mathrm{e}^{nux}\right] &\leqslant \mathrm{e}^{-nux}\left(\mathbb{E}\big[\mathrm{e}^{xZ_1}\big]\right)^n = \exp\Big(-n\big(ux - \Lambda(x)\big)\Big) \\ &\leqslant \exp\left(-n \min_{x \in [-1/2, 0]} \Big\{x\left(u - \mathbb{E}[Z_1]\right) - x^2\,\gamma/2\Big\}\right)\end{aligned} \tag{46}$$

which we will further upper bound depending on whether $u > \mathbb{E}[Z_1] - \gamma/2$ or $u \leqslant \mathbb{E}[Z_1] - \gamma/2$.

Proofs of the statements on Λ. That Λ is well-defined over $[-1, 1]$ follows from the inequality $\mathrm{e}^{xZ_1} \leqslant \mathrm{e}^{Z_1} + \mathrm{e}^{-Z_1}$, which is valid for all $x \in [-1, 1]$ and whose right-hand side is integrable as already noted above. That $\psi : x \mapsto \mathbb{E}\big[\mathrm{e}^{xZ_1}\big]$ is differentiable at least on $(-1, 1)$ follows from the fact that $x \in (-1, 1) \mapsto Z_1\,\mathrm{e}^{xZ_1}$ is locally dominated by an integrable random variable; indeed, for $x \in (-1, 1)$,

$$\big|Z_1\,\mathrm{e}^{xZ_1}\big| = Z_1\,\mathrm{e}^{xZ_1}\,\mathbb{1}_{\{Z_1 \geqslant 0\}} + Z_1\,\mathrm{e}^{xZ_1}\,\mathbb{1}_{\{Z_1 < 0\}} \leqslant a\,\mathrm{e}^a + \frac{1}{x} \sup_{(-\infty, 0)} f = a\,\mathrm{e}^a + \frac{1}{\mathrm{e}x}$$

where $f(t) = -t\,\mathrm{e}^t$. Similarly, $x \in (-1, 1) \mapsto Z_1^2\,\mathrm{e}^{xZ_1}$ is also locally dominated by an integrable random variable. Thus, ψ is twice differentiable at least on $(-1, 1)$, with first and second derivatives

$$\psi'(x) = \mathbb{E}\big[Z_1\,\mathrm{e}^{xZ_1}\big] \qquad \text{and} \qquad \psi''(x) = \mathbb{E}\big[Z_1^2\,\mathrm{e}^{xZ_1}\big]$$

and therefore, so is $\Lambda = \ln \psi$, with

$$\Lambda'(x) = \frac{\psi'(x)}{\psi(x)} = \frac{\mathbb{E}\big[Z_1\,\mathrm{e}^{xZ_1}\big]}{\mathbb{E}\big[\mathrm{e}^{xZ_1}\big]} \quad \text{and} \quad \Lambda''(x) = \frac{\psi''(x)\,\psi(x) - \big(\psi'(x)\big)^2}{\psi(x)^2} \leqslant \frac{\psi''(x)}{\psi(x)} = \frac{\mathbb{E}\big[Z_1^2\,\mathrm{e}^{xZ_1}\big]}{\mathbb{E}\big[\mathrm{e}^{xZ_1}\big]}$$

In particular, $\Lambda'(0) = \mathbb{E}[Z_1]$. As for the bound on $\Lambda''(x)$, we note first that $\mathrm{e}^{xZ_1} \geqslant \mathrm{e}^{xa} \geqslant 1/\sqrt{\mathrm{e}^a}$ as $Z_1 \leqslant a$ and $x \in [-1/2,\, 0]$. Second, using that (proof below)

$$\forall\, x \in [-1/2,\, 0],\ z \in (-\infty, a), \qquad z^2\,\mathrm{e}^{xz} \leqslant 16\,\mathrm{e}^{-2}\mathrm{e}^{-z} + a^2 \tag{47}$$

we get $\mathbb{E}\big[Z_1^2\,\mathrm{e}^{xZ_1}\big] \leqslant 16\,\mathrm{e}^{-2}b + a^2$. The claimed bound $\Lambda''(x) \leqslant \gamma = \sqrt{\mathrm{e}^a}\big(16\,\mathrm{e}^{-2}b + a^2\big)$ follows. We prove (47): if $z \geqslant 0$, since $x \leqslant 0$ we have $z^2\,\mathrm{e}^{xz} \leqslant z^2 \leqslant a^2$, while, if $z \leqslant 0$, using $z^2 \leqslant 16\mathrm{e}^{-2-z/2}$ in this case, we obtain $z^2\,\mathrm{e}^{xz} \leqslant 16\mathrm{e}^{-2}\,\mathrm{e}^{(x-1/2)z} \leqslant 16\mathrm{e}^{-2}\mathrm{e}^{-z}$ as $x \geqslant -1/2$.

Upper bounds on the minimum in (46). We rewrite

$$x\,(u - \mathbb{E}[Z_1]) - x^2\,\gamma/2 = \frac{\gamma x}{2}\left(x - 2\frac{u - \mathbb{E}[Z_1]}{\gamma}\right)$$

and deal with a second-order polynomial with roots 0 and $2\big(u-\mathbb{E}[Z_1]\big)/\gamma < 0$ and whose minimum over the entire real line $(-\infty, +\infty)$ is thus achieved at the midpoint $x^\star = \big(u - \mathbb{E}[Z_1]\big)/\gamma < 0$ between these roots. But the expression above is to be minimized over $[-1/2,\,0]$ only. In the case where $u > \mathbb{E}[Z_1] - \gamma/2$, then $x^\star$ belongs to the interval of interest and

$$\min_{x\in[-1/2,0]}\Big\{x\,\big(u - \mathbb{E}[Z_1]\big) - x^2\,\gamma/2\Big\} = \frac{\gamma x^\star}{2}\left(x^\star - 2\frac{u - \mathbb{E}[Z_1]}{\gamma}\right) = \frac{\big(u - \mathbb{E}[Z_1]\big)^2}{2\gamma}$$

Otherwise, $u - \mathbb{E}[Z_1] \leqslant -\gamma/2$ and the midpoint $x^\star$ is to the left of $-1/2$ and the considered expression is decreasing with x on $[-1/2, 0]$, so that the minimum is achieved at $-1/2$, that is,

$$\min_{x\in[-1/2,0]}\Big\{x\,\big(u - \mathbb{E}[Z_1]\big) - x^2\,\gamma/2\Big\} = -\frac{u - \mathbb{E}[Z_1]}{2} - \frac{\gamma}{8} \geqslant \frac{\gamma}{8}$$

which concludes the proof. ■

Appendix B. A simplified proof of the regret bounds for MOSS and MOSS anytime

The regret bounds proven here are not new all, see Audibert and Bubeck (2009) and Degenne and Perchet (2016) for, respectively, the case of a known horizon T and the anytime version of MOSS; however the proof exposed here is somewhat simpler and more direct than in these references. In previous works, attempts were made to simultaneously build the distribution-free and some type of distribution-dependent bounds. This raised technical difficulties because of the correlations between the choices of the arms and the observed rewards. The idea of this proof is to focus solely on the distribution-free regime, for which we notice that some crude boundings neglecting the correlations suffice (i.e., our analysis deals with all suboptimal arms in the same way, independently of how often they are played). We have also simplified the use of the peeling trick, by performing it only once on integrated quantities (instead of performing a different doubling trick for each deviation). All in all, our proof therefore consists entirely of fairly elementary and natural steps, with Hoeffding's maximal inequality in its integrated version (Corollary 8) as the only necessary technical ingredient.

