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

KL-UCB-Switch: Optimal Regret Bounds for Stochastic Bandits from Both a Distribution-Dependent and a Distribution-Free Viewpoints

Aurélien Garivier AURELIEN.GARIVIER@ENS-LYON.FR
Univ. Lyon, ENS de Lyon, UMPA UMR 5669, LIP UMR 5668, Lyon, France

Hédi Hadiji HEDI.HADIJI@GMAIL.COM
Laboratoire de mathématiques d'Orsay, Université Paris-Saclay, CNRS, Orsay, France

Pierre Ménard PIERRE.MENARD@INRIA.FR
Inria Lille Nord Europe, Lille, France

Gilles Stoltz GILLES.STOLTZ@UNIVERSITE-PARIS-SACLAY.FR
Laboratoire de mathématiques d'Orsay, Université Paris-Saclay, CNRS, Orsay, France

Abstract

We consider K–armed stochastic bandits and consider cumulative regret bounds up to time T. We are interested in strategies achieving *simultaneously* a distribution-free regret bound of optimal order $\sqrt{KT}$ and a distribution-dependent regret that is asymptotically optimal, that is, matching the $\kappa \ln T$ lower bound by Lai and Robbins (1985) and Burnetas and Katehakis (1996), where κ is the optimal problem-dependent constant. This constant κ depends on the model $\mathcal{D}$ considered (the family of possible distributions over the arms). Ménard and Garivier (2017) provided strategies achieving such a bi-optimality in the parametric case of models given by one-dimensional exponential families, while Lattimore (2016, 2018) did so for the family of (sub)Gaussian distributions with variance less than 1. We extend this result to the non-parametric case of all distributions over $[0, 1]$. We do so by combining the MOSS strategy by Audibert and Bubeck (2009), which enjoys a distribution-free regret bound of optimal order $\sqrt{KT}$, and the KL-UCB strategy by Cappé et al. (2013), for which we provide in passing the first analysis of an optimal distribution-dependent $\kappa \ln T$ regret bound in the model of all distributions over $[0, 1]$. We were able to obtain this non-parametric bi-optimality result while working hard to streamline the proofs (of previously known regret bounds and thus of the new analyses carried out); a second merit of the present contribution is therefore to provide a review of proofs of classical regret bounds for index-based strategies for K–armed stochastic bandits.

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

1. Introduction, Brief Literature Review, and Main Achievements

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 samples at each step $t \in \mathbb{N}^*$ one out of K independent sources of randomness, and receives the corresponding outcome as a reward. The most investigated challenge is to minimize the 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. 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, see (3). In the 1990s, Agrawal (1995) and Burnetas and Katehakis (1996) analyzed the UCB algorithm, a simple procedure that picks at step t the arm with the highest upper confidence bound constructed on the past observations. 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. First, a bandit strategy should not address only specific statistical models, but general and non-parametric families of probability distributions, e.g., bounded distributions. (Unless stated otherwise, results discussed below hold for the model of all distributions over a known bounded interval, e.g., $[0, 1]$.) Second, the regret analysis should not only be asymptotic, but should provide finite-time bounds (with closed-form expressions). 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 $\sqrt{KT}$).

We now summarize and put into perspective how the ideas listed above were implemented over the years. A note in passing is that the present contributions actually date back to Garivier et al. (2018).

1.1 Literature Review

Optimal finite-time distribution-free regret upper bounds. Classical UCB strategies enjoy finite-time distribution-free regret upper bounds of order $\sqrt{KT \ln T}$ (folklore knowledge) while strategies based on exponential weights have such bounds of order $\sqrt{KT \ln K}$, actually holding in the more challenging setting of adversarial rewards (Auer et al., 2002b). A modification of UCB named MOSS was proposed by Audibert and Bubeck (2009) and enjoys an optimal finite-time distribution-free regret upper bound of order $\sqrt{KT}$.

Optimal finite-time distribution-dependent regret upper bounds. The path towards such optimal bounds was longer; optimality refers to matching the lower bound (3).

The pioneering work of Lai (and Robbins—see Lai and Robbins, 1985 and Lai, 1987) revolved around the derivation of asymptotic expansions of Gittins' Bayesian-optimal strategy. These expansions for one-parameter exponential families of reward distributions suggested the introduction of upper-confidence bounds policies involving Kullback-Leibler divergence in Lai (1987). An optimal but (very) asymptotic distribution-dependent regret bound is proved therein, and the MOSS-flavor of the confidence intervals used there could already have led to $\sqrt{KT \ln K}$ minimax bounds. These strategies and asymptotic results were later extended by Burnetas and Katehakis (1996) to more general families of distributions.

Auer et al. (2002a) then took a different angle and exhibited an elegant, elementary, finite-time and non-parametric analysis of the UCB algorithm, at the price of a sub-optimal

distribution-dependent factor in the regret upper bounds (depending on the expectation gaps between distributions). In simple settings (for example, for binary rewards or more generally, for one-dimensional exponential families), finite-time *and* optimal distribution-dependent regret upper bounds were proved by Maillard et al. (2011) and Garivier and Cappé (2011), based on specific versions of the KL-UCB algorithm recalled in Section 2.1. Later on, Kaufmann et al. (2012) with the BayesUCB algorithm or Korda et al. (2013) with Thompson sampling obtained similar results.

The results of most interest for the present article (i.e., finite-time, optimal and non-parametric distribution-dependent regret bounds) were initiated by Honda and Takemura with an algorithm called IMED (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 analysis for IMED was provided for all (semi-)bounded distributions, while the analysis for KL-UCB was restricted to some classes of distributions (e.g., bounded distributions with finite supports). However, the regret bounds for IMED are still somewhat asymptotic and not fully in closed form.

In this respect, a *contribution in passing* of the present article is to finally provide finite-time, optimal and non-parametric distribution-dependent regret bounds for the KL-UCB algorithm.

Enjoying simultaneously distribution-dependent and distribution-free regret bounds. As indicated above, it is a folklore knowledge that classical UCB strategies (e.g., the UCB1 strategy by Auer et al., 2002a) enjoy finite-time distribution-free regret upper bounds of order $\sqrt{KT\ln T}$; these bounds are actually consequences of distribution-dependent regret bounds of the form: for all sub-optimal arms a, for all $T \geqslant 1$,

$$\mathbb{E}[N_a(T)] \leqslant c\,\frac{\ln T}{\Delta_a^2} + r_T\,, \qquad (\star)$$

where, e.g., $c = 8$ and $r_T = 2$ for UCB1. This is obtained via setting a threshold $\varepsilon \in (0,1)$ and upper-bounding the regret as

$$R_T = \sum_{a=1}^{K} \Delta_a\,\mathbb{E}\big[N_a(T)\big] \leqslant \varepsilon T + \sum_{a:\Delta_a>\varepsilon} \left(c\,\frac{\ln T}{\Delta_a} + \Delta_a\, r_T \right) \leqslant \varepsilon T + c\,K\frac{\ln T}{\varepsilon} + K\,r_T\,.$$

For T large enough, $\varepsilon = \sqrt{K(\ln T)/T}$ provides the claimed $\sqrt{KT\ln T}$ bound.

One may wonder whether any strategy with distribution-dependent regret bounds of the form $(\star)$, or of a sharper form like the one achieved by KL-UCB and IMED, automatically enjoys a distribution-free regret bound of order $\sqrt{KT}$ up to logarithmic factors. This is actually not the case in general: the argument above for UCB1 only works because the remainder term r_T is uniform. When this remainder term does depend on the underlying bandit problem, which is typically the case for sharper distribution-dependent regret bounds involving the optimal constants stated in (3), then no distribution-free guarantee follows from distribution-dependent regret bounds (see Lattimore, 2018 for more discussions).

The question now is: given that a strategy can simultaneously enjoy distribution-dependent and distribution-free regret bounds, can it simultaneously enjoy optimal such bounds?

Bi-optimal regret bounds. Lattimore (2016, 2018) and Ménard and Garivier (2017) proved that, in simple parametric settings, a strategy can indeed enjoy, at the same time, finite-time regret bounds that are optimal both from a distribution-dependent and a distribution-free viewpoints; they studied, respectively, (sub)Gaussian distributions with variance less than 1 and one-dimensional exponential families.

The *main contribution* of this article is to extend this result to the non-parametric case of all distributions over $[0,1]$, for an algorithm called KL-UCB-Switch. The latter is an index policy based on KL-UCB and MOSS: it uses the tighter KL-UCB upper confidence bounds whenever an arm has not been pulled often enough and switches otherwise to the looser MOSS upper confidence bounds.

This extension was possible without too many technicalities since we first streamlined and generalized earlier analyses of KL-UCB and MOSS; a second *contribution in passing* of the present article is therefore to provide a review of proofs of classical regret bounds for index-based strategies for K–armed stochastic bandits. Furthermore, our simplified analysis allowed us to derive similar bi-optimality results for the anytime version of this new KL-UCB-Switch algorithm, with little if any additional effort.

Another type of simultaneous regret bounds: "best-of-both-worlds" regret guarantees. A strengthening of the notion of distribution-free regret bounds is offered by (oblivious) adversarial regret bounds, which hold for individual sequences of rewards (not necessarily generated by some stochastic process but picked beforehand). A series of articles initiated by Bubeck and Slivkins (2012) and culminating so far in Zimmert and Seldin (2021) exhibits strategies that enjoy simultaneously finite-time non-parametric distribution-dependent regret bounds of order $\ln T$ and optimal finite-time (oblivious) adversarial regret bounds of order $\sqrt{KT}$. Such a simultaneous regret guarantee is called a "best-of-both-worlds" guarantee. However, so far, the distribution-dependent constant in front of the $\ln T$ in "best-of-both-worlds" guarantees is suboptimal and corresponds, up to some numerical constant, to the one of UCB, that is, to a sum of inverse gaps in expected means. This constant can be much larger than the optimal constant suggested by the lower bound (3) recalled below and which requires some care to be achieved. Put differently, for the time being, the individual-sequence guarantee (which is much stronger than the distribution-free regret guarantee) comes at the cost of a poorer distribution-dependent guarantee. Our stochastic bi-optimality results are thus incomparable with the "best-of-both-worlds" regret guarantees obtained so far, though both series of results have their own merits. It is somehow a matter of taste whether better distribution-dependent constants are preferable to individual-sequence guarantees. The latter are often praised for providing robustness and being able to deal with data that is not given by the realization of independent and identically distributed random draws.

This balance between two types of guarantees may be illustrated on simulations, see, e.g., the ones performed by Besson (2019). He considered, on top of KL-UCB-Switch and of the algorithms discussed later in Section 3, the best algorithm so far for "best-of-both-worlds" guarantees: Tsallis-INF, which was introduced by Audibert and Bubeck (2009) and further analyzed by Zimmert and Seldin (2019) and Zimmert and Seldin (2021). In particular, as expected, this algorithm performs significantly worse than KL-UCB-Switch on stochastic problems.

1.2 Organization of the Article

Section 2 presents the main contributions of this article: a description of the KL-UCB-Switch algorithm, statements of its optimality both from a distribution-free viewpoint (Theorem 1) and from a distribution-dependent viewpoint in the class of all distributions over $[0,1]$ (Theorem 2), and corresponding results (Theorems 3 and 4) for an anytime version of the KL-UCB-Switch algorithm. We actually go one step further by providing, as Honda and Takemura (2015) already achieved for IMED, a negative second-order term of the optimal order $-\ln\ln T$ in the distribution-dependent bound for the version of KL-UCB-Switch relying on the knowledge of the horizon T (Theorem 2).

Section 3 presents some (brief) numerical experiments comparing the performance of an empirically tuned version of the KL-UCB-Switch algorithm to competitors like IMED or KL-UCB. The focus is not only set on the growth of the regret with time, but also on its dependency with respect to the number K of arms.

Section 4 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 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), though sometimes in appendix, which makes our paper fully self-contained.

These results are used as building blocks in Section 5 and 6, where the main theorems of this article are proved: Section 5 is devoted to distribution-free bounds (Theorems 1 and 3), while Section 6 focuses on the anytime distribution-dependent bound (Theorem 4).

Section 7 provides some reflections on the distribution-dependent and distribution-free analyses of our new strategy KL-UCB-Switch. In particular, it explains why a switch between the two types of indices used is conceptually intuitive and handy from a technical viewpoint.

An appendix provides the proofs of the classical material presented in Section 4, whenever these proofs did not fit in a few lines. This includes an anytime analysis of the MOSS strategy (Appendix A) and proofs of the regularity and deviation results on the $\mathcal{K}_{\inf}$ quantity mentioned above (Appendix B, with the use of a variational formula for $\mathcal{K}_{\inf}$ re-proved in Appendix D). All these results might be of independent interest. The appendix also features the proof of the sophisticated distribution-dependent regret bound of Theorem 2, with an optimal second order term of order $-\ln\ln T$ in the case of a known T (Appendix C).

2. Description of the Setting and Statement of the Main Results

We consider the simplest case of a bounded stochastic bandit problem with finitely many arms indexed by $a \in \{1,\dots,K\}$ and with rewards in $[0,1]$. We denote by $\mathcal{P}[0,1]$ the set of probability distributions over $[0,1]$: each arm a is associated with an unknown probability distribution $\nu_a \in \mathcal{P}[0,1]$. 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}. The sequence of these rewards 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 reward 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 sub-optimal arms. The expected regret of a strategy equals

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

$$\text{where} \quad N_a(T) = \sum_{t=1}^T \mathbb{1}_{\{A_t=a\}}.$$

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 sub-optimal 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 $[0,1]$. Hence, a strategy is called optimal from a distribution-free viewpoint if there exists a numerical constant C such that for all $K \geqslant 2$, for all bandit problems $\underline{\nu}$ over $[0,1]$, for all $T \geqslant 1$, the regret is bounded by $R_T \leqslant C\sqrt{KT}$.

The key notion in distribution-dependent lower bounds is the Kullback-Leibler divergence KL between two probability distributions. We recall its definition: for two probability distributions ν, ν' over $[0,1]$, we write $\nu \ll \nu'$ whenever ν is absolutely continuous with respect to ν', and denote by $\mathrm{d}\nu/\,\mathrm{d}\nu'$ the density (the Radon-Nikodym derivative) of ν with respect to ν'. Then,

$$\mathrm{KL}(\nu,\nu') = \begin{cases} \displaystyle\int_{[0,1]} \ln\left(\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}\right) \mathrm{d}\nu & \text{if } \nu \ll \nu'; \\ +\infty & \text{otherwise.} \end{cases}$$

Now, the key information-theoretic quantity for stochastic bandit problems is given by an infimum of Kullback-Leibler divergences: for $\nu_a \in \mathcal{P}[0,1]$ and $x \in [0,1]$,

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

where $\mathrm{E}(\nu'_a)$ denotes the expectation of the distribution ν'_a and where by convention, the infimum of the empty set equals $+\infty$. Because of this convention, we may equivalently define $\mathcal{K}_{\inf}$ as

$$\mathcal{K}_{\inf}(\nu_a, x) = \inf\Big\{\mathrm{KL}(\nu_a,\nu'_a) : \quad \nu'_a \in \mathcal{P}[0,1] \;\text{ with }\; \nu_a \ll \nu'_a \;\text{ and }\; \mathrm{E}(\nu'_a) > x\Big\}\,. \tag{2}$$

As essentially proved by Lai and Robbins (1985) and Burnetas and Katehakis (1996)—see also Garivier et al. (2019)—, for any "reasonable" strategy, for any bandit problem $\underline{\nu}$ over $[0,1]$, for any sub-optimal arm a,

$$\liminf_{T\to\infty} \frac{\mathbb{E}\big[N_a(T)\big]}{\ln T} \geqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}\,. \tag{3}$$

A strategy is called optimal from a distribution-dependent viewpoint if the reverse inequality holds with a lim sup instead of a lim inf, for any bandit problem $\underline{\nu}$ over $[0,1]$ and for any sub-optimal arm a.

By a "reasonable" strategy above, we mean a strategy that (according to the terminology introduced by Burnetas and Katehakis, 1996) is uniformly fast convergent on $\mathcal{P}[0,1]$, that is, such that for all bandit problems $\underline{\nu}$ over $[0,1]$, for all sub-optimal arms a,

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

Such strategies exist, such as, for instance, the UCB strategy mentioned above. For uniformly super-fast convergent strategies, that is, strategies for which there actually exists a constant C such for all bandit problems $\underline{\nu}$ over $[0,1]$, for all sub-optimal arms a,

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

(again, UCB is such a strategy), the lower bound above can be strengthened into: for any bandit problem $\underline{\nu}$ over $[0,1]$, for any sub-optimal arm a,

$$\mathbb{E}\big[N_a(T)\big] \geqslant \frac{\ln T}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} - \Omega(\ln\ln T)\,, \tag{4}$$

see Garivier et al. (2019, 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 $t = K, \dots, T-1$ **do**
  `Pull an arm` $A_{t+1} \in \underset{a=1,\dots,K}{\arg\max}\, U_a(t)$
  `Get a reward` Y_{t+1} `drawn independently at random according to` $\nu_{A_{t+1}}$
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

$$U_a^{\mathrm{M}}(t) \stackrel{\mathrm{def}}{=} \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)} \ln_+\left(\frac{T}{KN_a(t)}\right)}, \tag{5}$$

where $\ln_+$ denotes the non-negative 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

$$U_a^{\mathrm{KL}}(t) \stackrel{\mathrm{def}}{=} \sup\left\{ \mu \in [0,1] \;\middle|\; \mathcal{K}_{\inf}\big(\widehat{\nu}_a(t), \mu\big) \leqslant \frac{1}{N_a(t)} \ln_+\left(\frac{T}{KN_a(t)}\right) \right\}. \tag{6}$$

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. The purpose is to capture the good properties of both algorithms. In the sequel we will take $f(T,K) = \lfloor (T/K)^{1/5} \rfloor$ for the sake of concreteness and of readability of the bounds, but Section 7.1 explains the (lack of) impact of this choice of $f(T,K)$ on the regret bounds and details which values lead to optimal bounds.

More precisely, we define the index functions

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

The reasons for the choice of a threshold $f(T,K) = \lfloor (T/K)^{1/5} \rfloor$ will become clear in the proof of Theorem 1. Note that asymptotically KL-UCB-Switch should behave like KL-UCB–type algorithm, as for large T we expect the number of pulls of a sub-optimal arm to be of order $N_a(t) \sim \ln(T)$ and optimal arms to have been played linearly many times, entailing $U_a^{\mathrm{M}}(t) \approx U_a^{\mathrm{KL}}(t) \approx \widehat{\mu}_a(t)$.

Since we are considering distributions over $[0,1]$, the data-processing inequality for Kullback-Leibler divergences ensures (see, e.g., Garivier et al., 2019, Lemma 1) that for all $\nu \in \mathcal{P}[0,1]$ and all $\mu \in \big(\mathrm{E}(\nu), 1\big)$,

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

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

$$\mathcal{K}_{\inf}(\nu,\mu) \geqslant 2\big(\mathrm{E}(\nu) - \mu\big)^2, \qquad \text{thus} \qquad U_a^{\mathrm{KL}}(t) \leqslant U_a^{\mathrm{M}}(t) \tag{7}$$

for all arms a and all rounds $t \geqslant K$. In particular, this actually shows that KL-UCB-Switch interpolates between KL-UCB and MOSS,

$$U_a^{\mathrm{KL}}(t) \leqslant U_a(t) \leqslant U_a^{\mathrm{M}}(t)\,. \tag{8}$$

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 two theorems below are provided in Section 5 and Appendix C, respectively.

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*

$$R_T \leqslant (K-1) + 23\sqrt{KT}\,.$$

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 distribution-free regret bound but its construction (relying on a sub-Gaussian assumption) prevents it from being optimal from a distribution-dependent viewpoint; MOSS can even be arbitrarily worse than a classical strategy like UCB in some situations (see Szepesvári and Lattimore, 2020, Section 9.2).

By considering the exact same algorithm, we may also obtain a (sophisticated) distribution-dependent regret bound. A simple analysis similar to the one for Theorem 4 would yield a second-order term in the regret bound below of the order of $\mathcal{O}_T\big((\ln T)^{6/7}\big)$. On the other hand, an extremely technical analysis (deferred to Appendix C) gets the improved second-order term $-\ln\ln T/\mathcal{K}_{\inf}(\nu_a,\mu^\star)$ stated below; it is partially built on the analysis of Honda and Takemura (2015).

We recall that the $\mathcal{O}_T(\,\cdot\,)$ symbol means the following: a quantity Q_T, possibly depending on other parameters than T, is a $\mathcal{O}_T\big(r(T)\big)$ for some positive rate function r if

$$\limsup_{T\to\infty} \frac{|Q_T|}{r(T)} < +\infty\,.$$

Theorem 2 (Distribution-dependent bound) *Given $T \geqslant 1$, the KL-UCB-Switch algorithm, tuned with the knowledge of T and the switch function $f(T,K) = \lfloor (T/K)^{1/5} \rfloor$, ensures that for all bandit problems $\underline{\nu}$ over $[0,1]$ with $\mu^\star \in (0,1)$, for all sub-optimal arms a, for all $T \geqslant K/\min\big\{1-\mu^\star,\,(\Delta_a/9)^{12}\big\}$,*

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

where a finite-time, closed-form expression of the $\mathcal{O}_T(1)$ term is provided in Equation (52) and in the comments following it.

KL-UCB-Switch thus enjoys a distribution-dependent regret bounds of optimal orders, see (3) and (4). This optimal order was already reached by the IMED strategy by Honda and Takemura (2015) on the same model $\mathcal{P}[0,1]$, though the regret bound exhibited for IMED is of a somewhat asymptotic nature. 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 probability 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, on the one hand, there are 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 on the other hand, 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 (which is non-decreasing) by

$$\varphi(x) = \ln_+\big(x(1+\ln_+^2 x)\big) \tag{9}$$

and the associated index functions by

$$U_a^{\text{KL-A}}(t) \stackrel{\text{def}}{=} \sup\left\{ \mu\in[0,1] \;\middle|\; \mathcal{K}_{\inf}\big(\widehat{\nu}_a(t),\mu\big) \leqslant \frac{1}{N_a(t)}\,\varphi\!\left(\frac{t}{KN_a(t)}\right)\right\} \tag{10}$$

$$\text{and}\qquad U_a^{\text{M-A}}(t) \stackrel{\text{def}}{=} \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)}\,\varphi\!\left(\frac{t}{KN_a(t)}\right)}\,. \tag{11}$$

For matters related to proofs, it will also be convenient to define the index function $U_a^{\text{M},\varphi}(t)$ by

$$U_a^{\text{M-A}}(t) \leqslant U_a^{\text{M},\varphi}(t) \stackrel{\text{def}}{=} \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)}\,\varphi\!\left(\frac{T}{KN_a(t)}\right)}\,. \tag{12}$$

The -A in the superscripts stands for "augmented" or for "anytime" as this augmented exploration gives rise to the anytime version of KL-UCB-Switch, which simply relies on the index

$$U_a^{\text{A}}(t) = \begin{cases} U_a^{\text{KL-A}}(t) & \text{if } N_a(t) \leqslant f(t,K), \\ U_a^{\text{M-A}}(t) & \text{if } N_a(t) > f(t,K), \end{cases} \tag{13}$$

where $f(t,K) = \lfloor (t/K)^{1/5} \rfloor$. Note that the thresholds $f(t,K)$ for the switches between the sub-indices $U_a^{\text{KL-A}}(t)$ and $U_a^{\text{M-A}}(t)$ now vary with t (and we cannot exclude that a switch back may occur).

For this anytime version of KL-UCB-Switch, the same ranking of (sub-)indexes holds as the one (8) for our first version of KL-UCB-Switch relying on the horizon T:

$$U_a^{\text{KL-A}}(t) \leqslant U_a^{\text{A}}(t) \leqslant U_a^{\text{M-A}}(t)\,. \tag{14}$$

The performance guarantees are indicated in the next two theorems, whose proofs may be found in Sections 5 and 6, respectively. 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.

Theorem 3 (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 $T \geqslant 1$,*

$$R_T \leqslant (K-1) + 44\sqrt{KT}\,.$$

Theorem 4 (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, for all $T \geqslant 1$,*

$$\mathbb{E}[N_a(T)] \leqslant \frac{\ln T}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} + \mathcal{O}_T\big((\ln T)^{6/7}\big)\,,$$

where a finite-time, closed-form expression of the $\mathcal{O}_T\big((\ln T)^{6/7}\big)$ term is given in Equation (32) and in the comments following it.

