Journal of Machine Learning Research 24 (2023) 1-33 Submitted 2/21; Revised 6/22; Published 1/23

Adaptation to the Range in K–Armed Bandits

Hédi Hadiji HEDI.HADIJI@MATH.U-PSUD.FR
Gilles Stoltz GILLES.STOLTZ@MATH.U-PSUD.FR
Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France

Editor: Ambuj Tewari

Abstract

We consider stochastic bandit problems with K arms, each associated with a distribution supported on a given finite range $[m, M]$. We do not assume that the range $[m, M]$ is known and show that there is a cost for learning this range. Indeed, a new trade-off between distribution-dependent and distribution-free regret bounds arises, which prevents from simultaneously achieving the typical $\ln T$ and $\sqrt{T}$ bounds. For instance, a $\sqrt{T}$ distribution-free regret bound may only be achieved if the distribution-dependent regret bounds are at least of order $\sqrt{T}$. We exhibit a strategy achieving the rates for regret imposed by the new trade-off.

Keywords: multiarmed bandits, adversarial learning, cumulative regret, information-theoretic proof techniques

1. Introduction

Stochastic multi-armed bandits form a standard setting to deal with sequential decision-making problems like the design of clinical trials—one of the first applications mentioned—or online advertisement and online revenue management.

Except for notable exceptions discussed below, virtually all articles on stochastic K–armed bandits either assume that distributions of the arms belong to some parametric family—often, one-dimensional exponential families—or are sub-Gaussian with a known parameter σ^2. Among the latter category, the case of the non-parametric family of distributions supported on a known range $[m, M]$ is of particular interest to us.

We show that the knowledge of the range $[m, M]$ is a crucial information and that facing bounded bandit problems but ignoring the bounds m and M is much harder. We do so by studying what may be achieved and what cannot be achieved anymore when the range $[m, M]$ is unknown and the strategies need to learn it. We call this problem adaptation to the range, or scale-free regret minimization. Why this problem is important and why we considered it is explained in Section 1.2.

More precisely, we prove that adaptation to the range is actually possible but that it has a cost: our most striking result (in Section 2.3) is a severe trade-off between the scale-free distribution-dependent and distribution-free regret bounds that may be achieved. For instance, no strategy adaptive to the range can simultaneously achieve distribution-dependent regret bounds of order $\ln T$ and distribution-free regret bounds of order $\sqrt{T}$ up to polynomial factors; this is in contrast with the case of a known range where simple strategies like UCB strategies (by Auer et al., 2002a) do so. Our general trade-off shows, for instance, that if one wants to keep the same $\sqrt{T}$ order of magnitude for the scale-free distribution-free regret bounds, then the best scale-free distribution-dependent rate that may be achieved is $\sqrt{T}$.

We also provide (in Section 4) a strategy, based on exponential weights, that adapts to the range and obtains optimal distribution-dependent and distribution-free regret bounds in the eyes of the exhibited trade-off: these are of respective orders $T^{1-\alpha}$ and T^{α}, where $\alpha \in [1/2, 1)$ is a parameter of the strategy.

1.1 Literature Review

Optimal scale-free regret minimization under full monitoring for adversarial sequences is offered by the AdaHedge strategy by De Rooij et al. (2014), which we will use as a building block in Section 4.

For stochastic bandits, the main difficulty in adaptation to the range is the adaptation to the upper end M (see Remark 4); this is why Honda and Takemura (2015) could provide optimal $\ln T$ distribution-dependent regret bounds for payoffs lying in ranges of the form $(-\infty, M]$, with a known M. Lattimore (2017) considers models of distributions with a known bound on their kurtosis, which is a scale-free measure of the skewness of the distributions; he provides a scale-free algorithm based on the median-of-means estimators, with $\ln T$ distribution-dependent regret bounds. However, bounded bandits can have an arbitrarily high kurtosis, so our settings are not directly comparable. Cowan and Katehakis (2015) study adaptation to the range but in the restricted case of uniform distributions over unknown intervals. They provide optimal $\ln T$ distribution-dependent regret bounds for that specific model: in their model, the cost for adaptation is mild and lies only in the multiplicative constant before the $\ln T$. In the setting of bounded bandits, we show that distribution-dependent regret bounds must be larger than $\ln T$, but the argument of Lattimore (2017, Remark 8) entails that any regret rate larger than $\ln T$, e.g., $(\ln T) \ln \ln T$, may be achieved. Similar results by Cowan et al. (2018) for Gaussian distributions with unknown means and variances were also obtained.

Finally, on the front of adversarial bandits, no prior work discussed adaptation to the range, to the best of our knowledge.

Additional important references performing adaptation in some other sense for stochastic and adversarial K–armed bandits are discussed now, including some follow-up work to this article.

Adaptation to the effective range or to unbounded ranges in adversarial bandits. Gerchinovitz and Lattimore (2016) show that it is impossible to adapt to the so-called effective range in adversarial bandits. A sequence of rewards has effective range smaller than b if for all rounds t, rewards $y_{t,a}$ at this round all lie in an interval of the form $[m_t, M_t]$ with $M_t - m_t \leqslant b$. The lower bound they exhibit relies on a sequence of changing intervals of fixed size. This problem is thus different from our setting. See also positive results—regret upper bounds under additional assumptions—by Cesa-Bianchi and Shamir (2018) and Thune and Seldin (2018) for adaptation to the effective range.

Allenberg et al. (2006) deal with unbounded ranges $[m_t, M_t]$ in adversarial bandits and other partial monitoring settings, where, e.g., $M_t = -m_t = t^{\beta}$ for some $\beta > 0$. They provide regret upper bounds scaling with $t^{\beta/2}$ when β is known, but do not detail the price to pay for not knowing β—though they suggest to resort to a doubling trick in that case.

Adaptation to the variance. Audibert et al. (2009) consider a variant of UCB called UCB-V, which adapts to the unknown variance. Its analysis assumes that rewards lie in a known range $[0, M]$. The results crucially use Bernstein's inequality, which we state as Reminder 3 in Appendix C. As Bernstein's inequality holds for random variables with supports in $(-\infty, M]$, the analysis of UCB-V might perhaps be extended to this case as well. Deviation bounds in Bernstein's inequality contain two terms, a main term scaling with the standard deviation, and a remainder term, scaling with M. This remainder term, which seems harmless, is actually a true issue when M is not known, as shown by the results of the present article.

Adaptation to other criteria. Wei and Luo (2018), Zimmert and Seldin (2019), Bubeck et al. (2018), and many more, provide strategies for adversarial bandits with rewards in a known range, say $[0, 1]$, and adapting to additional regularity in the data, like small variations or stochasticity of the data—but never to the range itself.

Follow-up works. Following an earlier version of this work, further research on range-adaptive bandit algorithms has been conducted. Baudry et al. (2021) bypass our lower bound by imposing minimal extra conditions on the reward distributions, avoiding the heavy-tail construction from Theorem 3; in this context, they provide a fully range-adaptive algorithm. For adversarial multi-armed bandits, Putta and Agrawal (2022) recover some small-loss bounds while being agnostic to the range, at the cost of degraded worst-case guarantees, and Huang et al. (2021) obtain similar results under delayed feedback.

1.2 Why Studying Adaptation to the Range for Finite-Armed Bandits

We encountered the problem of learning the range $[m, M]$ of bandits problems when designing bandit algorithms for continuum-armed problems, see Hadiji (2019). Therein, arms are indexed by some bounded interval $\mathcal{I}$, and the mean-payoff function $f : \mathcal{I} \to \mathbb{R}$ is assumed to be smooth enough, e.g., Hölderian-smooth, with unknown regularity parameters L and β. The mean-payoff function f has a bounded range, as it is continuous over a bounded interval. To optimally learn these smoothness parameters, histogram-reductions of continuum-armed bandit problems to finite-armed bandits problems, of proper bandwidth, are performed (à la Kleinberg, 2004), by zooming out. Any reasonable K–armed bandit algorithm may be used in this algorithmic scheme. However, given this reduction, it had to be assumed that the range $[m, M]$ of f is known, as all K–armed bandit algorithms we were aware of assumed that the range of the distributions over the arms was known. To get more complete adaptivity results in the continuum-armed case and be able to ignore the range of the mean-payoff function f, it was necessary and sufficient to deal with the similar issue of range adaptivity in the case of finitely many arms—which this article provides.

We also believe that exhibiting impossibility results, like the existence of a severe trade-off between distribution-dependent and distribution-free regret bounds in the case of the model of bounded distributions with an unknown range, has consequences beyond that model. This impossibility result holds in particular for all larger models, like non-parametric models containing all distributions over the entire real line $\mathbb{R}$ satisfying certain assumptions on their tails to make sure that they are not too large. We therefore provide some intrinsic limitation to learning in K–armed stochastic bandits.

The techniques introduced extend to more complex settings, like linear bandits; see an extended version of this article [arXiv:2006.03378], Appendix D.

2. Setting and Main Results

We consider finitely-armed stochastic bandits with bounded and possibly signed rewards. More precisely, $K \geqslant 2$ arms are available; we denote by $[K]$ the set $\{1, \ldots, K\}$ of these arms. With each arm a is associated a Borel probability distribution ν_a lying in some known model $\mathcal{D}$; a model is a set of Borel probability distributions over $\mathbb{R}$ with a first moment. The models of interest in this article are discussed below; in the sequel, we only consider Borel distributions even though we will omit this specification.

A bandit problem in $\mathcal{D}$ is a K–vector of probability distributions in $\mathcal{D}$, denoted by $\underline{\nu} = (\nu_a)_{a \in [K]}$. The player knows $\mathcal{D}$ but not $\underline{\nu}$. As is standard in this setting, we denote by $\mu_a = \mathrm{E}(\nu_a)$ the mean payoff provided by an arm a. An optimal arm and the optimal mean payoff are respectively given by $a^\star \in \mathrm{argmax}_{a \in [K]} \mu_a$ and $\mu^\star = \max_{a \in [K]} \mu_a$. Finally, $\Delta_a = \mu^\star - \mu_a$ denotes the gap of an arm a.

The online learning game goes as follows: at round $t \geqslant 1$, the player picks an arm $A_t \in [K]$, possibly at random according to a probability distribution $p_t = (p_{t,a})_{a \in [K]}$ based on an auxiliary randomization U_{t-1}, e.g., uniformly distributed over $[0,1]$, and then receives and observes a reward Z_t drawn independently at random according to the distribution ν_{A_t}, given A_t. More formally, a strategy of the player is a sequence of measurable mappings from the observations to the action set, $(U_0,\, Z_1, U_1,\, \ldots,\, Z_{t-1}, U_{t-1}) \mapsto A_t$. At each given time $T \geqslant 1$, we measure the performance of a strategy through its expected regret:

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

where we used the tower rule for the first equality and defined $N_a(T)$ as the number of times arm a was pulled between time rounds 1 and T.

Doob's optional skipping (see Doob, 1953, Chapter III, Theorem 5.2, page 145 for the original reference, see also Chow and Teicher, 1988, Section 5.3 for a more recent reference) shows that we may assume that i.i.d. sequences of rewards $(Y_{t,a})_{t \geqslant 1}$ are drawn beforehand, independently at random, for each arm a and that the obtained payoff at round $t \geqslant 1$ given the choice A_t equals $Z_t = Y_{t,A_t}$. We will use this second formulation in the rest of the paper as it is the closest to the one of oblivious individual sequences described later in Section 4.1. We may then assume that the auxiliary randomizations $U_0, U_1, \ldots$ are i.i.d. random variables independent from the $(Y_{t,a})_{t \geqslant 1}$ and distributed according to a uniform distribution over $[0,1]$.