To emphasize the similarity of the proofs in the anytime and non-anytime case, we present both of them in a unified fashion. The indexes used only differ by the replacement of T by t in the logarithmic exploration term in case T is unknown, see (4) and (26): compare

$$U_a^{\text{M}}(t) = \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)}\ln_+\left(\frac{T}{KN_a(t)}\right)} \quad \text{and} \quad U_a^{\text{M-A}}(t) = \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)}\ln_+\left(\frac{t}{KN_a(t)}\right)}$$

Note in particular that $U_a^{\text{M-A}}(t) \leqslant U_a^{\text{M}}(t)$ for all arms a and all steps $1 \leqslant t \leqslant T$. We will denote by

$$U_{a,\tau}^{\text{GM}}(t) = \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)} \ln_+\left(\frac{\tau_t}{KN_a(t)}\right)}$$

the index of generic MOSS (GM) strategy, so that $U_a^{\text{M}}(t) = U_{a,T}^{\text{GM}}(t)$ and $U_a^{\text{M-A}}(t) = U_{a,t}^{\text{GM}}(t)$. This GM strategy considers a sequence $(\tau_1, \dots, \tau_T)$ of integers, either $\tau_t \equiv T$ for MOSS or $\tau_t = t$ for MOSS anytime, and pick at each step $t \geqslant K+1$, an arm A_t with maximal index $U_{a,\tau}^{\text{GM}}(t)$.

Proof of Proposition 9 and of the claim after it The first step is standard, see Bubeck and Liu (2013). Using the fact that $U_{a^\star,\tau_t}^{\text{GM}}(t) \leqslant U_{A_t^{\text{GM}},\tau_t}^{\text{GM}}(t)$ by definition of the index policy, the regret is smaller than

$$R_T = \sum_{t=1}^{T} \mathbb{E}\big[\mu^\star - \mu_{A_t^{\text{GM}}}\big] \leqslant (K-1) + \sum_{t=K+1}^{T} \mathbb{E}\big[\mu^\star - U_{a^\star,\tau_t}^{\text{GM}}(t)\big] + \sum_{t=K+1}^{T} \mathbb{E}\Big[U_{A_t^{\text{GM}},\tau_t}^{\text{GM}}(t) - \mu_{A_t^{\text{GM}}}\Big] \quad (48)$$

Since $x \leqslant \delta + (x-\delta)^+$ for all x and δ, for the first inequality, by optional skipping (Section 5.1) for the second inequality, where we also use that pairs (a,n) such $A_t^{\text{GM}} = a$ and $N_a(t) = n$ correspond to at most one $t \in \{K+1, \dots, T\}$, and by using that $U_{a,\tau}^{\text{GM}}(t)$ is increasing with τ for the third inequality,

$$\begin{aligned}
\sum_{t=K+1}^{T} \mathbb{E}\Big[U_{A_t^{\text{GM}}}^{\text{GM}}(t) - \mu_{A_t^{\text{GM}}}\Big] &\leqslant \sqrt{KT} + \sum_{t=K+1}^{T} \mathbb{E}\left[\left(U_{A_t^{\text{GM}},\tau_t}^{\text{GM}}(t) - \mu_{A_t^{\text{GM}}} - \sqrt{\frac{K}{T}}\right)^+\right] \\
&\leqslant \sqrt{KT} + \sum_{a=1}^{K}\sum_{n=1}^{T} \mathbb{E}\left[\left(U_{a,\tau_t,n}^{\text{GM}} - \mu_a - \sqrt{\frac{K}{T}}\right)^+\right] \\
&\leqslant \sqrt{KT} + \sum_{a=1}^{K}\sum_{n=1}^{T} \mathbb{E}\left[\left(U_{a,T,n}^{\text{GM}} - \mu_a - \sqrt{\frac{K}{T}}\right)^+\right]
\end{aligned}$$

While this latter inequality may seem very crude, it turns out it is sharp enough to obtain the claimed distribution-free bounds. Moreover, it gets rid of the bothersome dependencies among the arms that are contained in the choice A_t^{GM}. Substituting in (48), we have shown the first inequality of Proposition 9, namely,

$$R_T \leqslant (K-1) + \sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star,\tau_t}^{\text{GM}}(t)\big)^+\Big] + \sqrt{KT} + \sum_{a=1}^{K}\sum_{n=1}^{T} \mathbb{E}\Big[\big(U_{a,T,n}^{\text{GM}} - \mu_a - \sqrt{K/T}\big)^+\Big] \quad (49)$$

This inequality actually holds for all choices of sequences $(\tau_t)_{t\leqslant T}$ with $\tau_t \leqslant T$. The first sum in the right-hand side of (49) depends on the specific value of $(\tau_t)_{t\leqslant T}$ but the second sum only depends on the bound T.

Control of the left deviations of the best arm, that is, of the first sum in (49). For each given round $t \geqslant K+1$, we decompose

$$\mathbb{E}\Big[\big(\mu^\star - U_{a^\star,\tau_t}^{\text{GM}}(t)\big)^+\Big] = \mathbb{E}\Big[\big(\mu^\star - U_{a^\star,\tau_t}^{\text{GM}}(t)\big)^+ \mathbb{1}_{\{N_{a^\star}(t) < \tau_t/K\}}\Big] + \mathbb{E}\Big[\big(\mu^\star - U_{a^\star,\tau_t}^{\text{GM}}(t)\big)^+ \mathbb{1}_{\{N_{a^\star}(t) \geqslant \tau_t/K\}}\Big]$$

The two pieces are handled differently. The second one is easily treated by optional skipping (Section 5.1) and by Corollary 8, using that $U^{\mathrm{GM}}_{a^\star,\tau_t}(t) \geqslant \widehat{\mu}_{a^\star}(t)$, which actually holds with equality given $N_{a^\star}(t) \geqslant \tau_t/K$:

$$\mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{GM}}_{a^\star,\tau_t}(t)\big)^+ \mathbb{1}_{\{N_{a^\star}(t)\geqslant \tau_t/K\}}\Big] \leqslant \mathbb{E}\Big[\max_{n\geqslant \tau_t/K} \big(\mu^\star - \widehat{\mu}_{a^\star,n}\big)^+\Big] \leqslant \sqrt{\frac{\pi}{8}}\sqrt{\frac{K}{\tau_t}} \tag{50}$$

When the arm has not been pulled often enough, we resort to a "peeling trick". We consider a real number $\beta > 1$ and further decompose the event $\big\{N_{a^\star}(t) < \tau_t/K\big\}$ along the geometric grid $x_\ell = \beta^{-\ell}\,\tau_t$, where $\ell = 0, 1, 2, \dots$ (the endpoints x_ℓ are not necessarily integers, and some intervals $[x_{\ell+1}, x_\ell)$ may contain no integer, but none of these facts is an issue):

$$\begin{aligned}
\mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{GM}}_{a^\star,\tau_t}(t)\big)^+ \mathbb{1}_{\{N_{a^\star}(t)<\tau_t/K\}}\Big] &\leqslant \sum_{\ell=0}^{+\infty} \mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{GM}}_{a^\star,\tau_t}(t)\big) \mathbb{1}_{\{x_{\ell+1}\leqslant N_{a^\star}(t)<x_\ell\}}\Big] \\
&\leqslant \sum_{\ell=0}^{+\infty} \mathbb{E}\Big[\max_{x_{\ell+1}\leqslant n<x_\ell} \big(\mu^\star - U^{\mathrm{GM}}_{a^\star,\tau_t,n}\big)\Big]
\end{aligned}$$

where in the second inequality, we applied optional skipping (Section 5.1) once again. Now for any ℓ, the summand can be controlled as follows, first by using $n < x_\ell$ and second by Corollary 8:

$$\begin{aligned}
\mathbb{E}\Big[\max_{x_{\ell+1}\leqslant n<x_\ell} \big(\mu^\star - U^{\mathrm{GM}}_{a^\star,\tau_t,n}\big)^+\Big] &= \mathbb{E}\left[\max_{x_{\ell+1}\leqslant n<x_\ell} \left(\mu^\star - \widehat{\mu}_{a^\star,\tau_t,n} - \sqrt{\frac{1}{2n}\ln\Big(\frac{\tau_t}{Kn}\Big)}\right)^+\right] \\
&\leqslant \mathbb{E}\left[\max_{x_{\ell+1}\leqslant n<x_\ell} \left(\mu^\star - \widehat{\mu}_{a^\star,n} - \sqrt{\frac{1}{2x_\ell}\ln\Big(\frac{\tau_t}{Kx_\ell}\Big)}\right)^+\right] \\
&\leqslant \sqrt{\frac{\pi}{8}}\sqrt{\frac{1}{x_{\ell+1}}}\exp\left(-\frac{x_{\ell+1}}{x_\ell}\ln\Big(\frac{\tau_t}{x_\ell}\Big)\right) \\
&= \sqrt{\frac{\pi}{8}}\sqrt{\frac{1}{x_{\ell+1}}}\big(\beta^{-\ell}\big)^{1/\beta} = \sqrt{\frac{\pi}{8}}\sqrt{\frac{1}{\tau_t}}\,\beta^{1/2+\ell(1/2-1/\beta)}\,.
\end{aligned}$$

The above series is summable whenever $\beta \in (1,2)$. For instance we may choose $\beta = 3/2$, for which

$$\sum_{\ell=0}^{+\infty}\Big(\frac{3}{2}\Big)^{1/2+\ell(1/2-2/3)} = \sqrt{\frac{3}{2}}\sum_{\ell=0}^{+\infty}\alpha^{-\ell} = \frac{1}{1-\alpha}\sqrt{\frac{3}{2}} \leqslant 19 \qquad \text{where} \qquad \alpha = \Big(\frac{3}{2}\Big)^{(1/2-2/3)}\,.$$

Therefore we have shown that

$$\mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{GM}}_{a^\star,\tau_t}(t)\big)^+ \mathbb{1}_{\{N_{a^\star}(t)<\tau_t/K\}}\Big] \leqslant 19\sqrt{\frac{\pi}{8}}\sqrt{\frac{K}{\tau_t}}\,. \tag{51}$$

Combining this bound with (50) and summing over t, we proved

$$\sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{GM}}_{a^\star,\tau_t}(t)\big)^+\Big] \leqslant 20\sqrt{\frac{\pi}{8}}\sum_{t=K+1}^{T}\sqrt{\frac{K}{\tau_t}}\,. \tag{52}$$

Control of the right deviations of all arms, that is, of the second sum in (49). As $(x+y)^+ \leqslant x^+ + y^+$ for all real numbers x, y, we have, for all a and $n \geqslant 1$,

$$\begin{aligned}\left(U^{\mathrm{GM}}_{a,T,n} - \mu_a - \sqrt{K/T}\right)^+ &\leqslant \left(\widehat{\mu}_{a,n} - \mu_a - \sqrt{K/T}\right)^+ + \sqrt{\frac{1}{2n}\ln_+\left(\frac{T}{Kn}\right)} \\ &= \left(\widehat{\mu}_{a,n} - \mu_a - \sqrt{K/T}\right)^+ + \begin{cases} 0 & \text{if } n \geqslant T/K \\ \sqrt{\dfrac{1}{2n}\ln\left(\dfrac{T}{Kn}\right)} & \text{if } n < T/K \end{cases}\end{aligned}$$

Therefore, for each arm a,

$$\sum_{n=1}^{T} \mathbb{E}\left[\left(U^{\mathrm{GM}}_{a,T,n} - \mu_a - \sqrt{K/T}\right)^+\right] \leqslant \sum_{n=1}^{T} \mathbb{E}\left[\left(\widehat{\mu}_{a,n} - \mu_a - \sqrt{K/T}\right)^+\right] + \sum_{n=1}^{\lfloor T/K \rfloor} \sqrt{\frac{1}{2n}\ln\left(\frac{T}{Kn}\right)} \tag{53}$$

We are left with two pieces to deal with separately. For the first sum in (53), we exploit the integrated version of Hoeffding's inequality (Corollary 8),

$$\begin{aligned}\sum_{n=1}^{T} \mathbb{E}\left[\left(\widehat{\mu}_{a,n} - \mu_a - \sqrt{K/T}\right)^+\right] &\leqslant \sqrt{\frac{\pi}{8}} \sum_{n=1}^{T} \sqrt{\frac{1}{n}} \mathrm{e}^{-2n\left(\sqrt{K/T}\right)^2} \leqslant \sqrt{\frac{\pi}{8}} \int_0^T \sqrt{\frac{1}{x}}\, \mathrm{e}^{-2xK/T}\, \mathrm{d}x \\ &= \sqrt{\frac{\pi}{8}} \sqrt{\frac{T}{2K}} \int_0^{+\infty} \frac{\mathrm{e}^{-u}}{\sqrt{u}}\, \mathrm{d}u = \frac{\pi}{4}\sqrt{\frac{T}{K}}, \end{aligned} \tag{54}$$

where we used the equalities $\displaystyle\int_0^{+\infty} \left(\mathrm{e}^{-u}/\sqrt{u}\right) \mathrm{d}u = \int_0^{+\infty} \mathrm{e}^{-v^2}\, \mathrm{d}v = \sqrt{\pi}$.

For the second sum in (53), we also use a sum–integral comparison: which can be handled by comparing it to an integral and performing the change of variable $u = T/(Kx)$:

$$\sum_{n=1}^{\lfloor T/K \rfloor} \sqrt{\frac{1}{2n}\ln\left(\frac{T}{Kn}\right)} \leqslant \int_0^{\lfloor T/K \rfloor} \sqrt{\frac{1}{2x}\ln\left(\frac{T}{Kx}\right)}\, \mathrm{d}x \leqslant \sqrt{\frac{T}{2K}} \int_1^{+\infty} u^{-3/2}\sqrt{\ln(u)}\, \mathrm{d}u = \sqrt{\pi}\sqrt{\frac{T}{K}}$$

as $\displaystyle\int_1^{+\infty} u^{-3/2}\sqrt{\ln(u)}\, \mathrm{d}u = 2\int_0^{+\infty} v^2\, \mathrm{e}^{-v^2/2}\, \mathrm{d}v = \sqrt{2\pi}$ by the change of variable $u = \mathrm{e}^{v^2}$.