3. Numerical Experiments

We provide here some numerical experiments comparing the different algorithms we refer to in this work. These simulations are only provided for the sake of illustration: their high-level message is exactly what we expected to see. Namely, we consider four benchmark algorithms, KL-UCB (yellow curves), MOSS (blue curves), IMED (purple curves), and Tsallis-INF (red curves). Among these, KL-UCB and IMED perform the best from a distribution-dependent point of view (see Figure 1) while MOSS performs the best from a distribution-free point of view (see Figure 2). We consider three instances of KL-UCB-Switch (green curves), with respective switch functions $f(t,K) = \lfloor t/K \rfloor^\alpha$ where $\alpha \in \{1/5,\, 1/2,\, 8/9\}$, and generally observe that well-calibrated versions of KL-UCB-Switch perform as well as, and even outperform, the best benchmark strategies.

We provide a more detailed analysis below but first indicate the exact specifications of the four benchmark algorithms. MOSS is implemented as in (11). KL-UCB is implemented based on the indices

$$\sup\left\{ \mu \in [0,1] \;\middle|\; \mathcal{K}_{\inf}\big(\widehat{\nu}_a(t), \mu\big) \leqslant \frac{\phi(t)}{N_a(t)} \right\}$$

with $\phi(t) = \ln t$; Cappé et al. (2013) recommended $\phi(t) = \ln t + \ln\ln t$ or $\phi(t) = \ln t + 3\ln\ln t$ depending on the model (distributions over $[0,1]$ with finite supports or exponential families), so it was not clear what exploration function $\phi(t)$ to use, which is why we pick the simplest choice $\phi(t) = \ln t$. Note also that unlike the definition (10), we do not define the exploration bonus in terms of $\phi\big((t/K)/N_a(t)\big)$. IMED, from Honda and Takemura (2015), picks the arm

$$A_t \in \arg\min \left\{ N_a(t)\, \mathcal{K}_{\inf}\Big(\widehat{\nu}_a(t), \max_{j \in \{1,\dots,K\}} \widehat{\mu}_j(t)\Big) + \ln N_a(t) \right\}.$$

Tsallis-INF was originally introduced by Audibert and Bubeck (2009) as a minimax optimal algorithm for adversarial rewards (and was later identified, in Audibert et al., 2011, as an instance of a follow-the-regularized-leader strategy). Zimmert and Seldin (2019) and Zimmert

and Seldin (2021) observed that Tsallis-INF also enjoys logarithmic distribution-dependent regret bounds in the stochastic setting, and provided details on an efficient implementation thereof. Tsallis-INF picks A_t at random according to the probability distribution $(p_{t,a})_{a\in\{1,\dots,K\}}$ with coordinates

$$p_{t,a} = 4\left(\eta_t \sum_{s=1}^{t-1} \widehat{L}_{s,a} - C_t\right)^{-2}, \qquad \text{where} \qquad \widehat{L}_{s,a} = \frac{1-Y_s}{p_{a,s}} \mathbb{1}_{\{A_s=a\}}\,, \qquad \eta_t = \frac{2}{\sqrt{t}}\,,$$

and $C_t \in \mathbb{R}$ is a normalization factor.

Distribution-dependent bounds. We compare in Figure 1 the distribution-dependent behaviors of the algorithms. We use a logarithmic scale on the x–axis as the regrets scale logarithmically; we indeed observe linear curves. IMED is the best-performing benchmark for the three situations considered, followed by KL-UCB. The regret of KL-UCB-Switch depends on α: for the small value $\alpha = 1/5$, the performance of KL-UCB-switch follows the one of MOSS; for the intermediate value $\alpha = 1/2$, it follows the one of KL-UCB in two out of the three situations; finally, the choice $\alpha = 8/9$ outperforms all four benchmarks.

Distribution-free bounds. Figure 2 reports the behavior of the normalized regret $R_T/\sqrt{KT}$, either as a function of T (top part of the figure) or of K (bottom part of the figure). This quantity should remain bounded as T or K increases. MOSS and the three versions of KL-UCB-Switch share the same performance and clearly outperform the three other benchmarks. The performance of KL-UCB seems to not scale optimally with T or K, while the one for IMED scales well with T but seem to be slightly suboptimal with K.

Illustration of the switching profiles. Figures 3 and 4 illustrate the switching profiles of optimal and suboptimal arms, in the case $\alpha = 1/5$. Therein, we provide, for each arm, an estimation of the probability, according to time, that it lies in the "KL-UCB mode" (10) or in the "MOSS mode" (11). We also provide an estimation of the distribution of the number of switches (back and forth) between the two modes.

In the first illustration, in Figure 3, we consider a Bernoulli bandit with $K = 2$ Bernoulli arms with close means, namely $\mu_1 = 0.9$ and $\mu_2 = 0.75$. Therein, for most of the runs, both arms switched only once and stayed in the MOSS mode the rest of the time. For the optimal arm, 92% of the runs had their switch exactly at time $t = 4$, and the switch always occurred before time $t = 13$ on the $1,000$ runs considered. For the suboptimal arm, the first switch occurred before time $t = 30$ in 90% of the runs, and before $t = 54$ in 99% of the runs. There were two outliers, with first-switch times at $t = 440$ and $t = 480$.

In the second illustration, in Figure 4, we consider another Bernoulli problem with larger suboptimality gaps in order to highlight the differences in behavior between the arms. We take $K = 5$ arms, associated with means

$$\mu_1 = 0.9, \qquad \mu_2 = \mu_3 = 0.6, \qquad \text{and} \qquad \mu_4 = \mu_5 = 0.3\,.$$

More diverse behaviors arise: while the optimal arm again quickly switches to a MOSS mode, the suboptimal arms have a large probability to switch four times. Also, at time $T = 5,000$, a significant fraction of the arms is again in the initial KL-UCB mode.

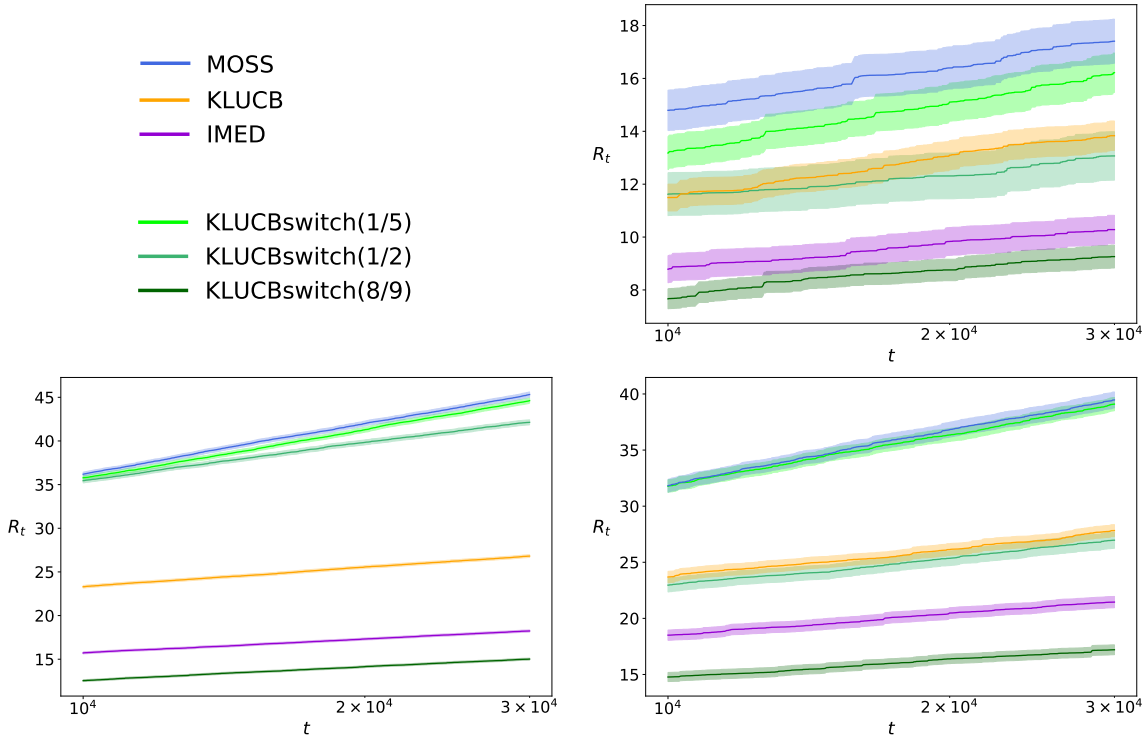


Figure 1: Regrets approximated over 100 runs, shown on a logarithmic scale for the x–axis; the shaded areas correspond to standard errors in the empirical means computed. Distributions of the arms consist of:
[*Top graph*] Bernoulli distributions with parameters (0.9, 0.8)
[*Bottom-left graph*] Exponential distributions with expectations (0.15, 0.12, 0.10, 0.05), truncated on $[0, 1]$
[*Bottom-right graph*] Gaussian distributions with means (0.7, 0.5, 0.3, 0.2) and same standard deviation $\sigma = 0.1$, truncated on $[0, 1]$
The performance of Tsallis-INF is outside of the range considered and is therefore not displayed.

4. 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). *Readers familiar with the material described here are urged to move to the next section.*

4.1 Optional Skipping—How to Go from Global Times t to Local Times n

The trick detailed here is standard in the bandit literature, see, e.g., its application in Auer et al. (2002a). It is sometimes called optional skipping, and sometimes, optional sampling;

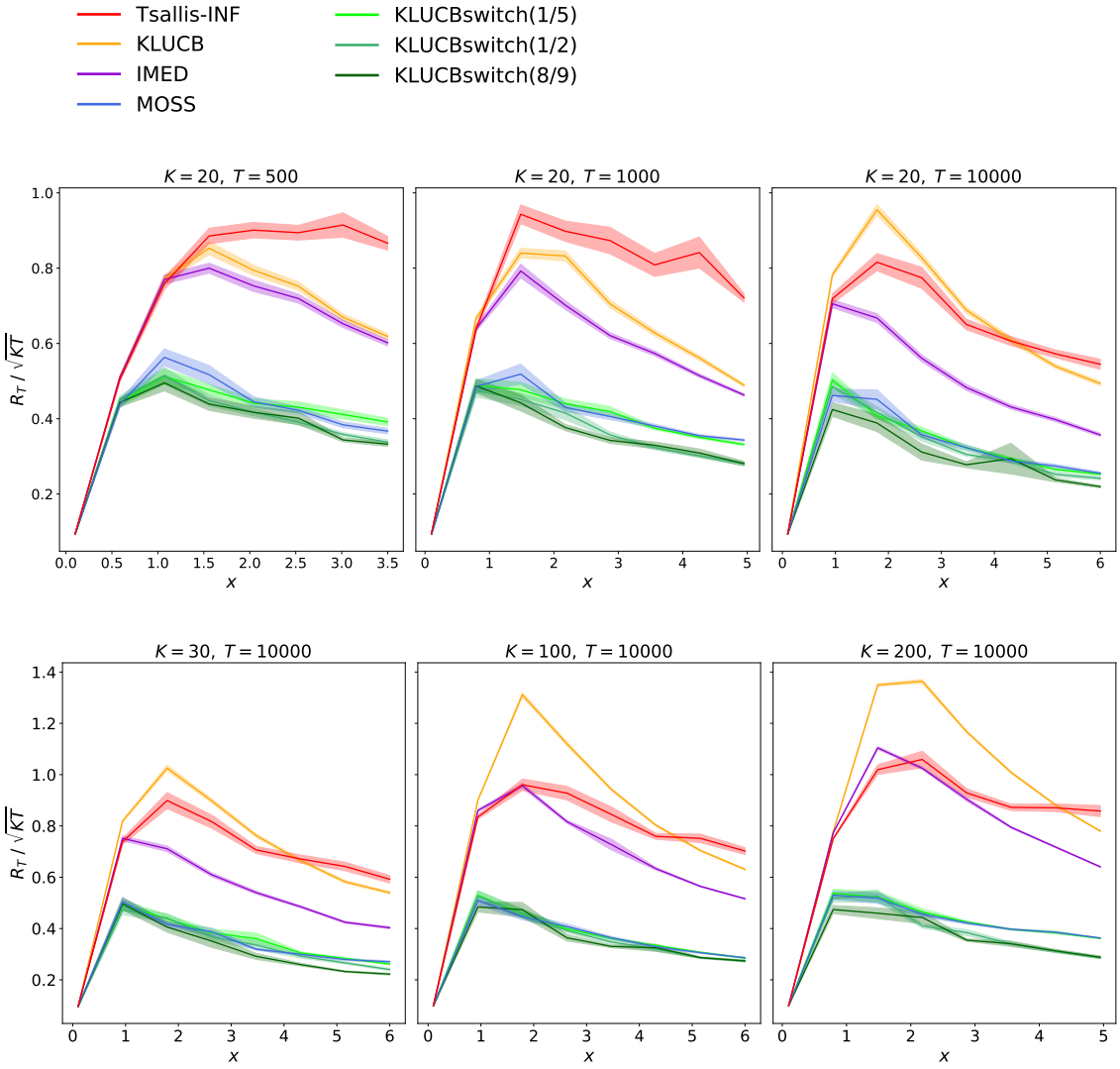
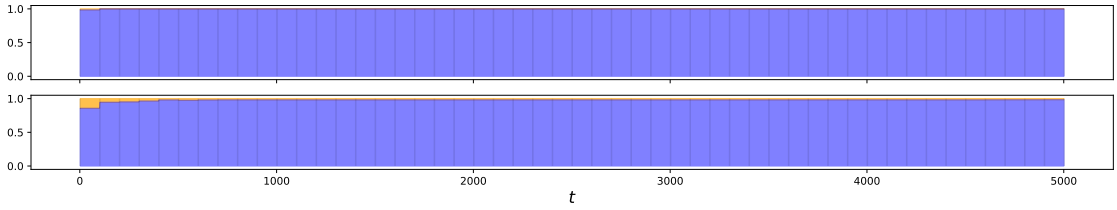


Figure 2: Expected regret $R_T/\sqrt{KT}$, approximated over 100 runs; the shaded areas correspond to standard errors in the empirical means computed.

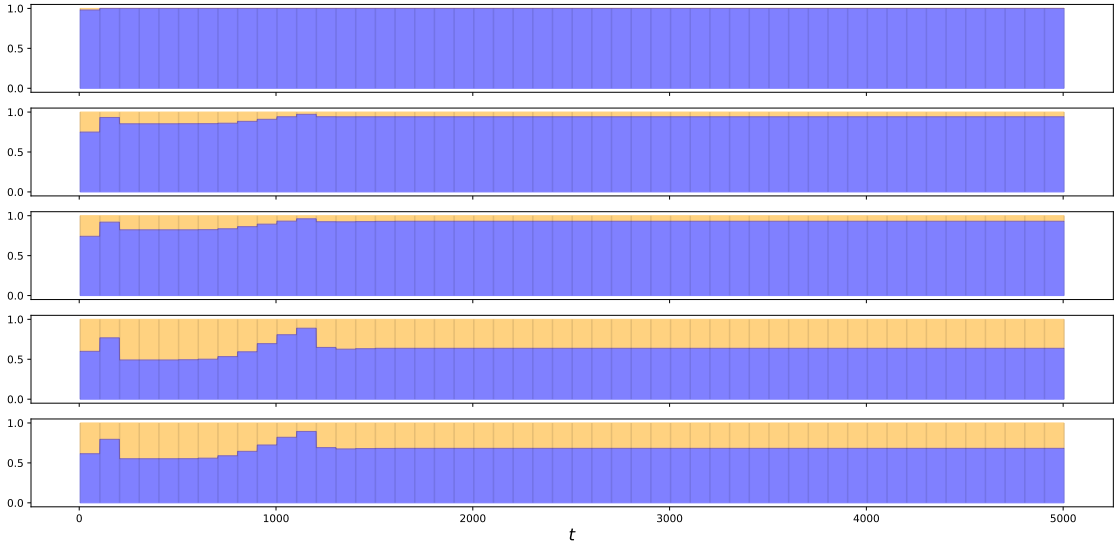
Top graphs: as a function of x, for a Bernoulli bandit problem with $K = 20$ arms, for time horizons $T \in \{500;\ 1,000;\ 10,000\}$, and for respective parameters $(0.8,\ 0.8 - x\sqrt{K/T},\ \ldots,\ 0.8 - x\sqrt{K/T})$

Bottom graphs: as a function of x, for a Bernoulli bandit problem with $K \in \{30,\ 100,\ 200\}$ arms, for a time horizon $T = 10,000$, and for parameters $(0.8,\ 0.8 - x\sqrt{K/T},\ \ldots,\ 0.8 - x\sqrt{K/T})$



Number of switches	0	1	2	3	4	$\geqslant 5$
Optimal arm	0	100	0	0	0	0
Suboptimal arm	0	94.8	0.3	3.3	1.4	0.2

Figure 3: KL-UCB-Switch with $f(t, K) = \lfloor t/K \rfloor^{1/5}$ is run on a Bernoulli bandit problem with $K = 2$ arms, of parameters (0.9, 0.75), and for $T = 5,000$ rounds; $N = 1,000$ runs are performed.
Top graphs: Each box depicts the proportion of runs for which the index of the corresponding arm was in MOSS mode (blue) or in KL mode (orange).
Bottom table: Distributions of the number of switches for each arm (from the KL-UCB mode to the MOSS mode, or the other way round).



Number of switches	0	1	2	3	4	$\geqslant 5$
Optimal arm, $\mu_1 = 0.9$	0	100	0	0	0	0
Suboptimal arms, $\mu_2 = \mu_3 = 0.6$	0	82.2	0.9	10.8	6.1	0
Suboptimal arms, $\mu_4 = \mu_5 = 0.3$	0	54.6	5.8	13.6	26.0	0

Figure 4: Same legend as for Figure 3, for the Bernoulli bandit problem with $K = 5$ arms, of parameters (0.9, 0.6, 0.6, 0.3, 0.3).

we pick the first terminology, following what seems to be the preferred terminology in probability theory[1]. In any case, the original reference is Theorem 5.2 of Doob (1953, Chapter III, p. 145); one can also check Chow and Teicher (1988, Section 5.3) for a more recent reference.

Doob's optional skipping enables the rewriting of 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 $N_a(t) = n$ that the specific arm considered has been pulled so far. 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

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

the round at which arm a was pulled for the n–th time. Now, Doob's optional skipping 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 $t \geqslant K$.

$$\text{on the event } \bigl\{N_a(t) = n\bigr\}, \qquad \widehat{\mu}_a(t) = \widehat{\mu}_{a,N_a(t)} = \widehat{\mu}_{a,n}\,.$$

Here is an example of how to use this rewriting.

Example 1 (Simple application) *In our initial example, we start with a simple application: we consider a subset $\mathcal{E} \subseteq [0,1]$ and are interested in bounding the probability*

$$\mathbb{P}\bigl[\widehat{\mu}_a(t) \in \mathcal{E}\bigr]\,.$$

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. We get

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

so that, by a union bound,

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

1. The abstract of a recent article by Simons et al. (2002) reads: "A general set of distribution-free conditions is described under which an i.i.d. sequence of random variables is preserved under optional skipping. This work is motivated by theorems of J.L. Doob (1936) and Z. Ignatov (1977), unifying and extending aspects of both."

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.

Actually, a more careful use of optional skipping would be

$$\mathbb{P}\big[\widehat{\mu}_a(t) \in \mathcal{E}\big] \leqslant \mathbb{P}\left[\bigcup_{n=1}^{t-K+1} \big\{\widehat{\mu}_{a,n} \in \mathcal{E}\big\}\right] = \mathbb{P}\Big[\exists\, n \in \{1,\dots,t-K+1\} :\ \widehat{\mu}_{a,n} \in \mathcal{E}\Big]\,.$$

There was no constraint on the number of times $N_a(t)$ arm a was pulled in the previous example, but imposing a lower bound $n_0 \geqslant 1$ on $N_a(t)$ leads to a summation over n starting not at 1 but at n_0. For instance (and considering expectations for a change), given a bounded function g,

$$\mathbb{E}\Big[f\big(\widehat{\mu}_a(t)\big)\,\mathbb{1}_{\{N_a(t)\geqslant n_0\}}\Big] = \sum_{n=n_0}^{t-K+1} \mathbb{E}\Big[f\big(\widehat{\mu}_a(t)\big)\,\mathbb{1}_{\{N_a(t)=n\}}\Big] = \sum_{n=n_0}^{t-K+1} \mathbb{E}\Big[f\big(\widehat{\mu}_{a,n}\big)\,\mathbb{1}_{\{N_a(t)=n\}}\Big]\,.$$

Example 2 (More complex application with random arms A_t) *Given a subset $\mathcal{E} \subseteq [0,1]$ and a strategy to sequentially pick arms A_t, we are now interested in bounding the sum of probabilities*

$$\sum_{t=1}^{T} \mathbb{P}\big[\widehat{\mu}_{A_t}(t) \in \mathcal{E}\big]\,.$$

We start with a decomposition according to the values a of A_t and n of $N_a(t)$, for each t:

$$\begin{aligned}\big\{\widehat{\mu}_{A_t}(t) \in \mathcal{E}\big\} &= \bigcup_{a=1}^{K} \bigcup_{n=1}^{t-K+1} \big\{\widehat{\mu}_a(t) \in \mathcal{E} \ \text{ and } \ A_t = a \ \text{ and } \ N_a(t) = n\big\} \\ &= \bigcup_{a=1}^{K} \bigcup_{n=1}^{t-K+1} \big\{\widehat{\mu}_{a,n} \in \mathcal{E} \ \text{ and } \ A_t = a \ \text{ and } \ N_a(t) = n\big\}\,.\end{aligned}$$

Therefore (since for a given t, the events above are disjoint as a and n vary),

$$\sum_{t=1}^{T} \mathbb{P}\big[\widehat{\mu}_{A_t}(t) \in \mathcal{E}\big] = \sum_{a=1}^{K} \sum_{n=1}^{t-K+1} \left(\sum_{t=1}^{T} \mathbb{P}\big[\widehat{\mu}_{a,n} \in \mathcal{E} \ \text{ and } \ A_t = a \ \text{ and } \ N_a(t) = n\big]\right).$$

Now, we observe that for a given pair (a,n), the events

$$\mathcal{N}_{a,n,t} = \big\{A_t = a \ \text{ and } \ N_a(t) = n\big\}$$

are disjoint as t varies from 1 to T (but their union is not necessarily the entire probability space). Indeed, if for a given t_0 we have $A_{t_0} = a$ and $N_a(t_0) = n$, then $N_a(t) \leqslant n-1$ for all $t \leqslant t_0 - 1$, while for $t \geqslant t_0 + 1$, if $A_t = a$ then $N_a(t) \geqslant n+1$. The combination of $A_t = a$ and $N_a(t) = n$ may therefore happen for at most one value of $t \in \{1,\dots,T\}$. Because of this, for a given pair (a,n), we get the upper bound

$$\sum_{t=1}^{T} \mathbb{P}\big[\widehat{\mu}_{a,n} \in \mathcal{E} \ \text{ and } \ A_t = a \ \text{ and } \ N_a(t) = n\big] \leqslant \mathbb{P}\big[\widehat{\mu}_{a,n} \in \mathcal{E}\big]\,.$$

All in all, collecting all inequalities, we have

$$\sum_{t=1}^{T} \mathbb{P}\big[\widehat{\mu}_{A_t}(t) \in \mathcal{E}\big] \leqslant \sum_{a=1}^{K} \sum_{n=1}^{t-K+1} \mathbb{P}\big[\widehat{\mu}_{a,n} \in \mathcal{E}\big] \,.$$

4.2 Maximal Version of Hoeffding's Inequality

The maximal version of Hoeffding's inequality (Proposition 5) is a standard result from Hoeffding (1963). It was already used in the original analysis of MOSS (Audibert and Bubeck, 2009). For our slightly simplified analysis of MOSS (see Section 4.3), we will rather rely on Corollary 6, a consequence of Proposition 5 obtained by integrating it.

Proposition 5 *Let $X_1, \ldots, X_n$ 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 $u \geqslant 0$ and for all $N \geqslant 1$:*

$$\mathbb{P}\left[\max_{n \geqslant N} \big(\widehat{\mu}_n - \mu\big) \geqslant u\right] \leqslant \mathrm{e}^{-2Nu^2} \,. \tag{15}$$

Corollary 6 *Under the same assumptions, for all $\varepsilon \geqslant 0$,*

$$\mathbb{E}\left[\left(\max_{n \geqslant N} \big(\mu - \widehat{\mu}_n - \varepsilon\big)\right)^+\right] \leqslant \sqrt{\frac{\pi}{8}}\sqrt{\frac{1}{N}}\,\mathrm{e}^{-2N\varepsilon^2} \,. \tag{16}$$

Of course, by symmetry Proposition 5 and Corollary 6 hold with $\mu - \widehat{\mu}_n$ instead of $\widehat{\mu}_n - \mu$.