Model: bounded signed rewards with unknown range. For a given range $[m, M]$, where $m < M$ are two real numbers, not necessarily nonnegative, we denote by $\mathcal{D}_{m,M}$ the set of probability distributions supported on $[m, M]$. Then, the model corresponding to distributions with a bounded but unknown range is the union of all such $\mathcal{D}_{m,M}$:

$$\mathcal{D}_{-,+} = \bigcup_{m,M \in \mathbb{R} : m<M} \mathcal{D}_{m,M} \,.$$

2.1 Adaptation to the Range: Concept of Scale-Free Regret Bounds

Regret scales with the range length $M-m$, thus regret bounds involve a multiplicative factor $M-m$. We therefore consider such bounds divided by the scale factor $M-m$ and call them scale-free regret bounds. We denote by $\mathbb{N}$ the set of natural integers; rates on regret bounds will be given by functions $\Phi : \mathbb{N} \to [0,+\infty)$. We define adaptation to the unknown range in Definitions 1 and 2 below.

Definition 1 (Scale-free distribution-free regret bounds) *A strategy for stochastic bandits is adaptive to the unknown range of payoffs with a scale-free distribution-free regret bound* $\Phi_{\mathrm{free}} : \mathbb{N} \to [0,+\infty)$ *if for all real numbers* $m < M$*, the strategy ensures, without the knowledge of* m *and* M*:*

$$\forall \underline{\nu} \text{ in } \mathcal{D}_{m,M}, \quad \forall T \geqslant 1, \qquad R_T(\underline{\nu}) \leqslant (M-m)\, \Phi_{\mathrm{free}}(T) \,.$$

We show in Section 4 that adaptation to the unknown range may indeed be performed in the sense of Definition 1, with a scale-free distribution-free regret bound of order $\sqrt{KT \ln K}$. The latter is optimal up to maybe a factor of $\sqrt{\ln K}$ as Auer et al. (2002b) provided a lower bound $(1/20)\min\bigl\{\sqrt{KT}, T\bigr\}$ on the regret of any strategy against individual sequences in $[0,1]^K$, thus for bandit problems in $\mathcal{D}_{0,1}$, thus for scale-free distribution-free regret bounds.

Definition 2 (Distribution-dependent rates for adaptation) *A strategy for stochastic bandits is adaptive to the unknown range of payoffs with a distribution-dependent rate* $\Phi_{\mathrm{dep}} : \mathbb{N} \to [0,+\infty)$ *if for all real numbers* $m < M$*, the strategy ensures, without the knowledge of* m *and* M*:*

$$\forall \underline{\nu} \text{ in } \mathcal{D}_{m,M}, \qquad \limsup_{T \to +\infty} \frac{R_T(\underline{\nu})}{\Phi_{\mathrm{dep}}(T)} < +\infty \,.$$

Put differently, the strategy ensures that $\limsup R_T(\underline{\nu})/\Phi_{\mathrm{dep}}(T) < +\infty$ *for all* $\underline{\nu} \in \mathcal{D}_{-,+}$.

Definition 2 does not add much to the classical notion of distribution-dependent rates on regret bounds, as the scale factor $M-m$ does not appear in the definition; it merely ensures that the strategy is not informed of the range. Also, we are only interested in rates of convergence here, not in the value of the finite limit of $R_T(\underline{\nu})/\Phi_{\mathrm{dep}}(T)$. This limit however heavily depends on $\underline{\nu}$, which justifies the terminology of distribution-dependent *rates* for adaptation Φ_{dep}.

In contrast, the bounds targeted in the distribution-free case have a finite-time, closed-form expression, which is why we did not speak of rates in that case and rather referred to scale-free distribution-free regret *bounds* Φ_{free}.

2.2 Scale-Free Distribution-Dependent Regret Bounds Considered in Isolation

We first explain the impact of ignoring the range on distribution-dependent regret bounds. What follows is discussed in greater detail in Appendix A as these results were already known or, at least, much expected.

When m and M are known, there exist several strategies ensuring

$$\forall \underline{\nu} \text{ in } \mathcal{D}_{m,M}, \qquad \limsup_{T \to +\infty} \frac{R_T(\underline{\nu})}{\ln T} < +\infty \,,$$

even with an optimal value of the limit; see the end of Section A.1.

Given Definition 2, one may therefore wonder whether $\Phi_{\text{dep}} = \ln$ is achievable as a distribution-dependent rate for adaptation to the range. Theorem 11 in Section A.2 and the comment before its statement provide a negative answer to this question.

However, a UCB-strategy with an increased exploration rate given by a non-decreasing function $\varphi \gg \ln$ was suggested by Lattimore (2017, Remark 8) in the context of Gaussian bandits. It also works well in the setting of bounded bandits: Theorem 13 in Section A.3 states that it is adaptive to the unknown range of payoffs with a distribution-dependent rate $\Phi_{\text{dep}} = \varphi$. That is, any rate that is larger than a logarithm may be achieved, including, for instance, $\varphi(t) = (\ln t) \ln \ln t$.

2.3 Simultaneous Scale-Free Regret Bounds

When the range $[m, M]$ of the payoffs is known, it is possible to simultaneously achieve optimal distribution-free bounds, of order $\sqrt{KT}$, and optimal distribution-dependent bounds, of order $\ln T$ with the optimal constant recalled in Reminder 1 of Appendix A.1; see the KL-UCB-switch strategy by Garivier et al. (2019a). Put differently, when the range of payoffs is known, one can achieve optimal asymptotic distribution-dependent regret bounds while not sacrificing finite-time guarantees. Simpler strategies like UCB strategies (see Auer et al., 2002a) also simultaneously achieve regret bounds of similar $\sqrt{T \ln T}$ and $\ln T$ orders of magnitude but with suboptimal constants. Zimmert and Seldin (2019) also provide a strategy, Tsallis-INF with $\alpha = 1/2$, that provides simultaneously distribution-dependent regret guarantees of order $\ln T$, with suboptimal constants though, and adversarial guarantees of order $\sqrt{KT}$, which are stronger than just distribution-free guarantees.

First main result: existence of a trade-off. Our first main result states that getting simultaneously these $\ln T$ and $\sqrt{T}$ rates is not possible anymore when the range of payoffs is unknown.

Theorem 3 *A strategy with a scale-free distribution-free regret bound $\Phi_{\text{free}}(T) = o(T)$ may only achieve distribution-dependent rates Φ_{dep} for adaptation satisfying*

$$\Phi_{\text{dep}}(T) \geqslant \frac{T}{\Phi_{\text{free}}(T)} \,.$$

More precisely, the regret of such a strategy is lower bounded as follows: for all $\underline{\nu}$ in $\mathcal{D}_{-,+}$,

$$\liminf_{T \to \infty} \frac{R_T(\underline{\nu})}{T / \Phi_{\text{free}(T)}} \geqslant \frac{1}{16} \sum_{a=1}^{K} \Delta_a \,. \tag{2}$$

The orders of magnitude of the scale-free distribution-free regret bounds $\Phi_{\text{free}}(T)$ range between the optimal $\sqrt{T}$ and the trivial T rates. The distribution-dependent rates Φ_{dep} for adaptation to the range are therefore at best $\sqrt{T}$ for strategies enjoying scale-free distribution-free regret bounds; $\ln T$ rates are excluded. More generally, Theorem 3 shows that there is a trade-off: to force faster distribution-dependent rates for adaptation, one must suffer worsened scale-free distribution-free regret bounds.

The proof of Theorem 3 is provided in Section 3. It actually provides a finite-time, but messy, lower bound on $R_T(\underline{\nu})/\bigl(T/\Phi_{\text{free}(T)}\bigr)$.

Second main result: achieving the trade-off. Our second main result consists of showing that the trade-off imposed by Theorem 3 may indeed be achieved. Section 4 will introduce a strategy, relying on a parameter $\alpha \in [1/2, 1)$ and called AHB—which stands for AdaHedge for K–armed Bandits with extra-exploration; see Algorithm 1. Theorems 7 and 9 show in particular that AHB adapts to the unknown range, satisfies a scale-free distribution-free regret bound

$$\Phi^{\text{AHB}}_{\text{free}}(T) = \left(3 + \frac{5}{\sqrt{1-\alpha}}\right)\sqrt{K \ln K}\; T^{\alpha} + 10K \ln K = \mathcal{O}(T^{\alpha})\,,$$

and achieves $\Phi^{\text{AHB}}_{\text{dep}}(T) = T/\Phi^{\text{AHB}}_{\text{free}}(T) = \mathcal{O}(T^{1-\alpha})$ as a distribution-dependent rate for adaptation. Like Zimmert and Seldin (2019), we are actually able to prove an adversarial regret bound, not only the mentioned distribution-free regret bound.

Even better, Theorem 9 states that for all $\underline{\nu}$ in $\mathcal{D}_{-,+}$,

$$\limsup_{T\to\infty} \frac{R_T(\underline{\nu})}{T/\Phi^{\text{AHB}}_{\text{free}(T)}} \leqslant \frac{12 \ln K}{1-\alpha} \sum_{a=1}^{K} \Delta_a\,. \tag{3}$$

Discussion. The distribution-dependent constants in the right-hand sides of (2) and (3) are proportional to the sums of the gaps,

$$G(\underline{\nu}) = \sum_{a=1}^{K} \Delta_a\,,$$

and differ from this sum only by distribution-free factors of $1/16$ and $(12 \ln K)/(1-\alpha)$. The quantity $G(\underline{\nu})$ appears as a new measure of the underlying geometry of information. We have no deep interpretation thereof, but may despite all underline a fundamental difference in our setting compared to the setting of a known range.

When the payoff range is unknown, the optimal distribution-dependent number of pulls of a suboptimal arm may be bounded independently of $\underline{\nu}$. The proof of Theorem 3 in Section 3 indeed shows that for all suboptimal arms $a \in [K]$,

$$\liminf_{T\to+\infty} \frac{\mathbb{E}_{\underline{\nu}}\bigl[N_a(T)\bigr]}{T/\Phi_{\text{free}}(T)} \geqslant \frac{1}{16}\,.$$

This is in contrast with the case of a known range, for which the bound of Reminder 1 of Appendix A.1 is optimal and strongly depends on ν_a and $\mu^\star$.

The reason for this is that when ignoring the range, the player needs to be a lot more conservative in the exploitation and explore more often. Indeed, to maintain the distribution-free regret bound, the player must avoid the catastrophic case in which an apparently suboptimal arm turns out to be good because of large rewards occurring with small probability, i.e., because of heavy-tail-like issues. For this reason, the player must pull suboptimal arms more frequently than in the case of a known range. This intuition is supported by the construction in the lower bound presented in Section 3: the alternative bounded problem $\underline{\nu}'$ against a problem $\underline{\nu}$ has an arm satisfying $\mathbb{P}_{Y \sim \nu'_a}[Y \geqslant \mu_a + 2\Delta_a/\varepsilon] = \varepsilon$. This behavior is indeed reminiscent of issues arising with heavy-tailed distributions.