Conclusion. Collecting all the bounds above, we showed so far

$$\begin{aligned} R_T &\leqslant (K-1) + \sum_{t=K+1}^{T} \mathbb{E}\left[\left(\mu^\star - U^{\mathrm{GM}}_{a^\star,\tau_t}(t)\right)^+\right] + \sqrt{KT} + \sum_{a=1}^{K}\sum_{n=1}^{T} \mathbb{E}\left[\left(U^{\mathrm{GM}}_{a,T,n} - \mu_a - \sqrt{K/T}\right)^+\right] \\ &\leqslant (K-1) + \sqrt{KT} + \underbrace{20\sqrt{\frac{\pi}{8}}}_{\leqslant 12.6} \sum_{t=K+1}^{T} \sqrt{\frac{K}{\tau_t}} + K\underbrace{\left(\frac{\pi}{4} + \sqrt{\pi}\right)}_{\leqslant 2.6}\sqrt{\frac{T}{K}} \end{aligned}$$

In the known horizon case $\sum 1/\sqrt{\tau_t} = T/\sqrt{T} = \sqrt{T}$ and we get $R_T \leqslant (K-1) + 17\sqrt{KT}$, whereas in the anytime case,

$$\sum_{t=1}^{T} 1/\sqrt{\tau_t} = \sum_{t=1}^{T} 1/\sqrt{t} \leqslant \int_0^T \frac{1}{\sqrt{u}}\,\mathrm{d}u = 2\sqrt{T}\,,$$

hence $R_T \leqslant (K-1) + 29\sqrt{KT}$. ■

Appendix C. Bounds for KL-UCB-Switch-Anytime

As a preliminary result to the distribution-free bound, we present an analysis of MOSS-anytime with the additional exploration φ. While we could have presented this result and Proposition 9 inside a more general result, we have chosen to separate the two to improve clarity. In the following all indices are *anytime versions with exploration function* φ.

Lemma 16 (MOSS anytime with extra-exploration)

$$\sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star}^{\text{M-A}}(t)\big)^+\Big] + \sum_{a=1}^{K}\sum_{n=1}^{T} \mathbb{E}\Big[\big(U_{a,n,T}^{\text{M-A}} - \mu_a - \sqrt{K/T}\big)^+\Big] \leqslant 29\sqrt{KT} \tag{55}$$

Proof We bound both sums separately. For the first one we may recycle the bound we obtained for MOSS-anytime without the extra exploration. Indeed, as $\varphi(x) \geqslant \ln_+(x)$

$$U_{a^\star}^{\text{M-A}}(t) \geqslant \widehat{\mu}_a(t) + \sqrt{\frac{1}{N_a(t)}\ln_+\!\left(\frac{t}{KN_a(t)}\right)}$$

which is the usual MOSS-anytime index. Therefore by extracting (52) from the previous proof

$$\sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star}^{\text{M-A}}(t)\big)^+\Big] \leqslant 20\sqrt{\frac{\pi}{2}}\sqrt{KT}$$

For the second sum we use once again the fact that the exploration vanishes at $N_a(t) \geqslant T/K$ and to bound for all arms a as in Appendix B, eq. (53)

$$\sum_{n=1}^{T} \mathbb{E}\Big[\big(U_{a,n,T}^{\text{M-A}} - \mu_a - \sqrt{K/T}\big)^+\Big] \leqslant \sum_{n=1}^{T} \mathbb{E}\left[\left(\widehat{\mu}_{a,n} - \mu_a - \sqrt{K/T}\right)^+\right] + \sum_{n=1}^{\lfloor T/K\rfloor} \sqrt{\frac{1}{n}\varphi\!\left(\frac{T}{Kn}\right)} \tag{56}$$

From (54) we recall that the first sum is smaller than $\pi/4\sqrt{T/K}$. The second sum is treated as before by comparison to an integral

$$\sum_{n=1}^{\lfloor T/K\rfloor} \sqrt{\frac{1}{2n}\varphi\!\left(\frac{T}{Kn}\right)} \leqslant \int_0^{T/K} \sqrt{\frac{1}{2x}\varphi\!\left(\frac{T}{Kx}\right)}\,\mathrm{d}x = \sqrt{\frac{T}{2K}}\int_1^{+\infty} \sqrt{u^{-3}\ln\big(u(1+\ln^2(u))\big)}\,\mathrm{d}u$$

This integral is smaller than 4. We conclude by summing over a. ■

We now have all elements to provide a very short proof (with references to other results in this paper) of the distribution-free anytime bound.

Proof of Theorem 4 Once again we begin with now usual boundings by distinguishing the value of the index depending on $N_a(t)$ for all t

$$R_T \leqslant (K-1) + \sum_{t=K+1}^{T} \underbrace{\mathbb{E}\Big[\big(\mu^\star - U_{a^\star}^{\text{KL-A}}(t)\big)^+ \mathbb{1}_{\{N_{a^\star}(t) < f(t,K)\}}\Big]}_{\text{we show below that } \leqslant 8\sqrt{K/t} \text{ for each } t} + \underbrace{\sqrt{KT} + \sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star}^{\text{M-A}}(t)\big)^+\Big] + \sum_{a=1}^{K}\sum_{n=1}^{T} \mathbb{E}\Big[\big(U_{a,n,T}^{\text{M-A}} - \mu_a - \sqrt{K/T}\big)^+\Big]}_{\leqslant 30\sqrt{KT} \quad \text{by (55)}}$$

And we are left to bound the first sum. Now since the exploration function verifies $\varphi(x) \geqslant \ln_+(x)$ we may see that the index is greater than the usual KL-UCB index. Therefore the bound from the proof of Theorem 1 can be re-derived replacing T by t

$$\mathbb{E}\Big[\big(\mu^\star - U_{a^\star}^{\text{KL-A}}(t)\big)^+ \mathbb{1}_{\{N_{a^\star}(t) < f(t,K)\}}\Big] \leqslant 8\sqrt{\frac{K}{t}} \tag{57}$$

and the bound follows since $\sum_{t=1}^{T} \sqrt{1/t} \leqslant 2\sqrt{T}$ ■

The distribution-dependent anytime bound is different from the known horizon case, as we do not aim for the finer second order bound.

Proof of Theorem 5

$$\begin{aligned}\mathbb{E}[N_a(T)] &= 1 + \sum_{t=K}^{T-1} \mathbb{P}[U_{a^\star}(t) \leqslant U_a(t) \text{ and } A_{t+1} = a] \\ &\leqslant 1 + \sum_{t=K}^{T-1} \mathbb{P}[U_{a^\star}(t) \leqslant \mu^\star - \delta] + \sum_{t=K}^{T-1} \mathbb{P}[U_a(t) \geqslant \mu^\star - \delta \text{ and } A_{t+1} = a]\end{aligned} \tag{58}$$

The first sum is bounded by optional skipping, Proposition 13 and Hoeffding's maximal inequality as

$$\begin{aligned}\mathbb{P}\big[U_{a^\star}(t) \leqslant \mu^\star - \delta\big] &\leqslant \sum_{n=1}^{\lfloor t/K \rfloor} \mathbb{P}\left[\mathcal{K}_{\inf}(\widehat{\nu}_{a,n}, \mu^\star - \delta) \leqslant \frac{1}{n}\varphi\left(\frac{t}{Kn}\right)\right] + \mathbb{P}\big[\exists n \geqslant \lfloor t/K \rfloor + 1 \ : \ \widehat{\mu}_{a,n} \leqslant \mu^\star - \delta\big] \\ &\leqslant \sum_{n=1}^{\lfloor t/K \rfloor} \mathrm{e}(2n+1)\mathrm{e}^{-\varphi(t/(Kn))}\mathrm{e}^{-2n\delta^2} + \mathrm{e}^{-t\delta^2/K}\end{aligned}$$

For the sake of clarity, we delay some straightforward calculations (detailed after the proof) that lead us to

$$\sum_{t=K}^{T-1}\sum_{n=1}^{\lfloor t/K \rfloor} \mathrm{e}(2n+1)\mathrm{e}^{-\varphi(t/(Kn))}\mathrm{e}^{-2n\delta^2} \leqslant \frac{5\mathrm{e}(1+\pi)}{2(1-\mathrm{e}^{-2})^3}\frac{K}{\delta^6} \tag{59}$$