Proof By the Fubini-Tonelli theorem, an integration of the maximal deviation inequality (15) yields

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

■

4.3 Distribution-Free Bound for the MOSS Algorithm

Such a distribution-free bound was already provided 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 only provide a slightly shorter and more focused proof of these results based on Corollary 6 and indicate an intermediate result—see (17)—that will be useful for us in the analysis of our new KL-UCB-Switch algorithm. We do not claim any improvement on the results themselves, just a clarification of the existing proofs.

Our proof is slightly shorter and more focused for two reasons. First, in the two references mentioned, the peeling trick was used on the probabilities of deviations (see Proposition 5) 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, based on Corollary 6 and at fixed times depending solely on T. Second, unlike the two mentioned references, we do not attempt to simultaneously build a distribution-free and some type of distribution-dependent bound. This raised technical difficulties because of the correlations between the choices of the arms and the observed rewards. The idea of our approach is to focus solely on the distribution-free regime, for which we notice that some crude bounding neglecting the correlations suffice (i.e., our analysis deals with all sub-optimal arms in the same way, independently of how often they are played).

For a known horizon T, we denote by A^{M}_{t+1} the arm played by the index strategy maximizing, at each step $t+1$ with $t \geqslant K$, the quantities (5):

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

The superscripts M in A^{M}_{t+1} and $U^{\mathrm{M}}_a(t)$ stand for MOSS. We do so not to mix it with the arm A_{t+1} played by the KL-UCB-Switch strategy (no superscript), but of course, once an arm a was sufficiently pulled, we have $A_{t+1} = A^{\mathrm{M}}_{t+1}$ by definition of the KL-UCB-Switch strategy.

Appendix A provides the proof of the following regret bound. We denote by $a^\star$ an optimal arm, i.e., an arm such that $\mu_a = \mu^\star$.

Proposition 7 *For a known horizon $T \geqslant 1$, for all bandit problems $\underline{\nu}$ over $[0,1]$, MOSS achieves a regret bound smaller than $R_T \leqslant (K-1) + 17\sqrt{KT}$. More precisely, with the notation of optional skipping (Section 4.1), we have the inequalities*

$$\begin{aligned} R_T &= T\mu^\star - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{A^{\mathrm{M}}_t}\right] \\ &\leqslant (K-1) + \overbrace{\sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{M}}_{a^\star}(t-1)\big)^+\Big]}^{\leqslant 13\sqrt{KT}} \\ &\qquad + \underbrace{\sqrt{KT} + \sum_{a=1}^{K}\sum_{n=1}^{T} \mathbb{E}\left[\left(\widehat{\mu}_{a,n} + \sqrt{\frac{\ln_+\big(T/(Kn)\big)}{2n}} - \mu_a - \sqrt{\frac{K}{T}}\right)^+\right]}_{\leqslant 4\sqrt{KT}}\,. \end{aligned} \tag{17}$$

Remark 8 *The proof (see Remark 20) actually reveals that for a known horizon $T \geqslant 1$, for all bandit problems $\underline{\nu}$ over $[0,1]$, and for all strategies (not only MOSS), the following bound holds:*

$$\sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{M}}_{a^\star}(t-1)\big)^+\Big] \leqslant 13\sqrt{KT}\,.$$

We will re-use this fact to state a similar remark below (Remark 10), which will be useful for Part 2 of the proof lying in Section 5.

Our proof in Appendix A reveals that designing an adaptive version of MOSS comes at no effort. For this adaptive version we will also want to possibly explore more. We will do so by considering an augmented exploration function φ, that is, a function $\varphi \geqslant \ln_+$ as in (9). We therefore define MOSS-anytime (M-A) as relying on the indexes defined in (11), which we copy here:

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

We denote by $A_{t+1}^{\text{M-A}}$ the arm picked as $\arg\max_{a=1,\dots,K} U_a^{\text{M-A}}(t)$.

Proposition 9 *For all horizons $T \geqslant 1$, for all bandit problems $\underline{\nu}$ over $[0,1]$, MOSS-anytime achieves a regret bound smaller than $R_T \leqslant (K-1) + c\sqrt{KT}$ where $c = 30$ for $\varphi = \ln_+$ and $c = 33$ for the augmented exploration function $\varphi(x) = \ln_+\big(x(1+\ln_+^2 x)\big)$ defined in (9). More precisely, with the notation of optional skipping (Section 4.1), we have the inequalities*

$$\begin{aligned} R_T &= T\mu^\star - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{A_t^{\text{M-A}}}\right] \\ &\leqslant (K-1) + \overbrace{\sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star}^{\text{M-A}}(t-1)\big)^+\Big]}^{\leqslant 26\sqrt{KT}} \\ &\qquad + \underbrace{\sqrt{KT} + \sum_{a=1}^{K}\sum_{n=1}^{T} \mathbb{E}\left[\left(\widehat{\mu}_{a,n} + \sqrt{\frac{\varphi\big(T/(Kn)\big)}{2n}} - \mu_a - \sqrt{\frac{K}{T}}\right)^+\right]}_{\leqslant 4\sqrt{KT} \text{ for } \varphi=\ln_+ \text{ and } 7\sqrt{KT} \text{ for } \varphi(x)=\ln_+(x(1+\ln_+^2 x))}\,. \end{aligned} \tag{18}$$

Remark 10 *Similarly to above, the proof (see Remark 20) actually reveals that for a known horizon $T \geqslant 1$, for all bandit problems $\underline{\nu}$ over $[0,1]$, and for all strategies (not only MOSS-anytime), the following bound holds:*

$$\sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star}^{\text{M-A}}(t-1)\big)^+\Big] \leqslant 26\sqrt{KT}\,.$$

This remark will be useful for Part 2 of the proof lying in Section 5.

4.4 Regularity and Deviation/Concentration Results on $\mathcal{K}_{\inf}$

We start with a quantification of the (left-)regularity of $\mathcal{K}_{\inf}$ and then provide a deviation and a concentration result on $\mathcal{K}_{\inf}$.

4.4.1 Regularity of $\mathcal{K}_{\inf}$

The lower left-semi-continuity (19) first appeared as Lemma 7 in Honda and Takemura (2015), see also Garivier et al. (2019, Lemma 3) for a later but simpler proof. The upper

left-semi-continuity (20) relies on the same arguments as (7), namely, the data-processing inequality for Kullback-Leibler divergences and Pinsker's inequality. These two inequalities are proved in detail in Appendix B; the proposed proofs are slightly simpler or lead to sharper bounds than in the mentioned references.

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

$$\forall \varepsilon \in [0, \mu]\,, \qquad \mathcal{K}_{\inf}(\nu, \mu) \leqslant \mathcal{K}_{\inf}(\nu, \mu - \varepsilon) + \frac{\varepsilon}{1-\mu}\,, \tag{19}$$

and

$$\forall \varepsilon \in \big[0, \mu - \mathrm{E}(\nu)\big]\,, \qquad \mathcal{K}_{\inf}(\nu, \mu) \geqslant \mathcal{K}_{\inf}(\nu, \mu - \varepsilon) + 2\varepsilon^2\,. \tag{20}$$

We draw two consequences from Lemma 11: the left-continuity of $\mathcal{K}_{\inf}$ and a useful inclusion in terms of level sets.

Corollary 12 *For all $\nu \in \mathcal{P}[0,1]$, the function $\mathcal{K}_{\inf}(\nu, \cdot) : \mu \in (0,1) \mapsto \mathcal{K}_{\inf}(\nu, \mu)$ is left-continuous. In particular, on the one hand, $\mathcal{K}_{\inf}\big(\nu, \mathrm{E}(\nu)\big) = 0$ whenever $\mathrm{E}(\nu) \in (0,1)$, and on the other hand, for all $\nu \in \mathcal{P}[0,1]$ and $\mu \in (0,1)$,*

$$\mathcal{K}_{\inf}(\nu, \mu) = \inf\Big\{\mathrm{KL}(\nu, \nu') : \quad \nu' \in \mathcal{P}[0,1] \quad \text{and} \quad \mathrm{E}(\nu') \geqslant \mu\Big\}\,.$$

Proof The left-continuity follows from a sandwich argument via the upper bound (19) and the lower bound $\mathcal{K}_{\inf}(\nu, \mu - \varepsilon) \leqslant \mathcal{K}_{\inf}(\nu, \mu)$ that holds for all $\varepsilon \in [0, \mu]$ by the very definition of $\mathcal{K}_{\inf}$. The fact that $\mathcal{K}_{\inf}\big(\nu, \mathrm{E}(\nu) - \varepsilon\big) = 0$ for all $\varepsilon \in \big(0, \mathrm{E}(\nu)\big]$ thus entails, in particular, that $\mathcal{K}_{\inf}\big(\nu, \mathrm{E}(\nu)\big) = 0$. ■

Corollary 13 *For all $\nu \in \mathcal{P}[0,1]$, all $\mu \in (0,1)$, all $u > 0$, and all $\varepsilon > 0$,*

$$\big\{\mathcal{K}_{\inf}(\nu, \mu - \varepsilon) > u\big\} \subseteq \big\{\mathcal{K}_{\inf}(\nu, \mu) > u + 2\varepsilon^2\big\}\,.$$

Proof We apply (20) and merely need to explain why the condition $\varepsilon \in \big[0, \mu - \mathrm{E}(\nu)\big]$ therein is satisfied. Indeed, $\mathcal{K}_{\inf}(\nu, \mu - \varepsilon) > u > 0$ indicates in particular that $\mu - \varepsilon > \mathrm{E}(\nu)$, or put differently, $\varepsilon < \mu - \mathrm{E}(\nu)$. ■

4.4.2 Deviation Results on $\mathcal{K}_{\inf}$

We provide two deviation results on $\mathcal{K}_{\inf}$: first, in terms of probabilities of deviations and next, in terms of expected deviations.

The first deviation inequality was essentially provided by Cappé et al. (2013, Lemma 6). For the sake of completeness, we recall its proof in Section B.

Proposition 14 (deviation result on $\mathcal{K}_{\inf}$) *Let $\widehat{\nu}_n$ denote the empirical distribution associated with a sequence of $n \geqslant 1$ i.i.d. random variables with distribution ν over $[0,1]$ with $\mathrm{E}(\nu) \in (0,1)$. Then, for all $u \geqslant 0$,*

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

A useful corollary in terms of expected deviations can now be stated.

Corollary 15 (integrated deviations for $\mathcal{K}_{\inf}$) *Under the same assumptions as in Proposition 14, for all $\varepsilon > 0$, the index*

$$U_{\varepsilon,n} = \sup\left\{ \mu \in [0,1] \;\middle|\; \mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mu\big) \leqslant \varepsilon \right\}$$

satisfies

$$\mathbb{E}\Big[\big(\mathrm{E}(\nu) - U_{\varepsilon,n}\big)^+\Big] \leqslant (2n+1)\,\mathrm{e}^{-n\varepsilon}\sqrt{\frac{\pi}{n}}\,.$$

Proof By the Fubini-Tonelli theorem, just as in the proof of Corollary 6 (for the first two equalities), and subsequently using the definition of $U_{\varepsilon,n}$ as a supremum (for the third equality, together with the left-continuity of $\mathcal{K}_{\inf}$ deriving from Lemma 11), we have

$$\begin{aligned}\mathbb{E}\Big[\big(\mathrm{E}(\nu) - U_{\varepsilon,n}\big)^+\Big] &= \int_0^{+\infty} \mathbb{P}\Big[\mathrm{E}(\nu) - U_{\varepsilon,n} > u\Big]\,\mathrm{d}u = \int_0^{+\infty} \mathbb{P}\Big[U_{\varepsilon,n} < \mathrm{E}(\nu) - u\Big]\,\mathrm{d}u \\ &= \int_0^{+\infty} \mathbb{P}\Big[\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu) - u\big) > \varepsilon\Big]\,\mathrm{d}u\,.\end{aligned}$$

Now, Corollary 13 (for the first inequality) and the deviation inequality of Proposition 14 (for the second inequality) indicate that for all $u > 0$,

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

Combining all elements, we get

$$\mathbb{E}\Big[\big(\mathrm{E}(\nu) - U_{\varepsilon,n}\big)^+\Big] \leqslant \mathrm{e}(2n+1)\,\mathrm{e}^{-n\varepsilon}\int_0^{+\infty} \mathrm{e}^{-2nu^2}\,\mathrm{d}u = \mathrm{e}(2n+1)\,\mathrm{e}^{-n\varepsilon}\;\frac{1}{2}\sqrt{\frac{\pi}{2n}}\,.$$

from which the stated bound follows, as $\mathrm{e}/\big(2\sqrt{2}\big) \leqslant 1$. ∎

4.4.3 Concentration Result on $\mathcal{K}_{\inf}$

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 B.

Proposition 16 (concentration result on $\mathcal{K}_{\inf}$) *With the same notation and assumptions as in the previous proposition, consider a real number $\mu \in \big(\mathrm{E}(\nu), 1\big)$ and define*

$$\gamma = \frac{1}{\sqrt{1-\mu}}\left(16\mathrm{e}^{-2} + \ln^2\left(\frac{1}{1-\mu}\right)\right). \tag{21}$$

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

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

5. Proofs of the Distribution-Free Bounds: Theorems 1 and 3

The two proofs are extremely similar; we prove Theorem 3 and then explain the adaptations to prove Theorem 1. The first steps of the proof(s) use the exact same arguments as in the proofs of the performance bounds of MOSS (Propositions 7 and 9, see Appendix A) in the exact same order. We explain below why we had to copy them and had to resort to the intermediary bounds for MOSS stated in the indicated propositions.

We recall that we denote by $a^\star$ an optimal arm, i.e., an arm such that $\mu_a = \mu^\star$. We first apply a trick introduced by Bubeck and Liu (2013): by definition of the index policy, for $t \geqslant K$,

$$U^{\mathrm{A}}_{a^\star}(t) \leqslant \max_{a=1,\dots,K} U^{\mathrm{A}}_a(t) = U^{\mathrm{A}}_{A_{t+1}}(t)\,,$$

so that the regret of KL-UCB-Switch is bounded by

$$R_T = \sum_{t=1}^{T} \mathbb{E}\big[\mu^\star - \mu_{A_t}\big] \leqslant (K-1) + \sum_{t=K+1}^{T} \mathbb{E}\big[\mu^\star - U^{\mathrm{A}}_{a^\star}(t-1)\big] + \sum_{t=K+1}^{T} \mathbb{E}\big[U^{\mathrm{A}}_{A_t}(t-1) - \mu_{A_t}\big]\,. \quad (22)$$

Part 1: We first deal with the second sum in (22) and successively use $x \leqslant \delta + (x-\delta)^+$ for all x and δ for the first inequality; the fact that $U^{\mathrm{A}}_a(t) \leqslant U^{\text{M-A}}_a(t) \leqslant U^{\mathrm{M},\varphi}_a(t)$ by (12) and (14), for the second inequality; and optional skipping (Section 4.1, Example 2) for the third inequality, keeping in mind that pairs (a,n) such $A_t = a$ and $N_a(t-1) = n$ correspond to at most one round $t \in \{K+1,\dots,T\}$:

$$\begin{aligned}
\sum_{t=K+1}^{T} \mathbb{E}\big[U^{\mathrm{A}}_{A_t}(t-1) - \mu_{A_t}\big] &\leqslant \sqrt{KT} + \sum_{t=K+1}^{T} \mathbb{E}\left[\left(U^{\mathrm{A}}_{A_t}(t-1) - \mu_{A_t} - \sqrt{\frac{K}{T}}\right)^+\right] \\
&\leqslant \sqrt{KT} + \sum_{t=K+1}^{T} \mathbb{E}\left[\left(U^{\mathrm{M},\varphi}_{A_t}(t-1) - \mu_{A_t} - \sqrt{\frac{K}{T}}\right)^+\right] \quad (23) \\
&\leqslant \sqrt{KT} + \sum_{a=1}^{K}\sum_{n=1}^{T} \mathbb{E}\left[\left(U^{\mathrm{M},\varphi}_{a,n} - \mu_a - \sqrt{\frac{K}{T}}\right)^+\right], \quad (24)
\end{aligned}$$

where we recall that

$$U^{\mathrm{M},\varphi}_{a,n} = \widehat{\mu}_{a,n} + \sqrt{\frac{1}{2n}\,\varphi\!\left(\frac{T}{Kn}\right)}\,.$$

We now apply one of the bounds of Proposition 9 to further bound the sum at hand by

$$\sum_{t=K+1}^{T} \mathbb{E}\big[U^{\mathrm{A}}_{A_t}(t-1) - \mu_{A_t}\big] \leqslant \sqrt{KT} + \sum_{a=1}^{K}\sum_{n=1}^{T} \mathbb{E}\left[\left(U^{\mathrm{M},\varphi}_{a,n} - \mu_a - \sqrt{\frac{K}{T}}\right)^+\right] \leqslant 7\sqrt{KT}\,.$$

Remark 17 *We may now explain why we copied the beginning of the proof of Proposition 9 and why we cannot just say that the ranking $U^{\mathrm{A}}_a(t) \leqslant U^{\text{M-A}}_a(t)$ entails that the regret of the anytime version of KL-UCB-Switch is bounded by the regret of the anytime version of MOSS. Indeed, it is difficult to relate*

$$\sum_{t=K+1}^{T} \mathbb{E}\big[U^{\text{M-A}}_{A_t}(t-1) - \mu_{A_t}\big] \qquad \textit{and} \qquad \sum_{t=K+1}^{T} \mathbb{E}\big[U^{\text{M-A}}_{A^{\text{M-A}}_t}(t-1) - \mu_{A^{\text{M-A}}_t}\big]$$

as the two series of arms A_t (picked by KL-UCB-Switch) and $A_t^{\text{M-A}}$ (picked by the adaptive version of MOSS) cannot be related. Hence, it is difficult to directly bound quantities like (23)*. However, the proof of the performance bound of MOSS relies on optional skipping and considers, in some sense, all possible values a for the arms picked: it controls the quantity* (24)*, which appears as a regret bound that is achieved by all index policies with indexes smaller than the ones of the anytime version of MOSS.*

Part 2: We now deal with the first sum in (22). We take positive parts, get back to the definition (13) of $U_{a^\star}^{\text{A}}(t-1)$, and add some extra non-negative terms:

$$\begin{aligned}
&\sum_{t=K+1}^{T} \mathbb{E}\big[\mu^\star - U_{a^\star}^{\text{A}}(t-1)\big] \leqslant \sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star}^{\text{A}}(t-1)\big)^+\Big] \\
&= \sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star}^{\text{KL-A}}(t-1)\big)^+ \mathbb{1}_{\{N_{a^\star}(t-1)\leqslant f(t-1,K)\}}\Big] \\
&\qquad + \sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star}^{\text{M-A}}(t-1)\big)^+ \underbrace{\mathbb{1}_{\{N_{a^\star}(t-1)> f(t-1,K)\}}}_{\leqslant 1}\Big] \\
&\leqslant \sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star}^{\text{KL-A}}(t-1)\big)^+ \mathbb{1}_{\{N_{a^\star}(t-1)\leqslant f(t-1,K)\}}\Big] + \sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star}^{\text{M-A}}(t-1)\big)^+\Big].
\end{aligned}$$

Now, the bound (18) of Proposition 9, together with the Remark 10, indicates that

$$\sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star}^{\text{M-A}}(t-1)\big)^+\Big] \leqslant 26\sqrt{KT}\,.$$

Note that Remark 10 exactly explains that for the sum above we do not bump into the issues raised in Remark 17 for the other sum in (22).

Part 3: Integrated deviations in terms of $\mathcal{K}_{\inf}$ divergence. We showed so far that the distribution-free regret bound of the anytime version of KL-UCB-Switch was given by the (intermediary) regret bound (18) of Proposition 9, which is smaller than $(K-1)+33\sqrt{KT}$, plus

$$\begin{aligned}
&\sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star}^{\text{KL-A}}(t-1)\big)^+ \mathbb{1}_{\{N_{a^\star}(t-1)\leqslant f(t-1,K)\}}\Big] \\
&\qquad = \sum_{t=K}^{T-1} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star}^{\text{KL-A}}(t)\big)^+ \mathbb{1}_{\{N_{a^\star}(t)\leqslant f(t,K)\}}\Big] \leqslant \sum_{t=K}^{T-1} \sum_{n=1}^{f(t,K)} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star,t,n}^{\text{KL-A}}\big)^+\Big], \quad (25)
\end{aligned}$$

where we applied optional skipping (Section 4.1, comments after Example 1) and where we denoted by

$$U_{a^\star,t,n}^{\text{KL-A}} = \sup\left\{\mu \in [0,1] \;\middle|\; \mathcal{K}_{\inf}\big(\widehat{\nu}_{a^\star,n},\mu\big) \leqslant \frac{1}{n}\,\varphi\Big(\frac{t}{Kn}\Big)\right\} \quad (26)$$

the counterpart of the quantity $U_{a^\star}^{\text{KL-A}}(t)$ defined in (10). Here, the additional subscript t in $U_{a^\star,t,n}^{\text{KL-A}}$ refers to the numerator of $t/(Kn)$ in the $\varphi(t/(Kn))$ term.

Now, Corollary 15 exactly indicates that for each given t and all $n \geqslant 1$,

$$\mathbb{E}\Big[\big(\mu^\star - U_{a^\star,t,n}^{\text{KL-A}}\big)^+\Big] \leqslant (2n+1)\sqrt{\frac{\pi}{n}}\,\exp\left(-\varphi\left(\frac{t}{Kn}\right)\right).$$

The t considered are such that $t \geqslant K$ and thus, $f(t,K) \leqslant (t/K)^{1/5} \leqslant t/K$. Therefore, the considered n are such that $1 \leqslant n \leqslant f(t,K)$ and thus, $t/(Kn) \geqslant 1$. Given that $\varphi \geqslant \ln_+$, we proved

$$\mathbb{E}\Big[\big(\mu^\star - U_{a^\star,t,n}^{\text{KL-A}}\big)^+\Big] \leqslant (2n+1)\sqrt{\frac{\pi}{n}}\,\frac{Kn}{t} = \frac{K\sqrt{\pi}}{t}\,(2n+1)\sqrt{n}\,.$$

We sum this bound over $n \in \big\{1,\dots,f(t/K)\big\}$, using again that $f(t,K) \leqslant (t/K)^{1/5}$:

$$\sum_{n=1}^{f(t,K)} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star,t,n}^{\text{KL-A}}\big)^+\Big] \leqslant \frac{K\sqrt{\pi}}{t}\sum_{n=1}^{f(t,K)} \underbrace{(2n+1)\sqrt{n}}_{\leqslant 3f(t,K)^{3/2}} \leqslant \frac{3K\sqrt{\pi}}{t}\,\underbrace{f(t,K)^{5/2}}_{\leqslant (t/K)^{1/2}} \leqslant 3\sqrt{\pi}\,\sqrt{\frac{K}{t}}\,.$$

We substitute this inequality into (25):

$$\begin{aligned}
&\sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star}^{\text{KL-A}}(t-1)\big)^+ \mathbb{1}_{\{N_{a^\star}(t-1)\leqslant f(t-1,K)\}}\Big] \\
&\qquad \leqslant \sum_{t=K}^{T-1}\sum_{n=1}^{f(t,K)} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star,t,n}^{\text{KL-A}}\big)^+\Big] \leqslant 3\sqrt{\pi}\quad \underbrace{\sum_{t=K}^{T-1}\sqrt{\frac{K}{t}}}_{\leqslant 2\sqrt{KT},\ \text{see (35)}} \quad \leqslant 6\sqrt{\pi}\,\sqrt{KT} \leqslant 11\sqrt{KT}\,.
\end{aligned}$$

The final regret bound is obtained as the sum of this $11\sqrt{KT}$ bound plus the $(K-1) + 33\sqrt{KT}$ bound obtained above. This concludes the proof of Theorem 3.