3. Proof of Theorem 3: Existence of a Trade-Off

We follow a proof technique introduced by Lai and Robbins (1985) and Burnetas and Katehakis (1996) and recently revisited by Garivier et al. (2019b). We fix some bandit problem $\underline{\nu}$ in $\mathcal{D}_{-,+}$ and construct an alternative bandit problem $\underline{\nu}'$ in $\mathcal{D}_{-,+}$ by modifying the distribution of a single suboptimal arm a to make it optimal. This is always possible, as there is no bound on the upper end on the ranges of the payoffs in the model. We apply a fundamental inequality that links the expectations of the numbers of times $N_a(T)$ that a is pulled under $\underline{\nu}$ and $\underline{\nu}'$. We then substitute inequalities stemming from the definition of distribution-free scale-free regret bounds Φ_{free}, and the result follows by rearranging all inequalities.

Step 1: Alternative bandit problem. The lower bound is trivial—it equals 0—when all arms of $\underline{\nu}$ are optimal. We therefore assume that at least one arm is suboptimal and fix such an arm a. For some $\varepsilon \in [0,1]$ to be defined later by the analysis, we introduce the alternative problem $\underline{\nu}' = (\nu'_k)_{k \in [K]}$ with $\nu'_k = \nu_k$ for $j \neq a$ and $\nu'_a = (1-\varepsilon)\nu_a + \varepsilon \delta_{\mu_a + 2\Delta_a/\varepsilon}$. This distribution ν'_a has a bounded range, so that $\underline{\nu}'$ lies indeed in $\mathcal{D}_{-,+}$. The expectation of ν'_a equals $\mu'_a = \mu_a + 2\Delta_a = \mu^\star + \Delta_a > \mu^\star$. Thus, a is the only optimal arm in $\underline{\nu}'$. Finally, for $\varepsilon < 2\Delta_a/(M - \mu_a)$, the point $\mu_a + 2\Delta_a/\varepsilon$ is larger than M and thus lies outside of the bounded support of ν_a. In that case, the density of ν_a with respect to ν'_a is given by $1/(1-\varepsilon)$ on the support of ν_a and 0 elsewhere, so that $\mathrm{KL}(\nu_a, \nu'_a) = \ln\bigl(1/(1-\varepsilon)\bigr)$.

Step 2: Application of a fundamental inequality. We denote by $\mathrm{kl}(p,q)$ the Kullback-Leibler divergence between Bernoulli distributions with parameters p and q. We also index expectations in the rest of this proof only by the bandit problem they are relative to: for instance, $\mathbb{E}_{\underline{\nu}}$ denotes the expectation of a random variable when the ambient randomness is given by the bandit problem $\underline{\nu}$. The fundamental inequality for lower bounds on the regret of stochastic bandits (Garivier et al., 2019b, Section 2, Equation 6), which is based on the chain rule for Kullback-Leibler divergence and on a data-processing inequality for expectations of $[0,1]$–valued random variables, reads:

$$\mathrm{kl}\!\left(\frac{\mathbb{E}_{\underline{\nu}}\bigl[N_a(T)\bigr]}{T}, \frac{\mathbb{E}_{\underline{\nu}'}\bigl[N_a(T)\bigr]}{T} \right) \leqslant \mathbb{E}_{\underline{\nu}}\bigl[N_a(T)\bigr]\ \mathrm{KL}(\nu_a, \nu'_a) = \mathbb{E}_{\underline{\nu}}\bigl[N_a(T)\bigr] \ln\bigl(1/(1-\varepsilon)\bigr)\,.$$

Now, since $u \in (-\infty,1) \mapsto -u^{-1}\ln(1-u)$ is increasing, we have $\ln\bigl(1/(1-\varepsilon)\bigr) \leqslant (2\ln 2)\varepsilon$ for $\varepsilon \leqslant 1/2$. For all $(p,q) \in [0,1]^2$ and with the usual measure-theoretic conventions,

$$\mathrm{kl}(p,q) = \underbrace{p\ln p + (1-p)\ln(1-p)}_{\geqslant -\ln 2} + \underbrace{p\ln\frac{1}{q}}_{\geqslant 0} + (1-p)\ln\frac{1}{1-q} \geqslant (1-p)\ln\frac{1}{1-q} - \ln 2\,,$$

so that, putting all inequalities together, we have proved

$$\left(1 - \frac{\mathbb{E}_{\underline{\nu}}\bigl[N_a(T)\bigr]}{T}\right)\ln\left(\frac{1}{1-\mathbb{E}_{\underline{\nu}'}\bigl[N_a(T)\bigr]/T}\right) - \ln 2 \leqslant (2\ln 2)\,\varepsilon\,\mathbb{E}_{\underline{\nu}}\bigl[N_a(T)\bigr]\,. \tag{4}$$

In this step, we only imposed the constraint $\varepsilon \in [0,1/2]$. We recall that in the previous step, we imposed $\varepsilon < 2\Delta_a/(M-\mu_a)$. Both conditions are implied by $\varepsilon \leqslant \Delta_a/\bigl(2(M-\mu_a)\bigr)$, which we will assume in the sequel.

Step 3: Inequalities stemming from the definition of scale-free distribution-free regret bounds. We denote by $[m,M]$ a range containing the supports of all distributions of $\underline{\nu}$. By definition of Φ_{free}, given that a is a suboptimal arm (i.e., $\Delta_a > 0$):

$$\Delta_a\,\mathbb{E}_{\underline{\nu}}[N_a(T)] \leqslant R_T(\underline{\nu}) \leqslant (M-m)\,\Phi_{\mathrm{free}}(T)\,.$$

We now prove a similar inequality for $\underline{\nu}'$, for which we recall that a is the unique optimal arm. We denote by $\Delta'_k = \mu'_a - \mu_k$ the gap of arm k in $\underline{\nu}'$. By the definition of ν'_a, the distributions of $\underline{\nu}'$ have supports within the range $[m, M_\varepsilon]$, where we defined the upper end $M_\varepsilon = \max\{M,\, \mu_a + 2\Delta_a/\varepsilon\} = \mu_a + 2\Delta_a/\varepsilon$, given the condition imposed on ε. Therefore, by definition of Φ_{free}, and given that all gaps Δ'_k are larger than the gap $\Delta'_a = \mu'_a - \mu^\star = \Delta_a$ between the unique optimal arm a of $\underline{\nu}'$ and the second best arm(s) of $\underline{\nu}'$ (which were the optimal arms of $\underline{\nu}$), we have

$$\begin{aligned}\Delta_a\bigl(T - \mathbb{E}_{\underline{\nu}'}[N_a(T)]\bigr) = \Delta'_a\bigl(T - \mathbb{E}_{\underline{\nu}'}[N_a(T)]\bigr) \leqslant \sum_{j\neq a}\Delta'_j\,\mathbb{E}_{\underline{\nu}'}[N_j(T)] = R_T(\underline{\nu}')& \\ \leqslant (M_\varepsilon - m)\,\Phi_{\mathrm{free}}(T)\,.&\end{aligned}$$

By rearranging the two inequalities above, we get

$$1 - \frac{\mathbb{E}_{\underline{\nu}}\bigl[N_a(T)\bigr]}{T} \geqslant 1 - \frac{(M-m)\,\Phi_{\mathrm{free}}(T)}{T\Delta_a} \qquad \text{and} \qquad 1 - \frac{\mathbb{E}_{\underline{\nu}'}\bigl[N_a(T)\bigr]}{T} \leqslant \frac{(M_\varepsilon - m)\,\Phi_{\mathrm{free}}(T)}{T\Delta_a}\,,$$

thus, after substitution into (4),

$$\left(1 - \frac{(M-m)\,\Phi_{\mathrm{free}}(T)}{T\Delta_a}\right)\ln\left(\frac{T\Delta_a}{(M_\varepsilon - m)\,\Phi_{\mathrm{free}}(T)}\right) - \ln 2 \leqslant (2\ln 2)\,\varepsilon\,\mathbb{E}_{\underline{\nu}}\bigl[N_a(T)\bigr]\,. \tag{5}$$

Step 4: Final calculations. We take $\varepsilon = \varepsilon_T = \alpha^{-1}\,\Phi_{\mathrm{free}}(T)/T$ for some constant $\alpha > 0$; we will pick $\alpha = 1/8$. By the assumption $\Phi_{\mathrm{free}}(T) = o(T)$, we have $\varepsilon_T \leqslant \Delta_a/\bigl(2(M-\mu_a)\bigr)$, as needed, for T large enough, as well as $M_{\varepsilon_T} = \mu_a + 2\Delta_a/\varepsilon_T = \mu_a + 2\alpha\Delta_a T/\Phi_{\mathrm{free}}(T)$.

Substituting these values into (5), a finite-time lower bound on the quantity of interest is finally given by

$$\frac{\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]}{T/\Phi_{\mathrm{free}}(T)} \geqslant \frac{\alpha}{2\ln 2}\left(-\ln 2 + \Bigg(1 - \underbrace{\frac{(M-m)\,\Phi_{\mathrm{free}}(T)}{T\Delta_a}}_{\to 0}\Bigg)\ln\Bigg(\underbrace{\frac{T\Delta_a}{2\alpha\Delta_a T + (\mu_a - m)\Phi_{\mathrm{free}}(T)}}_{\to 1/(2\alpha)}\Bigg)\right).$$

It entails the asymptotic lower bound

$$\liminf_{T\to+\infty} \frac{\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]}{T/\Phi_{\mathrm{free}}(T)} \geqslant \frac{\alpha}{2\ln 2}\big(\ln(1/\alpha) - 2\ln 2\big) = \frac{1}{16}$$

for the choice $\alpha = 1/8$. The claimed result follows by adding these lower bounds for each suboptimal arm a, with a factor Δ_a, following the rewriting (1) of the regret.

Remark 4 *The proof above only exploits the fact that the upper end M of the range is unknown: the alternative problems lie in $\mathcal{D}_{m,M'}$ for some M' that can be arbitrarily large. Yet, by definition of adaptation to the range, the strategy needs to guarantee $(M'-m)\,\Phi_{\mathrm{free}}(T)$ distribution-free regret bounds in that case.*

We may note that therefore, Theorem 3 also holds for the model of bounded distributions with a known lower end $m \in \mathbb{R}$ for the range:

$$\mathcal{D}_{m,+} = \bigcup_{\substack{M\in\mathbb{R}:\\ M>m}} \mathcal{D}_{m,M}\,. \tag{6}$$

Definitions 1 and 2 handle the case of $\mathcal{D}_{-,+}$ but can be adapted in an obvious way to $\mathcal{D}_{m,+}$ by fixing m, by having the strategy know m, and requiring the bounds to hold for all $M \in [m,+\infty)$ and all bandit problems in $\mathcal{D}_{m,M}$, thus leading to the concept of adaptation to the upper end of the range.

This observation is in line with the folklore knowledge that there is a difference in nature between dealing with nonnegative payoffs, i.e., gains, or dealing with nonpositive payoffs, i.e., losses, for regret minimization under bandit monitoring; see Cesa-Bianchi and Lugosi (2006, Remark 6.5, page 164) for an early reference and Kwon and Perchet (2016) for a more complete literature review. Actually, 0 plays no special role, the issue is rather whether one end of the payoff range is known.

4. Adaptation to Range Based on AdaHedge: The AHB Strategy

When the range of payoffs is known, Auer et al. (2002b) achieve a distribution-free regret bound of order $\sqrt{KT\ln K}$ with exponential weights—the Hedge strategy—on estimated payoffs and with extra-exploration, i.e., by mixing exponential weights with the uniform distribution over arms. Actually, it is folklore knowledge that the extra-exploration used in this case is unnecessary (see, among others, Stoltz, 2005). To deal with the case of an unknown payoff range, we consider a self-tuned version of Hedge called AdaHedge (De Rooij et al., 2014, see also an earlier work by Cesa-Bianchi et al., 2007) and do add extra-exploration. Just as Auer et al. (2002b), we will actually obtain regret guarantees for oblivious adversarial bandits, not only distribution-free regret bounds for stochastic bandits. We therefore introduce now the setting of oblivious adversarial bandits and define adaptation to the range in that case.