Hence the first sum in (58) is bounded by

$$\frac{5\mathrm{e}(1+\pi)}{2(1-\mathrm{e}^{-2})^3}\frac{K}{\delta^6}+\frac{1}{1-\mathrm{e}^{-1}}\frac{K}{\delta^2} \tag{60}$$

and we are now to treat the second sum. We proceed by a fine and exhaustive decomposition of the sum thanks to optional skipping. Define the event

$$\mathcal{E}_a(n,t)=\big\{N_a(t)=n \text{ and } A_{t+1}=a\big\}$$

We will use repeatedly the fact that for all n there is most one value of t such that $\mathcal{E}_a(n,t)$ holds. A direct consequence of this fact is that for any event $\mathcal{F}(n)$ that does not depend on t

$$\sum_{n=n_0}^{n_1}\sum_{t=t_0}^{t_1}\mathbb{1}_{\{\mathcal{E}_a(n,t) \text{ and } \mathcal{F}(n)\}}=\sum_{n=n_0}^{n_1}\mathbb{1}_{\{\mathcal{F}(n)\}}\underbrace{\sum_{t=t_0}^{t_1}\mathbb{1}_{\{\mathcal{E}_a(n,t)\}}}_{\leqslant 1}\leqslant\sum_{n=n_0}^{n_1}\mathbb{1}_{\{\mathcal{F}(n)\}} \tag{61}$$

Then by definition of the switch index

$$\begin{aligned}\sum_{t=K}^{T-1}\mathbb{P}\big[U_a(t)\geqslant\mu^\star-\delta \text{ and } A_{t+1}=a\big]&=\sum_{t=K}^{T-1}\sum_{n=1}^{t}\mathbb{P}\big[U_{a,n,t}\geqslant\mu^\star-\delta \text{ and } \mathcal{E}_a(n,t)\big]\\&=\sum_{t=K}^{T-1}\sum_{n=1}^{f(t,K)}\mathbb{P}\big[U^{\mathrm{KL}}_{a,n,t}\geqslant\mu^\star-\delta \text{ and } \mathcal{E}_a(n,t)\big]+\sum_{t=K}^{T-1}\sum_{n=f(t,K)+1}^{t}\mathbb{P}\big[U^{\mathrm{M}}_{a,n,t}\geqslant\mu^\star-\delta \text{ and } \mathcal{E}_a(n,t)\big]\end{aligned}$$

For the first sum, we may use similar bounds as in the known horizon case, as $U^{\mathrm{KL}}_{a,n,t}\leqslant U^{\mathrm{KL}}_{a,n,T}$, and then by invoking (61). By using the exact same calculations as in the known horizon case, see (22), replacing $\ln$ by φ, for $\delta^2\leqslant\gamma_\star(1-\mu^\star)^2/2$ we bound the first sum by

$$\frac{\varphi(T/K)}{\mathcal{K}_{\inf}(\nu_a,\mu^\star)-2\delta/(1-\mu^\star)}+\frac{1}{1-\mathrm{e}^{-\delta^2/(2\gamma_\star(1-\mu^\star)^2)}} \tag{62}$$

where $\gamma_\star$ is defined in (36). The second sum requires a more refined treatment. Define the varying threshold

$$n_1(t)=\left\lfloor\frac{8\,\varphi(t/K)}{\Delta_a^2}\right\rfloor$$

so that for $\delta\leqslant\Delta_a/2$ and $n>n_1(t)$

$$\big\{U^{\mathrm{M}}_{a,n,t}\geqslant\mu^\star-\delta\big\}\subseteq\left\{\widehat{\mu}_{a,n}\geqslant\mu_a+\Delta_a-\delta-\sqrt{\frac{\varphi\big(t/K\big)}{2n_1(t)}}\right\}\subseteq\big\{\widehat{\mu}_{a,n}\geqslant\mu_a+\Delta_a/4\big\} \tag{63}$$

We then decompose the sum as

$$\sum_{t=K}^{T-1}\sum_{n=f(t,K)+1}^{n_1(t)}\mathbb{P}\big[U^{\mathrm{M}}_{a,n,t}\geqslant\mu^\star-\delta \text{ and } \mathcal{E}_a(n,t)\big]+\sum_{t=K}^{T-1}\sum_{n=n_1(t)+1}^{t}\mathbb{P}\big[U^{\mathrm{M}}_{a,n,t}\geqslant\mu^\star-\delta \text{ and } \mathcal{E}_a(n,t)\big]$$

Our choice of $n_1(t)$, via (63), leads to

$$\sum_{t=K}^{T}\sum_{n=n_1(t)+1}^{t}\mathbb{P}\big[U^{\mathrm{M}}_{a,n,t}\geqslant\mu^\star-\delta\text{ and }\mathcal{E}_a(n,t)\big]\leqslant\sum_{t=K}^{T}\sum_{n=n_1(t)+1}^{t}\mathbb{P}\big[\widehat{\mu}_{a,n}\geqslant\mu_a+\Delta_a/4\text{ and }\mathcal{E}_a(n,t)\big]$$

Now the event does not depend on t anymore, and thanks to (61) and Hoeffding's inequality, we may bound it by

$$\sum_{n=1}^{T}\mathbb{P}\big[\widehat{\mu}_{a,n}\geqslant\mu_a+\Delta_a/4\big]\leqslant\frac{1}{1-\mathrm{e}^{-\Delta_a^2/8}}$$

The only piece that remains to be bounded is now

$$\sum_{t=K}^{T-1}\sum_{n=f(t,K)+1}^{n_1(t)}\mathbb{P}\big[U^{\mathrm{M}}_{a,n,t}\geqslant\mu^\star-\delta\text{ and }\mathcal{E}_a(n,t)\big]$$

which we will bound deterministically thanks to the events $\mathcal{E}_a(n,t)$. Indeed

$$\sum_{t=K}^{T-1}\sum_{n=f(t,K)}^{n_1(t)}\mathbb{1}_{\{\mathcal{E}_a(n,t)\}}\leqslant\sum_{t=K}^{T-1}\mathbb{1}_{\{f(t,K)\leqslant n_1(t)\}}\leqslant\min\big\{t\geqslant K:f(t,K)>n_1(t)\big\}\overset{\text{def}}{=}T_0 \tag{64}$$

since for all t, there is at most one n such that $N_a(t)=n$: hence the inside sum is at most 1, and is trivially zero whenever $f(t,K)>n_1(t)$. T_0 is a constant that depends solely on Δ_a and K.

All in all we have shown that for all T and $\delta\leqslant\min\big(\gamma_\star(1-\mu^\star)^2/2,\Delta_a/2\big)$

$$\mathbb{E}[N_a(T)]\leqslant\frac{\varphi(T/K)}{\mathcal{K}_{\inf}(\nu_a,\mu^\star)-2\delta/(1-\mu^\star)}+\frac{C_1}{\delta^6}+C_2 \tag{65}$$

where C_1 and C_2 are constants that do not depend on T and δ. Therefore as $T\to\infty$ we may choose $\delta=\varphi(T/K)^{-1/7}$ which gives the claimed result, remembering that $\varphi(x)=\ln(x)+\mathrm{o}(\ln(x))$. ■