Part 4: Adaptations needed for Theorem 1, i.e., to analyze the version of KL-UCB-Switch relying on the knowledge of the horizon T. Parts 1 and 2 of the proof remain essentially unchanged, up to the (intermediary) regret bound to be applied now: (17) of Proposition 7, which is smaller than $(K-1) + 17\sqrt{KT}$. The additional regret bound, accounting, as we did in Part 3, for the use of KL-UCB-indexes for small T, is no larger than

$$\begin{aligned}
&\sum_{t=K}^{T-1}\sum_{n=1}^{f(T,K)} (2n+1)\sqrt{\frac{\pi}{n}}\,\exp\left(-\ln_+\left(\frac{T}{Kn}\right)\right) \\
&= \sum_{t=K}^{T-1}\sum_{n=1}^{f(T,K)} (2n+1)\sqrt{\frac{\pi}{n}}\frac{Kn}{T} = K\sqrt{\pi}\sum_{n=1}^{f(T,K)} \underbrace{(2n+1)\sqrt{n}}_{\leqslant 3f(T,K)^{3/2}} \\
&\leqslant 3\sqrt{\pi}\,K\,f(T,K)^{5/2} \leqslant 3\sqrt{\pi}\,K\,\sqrt{\frac{T}{K}} \leqslant 6\sqrt{KT}\,.
\end{aligned}$$

This yields the claimed $(K-1) + 23\sqrt{KT}$ bound.

6. Proofs of the Distribution-Dependent Bound of Theorem 4

The proof below can be adapted (simplified) to provide an elementary analysis of performance of the KL-UCB algorithm on the class of all distributions over a bounded interval, by keeping only its Parts 1 and 2. The study of KL-UCB in Cappé et al. (2013) remained somewhat intricate and limited to finitely supported distributions.

The proof starts as in Cappé et al. (2013). We fix a sub-optimal arm a. Given $\delta \in (0, \mu^\star)$ sufficiently small (to be determined by the analysis), we first decompose $\mathbb{E}\big[N_a(T)\big]$ as

$$\begin{aligned}\mathbb{E}\big[N_a(T)\big] &= 1 + \sum_{t=K}^{T-1} \mathbb{P}\big[A_{t+1} = a\big] \\ &= 1 + \sum_{t=K}^{T-1} \mathbb{P}\big[U_a^{\mathrm{A}}(t) < \mu^\star - \delta \ \text{ and } \ A_{t+1} = a\big] + \sum_{t=K}^{T-1} \mathbb{P}\big[U_a^{\mathrm{A}}(t) \geqslant \mu^\star - \delta \ \text{ and } \ A_{t+1} = a\big] .\end{aligned}$$

We then use that by definition of the index policy, $A_{t+1} = a$ only if $U_a^{\mathrm{A}}(t) \geqslant U_{a^\star}^{\mathrm{A}}(t)$, where we recall that $a^\star$ denotes an optimal arm (i.e., an arm such that $\mu_a = \mu^\star$). We also use $U_{a^\star}^{\mathrm{A}}(t) \geqslant U_{a^\star}^{\text{KL-A}}(t)$, which was stated in (14). We get

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

Finally, by the definition (13) of $U_a^{\mathrm{A}}(t)$, we proved so far

$$\begin{aligned}\mathbb{E}\big[N_a(T)\big] \leqslant 1 &+ \sum_{t=K}^{T-1} \mathbb{P}\big[U_{a^\star}^{\text{KL-A}}(t) < \mu^\star - \delta\big] \\ &+ \sum_{t=K}^{T-1} \mathbb{P}\big[U_a^{\text{KL-A}}(t) \geqslant \mu^\star - \delta \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) \leqslant f(t,K)\big] \\ &+ \sum_{t=K}^{T-1} \mathbb{P}\big[U_a^{\text{M-A}}(t) \geqslant \mu^\star - \delta \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) > f(t,K)\big] . \qquad (27)\end{aligned}$$

We now deal with each of the three sums above.

Part 1: We first deal with the first sum in (27) and to that end, fix some $t \in \{K, \dots, T-1\}$. By the definition (10) of $U_{a^\star}^{\text{KL-A}}(t)$ as a supremum,

$$\mathbb{P}\big[U_{a^\star}^{\text{KL-A}}(t) < \mu^\star - \delta\big] \leqslant \mathbb{P}\left[\mathcal{K}_{\inf}\big(\widehat{\nu}_{a^\star}(t), \mu^\star - \delta\big) > \frac{1}{N_{a^\star}(t)}\, \varphi\Big(\frac{t}{K N_{a^\star}(t)}\Big)\right] .$$

By a careful application of optional skipping (see Section 4.1, final part of Example 1),

$$\mathbb{P}\left[\mathcal{K}_{\inf}\big(\widehat{\nu}_{a^\star}(t), \mu^\star - \delta\big) > \frac{1}{N_{a^\star}(t)}\,\varphi\Big(\frac{t}{KN_{a^\star}(t)}\Big)\right] \\ \leqslant \mathbb{P}\left[\exists n \in \{1,\dots,t-K+1\} : \;\; \mathcal{K}_{\inf}\big(\widehat{\nu}_{a^\star,n}, \mu^\star - \delta\big) > \frac{1}{n}\,\varphi\Big(\frac{t}{Kn}\Big)\right].$$

Now, for $n \geqslant \lfloor t/K \rfloor + 1$ and given the definition (9) of φ, we have $\varphi\big(t/(Kn)\big) = 0$. By definition, $\mathcal{K}_{\inf}(\widehat{\nu}_{a^\star,n}, \mu^\star - \delta) > 0$ requires in particular that the expectation $\widehat{\mu}_{a^\star,n}$ of $\widehat{\nu}_{a^\star,n}$ be smaller than $\mu^\star - \delta$. This fact, together with a union bound, implies

$$\mathbb{P}\left[\exists n \in \{1,\dots,t-K+1\} : \;\; \mathcal{K}_{\inf}\big(\widehat{\nu}_{a^\star,n}, \mu^\star - \delta\big) > \frac{1}{n}\,\varphi\Big(\frac{t}{Kn}\Big)\right] \\ \leqslant \mathbb{P}\Big[\exists n \geqslant \lfloor t/K \rfloor + 1 : \;\; \widehat{\mu}_{a^\star,n} \leqslant \mu^\star - \delta\Big] + \sum_{n=1}^{\lfloor t/K \rfloor} \mathbb{P}\left[\mathcal{K}_{\inf}\big(\widehat{\nu}_{a^\star,n}, \mu^\star - \delta\big) > \frac{1}{n}\,\varphi\Big(\frac{t}{Kn}\Big)\right].$$

Hoeffding's maximal inequality (Proposition 5) upper bounds the first term by $\exp(-2\delta^2 t/K)$, while Corollary 13 and Proposition 14 provide the upper bound

$$\mathbb{P}\left[\mathcal{K}_{\inf}\big(\widehat{\nu}_{a^\star,n}, \mu^\star - \delta\big) > \frac{1}{n}\,\varphi\Big(\frac{t}{Kn}\Big)\right] \leqslant \mathrm{e}(2n+1)\,\exp\Big(-n\Big(2\delta^2 + \varphi\big(t/(Kn)\big)/n\Big)\Big).$$

Collecting all inequalities, we showed so far that

$$\mathbb{P}\big[U_{a^\star}^{\text{KL-A}}(t) < \mu^\star - \delta\big] \leqslant \exp(-2\delta^2 t/K) + \sum_{n=1}^{\lfloor t/K \rfloor} \mathrm{e}(2n+1)\,\exp\Big(-2n\delta^2 - \varphi\big(t/(Kn)\big)\Big).$$

Summing over $t \in \{K,\dots,T-1\}$, using the formula for geometric series, on the one hand, and performing some straightforward (and uninteresting) calculation detailed below in Lemma 18 on the other hand, we finally bound the first sum in (27) by

$$\sum_{t=K}^{T-1} \mathbb{P}\big[U_{a^\star}^{\text{KL-A}}(t) < \mu^\star - \delta\big] \\ \leqslant \sum_{t=K}^{T-1} \exp(-2\delta^2 t/K) + \sum_{t=K}^{T-1} \sum_{n=1}^{\lfloor t/K \rfloor} \mathrm{e}(2n+1)\,\exp\Big(-2n\delta^2 - \varphi\big(t/(Kn)\big)\Big) \\ \leqslant \frac{1}{1-\mathrm{e}^{-2\delta^2/K}} + \frac{\mathrm{e}(3+8K)}{(1-\mathrm{e}^{-2\delta^2})^3}\,.$$

This concludes the first part of this proof.

Part 2: We then deal with the second sum in (27). We introduce

$$\widetilde{U}_a^{\text{KL-A}}(t) \stackrel{\text{def}}{=} \sup\left\{\mu \in [0,1] \;\middle|\; \mathcal{K}_{\inf}\big(\widehat{\nu}_a(t), \mu\big) \leqslant \frac{1}{N_a(t)}\,\varphi\Big(\frac{T}{KN_a(t)}\Big)\right\},$$

which only differs from the original index $U_a^{\text{KL-A}}(t)$ defined in (10) by the replacement of $t/(Kn)$ by $T/(Kn)$ as the argument of φ. Therefore, we have $\widetilde{U}_a^{\text{KL-A}}(t) \geqslant U_a^{\text{KL-A}}(t)$. Replacing also $f(t,K)$ by the larger quantity $f(T,K)$, the second sum in (27) is therefore bounded by

$$\begin{aligned}
&\sum_{t=K}^{T-1} \mathbb{P}\big[U_a^{\text{KL-A}}(t) \geqslant \mu^\star - \delta \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) \leqslant f(t,K)\big] \qquad (28)\\
&\leqslant \sum_{t=K}^{T-1} \mathbb{P}\Big[\widetilde{U}_a^{\text{KL-A}}(t) \geqslant \mu^\star - \delta \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) \leqslant f(T,K)\Big]\\
&\leqslant \sum_{n=1}^{f(T,K)} \sum_{t=K}^{T-1} \mathbb{P}\Big[\widetilde{U}_a^{\text{KL-A}}(t) \geqslant \mu^\star - \delta \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) = n\Big]\,.
\end{aligned}$$

Optional skipping (see Section 4.1, Example 2) indicates that for each value of n,

$$\begin{aligned}
&\sum_{t=K}^{T-1} \mathbb{P}\Big[\widetilde{U}_a^{\text{KL-A}}(t) \geqslant \mu^\star - \delta \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) = n\Big]\\
&\qquad\qquad = \sum_{t=K}^{T-1} \mathbb{P}\Big[U_{a^\star,T,n}^{\text{KL-A}} \geqslant \mu^\star - \delta \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) = n\Big]\,,
\end{aligned}$$

where $U_{a^\star,T,n}^{\text{KL-A}}$ was defined in (26). We now note that the events $\big\{A_{t+1} = a \ \text{ and } \ N_a(t) = n\big\}$ are disjoint as t varies in $\{K,\dots,T-1\}$. Therefore,

$$\sum_{t=K}^{T-1} \mathbb{P}\Big[U_{a^\star,T,n}^{\text{KL-A}} \geqslant \mu^\star - \delta \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) = n\Big] \leqslant \mathbb{P}\Big[U_{a^\star,T,n}^{\text{KL-A}} \geqslant \mu^\star - \delta\Big]\,.$$

All in all, we proved so far that

$$\begin{aligned}
&\sum_{t=K}^{T-1} \mathbb{P}\big[U_a^{\text{KL-A}}(t) \geqslant \mu^\star - \delta \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) \leqslant f(t,K)\big]\\
&\qquad\qquad \leqslant \sum_{n=1}^{f(T,K)} \mathbb{P}\Big[U_{a^\star,T,n}^{\text{KL-A}} \geqslant \mu^\star - \delta\Big]\,. \qquad (29)
\end{aligned}$$

Now, note that the supremum in (26) is taken over a closed interval, as $\mathcal{K}_{\inf}$ is non-decreasing in its second argument (by its definition as an infimum) and as $\mathcal{K}_{\inf}$ is left-continuous (Corollary 12). This supremum is therefore a maximum. Hence, by distinguishing the cases where $U_{a^\star,T,n}^{\text{KL-A}} = \mu^\star - \delta$ and $U_{a^\star,T,n}^{\text{KL-A}} > \mu^\star - \delta$, we have the equality of events

$$\Big\{U_{a^\star,T,n}^{\text{KL-A}} \geqslant \mu^\star - \delta\Big\} = \left\{\mathcal{K}_{\inf}\big(\widehat{\nu}_{a,n}, \mu^\star - \delta\big) \leqslant \frac{1}{n}\varphi\Big(\frac{T}{Kn}\Big)\right\}.$$

We assume that $\delta \in (0, \mu^\star)$ is sufficiently small for

$$\delta < \frac{1-\mu^\star}{2}\,\mathcal{K}_{\inf}(\nu_a, \mu^\star)$$

to hold, and introduce

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

For $n \geqslant n_1$, by definition of n_1,

$$\frac{1}{n}\varphi\Big(\frac{T}{Kn}\Big) \leqslant \underbrace{\frac{\varphi\big(T/(Kn)\big)}{\varphi(T/K)}}_{\leqslant 1} \left(\mathcal{K}_{\inf}(\nu_a, \mu^\star) - \frac{2\delta}{1-\mu^\star}\right) \leqslant \mathcal{K}_{\inf}(\nu_a, \mu^\star) - \frac{2\delta}{1-\mu^\star} \,,$$

while by the regularity property (19), we have $\mathcal{K}_{\inf}\big(\widehat{\nu}_{a,n}, \mu^\star - \delta\big) \geqslant \mathcal{K}_{\inf}\big(\widehat{\nu}_{a,n}, \mu^\star\big) - \delta/(1-\mu^\star)$. We therefore proved that for $n \geqslant n_1$,

$$\begin{aligned}\mathbb{P}\Big[U^{\text{KL-A}}_{a^\star,T,n} \geqslant \mu^\star - \delta\Big] &= \mathbb{P}\left[\mathcal{K}_{\inf}\big(\widehat{\nu}_{a,n}, \mu^\star - \delta\big) \leqslant \frac{1}{n}\varphi\Big(\frac{T}{Kn}\Big)\right] \\ &\leqslant \mathbb{P}\left[\mathcal{K}_{\inf}\big(\widehat{\nu}_{a,n}, \mu^\star\big) \leqslant \mathcal{K}_{\inf}(\nu_a, \mu^\star) - \frac{\delta}{1-\mu^\star}\right].\end{aligned}$$

Therefore we may resort to the concentration inequality on $\mathcal{K}_{\inf}$ stated as Proposition 16. We set $x = \mathcal{K}_{\inf}(\nu_a, \mu^\star) - \delta/(1-\mu^\star)$ and simply sum the bounds obtained in the two regimes considered therein:

$$\mathbb{P}\left[\mathcal{K}_{\inf}\big(\widehat{\nu}_{a,n}, \mu^\star - \delta\big) \leqslant \mathcal{K}_{\inf}(\nu_a, \mu^\star) - \frac{\delta}{1-\mu^\star}\right] \leqslant \mathrm{e}^{-n/4} + \exp\Big(-\frac{n\delta^2}{2\gamma_\star(1-\mu^\star)^2}\Big) \,,$$

where $\gamma_\star$ was defined in (21). For $n \leqslant n_1 - 1$, we bound the probability at hand by 1. Combining all these arguments together yields

$$\begin{aligned}\sum_{n=1}^{f(T,K)} \mathbb{P}\Big[U^{\text{KL-A}}_{a^\star,T,n} \geqslant \mu^\star - \delta\Big] &\leqslant n_1 - 1 + \sum_{n=n_1}^{f(T,K)} \mathrm{e}^{-n/4} + \sum_{n=n_1}^{f(T,K)} \exp\Big(-\frac{n\delta^2}{2\gamma_\star(1-\mu^\star)^2}\Big) \\ &\leqslant \frac{\varphi(T/K)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star) - 2\delta/(1-\mu^\star)} + \underbrace{\frac{1}{1-\mathrm{e}^{-1/4}}}_{\leqslant 5} + \underbrace{\frac{1}{1-\mathrm{e}^{-\delta^2/(2\gamma_\star(1-\mu^\star)^2)}}}_{=\mathcal{O}(1/\delta^2)} \,,\end{aligned}$$

where the second inequality follows from the formula for geometric series and from the definition of n_1.

Part 3: We then deal with the third sum in (27). This sum involves the indexes $U_a^{\text{M-A}}(t)$ only when $N_a(t) > f(t,K)$, that is, when $N_a(t) \geqslant f(t,K)+1$, where $f(t,K) = \lfloor (t/K)^{1/5} \rfloor$. Under the latter condition, the indexes are actually bounded by

$$U_a^{\text{M-A}}(t) \stackrel{\text{def}}{=} \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)}\,\varphi\Big(\frac{t}{KN_a(t)}\Big)} \leqslant \widehat{\mu}_a(t) + \underbrace{\sqrt{\frac{1}{2(t/K)^{1/5}}\,\varphi\big((t/K)^{4/5}\big)}}_{\to 0 \text{ as } t\to\infty} \,.$$

We denote by $T_0(\Delta_a, K)$ the smallest time T_0 such that for all $t \geqslant T_0$,

$$\sqrt{\frac{1}{2(t/K)^{1/5}}\,\varphi\big((t/K)^{4/5}\big)} \leqslant \frac{\Delta_a}{4}\,. \tag{30}$$

This time T_0 only depends on K and Δ_a; a closed-form upper bound on its value could be easily provided. With this definition, we already have that the sum of interest may be bounded by

$$\begin{aligned}
&\sum_{t=K}^{T-1} \mathbb{P}\big[U_a^{\text{M-A}}(t) \geqslant \mu^\star - \delta \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) > f(t,K)\big] \\
&\leqslant T_0(\Delta_a, K) + \sum_{t=T_0(\Delta_a,K)}^{T-1} \mathbb{P}\Big[\widehat{\mu}_a(t) + \Delta_a/4 \geqslant \mu^\star - \delta \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) > f(t,K)\Big] \\
&\leqslant T_0(\Delta_a, K) + \sum_{t=T_0(\Delta_a,K)}^{T-1} \mathbb{P}\Big[\widehat{\mu}_a(t) \geqslant \mu_a + \Delta_a/2 \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) > f(t,K)\Big]\,,
\end{aligned}$$

where for the second inequality, we assumed that $\delta \in (0, \mu^\star)$ is sufficiently small for

$$\delta < \frac{\Delta_a}{4}$$

to hold. Optional skipping using that the events $\big\{A_{t+1} = a \ \text{ and } \ N_a(t) = n\big\}$ are disjoint as t varies—see Section 4.1, Example 2 and see the treatment performed between (28) and (29)—provides the upper bound

$$\begin{aligned}
&\sum_{t=T_0(\Delta_a,K)}^{T-1} \mathbb{P}\Big[\widehat{\mu}_a(t) \geqslant \mu_a + \Delta_a/2 \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) > f(t,K)\Big] \\
&\qquad \leqslant \sum_{n \geqslant 1} \mathbb{P}\big[\widehat{\mu}_{a,n} \geqslant \mu_a + \Delta_a/2\big] \leqslant \sum_{n \geqslant 1} \mathrm{e}^{-n\Delta_a^2/2} = \frac{1}{1 - \mathrm{e}^{-\Delta_a^2/2}}\,,
\end{aligned}$$

where the second inequality is due to Hoeffding's inequality (in its non-maximal version, see Proposition 5). A summary of the bound thus provided in this part is:

$$\begin{aligned}
&\sum_{t=K}^{T-1} \mathbb{P}\big[U_a^{\text{M-A}}(t) \geqslant \mu^\star - \delta \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) > f(t,K)\big] \\
&\qquad\qquad \leqslant T_0(\Delta_a, K) + \frac{1}{1 - \mathrm{e}^{-\Delta_a^2/2}} = \mathcal{O}(1)\,,
\end{aligned}$$

where $T_0(\Delta_a, K)$ was defined in (30).

Part 4: Conclusion of the proof of Theorem 4. Collecting all previous bounds and conditions, we proved that when $\delta \in (0, \mu^\star)$ is sufficiently small for

$$\delta < \min\left\{\frac{1-\mu^\star}{2}\,\mathcal{K}_{\inf}(\nu_a, \mu^\star),\ \frac{\Delta_a}{4}\right\} \tag{31}$$

to hold, then

$$\mathbb{E}\big[N_a(T)\big] \leqslant \frac{\varphi(T/K)}{\mathcal{K}_{\inf}(\nu_a,\mu^\star) - 2\delta/(1-\mu^\star)} + \overbrace{\frac{\mathrm{e}(3+8K)}{(1-\mathrm{e}^{-2\delta^2})^3}}^{=\mathcal{O}(1/\delta^6)} + \underbrace{\frac{1}{1-\mathrm{e}^{-2\delta^2/K}} + \frac{1}{1-\mathrm{e}^{-\delta^2/(2\gamma_\star(1-\mu^\star)^2)}}}_{=\mathcal{O}(1/\delta^2)} + \underbrace{T_0(\Delta_a,K) + \frac{1}{1-\mathrm{e}^{-\Delta_a^2/2}} + 6}_{=\mathcal{O}(1)}, \quad (32)$$

where

$$\frac{\varphi(T/K)}{\mathcal{K}_{\inf}(\nu_a,\mu^\star) - 2\delta/(1-\mu^\star)} = \frac{\ln T + \ln\ln T + \mathcal{O}(1)}{\mathcal{K}_{\inf}(\nu_a,\mu^\star) - 2\delta/(1-\mu^\star)} = \frac{\ln T + \ln\ln T}{\mathcal{K}_{\inf}(\nu_a,\mu^\star)} + \mathcal{O}(\delta \ln T)\,.$$

The leading term in this regret bound is $\ln T/\mathcal{K}_{\inf}(\nu_a,\mu^\star)$, while the order of magnitude of the smaller-order terms is given by

$$\delta \ln T + \frac{1}{\delta^6} = \mathcal{O}\big((\ln T)^{6/7}\big)$$

for δ of the order of $(\ln T)^{-1/7}$. When T is sufficiently large, this value of δ is smaller than the required threshold (31).

It only remains to state and prove Lemma 18 (used at the very end of the first part of the proof above).

Lemma 18 *We have the bound*

$$\sum_{t=K}^{T-1} \sum_{n=1}^{\lfloor t/K \rfloor} \mathrm{e}(2n+1)\, \exp\Big(-2n\delta^2 - \varphi\big(t/(Kn)\big)\Big) \leqslant \frac{\mathrm{e}(3+8K)}{(1-\mathrm{e}^{-2\delta^2})^3}\,.$$

Proof The double sum can be rewritten, by permuting the order of summations, as

$$\begin{aligned}
&\sum_{t=K}^{T-1} \sum_{n=1}^{\lfloor t/K \rfloor} \mathrm{e}(2n+1)\, \exp\Big(-2n\delta^2 - \varphi\big(t/(Kn)\big)\Big) \\
&= \sum_{n=1}^{\lfloor T/K \rfloor} \sum_{t=Kn}^{T-1} \mathrm{e}(2n+1)\, \exp\Big(-2n\delta^2 - \varphi\big(t/(Kn)\big)\Big) \\
&= \sum_{n=1}^{\lfloor T/K \rfloor} \mathrm{e}(2n+1)\, \exp\big(-2n\delta^2\big) \sum_{t=Kn}^{T-1} \exp\Big(-\varphi\big(t/(Kn)\big)\Big)\,.
\end{aligned}$$

We first fix $n \geqslant 1$ and use that $t \mapsto \exp\big(-\varphi(t/(Kn)\big)$ is non-increasing to get

$$\begin{aligned}
\sum_{t=Kn}^{T-1} \exp\Big(-\varphi\big(t/(Kn)\big)\Big) &\leqslant 1 + \int_{Kn}^{T-1} \exp\Big(-\varphi\big(t/(Kn)\big)\Big)\,\mathrm{d}t \\
&= 1 + Kn \int_1^{(T-1)/(Kn)} \exp\big(-\varphi(u)\big)\,\mathrm{d}u\,,
\end{aligned}$$

where we operated the change of variable $u = t/(Kn)$. Now, by the change of variable $v = \ln(u)$,

$$\int_1^{(T-1)/(Kn)} \exp\bigl(-\varphi(u)\bigr)\,\mathrm{d}u \leqslant \int_1^{+\infty} \exp\bigl(-\varphi(u)\bigr)\,\mathrm{d}u = \int_1^{+\infty} \frac{1}{u\bigl(1+\ln^2(u)\bigr)}\,\mathrm{d}u$$
$$= \int_0^{+\infty} \frac{1}{1+v^2}\,\mathrm{d}v = \bigl[\arctan\bigr]_0^{+\infty} = \frac{\pi}{2}\,.$$

All in all, we proved so far that

$$\sum_{t=K}^{T-1} \sum_{n=1}^{\lfloor t/K \rfloor} \mathrm{e}(2n+1)\,\exp\Bigl(-2n\delta^2 - \varphi\bigl(t/(Kn)\bigr)\Bigr) \leqslant \sum_{n=1}^{\lfloor T/K \rfloor} \mathrm{e}(2n+1)\bigl(1+Kn\pi/2\bigr)\,\exp\bigl(-2n\delta^2\bigr)$$
$$\leqslant \sum_{n=1}^{+\infty} \mathrm{e}\bigl(1+(2+K\pi/2)n + K\pi n^2\bigr)\,\exp\bigl(-2n\delta^2\bigr)\,.$$

To conclude our calculation, we use that by differentiation of series, for all $\theta > 0$,

$$\sum_{m=0}^{+\infty} \mathrm{e}^{-m\theta} = \frac{1}{1-\mathrm{e}^{-\theta}}\,,$$
$$-\sum_{m=1}^{+\infty} m\,\mathrm{e}^{-m\theta} = \frac{-\mathrm{e}^{-\theta}}{(1-\mathrm{e}^{-\theta})^2} \qquad \text{thus} \qquad \sum_{m=1}^{+\infty} m\,\mathrm{e}^{-m\theta} \leqslant \frac{1}{(1-\mathrm{e}^{-\theta})^2}\,, \tag{33}$$
$$\sum_{m=1}^{+\infty} m^2\,\mathrm{e}^{-m\theta} = \frac{\mathrm{e}^{-\theta}(1+\mathrm{e}^{-\theta})}{(1-\mathrm{e}^{-\theta})^3} \leqslant \frac{2}{(1-\mathrm{e}^{-\theta})^3}\,. \tag{34}$$

Hence, taking $\theta = 2\delta^2$,

$$\sum_{n=1}^{+\infty} \mathrm{e}\bigl(1+(2+K\pi/2)n + K\pi n^2\bigr)\,\exp\bigl(-2n\delta^2\bigr)$$
$$\leqslant \frac{\mathrm{e}}{1-\mathrm{e}^{-2\delta^2}} + \frac{\mathrm{e}(2+K\pi/2)}{(1-\mathrm{e}^{-2\delta^2})^2} + \frac{2\mathrm{e}K\pi}{(1-\mathrm{e}^{-2\delta^2})^3} \leqslant \frac{\mathrm{e}(3+8K)}{(1-\mathrm{e}^{-2\delta^2})^3}\,,$$

which concludes the proof of this lemma. ■

7. Reflections on the Algorithm and on its Analysis

We gather here two series of reflections on the algorithm and on its analysis: first, we discuss the desirable values of switching thresholds $f(t,K)$. Second, we explain why we introduced, in the first place, such switches for the indices.