4.1 Oblivious Adversarial Bandits

In the setting of fully oblivious adversarial bandits (see Cesa-Bianchi and Lugosi, 2006; Audibert and Bubeck, 2009), a range $[m, M]$ is set by the environment, where m, M are real numbers, not necessarily nonnegative. The player is unaware of $[m, M]$ and will remain so. The environment also picks beforehand a sequence $y_1, y_2, \ldots$ of reward vectors in $[m, M]^K$. We denote by $y_t = (y_{t,a})_{a \in [K]}$ the components of these vectors. The player will observe a component of each of these reward vectors in a sequential fashion, as follows. Auxiliary randomizations $U_0, U_1, \ldots$ i.i.d. according to a uniform distribution over $[0, 1]$ are available. At each round $t \geqslant 1$, the player picks an arm $A_t \in [K]$, possibly at random (thanks to U_{t-1}) according to a probability distribution $p_t = (p_{t,a})_{a \in [K]}$, and then receives and observes y_{t,A_t}.

More formally, a strategy of the player is a sequence of mappings from the observations to the action set, $(U_0,\, y_{1,A_1}, U_1,\, \ldots,\, y_{t-1,A_{t-1}}, U_{t-1}) \mapsto A_t$. The strategy does not rely on m nor M.

At each given time $T \geqslant 1$, denoting by $y_{1:T} = (y_1, \ldots, y_T)$ the reward vectors, we measure the performance of a strategy through its expected regret:

$$R_T(y_{1:T}) = \max_{a \in [K]} \sum_{t=1}^{T} y_{t,a} - \mathbb{E}\left[\sum_{t=1}^{T} y_{t,A_t}\right], \tag{7}$$

where, as rewards are fixed beforehand, all randomness lies in the choice of the arms A_t only, i.e., where the expectation is only over the choice of the arms A_t.

The counterpart of Definition 1 in this setting is stated next.

Definition 5 (Scale-free adversarial regret bounds) *A strategy for oblivious adversarial bandits is adaptive to the unknown range of payoffs with a scale-free adversarial regret bound $\Phi_{\mathrm{adv}} : \mathbb{N} \to [0, +\infty)$ if for all real numbers $m < M$, the strategy ensures, without the knowledge of m and M:*

$$\forall y_1, y_2, \ldots \text{ in } [m, M]^K, \quad \forall T \geqslant 1, \qquad R_T(y_{1:T}) \leqslant (M - m)\, \Phi_{\mathrm{adv}}(T)\,.$$

Conversion of upper/lower bounds from one setting to the other. We recall that when applying Doob's optional skipping in Section 2, for each arm a, we denoted by $(Y_{t,a})_{t \geqslant 1}$ an i.i.d. sequence of rewards drawn beforehand, independently at random, according to the distribution ν_a associated with that arm. By the tower rule for the right-most equality below, we note that for all $m < M$ and for all $\underline{\nu}$ in $\mathcal{D}_{m,M}$,

$$\begin{aligned} R_T(\underline{\nu}) = \max_{a \in [K]} \mathbb{E}\left[\sum_{t=1}^{T} Y_{t,a}\right] - \mathbb{E}\left[\sum_{t=1}^{T} Y_{t,A_t}\right] \leqslant \mathbb{E}\left[\max_{a \in [K]} \sum_{t=1}^{T} Y_{t,a} - \sum_{t=1}^{T} Y_{t,A_t}\right] &= \mathbb{E}\big[R_T(Y_{1:T})\big] \\ &\leqslant \sup_{y_{1:T} \text{ in } [m,M]^K} R_T(y_{1:T})\,. \end{aligned}$$

In particular, lower bounds on the regret for stochastic bandits are also lower bounds on the regret for oblivious adversarial bandits, and strategies designed for oblivious adversarial bandits obtain the same distribution-free regret bounds for stochastic bandits when the individual payoffs y_{t,A_t} in their definition are replaced with the stochastic payoffs Y_{t,A_t}.

Algorithm 1 AHB: AdaHedge for K–armed Bandits, with extra-exploration

```
Inputs: a payoff estimation scheme, e.g., (8); a sequence $(\gamma_t)_{t \geqslant 1}$ in $[0,1]$ of extra-exploration rates
for rounds $t = 1, \ldots, K$ do
    Draw arm $A_t = t$
    Get and observe the payoff $y_{t,t}$
end for
AdaHedge initialization: $\eta_{K+1} = +\infty$ and $q_{K+1} = (1/K, \ldots, 1/K) \overset{\text{def}}{=} \mathbf{1}/K$
for rounds $t = K+1, \ldots$ do
    Define $p_t$ by mixing $q_t$ with the uniform distribution as $p_t = (1-\gamma_t) q_t + \gamma_t \mathbf{1}/K$
    Draw an arm $A_t \sim p_t$, i.e., independently at random according to the distribution $p_t$
    Get and observe the payoff $y_{t,A_t}$
    Compute estimates $\widehat{y}_{t,a}$ of all payoffs with the payoff estimation scheme, e.g., (8)
    Compute the mixability gap $\delta_t \geqslant 0$ based on the distribution $q_t$ and on these estimates:
        $$\underbrace{\delta_t = -\sum_{a=1}^{K} q_{t,a}\, \widehat{y}_{t,a} + \frac{1}{\eta_t} \ln\!\left( \sum_{a=1}^{K} q_{t,a} \mathrm{e}^{\eta_t \widehat{y}_{t,a}} \right),}_{\text{when } \eta_t \leqslant +\infty \text{ or } \eta_t = +\infty} \qquad \text{i.e.,} \qquad \underbrace{\delta_t = -\sum_{a=1}^{K} q_{t,a}\, \widehat{y}_{t,a} + \max_{a \in [K]} \widehat{y}_{t,a}}_{\text{when } \eta_t = +\infty}$$
    Compute the learning rate $\eta_{t+1} = \left( \sum_{s=K+1}^{t} \delta_s \right)^{-1} \ln K$
    Define $q_{t+1}$ component-wise as
        $$q_{t+1,a} = \exp\!\left( \eta_{t+1} \sum_{s=K+1}^{t} \widehat{y}_{a,s} \right) \Big/ \sum_{k=1}^{K} \exp\!\left( \eta_{t+1} \sum_{s=K+1}^{t} \widehat{y}_{k,s} \right)$$
end for
```

4.2 The AHB Strategy

We state our main strategy, AHB—which stands for AdaHedge for K–armed Bandits, with extra-exploration—, in the setting of oblivious adversarial bandits, see Algorithm 1. In a setting of stochastic bandits, it suffices to replace therein y_{t,A_t} with Y_{t,A_t}. The AHB strategy relies on a payoff estimation scheme, which we discuss now.

In Algorithm 1, some initial exploration lasting K rounds is used to get a rough idea of the location of the payoffs and to center the estimates used at an appropriate location. Following Auer et al. (2002b), we consider, for all rounds $t \geqslant K+1$ and arms $a \in [K]$,

$$\widehat{y}_{t,a} = \frac{y_{t,A_t} - C}{p_{t,a}} \mathbb{1}_{\{A_t = a\}} + C \qquad \text{where} \qquad C \overset{\text{def}}{=} \frac{1}{K} \sum_{s=1}^{K} y_{s,s}\,. \tag{8}$$

Note that all $p_{t,a} > 0$ for Algorithm 1 due to the use of exponential weights. As proved by Auer et al. (2002b), the estimates $\widehat{y}_{t,a}$ are conditionally unbiased. Indeed, the distributions q_t and p_t, as well as the constant C, are measurable functions of the information

$H_{t-1} = (U_0,\, y_{1,A_1}, U_1,\, \ldots, U_{t-2},\, y_{t-1,A_{t-1}})$ available at the beginning of round $t \geqslant K+1$, and the arm A_t is drawn independently at random according to p_t based on an auxiliary randomization denoted by U_{t-1}. Therefore, given that the payoffs are oblivious, the conditional expectation of $\widehat{y}_{t,a}$ with respect to H_{t-1} amounts to integrating over the randomness given by the random draw $A_t \sim p_t$: for $t \geqslant K+1$,

$$\mathbb{E}\big[\widehat{y}_{t,a} \,\big|\, H_{t-1}\big] = \frac{y_{t,a}-C}{p_{t,a}}\, \mathbb{P}\big(A_t = a \,\big|\, H_{t-1}\big) + C = \frac{y_{t,a}-C}{p_{t,a}}\, p_{t,a} + C = y_{t,a}\,. \tag{9}$$

These estimators are bounded: assuming that all $y_{t,a}$, thus also C, belong to the range $[m, M]$, and given that the distributions p_t were obtained by a mixing with the uniform distribution, with weight γ_t, we have $p_{t,a} \geqslant \gamma_t/K$, and therefore,

$$\forall t \geqslant K+1, \quad \forall a \in [K], \qquad \big|\widehat{y}_{t,a} - C\big| \leqslant \frac{|y_{t,a}-C|}{p_{t,a}} \leqslant \frac{M-m}{\gamma_t/K}\,. \tag{10}$$

Remark 6 *Algorithm 1 is invariant by affine changes, i.e., translations by real numbers and/or multiplications by positive factors, of the payoffs, given that AdaHedge (see De Rooij et al., 2014, Theorem 16) and the payoff estimation scheme* (8) *are so. This is key for adaptation to the range. This invariance is achieved, when ignoring the range* $[m, M]$, *thanks to any value* $C \in [m, M]$. *Here, we chose to have* K *rounds of exploration in Algorithm 1 and let* C *equal the average of the payoffs achieved. However, it would of course have been sufficient to pick one arm at random, observe a single reward* y_{1,A_1} *and let* $C = y_{1,A_1}$.

4.3 Regret Analysis, Part 1: Scale-Free Adversarial Regret Bound

Theorem 7 *AdaHedge for K–armed bandits (Algorithm 1) with a non-increasing extra-exploration sequence* $(\gamma_t)_{t\geqslant 1}$ *smaller than* $1/2$ *and the estimation scheme given by* (8) *ensures that for all bounded ranges* $[m, M]$, *for all oblivious individual sequences* $y_1, y_2, \ldots$ *in* $[m, M]^K$, *for all* $T \geqslant 1$,

$$R_T(y_{1:T}) \leqslant 3(M-m)\sqrt{KT\ln K} + 5(M-m)\frac{K\ln K}{\gamma_T} + (M-m)\sum_{t=K+1}^{T} \gamma_t\,.$$

In particular, given a parameter $\alpha \in (0,1)$, *the extra-exploration*

$$\gamma_t = \min\Big\{1/2,\ \sqrt{5(1-\alpha)K\ln K}\big/t^{\alpha}\Big\}$$

leads to the scale-free adversarial regret bound

$$\Phi_{\mathrm{adv}}(T) = \left(3 + \frac{5}{\sqrt{1-\alpha}}\right)(M-m)\sqrt{K\ln K}\; T^{\max\{\alpha,1-\alpha\}} + 10(M-m)K\ln K\,. \tag{11}$$

For $\alpha = 1/2$, *the bound reads* $\Phi_{\mathrm{adv}}(T) = 7(M-m)\sqrt{TK\ln K} + 10(M-m)K\ln K$.