Proof of (59) This is straightforward calculations : we permute the sums and compare them to the corresponding integrals

$$\begin{aligned}
\sum_{t=K}^{T-1}\sum_{n=1}^{\lfloor t/K\rfloor}\mathrm{e}(2n+1)e^{-\varphi(t/(Kn))}\mathrm{e}^{-2n\delta^2}&=\sum_{n=1}^{\lfloor(T-1)/K\rfloor}\sum_{t=Kn}^{T-1}\mathrm{e}(n+2)\mathrm{e}^{-\varphi(t/(Kn))}\mathrm{e}^{-2n\delta^2}\\
&=\sum_{n=1}^{\lfloor(T-1)/K\rfloor}\mathrm{e}(2n+1)\mathrm{e}^{-2n\delta^2}\sum_{t=Kn}^{T-1}\mathrm{e}^{-\varphi(t/(Kn))}\\
&\leqslant\sum_{n=1}^{\lfloor(T-1)/K\rfloor}\mathrm{e}(2n+1)\mathrm{e}^{-2n\delta^2}\int_{t=Kn-1}^{T-1}\mathrm{e}^{-\varphi(t/(Kn))}\,\mathrm{d}t\\
&\leqslant\sum_{n=1}^{\lfloor(T-1)/K\rfloor}\mathrm{e}(2n+1)\mathrm{e}^{-2n\delta^2}\int_{u=1-1/(Kn)}^{T-1/(Kn)}Kn\mathrm{e}^{-\varphi(u)}\mathrm{d}u
\end{aligned}$$

Now we have chosen φ so that for all n

$$\int_{u=1-1/(Kn)}^{T-1/(Kn)} \mathrm{e}^{-\varphi(u)}\mathrm{d}u \leqslant \int_{1/2}^{+\infty} \mathrm{e}^{-\varphi(u)}\mathrm{d}u = \frac{1}{2} + \int_{1}^{+\infty} \frac{\mathrm{d}u}{u\big(1+\ln^2(u)\big)} = \frac{1+\pi}{2} \tag{66}$$

Hence our sum is smaller than

$$K\mathrm{e}\frac{1+\pi}{2} \sum_{n=1}^{\lfloor (T-1)/K \rfloor} n(2n+1)\mathrm{e}^{-2n\delta^2} \leqslant \frac{5\mathrm{e}(1+\pi)}{2(1-\mathrm{e}^{-2})^3}\frac{K}{\delta^6}$$

as already detailed in (14). ■

Appendix D. Proofs of the other results of Section 5.4

Proposition 13 of Section 5.4 was already proved in Appendix A. We now prove the three remaining results of Section 5.4, namely, Lemmas 10 and 11, as well as Proposition 12.

D.1. Proof of Lemma 11

The proof of Honda and Takemura (2015, Theorem 2, Lemma 6) relies on the exhibiting the formula of interest for finitely supported distributions, via KKT conditions, and then taking limits to cover the case of all distributions. We propose a more direct approach. But before we do, we explain why it is natural to expect to rewrite $\mathcal{K}_{\inf}$, which is an infimum, as a maximum. Indeed, given that Kullback-Leibler divergences are given by a supremum, $\mathcal{K}_{\inf}$ appears as an $\inf\sup$, which under some conditions (this is Sion's lemma) is equal to a $\sup\inf$.

More precisely, a variational formula for the Kullback-Leibler divergence, see Boucheron et al. (2013, Chapter 4), has it that

$$\mathrm{KL}(\nu,\nu') = \sup\Big\{\mathbb{E}_\nu[Y] - \ln \mathbb{E}_{\nu'}\big[\mathrm{e}^Y\big] \,:\, Y \text{ s.t. } \mathbb{E}_\nu[\mathrm{e}^Y] < +\infty\Big\}$$

where we indexed the expectations with respect to the underlying probability. In particular, denoting by X the identity and considering, for $\lambda \in [0,1]$, the bounded variables

$$Y_\lambda = \ln\left(1 - \lambda\frac{X-\mu}{1-\mu}\right) \leqslant \ln\left(1 + \frac{\lambda\mu}{1-\mu}\right)$$

we have, for any probability measure ν' such that $\mathrm{E}(\nu') > \mu$:

$$\ln \mathbb{E}_{\nu'}\big[\mathrm{e}^{Y_\lambda}\big] = \ln\left(\mathbb{E}_{\nu'}\left[1 - \lambda\frac{X-\mu}{1-\mu}\right]\right) = \ln\left(1 - \lambda\frac{\mathrm{E}(\nu')-\mu}{1-\mu}\right) \leqslant 0$$

Hence, for these distributions ν',

$$\mathrm{KL}(\nu,\nu') \geqslant \max_{\lambda\in[0,1]} \mathbb{E}_\nu[Y_\lambda] - \ln \mathbb{E}_{\nu'}\big[\mathrm{e}^{Y_\lambda}\big] \geqslant \max_{\lambda\in[0,1]} \mathbb{E}_\nu\left[\ln\left(1 - \lambda\frac{X-\mu}{1-\mu}\right)\right]$$

and by taking the infimum over all distributions ν' with $\mathrm{E}(\nu') > \mu$:

$$\mathcal{K}_{\inf}(\nu, \mu) \geqslant \max_{0 \leqslant \lambda \leqslant 1} \mathbb{E}_\nu \left[\ln \left(1 - \lambda \frac{X - \mu}{1 - \mu} \right) \right] \tag{67}$$

We now only need to prove the converse inequality.

To do so, we define the function

$$H : \lambda \in [0, 1] \longmapsto \mathbb{E}_\nu \left[\ln \left(1 - \lambda \frac{X - \mu}{1 - \mu} \right) \right]$$

The function is well defined, except maybe at $\lambda = 1$ when $\nu\{1\} > 0$; we then take it equal to $-\infty$. We begin by a study of the function H.

Lemma 17 *Assume here that $\mu < \mathrm{E}(\nu) < 1$. The function H is twice differentiable on $(0, 1)$ and its derivative can be defined at 1. For all $\lambda \in (0, 1]$,*

$$H'(\lambda) = \frac{1}{\lambda} \left(1 - \mathbb{E}_\nu \left[\frac{1}{1 - \lambda \frac{X - \mu}{1 - \mu}} \right] \right) \tag{68}$$

Moreover, for $\lambda^\star \in \operatorname*{arg\,max}_{0 \leqslant \lambda \leqslant 1} H(\lambda)$, we have

$$\mathbb{E}_\nu \left[\frac{1}{1 - \lambda^\star \frac{X - \mu}{1 - \mu}} \right] = 1 \ \textit{if } \lambda^\star < 1 \qquad \textit{and} \qquad \mathbb{E}_\nu \left[\frac{1 - \mu}{1 - X} \right] \leqslant 1 \ \textit{if } \lambda^\star = 1 \, ;$$

in the case when $\lambda^\star = 1$, we have in particular $\nu\{1\} = 0$.

Proof For $\lambda \in (0, 1)$, we get, by legitimately differentiating under the expectation,

$$H'(\lambda) = \mathbb{E}_\nu \left[\left(\frac{X - \mu}{1 - \mu} \right) \frac{1}{1 - \lambda \frac{X - \mu}{1 - \mu}} \right] \quad \text{and} \quad H''(\lambda) = -\frac{1}{(1 - \mu)^2} \mathbb{E}_\nu \left[\frac{(X - \mu)^2}{\left(1 - \lambda \frac{X - \mu}{1 - \mu} \right)^2} \right] . \tag{69}$$

Indeed as long as $\lambda < 1$, both variables in the expectations are bounded and we may invoke a standard differentiation theorem under the integral sign. This proves that $H'' < 0$ and therefore that H is strictly concave on $(0, 1)$. Furthermore, H is continuous on $[0, 1]$, possibly by defining $H(1) = -\infty$, as by monotone convergence

$$\lim_{\lambda \to 1} \mathbb{E} \left[\ln \left(1 - \lambda \frac{X - \mu}{1 - \mu} \right) \mathbb{1}_{\{X < \mu\}} \right] = \mathbb{E} \left[\ln \left(\frac{1 - X}{1 - \mu} \right) \mathbb{1}_{\{X < \mu\}} \right]$$
$$\lim_{\lambda \to 1} \mathbb{E} \left[\ln \left(1 - \lambda \frac{X - \mu}{1 - \mu} \right) \mathbb{1}_{\{X \geqslant \mu\}} \right] = \mathbb{E} \left[\ln \left(\frac{1 - X}{1 - \mu} \right) \mathbb{1}_{\{X \geqslant \mu\}} \right]$$

where the first expectation is finite (but the second may equal $-\infty$). The same argument shows that H' is continuous on $[0, 1]$, and therefore (by a theorem on the limit of the derivatives) that H is

right-derivable at 0 with derivative $-(\mathbb{E}(\nu)-\mu)/(1-\mu) > 0$. Since H is strictly concave on $(0,1)$ and continuous, it reaches its maximum exactly once in $[0,1]$. The last disjunction comes from the fact that since $H'(0) > 0$ and H' is decreasing, either $H'(1) \geqslant 0$ and H reaches its maximum at 1, or $H'(1) < 0$ and H reaches its maximum inside $(0,1)$. Since H is continuously differentiable, the derivative at the maximum is 0 in that case, which implies the equality of the expectation. ■

We may now turn to the rest of the proof of Lemma 11.