7.1 On the (Lack of) Impact of the Switching Thresholds $f(t, K)$

First of all, note that the inequalities between the various indices stated in (8) and (14), namely, $U_a^{\mathrm{KL}}(t) \leqslant U_a(t) \leqslant U_a^{\mathrm{M}}(t)$ and $U_a^{\text{KL-A}}(t) \leqslant U_a^{\mathrm{A}}(t) \leqslant U_a^{\text{M-A}}(t)$, hold regardless of the values of the switching thresholds. A large portions of the proofs rely solely on these inequalities: Parts 1, 2, and the first half of Part 3 of Theorems 1 and 3 (in Section 5), and Parts 1, 2, and 4 of the proof of Theorem 4 (in Section 6). That being said, the switching threshold affects the results in two ways.

Concerning the distribution-dependent bounds. The impact comes in lower-order terms. The specific value of the switching threshold plays a role in Part 3 of the proof of Theorem 4 (in Section 7), in the definition of $T_0(\Delta_a, K)$; see (30). This term $T_0(\Delta_a, K)$ then comes as an additive $\mathcal{O}_T(1)$ term in the final bound on $\mathbb{E}[N_a(T)]$ for any reasonable choice of $f(t, K)$, and thus leaves the asymptotic statement unaffected.

More precisely, as long as $\varphi\big(t/(Kf(t,K))\big)/f(t,K) \to 0$ as $t \to \infty$, the time $T_0(\Delta_a, K)$ exists (takes a finite value); we may then follow the proof exactly as it is written. For example, if $\varphi = \ln$, then any positive power $(t/K)^\alpha$ with $\alpha \in (0,1)$ is suitable; this yields a value of $T_0(\Delta_a, K)$ of $K\Delta_a^{-2/\alpha}$ up to logarithmic factors in Δ_a and K. Note that the larger α, the lower $T_0(\Delta_a, K)$.

Concerning the distribution-free bounds. The value of the switching threshold affects Part 3 (and its non-anytime counterpart Part 4) in Section 5, in the expectations of the left-deviations of the index of the optimal arm when it is selected less than $f(t, K)$ times. The final regret bound actually consists of some $\sqrt{KT}$ term plus a term of order $K\, f(T,K)^{5/2}$. Values $f(t, K)$ of order $(t/K)^\alpha$ with $\alpha \in (0, 1/5]$ thus lead to a distribution-free bound of order $\sqrt{KT}$, as desired. We took the limit value $\alpha = 1/5$ in our analysis, but this is an arbitrary choice. Note that the larger α, the larger the distribution-free bound obtained.

7.2 Why Consider a Switch-Based Algorithm?

In the parametric case of one-dimensional exponential families, Ménard and Garivier (2017) could exhibit a bi-optimal strategy called kl-UCB++, a version of KL-UCB tailored to these exponential families. They provide a distribution-free analysis based on a deviation inequality of the form

$$\mathbb{P}\Big[\max_{n \geqslant N} \mathrm{kl}\big(\widehat{\mu}_n, \mu\big) \geqslant u\Big] \leqslant C\, \mathrm{e}^{-Nu}\,,$$

for some numerical constant C, where $\widehat{\mu}_n$ denotes the empirical mean of an n–sample whose distribution has expectation μ. This analysis mimics the distribution-free analysis of MOSS and in particular, the part thereof based on the peeling trick—see (38)–(40) in Section A. The fact that the deviation upper bound is of the order of e^{-Nu} and not of the form $N\,\mathrm{e}^{-Nu}$ is crucial to that end.

However, for KL-UCB in the non-parametric case of all distributions over $[0,1]$, the deviation result of Proposition 14 states

$$\mathbb{P}\Big[\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu)\big) \geqslant u\Big] \leqslant \mathrm{e}(2n+1)\,\mathrm{e}^{-nu}\,, \qquad \text{and not} \qquad \cancel{\mathbb{P}\Big[\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu)\big) \geqslant u\Big] \leqslant C\,\mathrm{e}^{-nu}}$$

for some numerical constant C. Intuitively, the extra polynomial term in n is the price for adaptivity (to the distribution) in the non-parametric setting. We do not know how to prove a refined inequality with an upper bound of the order of e^{-nu}, with no additional factor of the order of n. Actually, we are uncertain that this is possible: had the set $\big\{\nu' : \mathcal{K}_{\inf}\big(\nu', \mathrm{E}(\nu)\big) \geqslant u\big\}$ been convex, Sanov's bound

$$n^{-1} \ln \mathbb{P}\Big[\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu)\big) \geqslant u\Big] \to -u$$

could have been translated into a non-asymptotic inequality (see Csiszar, 1984). Unfortunately, this set is the complement of a convex set, for which we found no sufficiently good non-asymptotic inequality.

This difficulty is exactly the reason why we introduced a regime switch in the algorithm proposed in the present article. This switch is rather intuitive: the distribution-dependent lower bound (2) features the distributions of sub-optimal arms while for optimal arms only the expectation $\mu^\star$ matters. Therefore, it is not surprising that the indices of the optimal arms should be of a different nature than the indices of the suboptimal arms—namely, the "expensive" KL-UCB indices (that adapt to the whole distribution) are used for sub-optimal arms (arms not played often) while using the "cheaper" MOSS-indices (mean-based) are used for the near-optimal arms (arms played often). This is exactly what KL-UCB-Switch does, as sketched in the discussion after Equation (6).

Acknowledgments

This work was 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). Aurélien Garivier also acknowledges the support of the Project IDEXLYON of the University of Lyon, in the framework of the Programme Investissements d'Avenir (ANR-16-IDEX-0005), and of Chaire SeqALO (ANR-20-CHIA-0020-01).

A. A Simplified Proof of the Regret Bounds for MOSS(-Anytime)

This section provides the proofs of Propositions 7 and 9. To emphasize the similarity of the analyses in the anytime and non-anytime cases, 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 (5) and (11), which we both state with a generic exploration function φ. Indeed, compare

$$U_a^{\mathrm{M}}(t) = \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)}\,\varphi\Big(\frac{T}{KN_a(t)}\Big)} \quad \text{and} \quad U_a^{\text{M-A}}(t) = \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)}\,\varphi\Big(\frac{t}{KN_a(t)}\Big)}\,.$$

We will denote by

$$U_{a,\tau}^{\mathrm{GM}}(t) = \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)}\,\varphi\Big(\frac{\tau}{KN_a(t)}\Big)}$$

the index of the generic MOSS strategy (superscript GM), so that $U_a^{\mathrm{M}}(t) = U_{a,T}^{\mathrm{GM}}(t)$ and $U_a^{\text{M-A}}(t) = U_{a,t}^{\mathrm{GM}}(t)$. This GM strategy considers a sequence $(\tau_K, \dots, \tau_{T-1})$ of integers, either $\tau_t \equiv T$ for MOSS or $\tau_t = t$ for MOSS-anytime, and picks at each step $t+1$ with $t \geqslant K$, an arm A_{t+1}^{GM} with maximal index $U_{a,\tau_t}^{\mathrm{GM}}(t)$. For a given t, we denote by $U_{a,\tau_t,n}^{\mathrm{GM}}$ the quantities corresponding to $U_{a,\tau_t}^{\mathrm{GM}}(t)$ by optional skipping (see Section 4.1).

We provide below an analysis for increasing exploration functions $\varphi : (0, +\infty) \to [0, +\infty)$ such that φ vanishes on $(0, 1]$ and $\varphi \geqslant \ln_+$, properties that are all satisfied for the two exploration functions stated in Proposition 9. The general result is stated as the next proposition.

Proposition 19 *For all bandit problems $\underline{\nu}$ over $[0, 1]$, for all $T \geqslant 1$ and all sequences $(\tau_K, \dots, \tau_{T-1})$ bounded by T, the regret of the generic MOSS strategy described above, with an increasing exploration function $\varphi \geqslant \ln_+$ vanishing on $(0, 1]$, is smaller than*

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

where

$$U_{a,T,n}^{\mathrm{GM}} = \widehat{\mu}_{a,n} + \sqrt{\frac{1}{2n}\,\varphi\Big(\frac{T}{Kn}\Big)}.$$

In addition,

$$\sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star,\tau_{t-1}}^{\mathrm{GM}}(t-1)\big)^+\Big] \leqslant \underbrace{20\sqrt{\frac{\pi}{8}}}_{\leqslant 12.6} \sum_{t=K}^{T-1} \sqrt{\frac{K}{\tau_t}}$$

and

$$\sqrt{KT} + \sum_{a=1}^{K}\sum_{n=1}^{T} \mathbb{E}\Big[\big(U_{a,T,n}^{\mathrm{GM}} - \mu_a - \sqrt{K/T}\big)^+\Big] \leqslant \sqrt{KT}\left(1 + \frac{\pi}{4} + \frac{1}{\sqrt{2}}\int_1^{+\infty} u^{-3/2}\sqrt{\varphi(u)}\,\mathrm{d}u\right).$$

The bounds of Propositions 7 and 9, including the intermediary bounds (17) and (18), follow from this general result, up to the following straightforward calculation. On the one hand, in the known horizon case $\sum 1/\sqrt{\tau_t} \leqslant T/\sqrt{T} = \sqrt{T}$, whereas in the anytime case,

$$\sum_{t=K}^{T-1} 1/\sqrt{\tau_t} = \sum_{t=K}^{T-1} 1/\sqrt{t} \leqslant \int_0^T \frac{1}{\sqrt{u}}\,\mathrm{d}u = 2\sqrt{T}. \tag{35}$$

On the other hand, by the change of variable $u = \mathrm{e}^{v^2}$,

$$\int_1^{+\infty} u^{-3/2}\sqrt{\ln(u)}\,\mathrm{d}u = 2\int_0^{+\infty} v^2\,\mathrm{e}^{-v^2/2}\,\mathrm{d}v = \sqrt{2\pi}$$

and, using well-known inequalities like $\sqrt{x+x'} \leqslant \sqrt{x} + \sqrt{x'}$ and $\ln(1+x) \leqslant x$ for $x, x' \geqslant 0$,

$$\begin{aligned}\int_1^{+\infty} \sqrt{u^{-3}\ln\bigl(u(1+\ln^2(u)\bigr)}\,\mathrm{d}u &\leqslant \int_1^{+\infty} \sqrt{u^{-3}\ln(u)}\,\mathrm{d}u + \int_1^{+\infty} \sqrt{u^{-3}\ln\bigl(1+\ln^2(u)\bigr)}\,\mathrm{d}u \\ &\leqslant \int_1^{+\infty} \sqrt{u^{-3}\ln(u)}\,\mathrm{d}u + \int_1^{+\infty} \sqrt{u^{-3}\ln^2(u)}\,\mathrm{d}u \\ &= 2\int_0^{+\infty} v^2\,\mathrm{e}^{-v^2/2}\,\mathrm{d}v + 2\int_0^{+\infty} v^3\,\mathrm{e}^{-v^2/2}\,\mathrm{d}v = \sqrt{2\pi} + 4\,.\end{aligned}$$

The constant 17 of Proposition 7 (where $\tau_t \equiv T$ and $\varphi = \ln_+$) is obtained as an upper bound on the sum of $12.6 \leqslant 13$ and $1 + \pi/4 + \sqrt{\pi} \leqslant 3.6 \leqslant 4$. The constants 30 and 33 of Proposition 9 correspond to the cases where $\varphi = \ln_+$ and $\varphi : x \mapsto \ln_+\bigl(x(1+\ln_+^2 x)\bigr)$, respectively, together with $\tau_t = t$; they are obtained as upper bounds on the sum of $2\times 12.6 \leqslant 26$ and $1+\pi/4+\sqrt{\pi} \leqslant 4$, and on the sum of $2\times 12.6 \leqslant 26$ and $1+\pi/4+\sqrt{\pi}+4/\sqrt{2} \leqslant 6.4 \leqslant 7$, respectively.

Proof The beginning of this proof is completely similar to the beginning of the proof provided in Section 5.

The first step is standard, see Bubeck and Liu (2013). By definition of the index policy, for $t \geqslant K$,

$$U^{\mathrm{GM}}_{a^\star,\tau_t}(t) \leqslant \max_{a=1,\dots,K} U^{\mathrm{GM}}_{a,\tau_t}(t) = U^{\mathrm{GM}}_{A^{\mathrm{GM}}_{t+1},\tau_t}(t)\,,$$

so that the regret of the strategy is smaller than

$$\begin{aligned}R_T &= \sum_{t=1}^{T} \mathbb{E}\bigl[\mu^\star - \mu_{A^{\mathrm{GM}}_t}\bigr] \\ &\leqslant (K-1) + \sum_{t=K+1}^{T} \mathbb{E}\bigl[\mu^\star - U^{\mathrm{GM}}_{a^\star,\tau_{t-1}}(t-1)\bigr] + \sum_{t=K+1}^{T} \mathbb{E}\Bigl[U^{\mathrm{GM}}_{A^{\mathrm{GM}}_t,\tau_{t-1}}(t-1) - \mu_{A^{\mathrm{GM}}_t}\Bigr]\,. \quad (36)\end{aligned}$$

The term $K-1$ above accounts for the initial K rounds, when each arm is played once.

A preliminary transformation of the right-hand side of (36). We successively use the fact that the index $U^{\mathrm{GM}}_{a,\tau}(t-1)$ increases with τ since φ is increasing (for the first inequality below), $x \leqslant \delta + (x-\delta)^+$ for all x and δ (for the second inequality), and optional skipping (Section 4.1, Example 2, for the third inequality), keeping in mind that pairs (a,n) such $A^{\mathrm{GM}}_t = a$ and $N_a(t-1) = n$ correspond to at most one round $t \in \{K+1,\dots,T\}$:

$$\begin{aligned}\sum_{t=K+1}^{T} \mathbb{E}\Bigl[U^{\mathrm{GM}}_{A^{\mathrm{GM}}_t,\tau_{t-1}}(t-1) - \mu_{A^{\mathrm{GM}}_t}\Bigr] &\leqslant \sum_{t=K+1}^{T} \mathbb{E}\Bigl[U^{\mathrm{GM}}_{A^{\mathrm{GM}}_t,T}(t-1) - \mu_{A^{\mathrm{GM}}_t}\Bigr] \\ &\leqslant \sqrt{KT} + \sum_{t=K+1}^{T} \mathbb{E}\left[\left(U^{\mathrm{GM}}_{A^{\mathrm{GM}}_t,T}(t-1) - \mu_{A^{\mathrm{GM}}_t} - \sqrt{\frac{K}{T}}\right)^+\right] \\ &\leqslant \sqrt{KT} + \sum_{a=1}^{K}\sum_{n=1}^{T} \mathbb{E}\left[\left(U^{\mathrm{GM}}_{a,T,n} - \mu_a - \sqrt{\frac{K}{T}}\right)^+\right].\end{aligned}$$

While the last two inequalities may seem very crude, it turns out they are sharp enough to obtain the claimed distribution-free bounds. Moreover, they get rid of the bothersome dependencies among the arms that are contained in the choice of the arms A_t^{GM}. Therefore, we have shown that the right-hand side of (36) is bounded by

$$\begin{aligned}
&(K-1)+\sum_{t=K+1}^{T}\mathbb{E}\big[\mu^\star-U^{\mathrm{GM}}_{a^\star,\tau_{t-1}}(t-1)\big]+\sum_{t=K+1}^{T}\mathbb{E}\Big[U^{\mathrm{GM}}_{A_t^{\mathrm{GM}},\tau_{t-1}}(t-1)-\mu_{A_t^{\mathrm{GM}}}\Big]\\
\leqslant&(K-1)+\sum_{t=K+1}^{T}\mathbb{E}\Big[\big(\mu^\star-U^{\mathrm{GM}}_{a^\star,\tau_{t-1}}(t-1)\big)^+\Big]\\
&\qquad+\sqrt{KT}+\sum_{a=1}^{K}\sum_{n=1}^{T}\mathbb{E}\Big[\big(U^{\mathrm{GM}}_{a,T,n}-\mu_a-\sqrt{K/T}\big)^+\Big]\,. \qquad (37)
\end{aligned}$$

This inequality actually holds for all choices of sequences $(\tau_t)_{K\leqslant t\leqslant T-1}$ with $\tau_t\leqslant T$. The first sum in the right-hand side of (37) depends on the specific value of $(\tau_t)_{K\leqslant t\leqslant T-1}$, and thus, on the specific MOSS algorithm considered, but the second sum only depends on T.

This proves the first part of Proposition 19. We now bound each of the two sums in (36) and (37).

Control of the left deviations of the best arm, that is, of the first sum in (36) and (37). For each given round $t\in\{K,\dots,T-1\}$, we decompose

$$\begin{aligned}
&\mathbb{E}\Big[\big(\mu^\star-U^{\mathrm{GM}}_{a^\star,\tau_t}(t)\big)^+\Big]\\
&\qquad=\mathbb{E}\Big[\big(\mu^\star-U^{\mathrm{GM}}_{a^\star,\tau_t}(t)\big)^+\mathbb{1}_{\{N_{a^\star}(t)<\tau_t/K\}}\Big]+\mathbb{E}\Big[\big(\mu^\star-U^{\mathrm{GM}}_{a^\star,\tau_t}(t)\big)^+\mathbb{1}_{\{N_{a^\star}(t)\geqslant\tau_t/K\}}\Big]\,.
\end{aligned}$$

The two pieces are handled differently. The second one is dealt with by using $U^{\mathrm{GM}}_{a^\star,\tau_t}(t)\geqslant\widehat{\mu}_{a^\star}(t)$, which actually holds with equality given $N_{a^\star}(t)\geqslant\tau_t/K$, and by optional skipping (Section 4.1, comments after Example 1) and by the integrated version of Hoeffding's inequality (Corollary 6):

$$\begin{aligned}
\mathbb{E}\Big[\big(\mu^\star-U^{\mathrm{GM}}_{a^\star,\tau_t}(t)\big)^+\mathbb{1}_{\{N_{a^\star}(t)\geqslant\tau_t/K\}}\Big]&\leqslant\mathbb{E}\Big[\big(\mu^\star-\widehat{\mu}_{a^\star}(t)\big)^+\mathbb{1}_{\{N_{a^\star}(t)\geqslant\tau_t/K\}}\Big]\\
&=\sum_{n=\lceil\tau_t/K\rceil}^{T}\mathbb{E}\Big[\big(\mu^\star-\widehat{\mu}_{a^\star,n}\big)^+\mathbb{1}_{\{N_{a^\star}(t)=n\}}\Big]\\
&\leqslant\mathbb{E}\Big[\max_{n\geqslant\tau_t/K}\big(\mu^\star-\widehat{\mu}_{a^\star,n}\big)^+\Big]\leqslant\sqrt{\frac{\pi}{8}}\sqrt{\frac{K}{\tau_t}}\,. \qquad (38)
\end{aligned}$$

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 $x_\ell=\beta^{-\ell}\,\tau_t/K$, 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):

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

where in the second inequality, we applied optional skipping (Section 4.1, comments after Example 1) once again, as to get (38). Now for any ℓ, the summand can be controlled as follows, first, by $\varphi \geqslant \ln_+ = \ln$ on $[1,+\infty)$, second, by using $n < x_\ell$ and third, by Corollary 6:

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

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

$$\sum_{\ell=0}^{+\infty}\left(\frac{3}{2}\right)^{1/2+\ell(1/2-2/3)} = \sqrt{\frac{3}{2}}\sum_{\ell=0}^{+\infty}\alpha^\ell = \frac{1}{1-\alpha}\sqrt{\frac{3}{2}} \leqslant 19\,,$$

$$\text{where} \qquad \alpha = \left(\frac{3}{2}\right)^{(1/2-2/3)} \in (0,1)\,.$$

Therefore, we have shown that

$$\mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{GM}}_{a^\star,\tau_t}(t)\big)^+ \mathbb{1}_{\{N_{a^\star}(t)<\tau_t/K\}}\Big] \leqslant 19\sqrt{\frac{\pi}{8}}\sqrt{\frac{K}{\tau_t}}\,. \tag{39}$$

Combining this bound with (38) and summing over t, we proved that the first sum in (37) is bounded as

$$\sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{GM}}_{a^\star,\tau_{t-1}}(t-1)\big)^+\Big] \leqslant 20\sqrt{\frac{\pi}{8}}\sum_{t=K}^{T-1}\sqrt{\frac{K}{\tau_t}}\,. \tag{40}$$

Remark 20 *The proof technique reveals that the bound* (40) *obtained in this step of the proof actually holds even if the arms are pulled according to a strategy that is not a generic MOSS strategy. This is because we never used which specific arms A_t^{GM} were pulled: we only distinguished according to how many times $a^\star$ was pulled and resorted to optional skipping.*

Control of the right deviations of all arms, that is, of the second sum in (36) and (37). We use $(x+y)^+ \leqslant x^+ + y^+$ for all real numbers x, y, and the fact that φ vanishes on $(0,1]$ to get, for all a and $n \geqslant 1$,

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

Therefore, for each arm a,

$$\begin{aligned}&\sum_{n=1}^{T} \mathbb{E}\Big[\big(U_{a,T,n}^{\mathrm{GM}} - \mu_a - \sqrt{K/T}\big)^+\Big] \\ &\qquad\leqslant \sum_{n=1}^{T} \mathbb{E}\Big[\big(\widehat{\mu}_{a,n} - \mu_a - \sqrt{K/T}\big)^+\Big] + \sum_{n=1}^{\lfloor T/K\rfloor} \sqrt{\frac{1}{2n}\,\varphi\Big(\frac{T}{Kn}\Big)}\,. \qquad (41)\end{aligned}$$

We are left with two pieces to deal with separately. For the first sum in (41), we exploit the integrated version of Hoeffding's inequality (Corollary 6),

$$\begin{aligned}\sum_{n=1}^{T} \mathbb{E}\Big[\big(\widehat{\mu}_{a,n} - \mu_a - \sqrt{K/T}\big)^+\Big] &\leqslant \sqrt{\frac{\pi}{8}} \sum_{n=1}^{T} \sqrt{\frac{1}{n}}\mathrm{e}^{-2n\left(\sqrt{K/T}\right)^2} \leqslant \sqrt{\frac{\pi}{8}} \int_0^T \sqrt{\frac{1}{x}}\,\mathrm{e}^{-2xK/T}\,\mathrm{d}x \\ &= \sqrt{\frac{\pi}{8}}\sqrt{\frac{T}{2K}} \int_0^{+\infty} \frac{\mathrm{e}^{-u}}{\sqrt{u}}\,\mathrm{d}u = \frac{\pi}{4}\sqrt{\frac{T}{K}}\,, \qquad (42)\end{aligned}$$

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

For the second sum in (41), we also resort to a sum–integral comparison, which exploits the fact that $n \mapsto \varphi(T/Kn)$ is decreasing, and perform the change of variable $u = T/(Kx)$:

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

Collecting the bounds above, we showed, as desired,

$$\begin{aligned}\sum_{t=K+1}^{T} \mathbb{E}\Big[U_{A_t^{\mathrm{GM}},\tau_{t-1}}^{\mathrm{GM}}(t-1) - \mu_{A_t^{\mathrm{GM}}}\Big] &\leqslant \sqrt{KT} + \sum_{a=1}^{K}\sum_{n=1}^{T} \mathbb{E}\Big[\big(U_{a,T,n}^{\mathrm{GM}} - \mu_a - \sqrt{K/T}\big)^+\Big] \\ &\leqslant \sqrt{KT}\left(1 + \frac{\pi}{4} + \frac{1}{\sqrt{2}}\int_1^{+\infty} u^{-3/2}\sqrt{\varphi(u)}\,\mathrm{d}u\right).\end{aligned}$$

■

B. Proofs of the Regularity and Deviation/Concentration Results on $\mathcal{K}_{\inf}$

We provide here the proofs of all claims made in Section 4.4 about the $\mathcal{K}_{\inf}$ function. These proofs are all standard but we occasionally provide simpler or more direct arguments (or slightly refined bounds).