This value $\alpha = 1/2$ is the best one to consider if one is only interested in a distribution-free bound—i.e., if one is not interested in the distribution-dependent rates for the regret. The proof of Theorem 7 is detailed in Appendix B but we sketch its proof below.

Remark 8 *We strongly suspect that the $\sqrt{\ln K}$ factor in the bound of Theorem 7 is superfluous. In the case of a known range, the MOSS algorithm is known to be minimax optimal with a regret bound of order $\sqrt{KT}$. One idea could thus be to use a MOSS-type index, together with a Bernstein-type upper confidence bound to account for the unknown variance and range. A final ingredient would be to add initial extra-exploration, pulling every arm $\sqrt{T/K}$ times before running the standard phase of the algorithm; on a technical level, this automatically makes the sub-Poissonian term in Bernstein's inequality tractable. We have not managed yet to fill in the technical details in order to prove this, although we believe a variant of these ideas would get rid of the logarithmic factor. In contrast, the algorithm discussed here, based on AdaHedge, enjoys a simple distribution-free analysis—as sketched below—, as well as a distribution-dependent analysis (see Section 4.4), unlike an algorithm based on MOSS-type indices.*

Another promising approach would be to use the Tsallis-INF algorithm introduced by Audibert and Bubeck (2009) and further studied by Zimmert and Seldin (2019), which achieves a $(M-m)\sqrt{KT}$ adversarial regret bound when M and m are known. Unfortunately, current analyses of the algorithm rely crucially on the non-positivity of the reward estimates, or, equivalently on the knowledge of an upper bound on the rewards. Zimmert and Lattimore (2019) relax this requirement, but not enough for the relaxed version to be applied to our case. However, when M is known and m is unknown, i.e., only adaptation to m is needed, the reward estimates can be made non-positive by taking $C = M$ in the estimation scheme (8)*, and our techniques may be extended to show that Tsallis-INF indeed enjoys an adversarial regret bound of order $(M-m)\sqrt{KT}$ in this case. Details may be found in an extended version of this article [arXiv:2006.03378], Theorem 23 in Appendix F.*

Proof sketch A direct application of the AdaHedge regret bound (Lemma 3 and Theorem 6 of De Rooij et al., 2014), bounding the variance terms of the form $\mathbb{E}\big[(X-\mathbb{E}[X])^2\big]$ by $\mathbb{E}\big[(X-C)^2\big]$, ensures that

$$\max_{k\in[K]} \sum_{t=K+1}^{T} \widehat{y}_{t,k} - \sum_{\substack{t\geqslant K+1\\ a\in[K]}}^{T} q_{t,a}\,\widehat{y}_{t,a} \leqslant 2\sqrt{\sum_{\substack{t\geqslant K+1\\ a\in[K]}} q_{t,a}\big(\widehat{y}_{t,a}-C\big)^2 \ln K} + \frac{M-m}{\gamma_T/K}\left(2+\frac{4}{3}\ln K\right).$$

We take expectations, use the definition of the p_t in terms of the q_t in the left-hand side, and apply Jensen's inequality in the right-hand side to get

$$\mathbb{E}\left[\max_{k\in[K]} \sum_{t=K+1}^{T} \widehat{y}_{t,k} - \sum_{t=K+1}^{T} \overbrace{\sum_{a=1}^{K} p_{t,a}\,\widehat{y}_{t,a}}^{=y_{t,A_t}} + \sum_{t=K+1}^{T} \gamma_t \overbrace{\sum_{a=1}^{K} (1/K - q_{t,a})\,\widehat{y}_{t,a}}^{\mathbb{E}[\ldots]\in[m-M,M-m]}\right]$$

$$\leqslant 2\sqrt{\sum_{t=K+1}^{T}\sum_{a=1}^{K} \mathbb{E}\Big[q_{t,a}\big(\widehat{y}_{t,a}-C\big)^2\Big] \ln K} + \frac{M-m}{\gamma_T/K}\left(2+\frac{4}{3}\ln K\right).$$

Since $p_{t,a} \geqslant (1-\gamma_t)q_{t,a}$ with $\gamma_t \leqslant 1/2$ by assumption on the extra-exploration rate, we have the bound $q_{t,a} \leqslant 2p_{t,a}$. Together with standard calculations similar to (9), we have

$$\mathbb{E}\Big[q_{t,a}\big(\widehat{y}_{t,a}-C\big)^2\Big] \leqslant 2\,\mathbb{E}\Big[p_{t,a}\big(\widehat{y}_{t,a}-C\big)^2 \,\Big|\, H_{t-1}\Big] = 2\,\mathbb{E}\left[\frac{(y_{t,A_t}-C)^2}{p_{t,a}}\mathbb{1}_{\{A_t=a\}}\right] = 2\underbrace{(y_{t,a}-C)^2}_{\leqslant (M-m)^2}.$$

The proof of the first regret bound of the theorem is concluded by collecting all bounds and by taking care of the first K rounds. The second regret bound then follows from straightforward calculations. ■

4.4 Regret Analysis, Part 2: Distribution-Dependent Rates for Adaptation

Given the conversion explained in Section 4.1, Algorithm 1 tuned as in Corollary 7 for $\alpha \in [1/2,1)$ also enjoys the scale-free distribution-free regret bound $\Phi^{\mathrm{AHB}}_{\mathrm{free}}(T) = \Phi^{\mathrm{AHB}}_{\mathrm{adv}}(T)$ of order T^{α}. The theorem below entails that AHB is adaptive to the unknown range with a distribution-dependent regret rate $T/\Phi^{\mathrm{AHB}}_{\mathrm{free}(T)}$ of order $T^{1-\alpha}$ that is optimal given the lower bound stated by Theorem 3.

Theorem 9 *Consider AHB (Algorithm 1) tuned with some $\alpha \in [1/2,1)$ as in the second part of Theorem 7. For all distributions $\nu_1,\ldots,\nu_K$ in $\mathcal{D}_{-,+}$,*

$$\limsup_{T\to\infty} \frac{R_T(\underline{\nu})}{T/\Phi^{\mathrm{AHB}}_{\mathrm{free}(T)}} \leqslant \frac{12\ln K}{1-\alpha}\sum_{a=1}^{K}\Delta_a\,. \qquad (12)$$

The proof is provided in Appendix C. It follows quite closely that of Theorem 3 in Seldin and Lugosi (2017), where the authors study a variant of the Exp3 algorithm of Auer et al. (2002b) for stochastic rewards. It consists, in our setting, in showing that the number of times the algorithm chooses suboptimal arms is almost only determined by the extra-exploration. Our proof is simpler as we aim for cruder bounds. The main technical difference and issue to solve lies in controlling the learning rates η_t, which heavily depend on data in our case.

5. Numerical Illustrations

We provide some numerical experiments on synthetic data to illustrate the qualitative behavior of some popular algorithms like UCB strategies when they are incorrectly tuned, as opposed to strategies that are less sensitive to ignoring the range or to the AHB strategy which adapts to it. These experiments are only of an illustrative nature.

Bandit problems considered and UCB strategies. We consider stochastic bandit problems $\underline{\nu}^{(\alpha)} = (\nu^{(\alpha)}_a)_{a\in[K]}$ indexed by a scale parameter $\alpha \in \{0.01,\,1,\,100\}$. We take $K=10$ arms, each arm a being associated with a rectified Gaussian distribution. Precisely, the distribution $\nu^{(\alpha)}_a$ is the distribution of the variable

$$X^{\alpha}_a = \begin{cases} \alpha\,\max\big\{0,\min\{Y,\ 1.2\}\big\} & \text{with} \quad Y\sim\mathcal{N}(0.6\,,V) \quad \text{if } a=1, \\ \alpha\,\max\big\{0,\min\{Y,\ 1\}\big\} & \text{with} \quad Y\sim\mathcal{N}(0.5\,,V) \quad \text{if } a\neq 1, \end{cases}$$

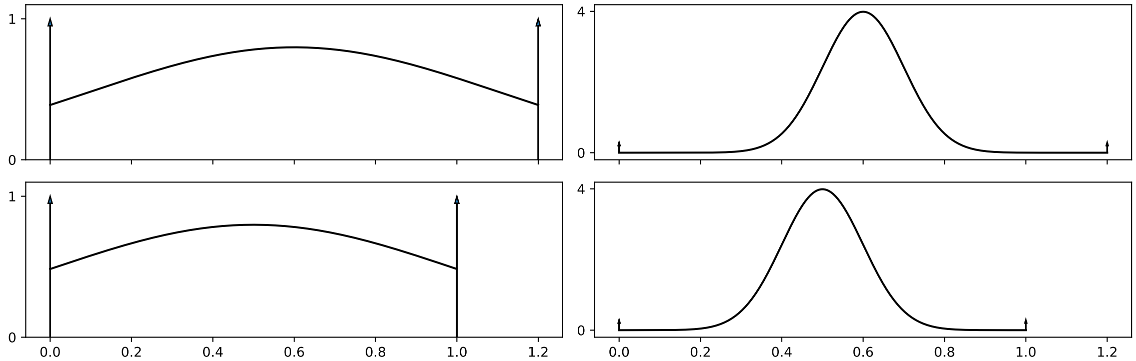

Figure 1: Probability density functions of the reward distributions with respect to the sum of the Lebesgue measure and Dirac masses at 0, 1, and 1.2. Left pictures: high-variance case; right pictures: low-variance case. Top pictures: first arm (optimal arm); bottom pictures: other arms. Arrows represent atoms and their lengths are only illustrative.

so that all distributions are commonly supported on $[m, M] = [0,\ 1.2\,\alpha]$, with arm 1 being the unique optimal arm. We will consider two values for V, namely $V = 0.01$ (low-variance case) and $V = 0.25$ (high-variance case). See Figure 1 for a plot of the corresponding probability density functions.

We denote by $\mu_1^{(\alpha)} = 0.6\,\alpha$ and $\mu_a^{(\alpha)} = 0.5\,\alpha$ if $a \neq 1$ the means associated with the distributions $\nu_1^{(\alpha)}$ and $\nu_a^{(\alpha)}$, respectively. The gaps therefore equal $\Delta_a^{(\alpha)} = 0.1\,\alpha$ for $a \geqslant 2$.

The main algorithm of interest is, of course, the AHB strategy with extra-exploration (Algorithm 1), which we tune as stated in Theorem 7 with parameter 1/2. We now present the competitors.

UCB strategies at different scales. We consider instances of UCB (Auer et al., 2002a) using indices of the form

$$\widehat{\mu}_a(t) + \sqrt{\frac{8\sigma^2 \ln T}{N_a(t)}}\,,$$

where $N_a(t)$ is the number of times arm a was pulled up to round t, and where $\widehat{\mu}_a(t)$ denotes the empirical average of payoffs obtained for arm a. We hesitated between setting σ^2 based on the range $M - m = 1.2\alpha$, namely, $\sigma^2 = (M-m)^2/4 = (1.2\alpha)^2$, or based on a sub-Gaussian parameter, which would be smaller. As distributions $\nu_a^{(\alpha)}$ are rectified Gaussians, it is not immediately clear whether they are sub-Gaussian, but we considered despite all the choice $\sigma^2 = V$. It turns out that this second choice outperformed the first one, which is why, in the rest of the study, we consider the following three instances of UCB:

$$\widehat{\mu}_a(t) + s\sqrt{\frac{8V \ln T}{N_a(t)}}\,, \qquad \text{where} \qquad s \in \{0.01,\ 1,\ 100\}\,.$$

When the scale parameter α is known, we would take $s = \alpha$.