Proof For the inequality converse to (67), it is enough to show that there exists one value of λ and one measure ν' such that $\mathbb{E}(\nu') > \mu$ and $\nu \ll \nu'$ and

$$\mathrm{KL}(\nu,\nu') \leqslant \mathbb{E}_\nu\left[\ln\left(1-\lambda\frac{X-\mu}{1-\mu}\right)\right] \tag{70}$$

Recalling the definition of the KL, it thus suffices to find λ and ν' that satisfy the above conditions and

$$\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}(x) = 1-\lambda\frac{x-\mu}{1-\mu} \qquad \nu\text{-a.s.} \tag{71}$$

We look for these by setting for $\lambda \in [0,1]$ the measure ν_λ defined by

$$\mathrm{d}\nu_\lambda = \frac{1}{1-\lambda\frac{x-\mu}{1-\mu}}\,\mathrm{d}\nu + \left(1-\mathbb{E}_\nu\left[\frac{1}{1-\lambda\frac{X-\mu}{1-\mu}}\right]\right)\mathrm{d}\delta_1 \tag{72}$$

where δ_1 is the Dirac delta measure at 1. This defines a probability measure if and only if the coefficient in front of $\mathrm{d}\delta_1$ is non-negative, i.e. if

$$\lambda H'(\lambda) \geqslant 0$$

Then for λ satifsfying this condition, ν_λ is a probability measure and $\nu \ll \nu_\lambda$. Furthermore, by (68):

$$\begin{aligned}
\mathrm{E}(\nu_\lambda) &= \int \frac{x}{1-\lambda(x-\mu)/(1-\mu)}\mathrm{d}\nu(x) + \lambda H'(\lambda) \\
&= \int \frac{x-\mu}{1-\lambda(x-\mu)/(1-\mu)}\mathrm{d}\nu(x) + \mu\big(1-\lambda H'(\lambda)\big) + \lambda H'(\lambda) \\
&= \mu - (1-\mu)H'(\lambda)(1-\lambda)
\end{aligned}$$

We wish to consider the case where $\mathrm{E}(\nu_\lambda) \geqslant \mu$ to use it to prove our inequality. The only value of λ that satisfies at the same time $H'(\lambda) \geqslant 0$ and $H'(\lambda)(1-\lambda) \leqslant 0$ is $\lambda^\star$, at which H reaches its maximum.

Now all that is left to prove is that

$$\frac{\mathrm{d}\nu}{\mathrm{d}\nu_{\lambda^\star}}(x) = 1-\lambda^\star\frac{x-\mu}{1-\mu} \qquad \nu\text{-a.s.}$$

We do so by distinguishing two cases. If $\lambda^\star < 1$, then by Lemma 17 the expectation in (72) is equal to 1, that is, the $\mathrm{d}\delta_1$ comes with a 0 factor. Hence, $\nu_{\lambda^\star}$ is absolutely continuous with respect to ν, with a positive density given by the inverse of what we read in (72).

If $\lambda^\star = 1$, then again by Lemma 17, we know that ν does not put any probability mass at 1, which guarantees once again the desired equality. ■

D.2. Proof of Lemma 10

The proof below is variations on the proofs that can be found in Honda and Takemura (2015) or earlier references.

Proof To prove (29) we upper bound $\mathcal{K}_{\inf}(\nu, \mu - \varepsilon)$. Let a probability distribution $\nu' \in \mathcal{P}[0,1]$ be such that

$$\mathrm{E}(\nu') > \mu - \varepsilon \quad \text{and} \quad \nu' \gg \nu \,.$$

Since ν' has a countable number of atoms, one can choose a real number $x > \mu$, arbitrary close to 1, such that $\delta_x \perp \nu'$, where δ_x is the Dirac distribution at x. Let the probability distribution ν'_α be the convex combination

$$\nu'_\alpha = \alpha\delta_x + (1-\alpha)\nu'$$

where,

$$\alpha = \frac{\varepsilon}{x - (\mu - \varepsilon)} \,,$$

this choice of α entails that:

$$\mathrm{E}(\nu'_\alpha) = (1-\alpha)\,\mathrm{E}(\nu') + \alpha x > (1-\alpha)(\mu - \varepsilon) + \alpha x = \mu \,.$$

Moreover, since $\nu'_\alpha \gg \nu' \gg \nu$ and $\delta_x \perp \nu'$, one obtains the following relations between the Radon-Nikodym derivative of ν over ν' and ν'_α :

$$\frac{\mathrm{d}\nu}{\mathrm{d}\nu'_\alpha} = \frac{1}{1-\alpha}\frac{\mathrm{d}\nu}{\mathrm{d}\nu'} \,.$$

This allows to compute explicitly the Kullback-Leibler divergence

$$\mathrm{KL}(\nu, \nu'_\alpha) = \int \ln\left(\frac{\mathrm{d}\nu}{\mathrm{d}\nu'_\alpha}\right) \mathrm{d}\nu = \mathrm{KL}(\nu, \nu') + \ln\frac{1}{1-\alpha} \,.$$

Since $\mathrm{E}(\nu'_\alpha) > \mu$ and by the definition of $\mathcal{K}_{\inf}$ we can lower bound the first term in the equality above

$$\mathcal{K}_{\inf}(\nu, \mu) \leqslant \mathrm{KL}(\nu, \nu') + \ln\frac{1}{1-\alpha} \,,$$

letting x go to 1, which implies α go to $\varepsilon/(1-\mu+\varepsilon)$ we have

$$\mathcal{K}_{\inf}(\nu, \mu) \leqslant \mathrm{KL}(\nu, \nu') + \ln\frac{1-\mu+\varepsilon}{1-\mu} = \mathrm{KL}(\nu, \nu') + \ln\left(1 + \frac{\varepsilon}{1-\mu}\right) \leqslant \mathrm{KL}(\nu, \nu') + \frac{\varepsilon}{1-\mu}$$

and thus taking the infimum over all the probability distributions ν' such that $\mathrm{E}(\nu') > \mu - \varepsilon$ entails that

$$\mathcal{K}_{\inf}(\nu, \mu) \leqslant \mathcal{K}_{\inf}(\nu, \mu - \varepsilon) + \frac{\varepsilon}{1-\mu} \,.$$

To prove the second part (30), we follow the same path as above. Let a probability distribution $\nu' \in \mathcal{P}[0,1]$ be such that

$$\mathrm{E}(\nu') > \mu \quad \text{and} \quad \nu' \gg \nu \,.$$

Let the probability distribution ν'_α be the convex combination $\nu'_\alpha = (1-\alpha)\nu' + \alpha\nu$, where

$$\alpha = \frac{\varepsilon}{\big(\mathrm{E}(\nu') - \mathrm{E}(\nu)\big)} \in (0,1) \text{ because } \mathrm{E}(\nu) < \mu - \varepsilon\,.$$