B.1 Proof of the Regularity Lemma (Lemma 11)

The proof below is a variation on the proofs that can be found in Honda and Takemura (2015) or earlier references of the same authors.

Proof To prove (19) we lower bound $\mathcal{K}_{\inf}(\nu, \mu - \varepsilon)$. To that end, given the definition (2), we lower bound $\mathrm{KL}(\nu, \nu')$ for any fixed probability distribution $\nu' \in \mathcal{P}[0,1]$ such that

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

Since ν' is a probability distribution, it has a countable number of atoms, and one can pick a real number $x > \mu$, arbitrary close to 1, such that $\delta_x \perp \nu'$ (such that the two probability measures δ_x and ν' are singular), where δ_x is the Dirac distribution at x. We define

$$\nu'_\alpha = (1-\alpha)\nu' + \alpha\delta_x \,, \qquad \text{where} \qquad \alpha = \frac{\varepsilon}{\varepsilon + (x-\mu)} \in (0,1) \,.$$

The expectation of ν'_α satisfies

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

Since $\alpha \in (0,1)$, we have $\nu'_\alpha \gg \nu'$; therefore, $\nu'_\alpha \gg \nu' \gg \nu$ and $\delta_x \perp \nu'$, which imply the following equalities involving densities (Radon-Nikodym derivatives): ν'_α–almost surely (and therefore also ν'– and ν–almost surely),

$$\frac{\mathrm{d}\nu'}{\mathrm{d}\nu'_\alpha} = \frac{1}{1-\alpha} \,, \qquad \text{thus} \qquad \frac{\mathrm{d}\nu}{\mathrm{d}\nu'_\alpha} = \frac{\mathrm{d}\nu'}{\mathrm{d}\nu'_\alpha}\,\frac{\mathrm{d}\nu}{\mathrm{d}\nu'} = \frac{1}{1-\alpha}\,\frac{\mathrm{d}\nu}{\mathrm{d}\nu'} \,. \tag{43}$$

This allows to compute explicitly the following Kullback-Leibler divergence:

$$\mathrm{KL}(\nu, \nu'_\alpha) = \int_{[0,1]} \ln\!\left(\frac{\mathrm{d}\nu}{\mathrm{d}\nu'_\alpha}\right) \mathrm{d}\nu = \mathrm{KL}(\nu, \nu') + \ln\frac{1}{1-\alpha} \,.$$

Since $\mathrm{E}(\nu'_\alpha) > \mu$ and by the definition of $\mathcal{K}_{\inf}$ as an infimum,

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

Letting x go to 1, which implies that α goes to $\varepsilon/(1-\mu+\varepsilon)$, yields

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

where we also used $\ln(1+u) \leqslant u$ for all $u > -1$. Finally, by taking the infimum in the right-most equation above over all probability distributions ν' such that $\mathrm{E}(\nu') > \mu - \varepsilon$ and $\nu' \gg \nu$, we obtain the desired inequality:

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

To prove the second part (20) of Lemma 11, we follow a similar path as above. We lower bound $\mathrm{KL}(\nu, \nu')$ for any fixed probability distribution $\nu' \in \mathcal{P}[0,1]$ such that

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

To that end, we introduce

$$\nu'_\alpha = (1-\alpha)\nu' + \alpha\nu \qquad \text{for} \qquad \alpha = \frac{\varepsilon}{\big(\mathrm{E}(\nu') - \mathrm{E}(\nu)\big)} \in (0,1)\,,$$

where $\alpha \in (0,1)$ since $\mathrm{E}(\nu) \leqslant \mu - \varepsilon$ by assumption and $\mathrm{E}(\nu') > \mu$. These two inequalities also indicate that

$$\mathrm{E}(\nu') - \mathrm{E}(\nu) > \varepsilon\,, \qquad \text{thus} \qquad \mathrm{E}(\nu'_\alpha) = \mathrm{E}(\nu') - \alpha\big(\mathrm{E}(\nu') - \mathrm{E}(\nu)\big) > \mu - \varepsilon\,, \tag{44}$$

so that $\mathrm{KL}(\nu, \nu'_\alpha) \geqslant \mathcal{K}_{\inf}(\nu, \mu - \varepsilon)$. Now, thanks to the absolute continuities $\nu' \gg \nu'_\alpha \gg \nu$, we have

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

Therefore, by Fubini's theorem, the Kullback-Leibler divergence between ν and ν' equals

$$\begin{aligned}
\mathrm{KL}(\nu, \nu') = \int_{[0,1]} \ln\bigg(\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}\bigg)\,\mathrm{d}\nu &= \int_{[0,1]} \ln\bigg(\frac{\mathrm{d}\nu}{\mathrm{d}\nu'_\alpha}\bigg)\,\mathrm{d}\nu + \int_{[0,1]} \ln\bigg((1-\alpha) + \alpha\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}\bigg)\,\mathrm{d}\nu \\
&\geqslant \int_{[0,1]} \ln\bigg(\frac{\mathrm{d}\nu}{\mathrm{d}\nu'_\alpha}\bigg)\,\mathrm{d}\nu + \alpha\int_{[0,1]} \ln\bigg(\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}\bigg)\,\mathrm{d}\nu \\
&= \mathrm{KL}(\nu, \nu'_\alpha) + \alpha\,\mathrm{KL}(\nu, \nu')\,,
\end{aligned}$$

where we used the concavity of logarithm for the inequality. By Pinsker's inequality together with the data-processing inequality for Kullback-Leibler divergences (see, e.g., Garivier et al., 2019, Lemma 1),

$$\mathrm{KL}(\nu, \nu') \geqslant \mathrm{KL}\Big(\mathrm{Ber}\big(\mathrm{E}(\nu)\big),\, \mathrm{Ber}\big(\mathrm{E}(\nu')\big)\Big) \geqslant 2\big(\mathrm{E}(\nu) - \mathrm{E}(\nu')\big)^2\,.$$

Substituting this inequality above, we proved so far

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

where we used the definition of α for the last inequality. By applying the bound (44) and its consequence $\mathrm{KL}(\nu, \nu'_\alpha) \geqslant \mathcal{K}_{\inf}(\nu, \mu - \varepsilon)$, we finally get

$$\mathrm{KL}(\nu, \nu') \geqslant \mathcal{K}_{\inf}(\nu, \mu - \varepsilon) + 2\varepsilon^2 \,.$$

The proof of (20) is concluded by taking the infimum in the left-hand side over the probability distributions ν' such that $\mathrm{E}(\nu') > \mu$ (and $\nu' \gg \nu$). ■

B.2 A Useful Tool: a Variational Formula for $\mathcal{K}_{\inf}$ (Statement)

The variational formula below appears in Honda and Takemura (2015) as Theorem 2 (and Lemma 6) and is an essential tool for deriving the deviation and concentration results for the $\mathcal{K}_{\inf}$. We state it here (and re-derive it in a direct way in Appendix D) for the sake of completeness.

Lemma 21 (variational formula for $\mathcal{K}_{\inf}$) *For all $\nu \in \mathcal{P}[0,1]$ and all $0 < \mu < 1$,*

$$\mathcal{K}_{\inf}(\nu, \mu) = \max_{0 \leqslant \lambda \leqslant 1} \mathbb{E}\left[\ln\left(1 - \lambda \frac{X - \mu}{1 - \mu}\right)\right] \qquad \text{where } X \sim \nu \,. \tag{45}$$

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

$$\mathbb{E}\left[\frac{1}{1 - \lambda^\star (X - \mu)/(1 - \mu)}\right] \leqslant 1 \,. \tag{46}$$

B.3 Proof of the Deviation Result (Proposition 14)

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

Proof We first upper bound $\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu)\big)$: as indicated by the variational formula of Lemma 21, it is a maximum of random variables indexed by $[0,1]$. We provide an upper bound that is a finite maximum. To that end, we fix a real number $\gamma \in (0,1)$, to be determined by the analysis, and let S_γ be the set below,

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

constructed as a finite grid of step size γ centered at $1/2$. The cardinality of this set S_γ is bounded by $1 + 1/\gamma$. Lemma 22 below (together with the consequence mentioned after its statement) indicates that for all $\lambda \in [0,1]$, there exists a $\lambda' \in S_\gamma$ such that for all $x \in [0,1]$,

$$\ln\left(1 - \lambda \frac{x - \mathrm{E}(\nu)}{1 - \mathrm{E}(\nu)}\right) \leqslant 2\gamma + \ln\left(1 - \lambda' \frac{x - \mathrm{E}(\nu)}{1 - \mathrm{E}(\nu)}\right). \tag{47}$$

(The small correction with respect to the original proof is the 2γ factor in the inequality above, instead of the claimed γ term therein; this is due to the constraint $\lambda \leqslant \lambda' \leqslant 1/2$

or $1/2 \leqslant \lambda' \leqslant \lambda$ in the statement of Lemma 22.) Now, a combination of the variational formula of Lemma 21 and of the inequality (47) yields a finite maximum as an upper bound on $\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu)\big)$:

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

In the second part of the proof, we control the deviations of the upper bound obtained. A union bound yields

$$\begin{aligned}\mathbb{P}\Big[\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu)\big) \geqslant u\Big] &\leqslant \mathbb{P}\left[\max_{\lambda'\in S_\gamma} \frac{1}{n}\sum_{k=1}^{n} \ln\left(1 - \lambda'\frac{X_k - \mathrm{E}(\mu)}{1-\mathrm{E}(\mu)}\right) \geqslant u - 2\gamma\right] \\ &\leqslant \sum_{\lambda'\in S_\gamma} \mathbb{P}\left[\frac{1}{n}\sum_{k=1}^{n} \ln\left(1 - \lambda'\frac{X_k - \mathrm{E}(\nu)}{1-\mathrm{E}(\nu)}\right) \geqslant u - 2\gamma\right]. \qquad (48)\end{aligned}$$

By the Markov–Chernov inequality, for all $\lambda' \in [0,1]$, we have

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

where we used the independence of the X_k. Substituting in (48) and using the bound $1+1/\gamma$ on the cardinality of S_γ, we get

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

Taking $\gamma = 1/(2n)$ concludes the proof. ∎

The proof above relies on the following lemma, which is extracted from Cappé et al. (2013, Lemma 7). Its elementary proof (not copied here) consists in bounding of derivative of $\lambda \mapsto \ln(1-\lambda c)$ and using a convexity argument.

Lemma 22 *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 $c \leqslant 1$,*

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

A consequence not drawn by Cappé et al. (2013) is that the lemma above actually also holds for $\lambda = 1$ and $\lambda' \in [1/2, 1)$. Indeed, by continuity and by letting $\lambda \to 1$, we get from this lemma that for all $\lambda' \in [1/2, 1)$ and for all real numbers $c < 1$,

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

The above inequality is also valid for $c = 1$ as the left-hand side equals $-\infty$.

B.4 Proof of the Concentration Result (Proposition 16)

We recall that Proposition 16—and actually most of its proof below—are similar in spirit to Honda and Takemura (2015, Proposition 11). However, they are tailored to our needs. The key ingredients in the proof will be the variational formula (45)—again—and Lemma 23 below. This lemma is a concentration result for random variables that are essentially bounded from one side only; it holds for possibly negative u (there is no lower bound on the u that can be considered).

Lemma 23 *Let $Z_1, \dots, Z_n$ be i.i.d. random variables such that there exist a, $b \geqslant 0$ with*

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

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

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

B.4.1 Proof of Proposition 16 Based on Lemma 23

We apply Lemma 21. We denote by $\lambda^\star \in [0,1]$ a real number achieving the maximum in the variational formula (45) for $\mathcal{K}_{\inf}(\nu, \mu)$. We then introduce the random variable

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

and i.i.d. copies $Z_1, \dots, Z_n$ of Z. Then, $\mathcal{K}_{\inf}(\nu, \mu) = \mathbb{E}\big[Z\big]$ and by the variational formula (45) again,

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

for all real numbers x. Now, $X \geqslant 0$ and $\lambda^\star \leqslant 1$, thus

$$Z \leqslant \ln\left(1 + \lambda^\star \frac{\mu}{1-\mu}\right) \leqslant \ln\left(\frac{1}{1-\mu}\right) \stackrel{\text{def}}{=} a\,.$$

On the other hand,

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

where the upper bound by 1 follows from (46). Using $b = 1$ and the value of a specified above, this proves Proposition 16 via Lemma 23, except for the inequality $\mathrm{e}^{-n\gamma/8} \leqslant \mathrm{e}^{-n/4}$ claimed therein. The latter is a consequence of $\gamma \geqslant 2$; indeed, as γ is an increasing function of $\mu > 0$,

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

Remark 24 *In the proof of Theorem 2 provided in Section C we will not use Proposition 16 as stated but a stronger result: the fact that for all $x < \mathcal{K}_{\inf}(\nu, \mu)$,*

$$\mathbb{P}\left[\sum_{i=1}^{n} Z_i \leqslant nx\right] \leqslant \begin{cases} \exp(-n\gamma/8) \leqslant \exp(-n/4) & \text{if } x \leqslant \mathcal{K}_{\inf}(\nu,\mu) - \gamma/2, \\ \exp\Big(-n\big(\mathcal{K}_{\inf}(\nu,\mu) - x\big)^2/(2\gamma)\Big) & \text{if } x > \mathcal{K}_{\inf}(\nu,\mu) - \gamma/2, \end{cases}$$

with the notation of Proposition 16. This is indeed what we proved above; Proposition 16 then followed from the inequality (also established above)

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

B.4.2 Proof of Lemma 23

This lemma is a direct application of the Crámer–Chernov method. We introduce the log-moment generating function Λ of Z_1:

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

Lemma 25 *Under the same assumptions $Z_1 \leqslant a$ and $\mathbb{E}\big[\mathrm{e}^{-Z_1}\big] \leqslant b$ as in Lemma 23, the log-moment generating function Λ is well-defined at least on the interval $[-1, 1]$ and twice differentiable at least on $(-1, 1)$, with $\Lambda'(0) = \mathbb{E}[Z_1]$ and $\Lambda''(x) \leqslant \gamma$ for $x \in [-1/2,\, 0]$, where $\gamma = \sqrt{\mathrm{e}^a}\big(16\,\mathrm{e}^{-2}b + a^2\big)$ denotes the same constant as in Lemma 23.*

Based on this lemma (proved below), we may resort to a Taylor expansion with a Lagrange remainder and get the bound:

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

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

$$\begin{aligned}\mathbb{P}\left[\sum_{i=1}^{n} Z_i \leqslant nu\right] = \mathbb{P}\left[\prod_{i=1}^{n} \mathrm{e}^{xZ_i} \geqslant \mathrm{e}^{nux}\right] &\leqslant \mathrm{e}^{-nux}\left(\mathbb{E}\big[\mathrm{e}^{xZ_1}\big]\right)^n = \exp\Big(-n\big(ux - \Lambda(x)\big)\Big) \\ &\leqslant \exp\left(n\Big(x^2\,\gamma/2 - x\,\big(u - \mathbb{E}[Z_1]\big)\Big)\right). \qquad (49)\end{aligned}$$

That is,

$$\mathbb{P}\left[\sum_{i=1}^{n} Z_i \leqslant nu\right] \leqslant \exp\Big(n \min_{x\in[1/2,\,0]} P(x)\Big)\,,$$

where we introduced the second-order polynomial function

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

The claimed bound is obtained by minimizing P over $[-1/2,\,0]$ depending on whether $u > \mathbb{E}[Z_1] - \gamma/2$ or $u \leqslant \mathbb{E}[Z_1] - \gamma/2$, which we do now.

We recall that by assumption, $u < \mathbb{E}[Z_1]$. We note that P is a second-order polynomial function with positive leading coefficient and roots 0 and $2\big(u-\mathbb{E}[Z_1]\big)/\gamma < 0$. Its 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 P is to be minimized over $[-1/2,\,0]$ only. In the case where $u > \mathbb{E}[Z_1] - \gamma/2$, the midpoint $x^\star$ belongs to the interval of interest and

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

Otherwise, $u - \mathbb{E}[Z_1] \leqslant -\gamma/2$ and the midpoint $x^\star$ is to the left of $-1/2$. Therefore, P is increasing on $[-1/2, 0]$, so that its minimum on this interval is achieved at $-1/2$, that is,

$$\min_{[-1/2,0]} P = P(-1/2) = \frac{\gamma}{8} + \frac{1}{2}\big(u - \mathbb{E}[Z_1]\big) \leqslant \frac{\gamma}{8} - \frac{\gamma}{4} = -\frac{\gamma}{8}\,.$$

This concludes the proof of Lemma 23. We end this section by proving Lemma 25, which stated some properties of the Λ function.

Proof of Lemma 25 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,

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

First, 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.

Second, 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 $x \in (-1, 1)$, using $y \leqslant \mathrm{e}^y$ for $y \geqslant 0$,

$$\begin{aligned}\big|Z_1\,\mathrm{e}^{xZ_1}\big| &= Z_1\,\mathrm{e}^{xZ_1}\,\mathbb{1}_{\{Z_1 \geqslant 0\}} + \frac{1}{x+1}\,\big(-Z_1(x+1)\big)\,\mathrm{e}^{xZ_1}\,\mathbb{1}_{\{Z_1<0\}} \\ &\leqslant a\,\mathrm{e}^a + \frac{1}{x+1}\mathrm{e}^{-Z_1(x+1)}\mathrm{e}^{xZ_1} = a\,\mathrm{e}^a + \frac{1}{x+1}\mathrm{e}^{-Z_1}\,.\end{aligned}$$

Given that $y^2 \leqslant \mathrm{e}^y$ for $y \geqslant 0$, we show similarly that $x \in (-1, 1) \mapsto Z_1^2\,\mathrm{e}^{xZ_1}$ is also locally dominated by an integrable random variable.

Thus, ψ is twice differentiable at least on $(-1, 1)$, with first and second derivatives

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

Therefore, so is $\Lambda = \ln \psi$, with derivatives

$$\Lambda'(x) = \frac{\psi'(x)}{\psi(x)} = \frac{\mathbb{E}\big[Z_1 \, \mathrm{e}^{xZ_1}\big]}{\mathbb{E}\big[\mathrm{e}^{xZ_1}\big]}$$

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

In particular, $\Lambda'(0) = \mathbb{E}[Z_1]$.

Finally, for the bound on $\Lambda''(x)$, we note first that $Z_1 \leqslant a$ (with $a \geqslant 0$) and $x \in [-1/2, \, 0]$ entail that $\mathrm{e}^{xZ_1} \geqslant \mathrm{e}^{xa} \geqslant 1/\sqrt{\mathrm{e}^a}$. Second, $\mathbb{E}\big[Z_1^2 \, \mathrm{e}^{xZ_1}\big] \leqslant 16 \, \mathrm{e}^{-2} b + a^2$ follows from replacing z by Z_1 and taking expectations in the inequality (proved below)

$$\forall \, x \in [-1/2, \, 0], \ z \in (-\infty, a], \qquad z^2 \, \mathrm{e}^{xz} \leqslant 16 \, \mathrm{e}^{-2} \mathrm{e}^{-z} + a^2 \, . \tag{50}$$

Collecting all elements together, we proved

$$\Lambda''(x) \leqslant \frac{\mathbb{E}\big[Z_1^2 \, \mathrm{e}^{xZ_1}\big]}{\mathbb{E}\big[\mathrm{e}^{xZ_1}\big]} \leqslant \sqrt{\mathrm{e}^a}\big(16 \, \mathrm{e}^{-2} b + a^2\big) = \gamma \, .$$

To see why (50) holds, note that in the case $z \geqslant 0$, since $x \leqslant 0$, we have the chain of inequalities $z^2 \, \mathrm{e}^{xz} \leqslant z^2 \leqslant a^2$. In the case $z \leqslant 0$, we have (by function study) $z^2 \leqslant 16 \mathrm{e}^{-2-z/2}$, so that $z^2 \, \mathrm{e}^{xz} \leqslant 16 \mathrm{e}^{-2} \, \mathrm{e}^{(x-1/2)z} \leqslant 16 \mathrm{e}^{-2} \mathrm{e}^{-z}$, where we used $x \geqslant -1/2$ for the final inequality. ■

C. Proof of Theorem 2 (with the $-\ln\ln T$ Term in the Regret Bound)

We incorporate two refinements to the proof of Theorem 4 in Section 6 to obtain Theorem 2 with this improved $-\ln\ln T$ term., with occasional simplifications due to not having to deal with varying values of t (e.g., the initial manipulations in Part 2 of the proof of Theorem 4 are unnecessary). The first refinement is that the left deviations of the index are controlled with an additional cut on the value of $U_a(t)$ *before* using the bound $U_a(t) \geqslant U_{a^\star}(t)$ that holds when $A_{t+1} = a$. This improves the dependency on the parameter δ used in the proof; as a consequence, $\delta = T^{-1/8}$ will be set instead of $\delta = (\ln T)^{-1/3}$, which will improve the order of magnitude of second-order terms. Second, to sharpen the bound on the quantity (55), which contains the main logarithmic term, we use a trick introduced in the analysis of the IMED policy by Honda and Takemura (2015, Theorem 5). Their idea was to deal with the deviations in a more careful way and relate the sum (55) to the behaviour of a biased random walk. Doing so, we obtain a bound of the form $\kappa \, W(cT)$, where W is Lambert's function, instead of the bound of the form $\kappa \ln(cT)$ stated in Theorem 4.