Range-estimating UCB. We also study a version of UCB estimating the range, namely, using indices

$$\widehat{\mu}_a(t) + \hat{r}_t \sqrt{\frac{2 \ln T}{N_a(t)}}\,, \qquad \text{where} \qquad \hat{r}_t = \max_{s \leqslant t} Y_{A_s,s} - \min_{s \leqslant t} Y_{A_s,s}$$

estimates the range $M - m$. We were unable to provide theoretical guarantees that match our lower bounds, and this algorithm does not perform particularly well in practice as we will discuss below.

ε–greedy. Finally, we also consider the ε–greedy strategy, which, at round $t \geqslant K+1$, picks with probability $1 - \varepsilon_t$ the arm with the best empirical mean, and otherwise, selects an arm uniformly at random. Following Auer et al. (2002a), we used the tuning

$$\varepsilon_t = \min\left\{1,\ \frac{5K}{d^2 t}\right\} \quad \text{with} \quad d = 1/12\,.$$

Indeed, Auer et al. (2002a) exhibit theoretical guarantees for distributions over $[0,1]$ in the case where d is smaller than or equal to the smallest gap. When rescaled on $[0,1]$, the smallest gap equals $0.1\alpha(M-m)/(1.2\alpha) = 1/12$ in our setting; this explains our choice $d = 1/12$, but note that the ε–greedy strategy defined above relies on some extra knowledge encompassed in the choice $d = 1/12$, compared to the completely agnostic AHB strategy. Interestingly, for any fixed-in-advance sequence of ε_t, the ε–greedy strategy is scale-free. Of course, its strong downside is that a proper tuning of the ε_t requires knowledge of a scaled lower bound on the gaps.

Experimental setting. Each algorithm is run $N = 100$ times, on a time horizon $T = 100{,}000$. We plot estimates of the rescaled regret $R_T(\underline{\nu}^{(\alpha)})/\alpha$ to have a meaningful comparison between the bandit problems. These estimates are constructed as follows. We index the arms picked in the n–th run by an additional subscript n, so that $A_{T,n}$ refers to the arm picked by some strategy at time t in the n–th run. The expected regret of a given strategy can be rewritten as

$$R_T(\underline{\nu}^{\alpha}) = T \max_{a \in [K]} \mu_a^{(\alpha)} - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{A_t}^{(\alpha)}\right] = T \times (0.6\,\alpha) - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{A_t}^{(\alpha)}\right]$$

and is estimated by

$$\widehat{R}_T(\alpha) = \frac{1}{N}\sum_{n=1}^{N} \widehat{R}_T(\alpha,n) \qquad \text{where} \qquad \widehat{R}_T(\alpha,n) = T \times (0.6\,\alpha) - \sum_{t=1}^{T} \mu_{A_{t,n}}^{(\alpha)}\,.$$

On Figures 2 and 3 we plot the estimates $\widehat{R}_T(\alpha)/\alpha$ of the rescaled regret as solid lines. The shaded areas correspond to ±2 standard errors of the sequences $\big(\widehat{R}_T(\alpha,n)/\alpha\big)_{n \in [N]}$.

Discussion of the results. An initial observation is that, as expected, the performance of AHB, the range-estimating UCB, and ε-greedy is unaffected by the scale of the problems (see the second lines of Figures 2 and 3). It turns out that out of these three algorithm, AHB performs best.

A second observation is that the performance of UCB depends dramatically on the value of the parameter s. UCB performs like follow-the-leader when s is too small, and

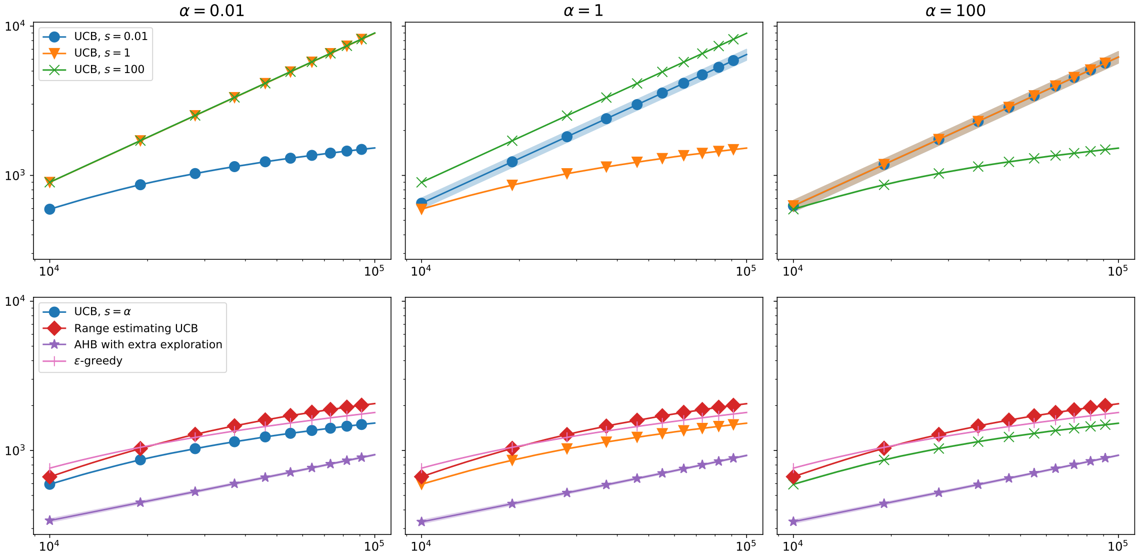

Figure 2: Comparison of the (estimated) regrets of various strategies over bandit problems $\underline{\nu}^{(\alpha)}$ in the high variance case, where α ranges in $\{0.01, 1, 100\}$ and $V = 0.25$. Each algorithm was run $N = 100$ times on every problem for $T = 100{,}000$ time steps. Solid lines report the values of the estimated regrets, while shaded areas correspond to ±2 standard errors of the estimates.

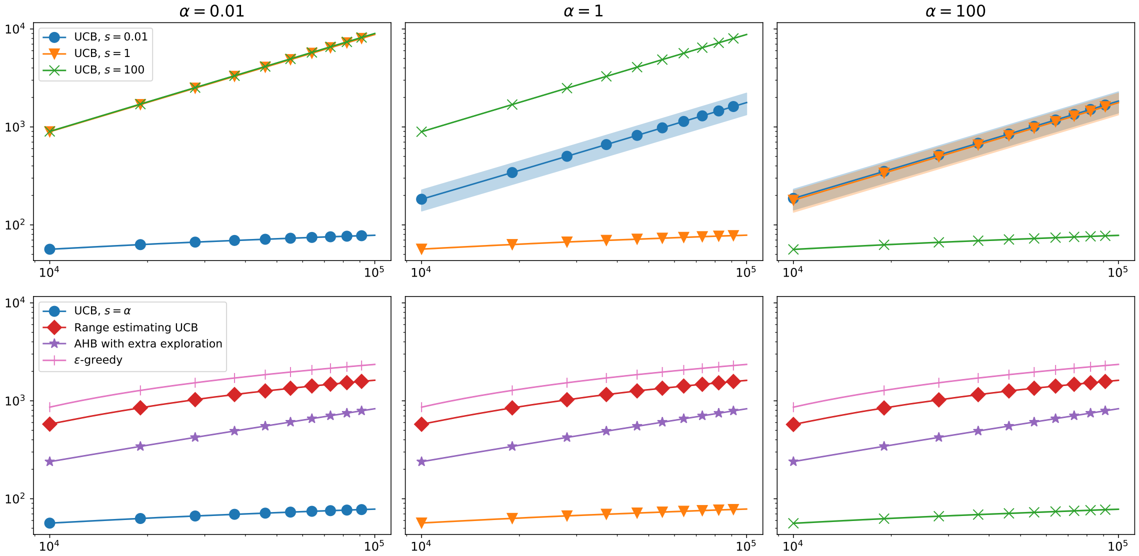

Figure 3: Same legend as for Figure 2, but in the low variance case.

like random play when s is too large; both of these strategies suffer linear regret and UCB incorrectly scaled also does so (see the first lines of Figures 2 and 3).

It remains to compare AHB to UCB tuned with the correct scale: the ranking between the two depends on the value of V, with AHB outperforming UCB tuned with the correct scale in the high-variance case and vice versa in the low variance case.

Our last observation is that in the low-variance case, the range-estimating version of UCB is far off from UCB tuned with the correct scale. This is because of the large difference between the sub-Gaussian parameter and its upper bound given by the squared half-range, which the range-estimating version of UCB is targeting.

Acknowledgments

Hédi Hadiji and Gilles Stoltz have no direct funding to acknowledge other than the salaries paid by their employers (Université Paris-Saclay, CNRS, and HEC Paris). They have no competing interests to declare.

Appendix A. More on Scale-Free Distribution-Dependent Regret Bounds Considered in Isolation

This section details the claims of Section 2.2: no strategy may be adaptive to the range and achieve $\Phi_{\text{dep}} = \ln T$ (Section A.2) but we may construct a strategy adaptive to the range and achieving $\Phi_{\text{dep}} \gg \ln T$ (Section A.3). Before we do so, we provide a reminder on a general, and optimal, distribution-dependent regret lower bound for K–armed stochastic bandits (Section A.1).

A.1 Reminder of a General Regret Lower Bound for K–Armed Bandits

This section considers some general model $\mathcal{D}$. It also rules out poor strategies by restricting its attention to so-called consistent strategies—according to the terminology introduced by Lai and Robbins (1985), while Burnetas and Katehakis, 1996 rather speak of uniformly fast convergent strategies.

Definition 10 *A strategy is consistent on a model $\mathcal{D}$ if for all bandit problems $\underline{\nu}$ in $\mathcal{D}$, it achieves a subpolynomial regret bound, that is, $R_T(\underline{\nu})/T^\alpha \to 0$ for all $\alpha \in (0,1]$.*

A lower bound on the distribution-dependent rates that such a strategy may achieve is provided by a general, and optimal, result of Lai and Robbins (1985) and Burnetas and Katehakis (1996); see also its rederivation by Garivier et al. (2019b). It involves a quantity defined as an infimum of Kullback-Leibler divergences: we recall that for two probability distributions ν, ν' defined on the same probability space $(\Omega, \mathcal{F})$,

$$\mathrm{KL}(\nu,\nu') = \begin{cases} \displaystyle\int_\Omega \ln\!\left(\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}\right)\mathrm{d}\nu & \text{if } \nu \ll \nu', \\ +\infty & \text{otherwise,} \end{cases}$$

where $\nu \ll \nu'$ means that ν is absolutely continuous with respect to ν' and $\mathrm{d}\nu/\mathrm{d}\nu'$ then denotes the Radon-Nikodym derivative. Now, for any probability distribution ν, any real number x, and any model $\mathcal{D}$, we define

$$\mathcal{K}_{\mathrm{inf}}(\nu, x, \mathcal{D}) = \inf\bigl\{\mathrm{KL}(\nu, \nu') : \nu' \in \mathcal{D} \text{ and } \mathrm{E}(\nu') > x\bigr\},$$

where by convention, the infimum of an empty set equals $+\infty$ and where we denoted by $\mathrm{E}(\nu')$ the expectation of ν'. The quantity $\mathcal{K}_{\mathrm{inf}}(\nu, x, \mathcal{D})$ can be null. With the usual measure-theoretic conventions, in particular, $0/0 = 0$, we then have the following lower bound.