By definition, we have $\mathrm{E}(\nu'_\alpha) = \mathrm{E}(\nu') - \alpha\big(\mathrm{E}(\nu') - \mathrm{E}(\nu)\big)$, therefore $\mathrm{E}(\nu'_\alpha) > \mu - \varepsilon$. Thanks to the following order of absolute continuity $\nu' \gg \nu'_\alpha \gg \nu$, we can easily compute the Radon-Nikodym derivative

$$\frac{\mathrm{d}\nu}{\mathrm{d}\nu'} = \frac{\mathrm{d}\nu}{\mathrm{d}\nu'_\alpha}\frac{\mathrm{d}\nu'_\alpha}{\mathrm{d}\nu} = \frac{\mathrm{d}\nu}{\mathrm{d}\nu'_\alpha}\Big((1-\alpha) + \frac{\mathrm{d}\nu}{\mathrm{d}\nu'}\Big)\,,$$

and the Kullback-Leibler divergence between ν and ν':

$$\begin{aligned}
\mathrm{KL}(\nu,\nu') &= \int \ln\Big(\frac{\mathrm{d}\nu}{\mathrm{d}\nu'_\alpha}\Big)\,\mathrm{d}\nu + \int \ln\Big((1-\alpha) + \alpha\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}\Big)\,\mathrm{d}\nu \\
&\geqslant \int \ln\Big(\frac{\mathrm{d}\nu}{\mathrm{d}\nu'_\alpha}\Big)\,\mathrm{d}\nu + \alpha\int \ln\Big(\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}\Big)\,\mathrm{d}\nu \\
&= \mathrm{KL}(\nu,\nu'_\alpha) + \alpha\mathrm{KL}(\nu,\nu')\,.
\end{aligned}$$

where we use the concavity of logarithm. Now to recover the term $\mathcal{K}_{\inf}(\nu,\mu-\varepsilon)$ we use in this order: the Pinsker inequality, the fact that $\mathrm{KL}(\nu,\nu'_\alpha) \geqslant \mathcal{K}_{\inf}(\nu,\mu-\varepsilon)$ and $\mathrm{E}(\nu') - \mathrm{E}(\nu) \geqslant \varepsilon$,

$$\begin{aligned}
\mathrm{KL}(\nu,\nu') &\geqslant \mathrm{KL}(\nu,\nu'_\alpha) + \alpha\mathrm{KL}(\nu,\nu') \\
&\geqslant \mathrm{KL}(\nu,\nu'_\alpha) + 2\alpha\big(\mathrm{E}(\nu') - \mathrm{E}(\nu)\big)^2 \\
&\geqslant \mathcal{K}_{\inf}(\nu,\mu-\varepsilon) + 2\varepsilon\big(\mathrm{E}(\nu') - \mathrm{E}(\nu)\big) \\
&\geqslant \mathcal{K}_{\inf}(\nu,\mu-\varepsilon) + 2\varepsilon^2\,.
\end{aligned}$$

To conclude it remains to take the infimum in the last inequality over the probability distributions ν' such that $\mathrm{E}(\nu') > \mu$.

■

D.3. Proof of Proposition 12

The following proof is exactly the same as that of Cappé et al. (2013, Lemma 6), except that we correct a small mistake in the constant.

Proof Fix a real number $\gamma \in (0,1)$ and let S_γ be the set

$$S_\gamma = \left\{\frac{1}{2} - \left\lfloor\frac{1}{2\gamma}\right\rfloor\gamma, \dots, \frac{1}{2} - \gamma, \frac{1}{2}, \frac{1}{2} + \gamma, \dots, \frac{1}{2} + \left\lfloor\frac{1}{2\gamma}\right\rfloor\gamma\right\},$$

which has at most $1 + 1/\gamma$ elements. Thanks to Lemma 18 below, for all $\widetilde{\lambda} \in [0,1]$ there exists a $\widetilde{\lambda}' \in S_\gamma$ such that for all $x \in [0,1]$

$$\ln\left(1 - \widetilde{\lambda}\frac{x - \mathrm{E}(\mu)}{1 - \mathrm{E}(\mu)}\right) \leqslant 2\gamma + \ln\left(1 - \widetilde{\lambda}'\frac{x - \mathrm{E}(\mu)}{1 - \mathrm{E}(\mu)}\right),$$

$$\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu)\big) = \max_{0\leqslant\widetilde{\lambda}\leqslant 1} \frac{1}{n}\sum_{k=1}^{n} \ln\left(1-\widetilde{\lambda}'\frac{X_k-\mathrm{E}(\nu)}{1-\mathrm{E}(\nu)}\right) \leqslant 2\gamma + \max_{\widetilde{\lambda}\in S_\gamma} \frac{1}{n}\sum_{k=1}^{n} \ln\left(1-\widetilde{\lambda}'\frac{X_k-\mathrm{E}(\mu)}{1-\mathrm{E}(\mu)}\right), \tag{73}$$

thanks to the variational representation of $\mathcal{K}_{\inf}$ (Lemma 11). It remains to apply the Markov's inequality and the union bound. Using the upper bound in Lemma 11 and the union bound we obtain

$$\mathbb{P}\Big[\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu)\big) \geqslant u\Big] \leqslant \sum_{\widetilde{\lambda}\in S_\gamma} \mathbb{P}\left[\frac{1}{n}\sum_{k=1}^{n} \ln\left(1-\widetilde{\lambda}'\frac{X_k-\mathrm{E}(\nu)}{1-\mathrm{E}(\nu)}\right) \geqslant u-2\gamma\right], \tag{74}$$

By Markov's inequality, for all $\widetilde{\lambda}\in[0,1]$ we have

$$\begin{aligned}\mathbb{P}\left[\frac{1}{n}\sum_{k=1}^{n} \ln\left(1-\widetilde{\lambda}'\frac{X_k-\mathrm{E}(\nu)}{1-\mathrm{E}(\nu)}\right) \geqslant u-2\gamma\right] &\leqslant \mathrm{e}^{-n(u-\gamma)}\,\mathbb{E}\left[\prod_{k=1}^{n}\left(1-\widetilde{\lambda}'\frac{X_k-\mathrm{E}(\nu)}{1-\mathrm{E}(\nu)}\right)\right] \\ &= \mathrm{e}^{-n(u-2\gamma)},\end{aligned}$$

using the independence of the X_k, thus plugging it in (74), we obtain

$$\mathbb{P}\Big[\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu)\big) \geqslant u\Big] \leqslant \sum_{\widetilde{\lambda}\in S_\gamma} \mathrm{e}^{-n(u-2\gamma)} \leqslant (1+1/\gamma)\,\mathrm{e}^{-n(u-2\gamma)}$$

since the cardinality of S_γ is at most $1+1/\gamma$. Taking $\gamma = 1/(2n)$ allows us to conclude. ■

The proof above relied on the following lemma, which is extracted from Cappé et al. (2013, Lemma 7) Its elementary proof consists in bounding of derivative of $\lambda \mapsto \ln(1-\lambda c)$ and using a convexity argument.

Lemma 18 *For all $\lambda, \lambda' \in [0,1)$ such that either $\lambda \leqslant \lambda' \leqslant 1/2$ or $1/2 \leqslant \lambda' \leqslant \lambda$, for all real numbers $c \leqslant 1$,*

$$\ln(1-\lambda c) - \ln(1-\lambda' c) \leqslant 2|\lambda - \lambda'|$$