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 $w > 0$. It is an increasing function satisfying (see, e.g., Hoorfar and Hassani, 2008, Corollary 2.4)

$$\forall \, x > \mathrm{e}, \qquad \ln x - \ln\ln x \leqslant W(x) \leqslant \ln x - \ln\ln x + \ln\big(1 + \mathrm{e}^{-1}\big) \, . \tag{51}$$

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

What we will exactly prove below is the following. We recall that we assume here $\mu^\star \in (0,1)$. Given $T \geqslant K/\min\{1-\mu^\star, (\Delta_a/9)^{12}\}$, the KL-UCB-Switch algorithm, tuned with the knowledge of T 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, and for all $\delta > 0$ satisfying

$$\delta < \min\left\{\mu^\star,\ \frac{\Delta_a}{2},\ \frac{1-\mu^\star}{2}\,\mathcal{K}_{\inf}(\nu_a,\mu^\star)\right\},$$

we have

$$\begin{aligned}
\mathbb{E}[N_a(T)] \leqslant 1 &\\
&+ \frac{5\mathrm{e}K}{\left(1-\mathrm{e}^{-\Delta_a^2/2}\right)^3} + T\mathrm{e}^{-\Delta_a^2 T/(2K)} \\
&+ \frac{K/T}{1-\mathrm{e}^{-\Delta_a^2/8}} \\
&+ \left\lceil \frac{8}{\Delta_a^2}\,\ln\left(\frac{T}{K}\right)\right\rceil \left(\frac{5\mathrm{e}K/T}{(1-\mathrm{e}^{-2\delta^2})^3} + \mathrm{e}^{-2\delta^2 T/K}\right) \\
&+ \frac{1}{\mathcal{K}_{\inf}(\nu_a,\mu^\star) - \delta/(1-\mu^\star)}\left(W\left(\frac{\ln\big(1/(1-\mu^\star)\big)}{K}\,T\right) + \ln\big(2/(1-\mu^\star)\big)\right) \\
&\qquad\qquad + 5 + \frac{1}{1-\mathrm{e}^{-\mathcal{K}_{\inf}(\nu,\mu^\star)^2/(8\gamma_\star)}} \\
&+ \frac{1}{1-\mathrm{e}^{-\Delta_a^2/8}}\,.
\end{aligned} \tag{52}$$

We write the bound in this way to match the decomposition of $\mathbb{E}[N_a(T)]$ appearing in the proof (see page 49). For a choice $\delta \to 0$ as $T \to +\infty$, the previous bound is of the form

$$\mathbb{E}[N_a(T)] \leqslant \frac{W\big(c_{\mu^\star} T\big)}{\mathcal{K}_{\inf}(\nu_a,\mu^\star) - \delta/(1-\mu^\star)} + \mathcal{O}_T\left(\frac{\ln T}{\delta^6 T}\right) + \mathcal{O}_T\big((\ln T)\,\mathrm{e}^{-2\delta^2 T/K}\big) + \mathcal{O}_T(1)\,,$$

where $c_{\mu^\star} = \ln\big(1/(1-\mu^\star)\big)/K$. Based on the inequalities (51) and on the first-order approximation $1/(1-\varepsilon) = 1 + \varepsilon + \mathcal{O}(\varepsilon)$ as $\varepsilon \to 0$, we get

$$\mathbb{E}[N_a(T)] \leqslant \frac{\ln T - \ln\ln T}{\mathcal{K}_{\inf}(\nu_a,\mu^\star)}\big(1 + \mathcal{O}_T(\delta)\big) + \mathcal{O}_T\left(\frac{\ln T}{\delta^6 T}\right) + \mathcal{O}_T\big((\ln T)\,\mathrm{e}^{-2\delta^2 T/K}\big) + \mathcal{O}_T(1)\,.$$

The choice $\delta = T^{-1/8}$ leads to the bound stated in Theorem 2, namely,

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

Proof structure of the closed-form bound (52) As in the proof of Theorem 4, given $\delta > 0$ sufficiently small, we decompose $\mathbb{E}\big[N_a(T)\big]$. However, this time we refine the decomposition quite a bit. Instead of simply distinguishing whether $U_a(t)$ is greater or smaller

than $\mu^\star - \delta$, we add a cutting point at $(\mu^\star + \mu_a)/2$. 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 $N_a(t) \leqslant n_0 - 1$ when $U_a(t) < \mu^\star - \delta$, while we keep the integer threshold $f(T,K)$ in the case $U_a(t) \geqslant \mu^\star - \delta$. More precisely,

$$\begin{aligned}
\big\{U_a(t) < \mu^\star - \delta\big\} \cup \big\{U_a(t) \geqslant \mu^\star - \delta\big\} & \\
&= \quad \big\{U_a(t) < \mu^\star - \delta \text{ and } N_a(t) \geqslant n_0\big\} \\
&\quad \cup \big\{U_a(t) < \mu^\star - \delta \text{ and } N_a(t) \leqslant n_0 - 1\big\} \\
&\quad \cup \big\{U_a(t) \geqslant \mu^\star - \delta \text{ and } N_a(t) \leqslant f(T,K)\big\} \\
&\quad \cup \big\{U_a(t) \geqslant \mu^\star - \delta \text{ and } N_a(t) \geqslant f(T,K) + 1\big\} \\
&\subseteq \quad \big\{U_a(t) < (\mu^\star + \mu_a)/2 \text{ and } N_a(t) \geqslant n_0\big\} \\
&\quad \cup \big\{(\mu^\star + \mu_a)/2 \leqslant U_a(t) < \mu^\star - \delta \text{ and } N_a(t) \geqslant n_0\big\} \\
&\quad \cup \big\{U_a(t) < \mu^\star - \delta \text{ and } N_a(t) \leqslant n_0 - 1\big\} \\
&\quad \cup \big\{U_a^{\mathrm{KL}}(t) \geqslant \mu^\star - \delta \text{ and } N_a(t) \leqslant f(T,K)\big\} \\
&\quad \cup \big\{U_a^{\mathrm{M}}(t) \geqslant \mu^\star - \delta \text{ and } N_a(t) \geqslant f(T,K) + 1\big\},
\end{aligned}$$

where, to get the inclusion, we further cut the first event into two events and we used the definition of the index $U_a(t)$ to replace it by $U_a^{\mathrm{KL}}(t)$ or $U_a^{\mathrm{M}}(t)$ in the last two events.

Hence, by intersecting this partition of the space with the event $\{A_{t+1} = a\}$ and by slightly simplifying the first and second events of the partition:

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

Only now do we inject the bound $U_{a^\star}(t) \leqslant U_a(t)$, valid when $A_{t+1} = a$, as well as a union bound, to obtain our working decomposition of $\mathbb{E}[N_a(t)]$:

$$\begin{aligned}
\mathbb{E}\big[N_a(T)\big] \leqslant & \ 1 \\
&+ \sum_{t=K}^{T-1} \mathbb{P}\big[U_{a^\star}(t) < (\mu^\star + \mu_a)/2\big] && (S_1) \\
&+ \sum_{t=K}^{T-1} \mathbb{P}\big[U_a(t) \geqslant (\mu^\star + \mu_a)/2 \text{ and } A_{t+1} = a \text{ and } N_a(t) \geqslant n_0\big] && (S_2) \\
&+ \sum_{t=K}^{T-1} \mathbb{P}\big[U_{a^\star}(t) < \mu^\star - \delta \text{ and } A_{t+1} = a \text{ and } N_a(t) \leqslant n_0 - 1\big] && (S_3) \\
&+ \sum_{t=K}^{T-1} \mathbb{P}\big[U_a^{\mathrm{KL}}(t) \geqslant \mu^\star - \delta \text{ and } A_{t+1} = a \text{ and } N_a(t) \leqslant f(T,K)\big] && (S_4) \\
&+ \sum_{t=K}^{T-1} \mathbb{P}\big[U_a^{\mathrm{M}}(t) \geqslant \mu^\star - \delta \text{ and } A_{t+1} = a \text{ and } N_a(t) \geqslant f(T,K) + 1\big]. && (S_5)
\end{aligned}$$

We call S_1, S_2, S_3, S_4, S_5 the five sums appearing in the right-hand side of the display above, and will now bound them separately. Most of the efforts will be dedicated to bounding the sum S_4. ■

C.1 Bound on S_5

The sum S_5 involves the indexes $U_a^{\mathrm{M}}(t)$ only under the condition $N_a(t) \geqslant f(T,K)+1$, in which case $N_a(t) \geqslant (T/K)^{1/5}$ and

$$U_a^{\mathrm{M}}(t) \stackrel{\mathrm{def}}{=} \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)}\,\ln_+\left(\frac{T}{KN_a(t)}\right)} \leqslant \widehat{\mu}_a(t) + \sqrt{\frac{1}{2\,(T/K)^{1/5}}\,\ln_+\left((T/K)^{4/5}\right)}\,.$$

We mimic the proof scheme of Part 3 of the proof of Theorem 4 (see around page 30). Since $T \geqslant K/(1-\mu^\star)$ by assumption, it holds $T/K \geqslant 1$. Using that $x \mapsto x^{1/24}/\ln(x)$ takes it minimum over $[1,+\infty)$ at e^{-24}, with value larger than 0.113, and since we assumed $T \geqslant K(9/\Delta_a)^{12}$, we obtain

$$\begin{aligned}\sqrt{\frac{1}{2\,(T/K)^{1/5}}\,\ln_+\left((T/K)^{4/5}\right)} &\leqslant \sqrt{\frac{1}{2\times 0.113\,(T/K)^{1/5}}\,(T/K)^{1/30}} \\ &= \frac{1}{\sqrt{0.226}}\left(\frac{K}{T}\right)^{1/12} \leqslant \frac{\Delta_a}{4}\,.\end{aligned}$$

Under the same condition $\delta < \Delta_a/4$ as therein, we get, by a careful application of optional skipping (Section 4.1, Example 2) using that the events $\big\{A_{t+1}=a \ \text{ and } \ N_a(t)=n\big\}$ are disjoint as t varies, and by Hoeffding's inequality,

$$\begin{aligned}S_5 &= \sum_{t=K}^{T-1} \mathbb{P}\big[U_a^{\mathrm{M}}(t) \geqslant \mu^\star - \delta \ \text{ and } \ A_{t+1}=a \ \text{ and } \ N_a(t) \geqslant f(T,K)+1\big] \\ &\leqslant \sum_{n=f(T,K)+1}^{T-1} \mathbb{P}\big[\widehat{\mu}_{a,n} \geqslant \mu_a + \Delta_a/2\big] \leqslant \sum_{n \geqslant f(T,K)+1} \mathrm{e}^{-n\Delta_a^2/2} \leqslant \frac{1}{1-\mathrm{e}^{-\Delta_a^2/2}}\,.\end{aligned}$$

C.2 Bound on S_2

Let

$$n_0 = \left\lceil \frac{8}{\Delta_a^2}\,\ln\left(\frac{T}{K}\right)\right\rceil. \tag{53}$$

By Pinsker's inequality (8), by definition of the MOSS index, and by our choice of n_0, we have, when $N_a(t) \geqslant n_0$,

$$U_a(t) \leqslant U_a^{\mathrm{M}}(t) = \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)}\ln_+\left(\frac{T}{KN_a(t)}\right)} \leqslant \widehat{\mu}_a(t) + \underbrace{\sqrt{\frac{1}{2n_0}\ln_+\left(\frac{T}{Kn_0}\right)}}_{\leqslant \Delta_a/4}\,. \tag{54}$$

In particular, we get the inclusion

$$\begin{aligned}\big\{U_a(t) \geqslant (\mu^\star + \mu_a)/2 \ \text{ and } \ N_a(t) \geqslant n_0\big\} &= \big\{U_a(t) \geqslant \mu_a + \Delta_a/2 \ \text{ and } \ N_a(t) \geqslant n_0\big\} \\ &\subseteq \big\{\widehat{\mu}_a(t) \geqslant \mu_a + \Delta_a/4 \ \text{ and } \ N_a(t) \geqslant n_0\big\}\,.\end{aligned}$$

Thus

$$S_2 \leqslant \sum_{t=K}^{T-1} \mathbb{P}\bigg[\widehat{\mu}_a(t) \geqslant \mu_a + \frac{\Delta_a}{4} \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) \geqslant n_0\bigg]\,.$$

We now proceed again similarly to what we already did on page 30. By a careful application of optional skipping (see Section 4.1, Example 2), using the fact that, as t varies, all the events $\{A_{t+1} = a \ \text{ and } \ N_a(t) = n\}$ are disjoint, the sum above may be bounded by

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

By a final application of Hoeffding's inequality (Proposition 5, actually not using the maximal form):

$$S_2 \leqslant \sum_{n=n_0}^{T} \mathbb{P}\bigg[\widehat{\mu}_{a,n} \geqslant \mu_a + \frac{\Delta_a}{4}\bigg] \leqslant \sum_{n=n_0}^{T} \mathrm{e}^{-n\Delta_a^2/8} = \frac{\mathrm{e}^{-n_0\Delta_a^2/8}}{1 - \mathrm{e}^{-\Delta_a^2/8}} \leqslant \frac{K/T}{1 - \mathrm{e}^{-\Delta_a^2/8}}\,,$$

where we substituted the value (53) of n_0.

C.3 Bounds on S_1 and S_3

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

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

allowing us to upper bound the probabilities in terms of events that do not depend on t:

$$\{U_{a^\star}(t) < (\mu^\star + \mu_a)/2\} \subseteq \mathcal{E}_\star\big((\mu^\star + \mu_a)/2\big) \qquad \text{and} \qquad \{U_{a^\star}(t) < \mu^\star - \delta\} \subseteq \mathcal{E}_\star\big(\mu^\star - \delta\big)\,.$$

Summing directly the first inclusion above yields an upper bound on S_1:

$$S_1 \leqslant T\ \mathbb{P}\Big(\mathcal{E}_\star\big((\mu^\star + \mu_a)/2\big)\Big)\,.$$

Using the deterministic control

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

together with the second inclusion above, we get (and this is where it is handy that the $\mathcal{E}_\star$ do not depend on a particular t)

$$\begin{aligned}\sum_{t=K}^{T-1} \mathbb{1}_{\{U_{a^\star}(t) < \mu^\star - \delta \ \text{ and } \ A_{t+1}=a \ \text{ and } \ N_a(t) \leqslant n_0 - 1\}} &\leqslant \mathbb{1}_{\mathcal{E}_\star(\mu^\star - \delta)} \sum_{t=K}^{T-1} \mathbb{1}_{\{A_{t+1}=a \ \text{ and } \ N_a(t) \leqslant n_0 - 1\}} \\ &\leqslant n_0\, \mathbb{1}_{\mathcal{E}_\star(\mu^\star - \delta)}\,,\end{aligned}$$

which in turn yields

$$S_3 \leqslant n_0 \, \mathbb{P}\big(\mathcal{E}_\star(\mu^\star - \delta)\big) \, .$$

We recall that n_0 was defined in (53). The lemma right below, respectively with $x = \Delta_a/2$ and $x = \delta$, yields the final bounds

$$S_1 \leqslant \frac{5\mathrm{e}K}{\big(1 - \mathrm{e}^{-\Delta_a^2/2}\big)^3} + T\mathrm{e}^{-\Delta_a^2 T/(2K)}$$

and

$$S_3 \leqslant \left\lceil \frac{8}{\Delta_a^2} \ln\!\left(\frac{T}{K}\right)\right\rceil \left(\frac{5\mathrm{e}K/T}{(1-\mathrm{e}^{-2\delta^2})^3} + \mathrm{e}^{-2\delta^2 T/K}\right).$$

Lemma 26 *For all $x \in (0, \mu^\star)$,*

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

Proof We first lower bound $U_{a^\star}(\tau)$ depending on whether $N_{a^\star}(\tau) < T/K$ or $N_{a^\star}(\tau) \geqslant T/K$. In the first case, we will simply apply Pinsker's inequality (8) to get $U^{\mathrm{KL}}_{a^\star}(\tau) \leqslant U_{a^\star}(\tau)$. In the second case, since $T \geqslant K/(1-\mu^\star) \geqslant K$, we have, by definition of $f(T,K)$, that $T/K \geqslant (T/K)^{1/5} \geqslant f(T,K)$ and thus, by definition of the $U_{a^\star}(\tau)$ index, $U_{a^\star}(\tau) = U^{\mathrm{M}}_{a^\star}(\tau)$. Now, the $\ln_+$ in the definition of $U^{\mathrm{M}}_{a^\star}(\tau)$ vanishes when $N_{a^\star}(\tau) \geqslant T/K$, so all in all we have $U_{a^\star}(\tau) = \widehat{\mu}_{a^\star}(\tau)$ when $N_{a^\star}(\tau) \geqslant T/K$. Therefore, by a careful application of optional skipping (see Section 4.1, end of Example 1),

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

As in the proof of Corollary 15, by the definition of the $U^{\mathrm{KL}}_{a^\star,m}$ index as some supremum (together with the left-continuity of $\mathcal{K}_{\inf}$ deriving from Lemma 11), we finally get

$$\begin{aligned}\mathbb{P}\Big(\mathcal{E}_\star\big(\mu^\star - x\big)\Big) &\leqslant \mathbb{P}\left[\exists\, m \in \big\{1, \dots, \lfloor T/K \rfloor\big\} : \; \mathcal{K}_{\inf}\big(\widehat{\nu}_{a^\star,m}, \mu^\star - x\big) > \frac{1}{m}\ln\left(\frac{T}{Km}\right)\right] \\ &\quad + \mathbb{P}\Big[\exists\, m \in \big\{\lceil T/K \rceil, \dots, T\big\} : \; \widehat{\mu}_{a^\star,m} < \mu^\star - x\Big] \, .\end{aligned}$$

The proof continues by bounding each probability separately. First, again as in the proof of Corollary 15, we apply Corollary 13 (for the first inequality below) and the deviation inequality of Proposition 14 (for the second inequality below), to see that for all $x \in (0, \mu^\star)$ and $\varepsilon > 0$,

$$\mathbb{P}\Big[\mathcal{K}_{\inf}\big(\widehat{\nu}_{a^\star,m}, \mu^\star - x\big) > \varepsilon\Big] \leqslant \mathbb{P}\Big[\mathcal{K}_{\inf}\big(\widehat{\nu}_{a^\star,m}, \mu^\star\big) > \varepsilon + 2x^2\Big] \leqslant \mathrm{e}(2n+1)\,\mathrm{e}^{-n(\varepsilon+2x^2)}\,.$$

Therefore, by a union bound, the above equation, and the calculations on geometric sums (33) and (34),

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

Second, by Hoeffding's maximal inequality (Proposition 5),

$$\begin{aligned}\mathbb{P}\Big[\exists\, m \in \big\{\lceil T/K \rceil, \dots, T\big\} :\ \widehat{\mu}_{a^\star,m} < \mu^\star - x\Big] \\ = \mathbb{P}\left[\max_{\lceil T/K \rceil \leqslant m \leqslant T}\Big(\big(1-\widehat{\mu}_{a^\star,m}\big) - \big(1-\mu^\star\big)\Big) > x\right] \leqslant \mathrm{e}^{-2\,\lceil T/K \rceil\, x^2} \leqslant \mathrm{e}^{-2x^2 T/K}\,.\end{aligned}$$

The proof is concluded by collecting the last two bounds. ■

C.4 Bound on S_4

We begin with a now standard use of optional skipping (see Section 4.1, Example 2), relying on the fact that the events $\{A_{t+1} = a \text{ and } N_a(t) = n\}$ are disjoint as t varies:

$$S_4 = \sum_{t=K}^{T-1} \mathbb{P}\big[U_a^{\mathrm{KL}}(t) \geqslant \mu^\star - \delta \text{ and } A_{t+1} = a \text{ and } N_a(t) \leqslant f(T,K)\big] \leqslant \sum_{n=1}^{f(T,K)} \mathbb{P}\big[U_{a,n}^{\mathrm{KL}} \geqslant \mu^\star - \delta\big]\,.$$

We show in this section that

$$\begin{aligned}\sum_{n=1}^{f(T,K)} \mathbb{P}\big[U_{a,n}^{\mathrm{KL}} \geqslant \mu^\star - \delta\big] \leqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star) - \dfrac{\delta}{1-\mu^\star}}\left(W\left(\frac{\ln\big(1/(1-\mu^\star)\big)}{K}\,T\right) + \ln\big(2/(1-\mu^\star)\big)\right) \\ + 5 + \frac{1}{1-\mathrm{e}^{-\mathcal{K}_{\inf}(\nu,\mu^\star)^2/(8\gamma_\star)}}\,, \quad (55)\end{aligned}$$

where, as in the statement of Proposition 16,

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

To do so, we 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 (55) to the behaviour of a biased random walk.

We start by rewriting the events of interest as

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

where, as in one step of the proof of Lemma 26, we used the definition of $U_{a,n}^{\mathrm{KL}}$ as well as the left-continuity of $\mathcal{K}_{\inf}$. We then follow the same steps as in the proof of Proposition 16 (see Section B.4) and link the deviations in $\mathcal{K}_{\inf}$ divergence to the ones of a random walk. The variational formula (Lemma 21) 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 (7) given that we imposed $\delta \leqslant \Delta_a/2$. We consider i.i.d. copies $X_{a,1}, \dots, X_{a,n}$ of X and form the random variables

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

By the variational formula (Lemma 21) again, applied this time to $\mathcal{K}_{\inf}(\widehat{\nu}_{a,n}, \mu^\star - \delta)$, we see

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

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

$$\left\{\mathcal{K}_{\inf}\big(\widehat{\nu}_{a,n}, \mu^\star - \delta\big) \leqslant \frac{1}{n}\ln\left(\frac{T}{Kn}\right)\right\} \subseteq \left\{\sum_{i=1}^{n} Z_{a,i} \leqslant \ln\left(\frac{T}{Kn}\right)\right\}. \tag{56}$$

Collecting all previous bounds and inclusions, we proved that the sum of interest (55) is bounded by

$$\begin{aligned}
S_4 &\leqslant \sum_{n=1}^{f(T,K)} \mathbb{P}\big[U_{a,n}^{\mathrm{KL}} \geqslant \mu^\star - \delta\big] = \sum_{n=1}^{f(T,K)} \mathbb{P}\left[\mathcal{K}_{\inf}\big(\widehat{\nu}_{a,n}, \mu^\star - \delta\big) \leqslant \frac{1}{n}\ln\left(\frac{T}{Kn}\right)\right] \\
&\leqslant \sum_{n=1}^{f(T,K)} \mathbb{P}\left[\sum_{i=1}^{n} Z_{a,i} \leqslant \ln\left(\frac{T}{Kn}\right)\right] = \mathbb{E}\left[\sum_{n=1}^{f(T,K)} \mathbb{1}_{\left\{\sum_{i=1}^{n} Z_{a,i} \leqslant \ln(T/(Kn))\right\}}\right] \\
&\leqslant \mathbb{E}\left[\sum_{n=1}^{T} \mathbb{1}_{\left\{\sum_{i=1}^{n} Z_{a,i} \leqslant \ln(T/(Kn))\right\}}\right].
\end{aligned}$$

The last upper bound may seem crude but will be good enough for our purpose.

We may reinterpret

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

as the expected number of times a random walk with positive drift 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. In what follows, $\wedge$ denotes the minimum of two numbers. We define the first hitting time τ_a of the barrier, if it exists, as

$$\tau_a = \inf\left\{n \geqslant 1 : \sum_{i=1}^{n} Z_{a,i} > \ln\left(\frac{T}{Kn}\right)\right\} \wedge T\,.$$

The time τ_a is bounded by T and 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_a is met for some $1 \leqslant n \leqslant T$, i.e., whether or not the barrier is hit for $1 \leqslant n \leqslant T$, we get

$$S_4 \leqslant \mathbb{E}\left[\sum_{n=1}^{T} \mathbb{1}_{\left\{\sum_{i=1}^{n} Z_{a,i} \leqslant \ln(T/(Kn))\right\}}\right] \leqslant \mathbb{E}[\tau_a] + \mathbb{E}\left[\sum_{n=\tau_a+1}^{T} \mathbb{1}_{\left\{\sum_{i=1}^{n} Z_{a,i} \leqslant \ln(T/(Kn))\right\}}\right], \quad (57)$$

where the sum from τ_a+1 to T is void thus null when $\tau_a = T$ (this is the case, in particular, when the barrier is hit for no $n \leqslant T$). We now state a lemma, in the spirit of Honda and Takemura (2015, Lemma 18), and will prove it later at the end of this section.