Reminder 1 *For all models $\mathcal{D}$, for all consistent strategies on $\mathcal{D}$, for all bandit problems $\underline{\nu}$ in $\mathcal{D}$,*

$$\liminf_{T \to +\infty} \frac{R_T(\underline{\nu})}{\ln T} \geqslant \sum_{a \in [K]} \frac{\Delta_a}{\mathcal{K}_{\mathrm{inf}}(\nu_a, \mu^\star, \mathcal{D})}\,.$$

The case of a known payoff range $[m, M]$. When the payoff range $[m, M]$ is known, i.e., when the model is $\mathcal{D}_{m,M}$, there exist strategies achieving the lower bound of Reminder 1, like the DMED strategy of Honda and Takemura (2011, 2015) or the KL–UCB strategy of Cappé et al. (2013) and Garivier et al. (2019a).

The case of a known payoff upper bound M. The DMED strategy of Honda and Takemura (2015) actually achieves the lower bound of Reminder 1 even for the model

$$\mathcal{D}_{-,M} = \bigcup_{\substack{m \in \mathbb{R}:\\ m < M}} \mathcal{D}_{m,M}$$

and for the model $\mathcal{D}_{-\infty,M}$ of all distributions upper bounded by M but not necessarily lower bounded. This suggests that adaptation to M is much more difficult than adaptation to m as far as distribution-dependent regret bounds are considered, and is in line with Remark 4.

That M is more important than m for distribution-dependent bounds is also reflected in the lower bound of Reminder 1: this lower bound does not depend on whether $\mathcal{D}$ equals some $\mathcal{D}_{m,M}$, or $\mathcal{D}_{-,M}$, or even $\mathcal{D}_{-\infty,M}$. We may indeed easily show (see an extended version of this article [arXiv:2006.03378], Appendix E) that given $M \in \mathbb{R}$, for all $m \leqslant M$, for all $\nu \in \mathcal{D}_{m,M}$ and all $\mu > \mathrm{E}(\nu)$,

$$\mathcal{K}_{\mathrm{inf}}\bigl(\nu, \mu, \mathcal{D}_{m,M}\bigr) = \mathcal{K}_{\mathrm{inf}}\bigl(\nu, \mu, \mathcal{D}_{-\infty,M}\bigr)\,.$$

A.2 Adaptation to the Range Impossible at Logarithmic Distribution-Dependent Rate

A strategy that would be adaptive to the range with a distribution-dependent rate $\Phi_{\mathrm{dep}} = \ln$ would, by definition and in particular, be consistent on $\mathcal{D}_{-,+}$. The following theorem therefore shows, by contradiction, that no strategy may be adaptive to the range with a distribution-dependent rate $\Phi_{\mathrm{dep}} = \ln$. A similar phenomenon was discussed by Lattimore (2017) in the case of stochastic bandits with Gaussian distributions.

Theorem 11 *For all distributions $\nu_a \in \mathcal{D}_{-,+}$ with expectation μ_a, and all $\mu^\star > \mu_a$,*

$$\mathcal{K}_{\mathrm{inf}}(\nu_a, \mu^\star, \mathcal{D}_{-,+}) = 0\,.$$

As a consequence, all consistent strategies on $\mathcal{D}_{-,+}$ are such that, for all bandit problems $\underline{\nu}$ in $\mathcal{D}_{-,+}$ with at least one suboptimal arm a,

$$\liminf_{T\to+\infty} \frac{R_T(\underline{\nu})}{\ln T} = +\infty\,.$$

Interestingly, Cowan and Katehakis (2015) observe that for the model of uniform distributions over bounded intervals, the $\mathcal{K}_{\inf}$ is positive, and thus the lower bound of Reminder 1 does not prevent logarithmic regret bounds. In fact, they also provide an algorithm enjoying optimal distribution-dependent bounds—thus being, in a sense, adaptive to the range in that very restricted model.

Proof We denote by $[m,M]$ an interval containing the support of ν_a. We remind the reader of the model $\mathcal{D}_{m,+}$ defined in (6), composed of all bounded distributions with unknown upper end on the range but known lower end m on the range. As $\mathcal{D}_{m,+} \subset \mathcal{D}_{-,+}$ and by definition of $\mathcal{K}_{\inf}$,

$$\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D}_{-,+}) \leqslant \mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D}_{m,+})\,,$$

so that it suffices to show that $\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D}_{m,+}) = 0$.

We have in particular $\mu_a \geqslant m$. We use the same construction as in the proof of Theorem 3. Let $\nu'_\varepsilon = (1-\varepsilon)\nu_a + \varepsilon\delta_{\mu_a+2\Delta_a/\varepsilon}$ for $\varepsilon \in (0,1)$: it is a bounded probability distribution, with lower end of support larger than m, that is, $\nu'_\varepsilon \in \mathcal{D}_{m,+}$. For ε small enough, $\mu_a + 2\Delta_a/\varepsilon$ lies outside of the bounded support of ν_a. In that case, the density of ν_a with respect to ν'_ε is given by $1/(1-\varepsilon)$ on the support of ν_a and 0 elsewhere, so that

$$\mathrm{KL}\big(\nu_a, \nu'_\varepsilon\big) = \ln\!\left(\frac{1}{1-\varepsilon}\right).$$

Moreover, $\mathrm{E}\big(\nu'_\varepsilon\big) = (1-\varepsilon)\mu_a + \varepsilon\big(\mu_a + 2\Delta_a/\varepsilon\big) = \mu_a + 2\Delta_a = \mu^\star + \Delta_a > \mu^\star$. Therefore, by definition of $\mathcal{K}_{\inf}$ as an infimum,

$$\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D}_{m,+}) \leqslant \mathrm{KL}\big(\nu_a, \nu'_\varepsilon\big) = \ln\!\left(\frac{1}{1-\varepsilon}\right).$$

This upper bound holds for all $\varepsilon > 0$ small enough and thus shows $\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D}_{m,+}) = 0$.

The second part of the theorem follows from Reminder 1, from the existence of an arm a with $\Delta_a = \mu^\star - \mu_a > 0$, and from the fact that $\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D}_{-,+}) = 0$, as we established above. ■

Remark 12 *Recall that Remark 4 defined a notion of adaptation to the upper end M of the payoff range. The proof above reveals that Theorem 11 holds with all occurrences of $\mathcal{D}_{-,+}$ replaced by $\mathcal{D}_{m,+}$, for some $m \in \mathbb{R}$. We may therefore similarly exclude a $\ln T$ distribution-dependent rate for adaptation to the upper end M of the payoff range.*

This observation is yet another example that the knowledge of the lower end m of the payoff range does not critically change the picture, and the difficulty in ignoring a payoff range lies in ignoring the upper end thereof.

A.3 UCB with an Increased Exploration Rate Adapts to the Range

The impossibility result implied by Theorem 11 does not prevent distribution-dependent rates for adaptation that are larger than a logarithm. Let φ be a non-decreasing function such that $\varphi(t) \gg \ln t$, like $\varphi(t) = (\ln t)^2$ or even $\varphi(t) = (\ln t) \ln\ln t$. Lattimore (2017, Remark 8) introduced and studied, in the case of Gaussian bandits with unknown variances, the following variant of UCB, which we refer to in this section as UCB with an increased exploration rate φ:

$$\widehat{\mu}_a(t) + \sqrt{\frac{\varphi(t)}{N_a(t)}} \qquad \text{where} \quad \frac{\varphi(t)}{\ln t} \to +\infty \quad \text{and} \quad \frac{\varphi(t)}{t} \to 0\,,$$

and where $\widehat{\mu}_a(t)$ denotes the empirical average of payoffs obtained till round t when playing arm a. The (asymptotic only) analysis of Lattimore (2017, Remark 8) relies on the fact that $\varphi(t) \geqslant 2(M-m)\ln t$ for t larger than some unknown threshold T_0, and that after T_0, the indexes are thus larger than the ones of the original version of UCB based on the knowledge of m and M. This argument readily extends to the case of sub-Gaussian distributions, where we recall that a distribution ν with expectation μ is v–sub-Gaussian, with $v > 0$, if

$$\forall t \in \mathbb{R}, \qquad \int \mathrm{e}^{t(x-\mu)} \mathrm{d}\nu(x) \leqslant \mathrm{e}^{vt^2/2}\,.$$

Hoeffding's lemma proves that distributions over a bounded range $[m, M]$ are $(M-m)^2/4$–sub-Gaussian. Based on a slightly different proof than the one of Lattimore (2017, Remark 8), one can prove the following finite-time result—where we did not aim for tight numerical constants.

Theorem 13 *UCB with an increased exploration rate given by a non-decreasing function φ ensures that for all $v > 0$, for all distributions $\nu_1, \ldots, \nu_K$ that are v–sub-Gaussian, for all $T \geqslant K+1$,*

$$R_T(\underline{\nu}) \leqslant \underbrace{\sum_{a \in [K]:\Delta_a > 0} \frac{4}{\Delta_a}\varphi(T)}_{\text{main term}} + \underbrace{\sum_{a \in [K]} 2\Delta_a \max\left\{\frac{32v}{\Delta_a^2},\, 1\right\} \left(1 + \sum_{t=K}^{T-1} \mathrm{e}^{-\varphi(t)/(2v)}\right)}_{\text{smaller-order term: typically, a } \mathcal{O}(1)}$$

Whenever $\varphi \gg \ln$, this strategy is therefore adaptive to the unknown range of payoffs with a distribution-dependent rate $\Phi_{\mathrm{dep}} = \varphi$.

The second part of the statement follows from the claimed bound given that $\varphi \gg \ln$ entails $\varphi(t) \geqslant 4v \ln t$ for t large enough, and therefore, $\mathrm{e}^{-\varphi(t)/(2v)} \leqslant 1/t^2$. As a consequence, the sum tagged as smaller-order term in the bound is finite. Possible such choices are $\varphi : t \mapsto (\ln t)^2$, or even $\varphi : t \mapsto (\ln t)(\ln\ln t)$.

However, as already mentioned in Lattimore (2017), as the distribution-dependent rate approaches $\ln t$, the smaller-order term blows up. For example, if $\varphi(t) = (\ln t)^2$, the summands $\mathrm{e}^{-(\ln t)^2/(2v)}$ in the smaller-order term are larger than e^{-1} for all $t \leqslant \mathrm{e}^{\sqrt{2v}}$: the smaller-order term is at least of the order of $\mathrm{e}^{\sqrt{2v}}$, and the regret thus carries an exponential dependence on $\sqrt{v}$. In the case of a bounded range, this means an exponential dependence on the range $M-m$. This is probably not an artifact of the proof: in the case of a bounded range, as long as $\varphi(t) \ll (M-m)\ln t$, the lack of exploration bonus entails that the strategy behaves similarly to a follow-the-leader strategy, which is known to suffer catastrophic, i.e., linear, regret.

Proof As indicated above, we did not aim for tight numerical constants here and we somehow simplified the standard analysis of UCB by not considering thresholds of the form $\mu^\star - \varepsilon$ but rather $\mu^\star - \Delta_a/4$. Hence the non-standard (much increased) numerical factor in front of $\sum_a \ln T/\Delta_a$ when we specify $\varphi : t \mapsto 2(M-m)^2 \ln t$ into the bound.