Lemma 27 *Let $(Z_i)_{i\geqslant 1}$ 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*

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

and denote by W Lambert's function. Then, for all $T \geqslant K\mathrm{e}^{\alpha}$,

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

The random variables $Z_{a,i}$ have positive expectation $\mathcal{K}_{\inf}(\nu_a, \mu^\star - \delta) > 0$ and are bounded by $\alpha = \ln\bigl(1/(1-\mu^\star)\bigr)$; indeed, since $X_{a,i} \geqslant 0$ and $\lambda_{a,\delta} \in [0,1]$, we have

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

In addition, we imposed that $T > K/(1-\mu^\star) = K\mathrm{e}^\alpha$. Therefore, Lemma 27 applies and yields the bound

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

where the second inequality follows by the regularity inequality (19) on $\mathcal{K}_{\inf}$ (and the denominator therein is still positive thanks to our assumption on δ). All in all, we obtained the first part of the bound (55) and conclude the proof of the latter based on the decomposition (57) by showing that

$$\mathbb{E}\left[\sum_{n=\tau_a+1}^{T} \mathbb{1}_{\left\{\sum_{i=1}^n Z_{a,i} \leqslant \ln(T/(Kn))\right\}} \right] \leqslant \beta \overset{\text{def}}{=} 5 + \frac{1}{1 - \mathrm{e}^{-\mathcal{K}_{\inf}(\nu_a,\mu^\star)^2/(8\gamma_\star)}}\,. \tag{58}$$

To that end, note that when $\tau_a < T$, we have by definition of τ_a,

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

The following implication thus holds for any $n \geqslant \tau_a$:

$$\sum_{i=1}^{n} Z_{a,i} \leqslant \ln\left(\frac{T}{Kn}\right) \qquad \text{implies} \qquad \sum_{i=1}^{n} Z_{a,i} \leqslant \ln\left(\frac{T}{Kn}\right) \leqslant \ln\left(\frac{T}{K\tau_a}\right) \leqslant \sum_{i=1}^{\tau_a} Z_{a,i}\,. \tag{59}$$

Hence, in this case,

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

This, together with a breakdown according to the values of τ_a (the case $\tau_a = T$ does not contribute to the expectation) and the independence between $\{\tau_a = k\}$ and $Z_{a,k+1}, \ldots, Z_{a,T}$, yields

$$\begin{aligned}&\mathbb{E}\left[\sum_{n=\tau_a+1}^{T} \mathbb{1}_{\left\{\sum_{i=1}^n Z_{a,i} \leqslant \ln(T/(Kn))\right\}} \right] = \mathbb{E}\left[\mathbb{1}_{\{\tau_a < T\}} \sum_{n=\tau_a+1}^{T} \mathbb{1}_{\left\{\sum_{i=1}^n Z_{a,i} \leqslant \ln(T/(Kn))\right\}} \right] \\ &\leqslant \mathbb{E}\left[\mathbb{1}_{\{\tau_a < T\}} \sum_{n=\tau_a+1}^{T} \mathbb{1}_{\left\{\sum_{i=\tau_a+1}^n Z_{a,i} < 0\right\}} \right] = \sum_{k=1}^{T-1} \mathbb{E}\left[\mathbb{1}_{\{\tau_a = k\}} \sum_{n=k+1}^{T} \mathbb{1}_{\left\{\sum_{i=k+1}^n Z_{a,i} < 0\right\}} \right] \\ &= \sum_{k=1}^{T-1} \sum_{n=k+1}^{T} \mathbb{P}[\tau_a = k]\ \mathbb{P}\left[\sum_{i=k+1}^{n} Z_{a,i} < 0 \right] \\ &= \sum_{k=1}^{T-1} \mathbb{P}[\tau_a = k] \left(\underbrace{\sum_{n=k+1}^{T} \mathbb{P}\left[\sum_{i=k+1}^{n} Z_{a,i} < 0 \right]}_{\text{we show below } \leqslant \beta, \text{ see (62)}} \right) \leqslant \beta\,,\end{aligned} \tag{60}$$

where β was defined in (58).

Indeed, we resort to Remark 24 of Section B.4, for the $n-k$ variables $Z_{a,k+1},\dots,Z_{a,n}$ and $x=0$; we legitimately do so as $\mu^\star-\delta>\mu_a$ by the imposed condition $\delta<\Delta_a/2$. Thus, 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 have

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

where the third inequality follows from (19) and the condition $\delta\leqslant(1-\mu^\star)\,\mathcal{K}_{\inf}(\nu_a,\mu^\star)/2$ that was imposed:

$$\mathcal{K}_{\inf}(\nu_a,\mu^\star-\delta)\geqslant\mathcal{K}_{\inf}(\nu_a,\mu^\star)-\frac{\delta}{1-\mu^\star}\geqslant\frac{\mathcal{K}_{\inf}(\nu_a,\mu^\star)}{2}\,.\tag{61}$$

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

$$\sum_{n=k+1}^{T}\mathbb{P}\left[\sum_{i=k+1}^{n}Z_{a,i}\leqslant 0\right]\leqslant\underbrace{\frac{1}{1-\mathrm{e}^{-1/4}}}_{\leqslant 5}+\frac{1}{1-\mathrm{e}^{-\mathcal{K}_{\inf}(\nu_a,\mu^\star)^2/(8\gamma_\star)}}\,,\tag{62}$$

which is the inequality claimed in (60).

It only remains to prove Lemma 27.

Proof of Lemma 27 This lemma 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 complete proof, with no significant additional merit compared to the original 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 entails 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 exactly 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, that

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

and thus,

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

In addition, when $\tau < T/K$, and again by definition of τ,

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

Applying the increasing function W to both 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

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

which concludes the proof. ■

D. Proof of the Variational Formula (Lemma 21)

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 that does not rely on discrete approximations of general distributions.

But before we do so, 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

$$\mathrm{KL}(\nu,\nu') = \sup\Big\{\mathbb{E}_\nu[Y] - \ln\mathbb{E}_{\nu'}\big[\mathrm{e}^Y\big] :\ Y \text{ s.t. } \mathbb{E}_{\nu'}[\mathrm{e}^Y] < +\infty\Big\}\,, \tag{63}$$

where (only here and in the next few lines) we index the expectation with respect to the assumed distribution of the random variable Y. In particular, denoting by X the identity over $[0,1]$ and considering, for $\lambda \in [0,1]$, the variables bounded from above

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

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

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

Hence, for these distributions ν',

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

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

$$\mathcal{K}_{\inf}(\nu,\mu) \geqslant \sup_{\lambda\in[0,1]} \mathbb{E}_\nu\left[\ln\left(1 - \lambda\frac{X-\mu}{1-\mu}\right)\right]. \tag{64}$$

Outline. We now only need to prove the converse inequality to get the rewriting (45) of Lemma 21, which we will do in Section D.2. Before that, in Section D.1, we prove the second statement of Lemma 21 together with several useful facts for the proof provided in Section D.2, including the fact that the supremum in the right-hand side of (64) is achieved. We conclude in Section D.3 with an alternative (sketch of) proof of the inequality (64), not relying on the variational formula (63) for the Kullback-Leibler divergences.

D.1 A Function Study

Let X denote a random variable with distribution $\nu \in \mathcal{P}[0,1]$. We recall that $\mu \in (0,1)$. The following function is well defined:

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

Indeed, since $X \in [0,1]$, the random variable $\ln\big(1 - \lambda(X-\mu)/(1-\mu)\big)$ is bounded from above by $\ln\big(1+\lambda\mu/(1-\mu)\big)$. Hence, H is well defined. For $\lambda \in [0,1)$, the considered random variable is bounded from below by $\ln(1-\lambda)$, hence H takes finite values. For $\lambda = 1$, we possibly have that $H(1)$ equals $-\infty$ (this is the case in particular when $\nu\{1\} > 0$).

We begin by a study of the function H.

Lemma 28 *Assume $\mu \in (0,1)$. The function H is continuous and strictly concave on $[0,1]$, differentiable at least on $[0,1)$, and its derivative $H'(1)$ can be defined at 1, with $H'(1) \in \mathbb{R} \cup \{-\infty\}$. We have the closed-form expression: for all $\lambda \in [0,1]$,*

$$H'(\lambda) = -\mathbb{E}\left[\left(\frac{X-\mu}{1-\mu}\right)\frac{1}{1-\lambda\frac{X-\mu}{1-\mu}}\right] = \frac{1}{\lambda}\left(1 - \mathbb{E}\left[\frac{1}{1-\lambda\frac{X-\mu}{1-\mu}}\right]\right). \tag{65}$$

It reaches a unique maximum over $[0,1]$, denoted by $\lambda^\star$,

$$\operatorname*{arg\,max}_{0\leqslant\lambda\leqslant 1} H(\lambda) = \{\lambda^\star\},$$

that satisfies $\lambda^\star > 0$ and at which $H'(\lambda^\star) = 0$ if $\lambda^\star \in (0,1)$ and $H'(\lambda^\star) \geqslant 0$ if $\lambda^\star = 1$.

Moreover, under the additional condition $\mathrm{E}(\nu) < \mu$,

$$\mathbb{E}\left[\frac{1}{1-\lambda^\star\frac{X-\mu}{1-\mu}}\right] = 1 \text{ if } \lambda^\star \in (0,1) \qquad \text{and} \qquad \mathbb{E}\left[\frac{1}{1-\lambda^\star\frac{X-\mu}{1-\mu}}\right] = \mathbb{E}\left[\frac{1-\mu}{1-X}\right] \leqslant 1 \text{ if } \lambda^\star = 1\,.$$

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

Note that $\mathcal{K}_{\inf}(\nu,\mu) = 0$ when $\mu \leqslant \mathrm{E}(\nu)$. In this case, necessarily $\lambda^\star = 0$ (there is a unique maximum) and we still have

$$\mathbb{E}\left[\frac{1}{1-\lambda^\star\frac{X-\mu}{1-\mu}}\right] = 1\,.$$

This concludes the proof of the statement (46) of Lemma 21.

Proof For the continuity of H, we note that the discussion before the statement of the lemma entails that the random variables $\ln\bigl(1-\lambda(X-\mu)/(1-\mu)\bigr)$ are uniformly bounded on ranges of the form $[0,\lambda_0]$ for $\lambda_0 < 1$. By a standard continuity theorem under the integral sign, this proves that H is continuous on $[0,1)$. For the continuity at 1, we separate the $H(\lambda)$ and $H(1)$ into two pieces, for which monotone convergences take place:

$$\lim_{\lambda\to 1}\mathbb{E}\left[\ln\left(1-\lambda\frac{X-\mu}{1-\mu}\right)\mathbb{1}_{\{X\in[0,\mu]\}}\right] = \mathbb{E}\left[\ln\left(\frac{1-X}{1-\mu}\right)\mathbb{1}_{\{X\in[0,\mu]\}}\right],$$
$$\lim_{\lambda\to 1}\mathbb{E}\left[\ln\left(1-\lambda\frac{X-\mu}{1-\mu}\right)\mathbb{1}_{\{X\in(\mu,1]\}}\right] = \mathbb{E}\left[\ln\left(\frac{1-X}{1-\mu}\right)\mathbb{1}_{\{X\in(\mu,1]\}}\right],$$

where the first expectation is finite (but the second may equal $-\infty$).

The strict concavity of H on $[0,1]$ follows from the one of ln on $(0,1]$ and from the continuity of H on $[0,1]$.

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

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

Indeed as long as $\lambda < 1$, the random variables in the expectations above are uniformly bounded on ranges of the form $[0, \lambda_0]$ for $\lambda_0 < 1$, so that we may invoke a standard differentiation theorem under the integral sign. A similar argument of double monotone convergences as above shows that $H'(\lambda)$ has a limit value as $\lambda \to 1$, with

$$\lim_{\lambda \to 1} H'(\lambda) = -\mathbb{E}\left[\frac{X-\mu}{1-X}\right].$$

By a standard limit theorem on derivatives, when the above value is finite, H is differentiable at 1 and $H'(1)$ equals the limit above; otherwise, H is not differentiable at 1 but we still denote $H'(1) = -\infty$.

Since H is strictly concave on $[0,1]$ and continuous, it reaches its maximum exactly once on $[0,1]$. Now, given the condition $\mathrm{E}(\nu) < \mu$, we have

$$H'(0) = -\frac{\mathrm{E}(\nu) - \mu}{1-\mu} > 0\,.$$

As H is concave, H' is decreasing: either $H'(1) \geqslant 0$ and H reaches its maximum at $\lambda^\star = 1$, or $H'(1) < 0$ and H reaches its maximum on the open interval $(0,1)$. It may be proved (by a standard continuity theorem under the integral sign) that H' is continuous on $[0,1)$, that is, that H is continuously differentiable on $[0,1)$. In the case $H'(1) < 0$, the derivative at the maximum therefore satisfies $H'(\lambda^\star) = 0$.

Substituting the expressions (65) for $H'(\lambda^\star)$ provides the final equality or inequality to 1 stated (depending on whether $\lambda^\star < 1$ or $\lambda^\star = 1$). In the case $\lambda^\star = 1$, we thus have $1 - \mu \in (0,1)$ and $1 - X \in [0,1]$ with

$$\mathbb{E}\left[\frac{1-\mu}{1-X}\right] \leqslant 1\,;$$

this prevents X from taking the value 1 with positive probability (otherwise, the expectation would be $+\infty$). Put differently, $\nu\{1\} = 0$. ■

D.2 Proof of $\leqslant$ in Equality (45)

We keep the notation introduced in the previous section. To prove this inequality, by the rewriting of $\mathcal{K}_{\inf}(\nu, \mu)$ stated in Corollary 12, it is enough to show that there exists a probability measure ν' on $[0,1]$ such that $\mathrm{E}(\nu') \geqslant \mu$ and $\nu \ll \nu'$ and

$$\mathrm{KL}(\nu, \nu') \leqslant \mathbb{E}\left[\ln\left(1 - \lambda^\star \frac{X-\mu}{1-\mu}\right)\right]. \tag{66}$$

Given the definition of the KL divergence, it suffices to find a probability measure ν' on $[0,1]$ such that $\mathrm{E}(\nu') \geqslant \mu$ and $\nu \ll \nu'$ and

$$\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}(x) = 1 - \lambda^\star \frac{x-\mu}{1-\mu} \qquad \nu\text{–a.s.} \tag{67}$$

It can be shown (proof omitted as this statement is only given to explain the intuition behind the proof) that

$$\frac{\mathrm{d}\nu}{\mathrm{d}\nu'} > 0 \quad \nu\text{–a.s.}, \qquad \text{with} \qquad \frac{\mathrm{d}\nu'_{\mathrm{ac}}}{\mathrm{d}\nu} = \left(\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}\right)^{-1} \quad \nu\text{–a.s.}, \tag{68}$$

where ν'_{ac} denotes the absolute part of ν' with respect to ν. This is why we introduce the measure ν' on $[0,1]$ defined by

$$\mathrm{d}\nu'(x) = \underbrace{\frac{1}{1-\lambda^\star \frac{x-\mu}{1-\mu}}}_{\geqslant 0} \mathrm{d}\nu(x) + \left(1 - \mathbb{E}\left[\frac{1}{1-\lambda^\star \frac{X-\mu}{1-\mu}}\right]\right) \mathrm{d}\delta_1(x)\,, \tag{69}$$

where δ_1 denotes the Dirac point-mass distribution at 1 and where X denotes a random variable with distribution ν. The measure ν' is a probability measure as by Lemma 28,

$$\mathbb{E}\left[\frac{1}{1-\lambda^\star \frac{X-\mu}{1-\mu}}\right] \leqslant 1\,.$$

Now, we show first that $\nu \ll \nu'$ with the density (67). We do so by distinguishing two cases. If $\lambda^\star \in [0,1)$, then by the last statement of Lemma 28, the probability measure ν' is actually defined by

$$\mathrm{d}\nu'(x) = \underbrace{\frac{1}{1-\lambda^\star \frac{x-\mu}{1-\mu}}}_{>0} \mathrm{d}\nu(x)\,,$$

and the strict positivity underlined in the equality above ensures the desired result by a standard theorem on Radon-Nikodym derivatives. In that case, ν and ν' are actually equivalent measures: $\nu \ll \nu'$ and $\nu' \ll \nu$. If $\lambda^\star = 1$, then again by Lemma 28, we know that ν does not put any probability mass at 1. The strict positivity of $f(x) = 1-(x-\mu)/(1-\mu)$ on $[0,1)$ and the fact that $\nu\{1\} = 0$ ensure the first equality below: for all Borel subsets A of $[0,1]$,

$$\nu(A) = \int \mathbb{1}_A\, f \frac{1}{f}\, \mathrm{d}\nu = \int \mathbb{1}_A\, f \left(\frac{1}{f}\, \mathrm{d}\nu + r\, \mathrm{d}\delta_1\right) = \int \mathbb{1}_A\, f\, \mathrm{d}\nu'$$

while the second equality follows from $f(1) = 0$ and the third equality is by definition of ν'. Put differently, $\nu \ll \nu'$ with the density f claimed in (67). In that case, $\nu \ll \nu'$ but ν' is not necessarily absolutely continuous with respect to ν.

We conclude this proof by showing that $\mathrm{E}(\nu') \geqslant \mu$. We recall that Lemma 28 ensures

$$\mathbb{E}\left[\left(\frac{X-\mu}{1-\mu}\right)\frac{1}{1-\lambda^\star \frac{X-\mu}{1-\mu}}\right] = -H'(\lambda^\star)$$

$$\text{and} \qquad \mathbb{E}\left[\frac{1}{1-\lambda^\star \frac{X-\mu}{1-\mu}}\right] = 1 - \lambda^\star\, H'(\lambda^\star)\,,$$

where X denotes a random variable with distribution ν and where both expectations are well defined (possibly with values $+\infty$ when $\lambda^\star = 1$). Therefore,

$$\begin{aligned}
\mathrm{E}(\nu') &= \overbrace{\mathbb{E}\left[\frac{X}{1-\lambda^\star\frac{X-\mu}{1-\mu}}\right]}^{\text{“}\nu\text{ part of }\nu'\text{”}} + \overbrace{\left(1-\mathbb{E}\left[\frac{1}{1-\lambda^\star\frac{X-\mu}{1-\mu}}\right]\right)}^{\text{“}\delta_1\text{ part of }\nu'\text{”}} \\
&= (1-\mu)\,\mathbb{E}\left[\left(\frac{X-\mu}{1-\mu}\right)\frac{1}{1-\lambda^\star\frac{X-\mu}{1-\mu}}\right] + \mu\,\mathbb{E}\left[\frac{1}{1-\lambda^\star\frac{X-\mu}{1-\mu}}\right] + \left(1-\mathbb{E}\left[\frac{1}{1-\lambda^\star\frac{X-\mu}{1-\mu}}\right]\right) \\
&= -(1-\mu)\,H'(\lambda^\star) + \mu\big(1-\lambda^\star\,H'(\lambda^\star)\big) + \lambda^\star\,H'(\lambda^\star) \\
&= \mu - \big((1-\mu)\,(1-\lambda^\star)\,H'(\lambda^\star)\big)\,,
\end{aligned}$$

where the first equality is justified in the case $\lambda^\star = 1$ by the same arguments of monotone convergence as in the proof of Lemma 28. All in all, we have $\mathrm{E}(\nu') \geqslant \mu$ as desired if and only if $(1-\lambda^\star)\,H'(\lambda^\star) \leqslant 0$. This is the case as we actually have $(1-\lambda^\star)\,H'(\lambda^\star) = 0$ in all cases, i.e., whether $\lambda^\star = 1$ or $\lambda^\star \in [0,1)$.

D.3 Alternative Proof of $\geqslant$ in Equality (45)

We use the notation of Sections D.1 and D.2 and prove the desired inequality (64), that is, the $\geqslant$ part of the equality (45), without resorting to the variational formula (63) for the Kullback-Leibler divergences. Actually, we only provide a sketch of proof and omit proofs of some facts about Radon-Nikodym derivatives.

Let $\nu'' \in \mathcal{P}[0,1]$ be such that $\mathrm{E}(\nu'') > \mu$ and $\nu \ll \nu''$; with no loss of generality, we assume that $\mathrm{KL}(\nu,\nu'') < +\infty$. By the definition (69) of ν' and the discussion following this definition, the divergence $\mathrm{KL}(\nu,\nu')$ equals the maximum of the continuous function H over $[0,1]$ and therefore also satisfies $\mathrm{KL}(\nu,\nu') < +\infty$. We denote by $\mathbb{L}_1(\nu)$ the set of ν–integrable random variables. That the divergences $\mathrm{KL}(\nu,\nu'')$ and $\mathrm{KL}(\nu,\nu')$ are finite exactly means that

$$\left|\ln\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}\right| \in \mathbb{L}_1(\nu) \qquad \text{and} \qquad \left|\ln\frac{\mathrm{d}\nu}{\mathrm{d}\nu''}\right| \in \mathbb{L}_1(\nu)\,.$$

Hence,

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

Now, by (67),

$$\ln\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}(x) = \ln\left(1-\lambda^\star\frac{x-\mu}{1-\mu}\right) \qquad \nu\text{–a.s.},$$

and by (68),

$$-\ln\frac{\mathrm{d}\nu}{\mathrm{d}\nu''} = \ln\frac{\mathrm{d}\nu''_{\mathrm{ac}}}{\mathrm{d}\nu}(x) \qquad \nu\text{–a.s.},$$

so that

$$\begin{aligned}
\mathrm{KL}(\nu,\nu'') - \mathrm{KL}(\nu,\nu') &= -\int \ln\left(\left(1-\lambda^\star\frac{x-\mu}{1-\mu}\right)\frac{\mathrm{d}\nu''_{\mathrm{ac}}}{\mathrm{d}\nu}(x)\right)\mathrm{d}\nu(x) \\
&\geqslant -\ln\left(\int \underbrace{\left(1-\lambda^\star\frac{x-\mu}{1-\mu}\right)}_{\geqslant 0}\underbrace{\frac{\mathrm{d}\nu''_{\mathrm{ac}}}{\mathrm{d}\nu}(x)\,\mathrm{d}\nu(x)}_{\mathrm{d}\nu''_{\mathrm{ac}}(x)}\right) \\
&\geqslant -\ln\left(\underbrace{\int\left(1-\lambda^\star\frac{x-\mu}{1-\mu}\right)\mathrm{d}\nu''(x)}_{\leqslant 1 \text{ as } \mathrm{E}(\nu'')>\mu}\right) \geqslant 0
\end{aligned}$$

where Jensen's inequality provided the first inequality, while the second one followed by increasing the integral in the logarithm. Taking the infimum over distributions $\nu'' \in \mathcal{P}[0,1]$ with $\mathrm{E}(\nu'') > \mu$ and $\nu \ll \nu''$ and $\mathrm{KL}(\nu,\nu'') < +\infty$, we proved

$$\mathcal{K}_{\inf}(\nu,\mu) - \mathrm{KL}(\nu,\nu') \geqslant 0\,,$$

which was the desired result.

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.

J.-Y Audibert, S. Bubeck, and G. Lugosi. Minimax policies for combinatorial prediction games. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT'2011)*, volume 19 of *PMLR*, pages 107–132, 2011.

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.

L. Besson. SMPyBandits: Open-source Python package for Single- and Multi-Players multi-armed Bandits algorithms, 2019. Version 140, see `https://github.com/SMPyBandits/SMPyBandits/issues/140`.

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. In *Advances in Neural Information Processing Systems (NeurIPS'13)*, volume 26, pages 638–646, 2013.

S. Bubeck and A. Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT'2012)*, volume 23 of *PMLR*, pages 42.1–42.23, 2012.

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.

I. Csiszar. Sanov property, generalized I-projection and a conditional limit theorem. *The Annals of Probability*, 12(3):768–793, 1984.

R. Degenne and V. Perchet. Anytime optimal algorithms in stochastic multi-armed bandits. In *Proceedings of the 33rd International Conference on Machine Learning (ICML'2016)*, volume 48 of *PMLR*, pages 1587–1595, 2016.

J.L. Doob. *Stochastic Processes*. Wiley Publications in Statistics. John Wiley & Sons, 1953.

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'2011)*, volume 19 of *PMLR*, pages 359–376, 2011.

A. Garivier, H. Hadiji, P. Ménard, and G. Stoltz. KL-UCB-switch: optimal regret bounds for stochastic bandits from both a distribution-dependent and a distribution-free viewpoints, 2018. Preprint, arXiv:1805.05071v1, May 2018.

A. Garivier, P. Ménard, and G. Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

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.

E. Kaufmann, O. Cappé, and A. Garivier. On Bayesian upper confidence bounds for bandit problems. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AIStats'2012)*, volume 22 of *PMLR*, pages 592–600, 2012.

N. Korda, E. Kaufmann, and R. Munos. Thompson sampling for 1–dimensional exponential family bandits. In *Advances in Neural Information Processing Systems (NeurIPS'13)*, volume 26, pages 1448–1456, 2013.

T.L. Lai. Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, 15(3):1091–1114, 1987.

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, 2016. Preprint, arXiv:1603.08661.

T. Lattimore. Refining the confidence level for optimistic bandit strategies. *Journal of Machine Learning Research*, 19(20):1–32, 2018.

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'2011)*, volume 19 of *PMLR*, pages 497–514, 2011.

P. Ménard and A. Garivier. A minimax and asymptotically optimal algorithm for stochastic bandits. In *Procedings of the 28th International Conference on Algorithmic Learning Theory (ALT'2017)*, volume 76 of *PMLR*, pages 223–237, 2017.

G. Simons, L. Yang, and Y.-C. Yao. Doob, Ignatov and optional skipping. *Annals of Probability*, 30(4):1933–1958, 2002.

C. Szepesvári and T. Lattimore. *Bandit Algorithms*. Cambridge University Press, 2020.

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.

J. Zimmert and Y. Seldin. An optimal algorithm for stochastic and adversarial bandits. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AIStats'19)*, volume 89 of *PMLR*, pages 467–475, 2019.

J. Zimmert and Y. Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(28):1–49, 2021.