In this proof, we repeatedly use that i.i.d. random variables $X_1, \ldots, X_n$ with a v–sub-Gaussian distribution with expectation μ satisfy, by the Cramér-Chernoff inequality, that for all $\varepsilon > 0$,

$$\begin{aligned}\mathbb{P}\!\left[\frac{1}{n}\sum_{i=1}^n X_i \geqslant \mu + \varepsilon\right] &\leqslant \inf_{\lambda>0}\ \mathrm{e}^{-n\lambda\varepsilon}\ \mathbb{E}\!\left[\exp\!\left(-\lambda\sum_{i=1}^n (X_i-\mu)\right)\right] \\ &\leqslant \inf_{\lambda>0} \mathrm{e}^{-n\lambda\varepsilon}\left(\mathrm{e}^{v\lambda^2/2}\right)^n = \mathrm{e}^{-n\varepsilon^2/(2v)}\,;\end{aligned}$$

and we obtain a similar inequality for deviations of the form "$\leqslant \mu - \varepsilon$".

Let $a^\star$ and a be an optimal and a suboptimal arm, respectively. Each arm is pulled once in the first K round. We bound $\mathbb{E}\big[N_a(T)\big]$ by using that for $t \geqslant K$, an arm A_{t+1} is pulled only if its index has the highest value, and then introduce the threshold $\mu^\star - \Delta_a/4$ to separate the $U_a(t)$ and the $U_{a^\star}(t)$:

$$\begin{aligned}
&\mathbb{E}\big[N_a(T)\big] \\
&\leqslant 1 + \sum_{t=K}^{T-1} \mathbb{P}\big[U_a(t) \geqslant U_{a^\star}(t) \text{ and } A_{t+1} = a\big] \\
&\leqslant 1 + \sum_{t=K}^{T-1} \mathbb{P}\!\left[U_a(t) \geqslant \mu^\star - \frac{\Delta_a}{4} \text{ and } A_{t+1} = a\right] + \sum_{t=K}^{T-1} \mathbb{P}\!\left[U_{a^\star}(t) \leqslant \mu^\star - \frac{\Delta_a}{4}\right] \\
&\leqslant 1 + \sum_{t=K}^{T-1} \mathbb{P}\!\left[\hat{\mu}_a(t) + \sqrt{\frac{\varphi(t)}{N_a(t)}} \geqslant \mu^\star - \frac{\Delta_a}{4} \text{ and } A_{t+1} = a\right] \\
&\qquad + \sum_{t=K}^{T-1} \mathbb{P}\!\left[\hat{\mu}_{a^\star}(t) \leqslant \mu^\star - \frac{\Delta_a}{4} - \sqrt{\frac{\varphi(t)}{N_{a^\star}(t)}}\right] \\
&\leqslant 1 + \sum_{t=K}^{T-1} \mathbb{P}\!\left[\hat{\mu}_a(t) + \sqrt{\frac{\varphi(T)}{N_a(t)}} \geqslant \mu^\star - \frac{\Delta_a}{4} \text{ and } A_{t+1} = a\right] \\
&\qquad + \sum_{t=K}^{T-1} \mathbb{P}\!\left[\hat{\mu}_{a^\star}(t) \leqslant \mu^\star - \frac{\Delta_a}{4} - \sqrt{\frac{\varphi(t)}{N_{a^\star}(t)}}\right] \\
&\leqslant 1 + \underbrace{\sum_{n=1}^{T-K+1} \mathbb{P}\!\left[\hat{\mu}_{a,n} \geqslant \mu^\star - \frac{\Delta_a}{4} - \sqrt{\frac{\varphi(T)}{n}}\right]}_{\mathrm{Sum}(a)} + \underbrace{\sum_{t=K}^{T-1}\sum_{n=1}^{t-K+1} \mathbb{P}\!\left[\hat{\mu}_{a^\star,n} \leqslant \mu^\star - \frac{\Delta_a}{4} - \sqrt{\frac{\varphi(t)}{n}}\right]}_{\mathrm{Sum}(a^\star)}.
\end{aligned}$$

Note that we used the fact that φ is non-decreasing to get to the last but one inequality, and we used optional skipping for the last one; we denote by $\hat{\mu}_{a,n}$ and $\hat{\mu}_{a^\star,n}$ the average of n i.i.d. rewards distributed according to ν_a and $\nu_{a^\star}$, respectively.

We first deal with $\mathrm{Sum}(a)$. Let $N_0 = \lceil 4\,\varphi(T)/\Delta_a^2 \rceil$. For $n \geqslant N_0$,

$$\mathbb{P}\!\left[\hat{\mu}_{a,n} \geqslant \mu^\star - \frac{\Delta_a}{4} - \sqrt{\frac{\varphi(T)}{n}}\right] \leqslant \mathbb{P}\!\left[\hat{\mu}_{a,n} \geqslant \mu_a + \frac{\Delta_a}{4}\right] \leqslant \mathrm{e}^{-n\Delta_a^2/(32v)} \,.$$

Therefore,

$$\mathrm{Sum}(a) \leqslant N_0 - 1 + \sum_{n \geqslant N_0} \mathrm{e}^{-n\Delta_a^2/(32v)} \leqslant \frac{4\varphi(T)}{\Delta_a^2} + \frac{1}{1-\mathrm{e}^{-\Delta_a^2/(32v)}} \leqslant \frac{4\varphi(T)}{\Delta_a^2} + 2\max\left\{\frac{32v}{\Delta_a^2},\,1\right\},$$

where we used[1] in the last step $1/(1-\mathrm{e}^{-x}) \leqslant 2/x$ for $x \in (0,1]$ and $1/(1-\mathrm{e}^{-x}) \leqslant 2$ for $x \geqslant 1$.

For $\mathrm{Sum}(a^\star)$, we apply the Cramér-Chernoff inequality, then use $(x+y)^2 \geqslant x^2+y^2$ for $x, y \geqslant 0$, and finally apply the same inequalities on $1/(1-\mathrm{e}^{-x})$ as for the other sum:

$$\sum_{t=K}^{T-1}\sum_{n=1}^{t-K+1} \mathbb{P}\!\left[\hat{\mu}_{a^\star,n} \leqslant \mu^\star - \frac{\Delta_a}{4} - \sqrt{\frac{\varphi(t)}{n}}\right] \leqslant \sum_{t=K}^{T-1}\sum_{n=1}^{t-K+1} \mathrm{e}^{-n\left(\Delta_a/4+\sqrt{\varphi(t)/n}\right)^2/(2v)}$$

$$\leqslant \sum_{t=K}^{T-1}\sum_{n=1}^{t-K+1} \mathrm{e}^{-n\Delta_a^2/(32v)}\ \mathrm{e}^{-\varphi(t)/(2v)} \leqslant \sum_{t=K}^{T-1} \mathrm{e}^{-\varphi(t)/(2v)}\,\frac{1}{1-\mathrm{e}^{-\Delta_a^2/(32v)}}$$

$$\leqslant 2\max\left\{\frac{32v}{\Delta_a^2},\,1\right\}\sum_{t=K}^{T-1} \mathrm{e}^{-\varphi(t)/(2v)} \,.$$

The proof is concluded by substituting the bounds in $R_T(\underline{\nu}) = \sum_{a\in[K]} \Delta_a\,\mathbb{E}\big[N_a(T)\big]$. ■

Appendix B. Proof of Theorem 7

How the second regret bound follows from the first one. We substitute the stated values of the γ_t. We have, first,

$$\sum_{t=K+1}^{T} \gamma_t \leqslant \sqrt{5(1-\alpha)K\ln K}\sum_{t=K+1}^{T} t^{-\alpha} \leqslant \sqrt{5(1-\alpha)K\ln K}\int_0^T \frac{1}{t^\alpha}\,\mathrm{d}t = \sqrt{\frac{5K\ln K}{1-\alpha}}\,T^{1-\alpha}\,, \tag{13}$$

second, using the definition of γ_T as a minimum,

$$\frac{K\ln K}{\gamma_T} \leqslant \frac{K\ln K}{1/2} + \frac{T^\alpha K\ln K}{\sqrt{5(1-\alpha)K\ln K}} = 2K\ln K + \sqrt{\frac{K\ln K}{5(1-\alpha)}}\,T^\alpha\,,$$

and third, $\sqrt{T} \leqslant T^{\max\{\alpha,1-\alpha\}}$, so that the first regret bound of Theorem 7 is further bounded by

$$(M-m)\sqrt{K\ln K}\left(3+2\sqrt{\frac{5}{1-\alpha}}\right)T^{\max\{\alpha,1-\alpha\}} + 10(M-m)K\ln K\,.$$

[1] For a bounded distribution, the case $x > 1$ does not occur as $x = \Delta_a^2/(32v) = \Delta_a^2/\big(8(M-m)^2\big) \leqslant 1/8$; but it may occur for other sub-Gaussian distributions.

The claimed expression for $\Phi_{\mathrm{adv}}(T)$ is obtained by bounding $2\sqrt{5}$ by 5.

First regret bound. In Algorithm 1, for time steps $t \geqslant K+1$, the weights q_t are obtained by using the AdaHedge algorithm of De Rooij et al. (2014) on the payoff estimates $\widehat{y}_{t,a}$. AdaHedge is designed for the case of a full monitoring—not of a bandit monitoring—, but the use of these estimates emulates a full monitoring. Section 2.2 of De Rooij et al. (2014)—see also an earlier analysis by Cesa-Bianchi et al. (2007)—ensures the bound stated next in Reminder 2.

We call pre-regret the quantity at hand in Reminder 2: it corresponds to some regret defined in terms of the payoff estimates.

Reminder 2 (Application of Lemma 3 and Theorem 6 of De Rooij et al., 2014) *For all sequences of payoff estimates $\widehat{y}_{t,a}$ lying in some bounded real-valued interval, denoted by $[b,B]$, for all $T \geqslant K+1$, the pre-regret of AdaHedge satisfies*

$$\max_{k\in[K]} \sum_{t=K+1}^{T} \widehat{y}_{t,k} - \sum_{t=K+1}^{T}\sum_{a=1}^{K} q_{t,a}\,\widehat{y}_{t,a} \leqslant 2 \sum_{t=K+1}^{T} \delta_t$$

$$\text{where} \qquad \sum_{t=K+1}^{T} \delta_t \leqslant \underbrace{\sqrt{\sum_{t=K+1}^{T}\sum_{a=1}^{K} q_{t,a}\left(\widehat{y}_{t,a} - \sum_{k\in[K]} q_{t,k}\,\widehat{y}_{t,k}\right)^2 \ln K}}_{\leqslant \sqrt{\sum_{t=K+1}^{T}\sum_{a=1}^{K} q_{t,a}(\widehat{y}_{t,a}-c)^2 \ln K} \quad \textit{for any } c\in\mathbb{R}} + (B-b)\left(1+\frac{2}{3}\ln K\right)$$

and AdaHedge does not require the knowledge of $[b,B]$ to achieve this bound.

The bound of Reminder 2 will prove itself particularly handy for three reasons: first, it is valid for real-valued payoffs; second, it is adaptive to the range of payoffs; third, the right-hand side looks at first sight not intrinsic enough a bound, as it also depends on the weights q_t, but we will see later that this dependency is particularly useful in our specific case. To the best of our knowledge, this is the first direct application of the AdaHedge bound depending on the weights q_t (previous applications were rather solving inequations on the regret, e.g., to get improvements for small losses; see Cesa-Bianchi et al., 2007 and De Rooij et al., 2014).

We recall that we start the summation in Reminder 2 at $t = K+1$ because the AdaHedge algorithm is only started at this time, after the initial exploration. The bound holding "for any $c \in \mathbb{R}$" is obtained by a classical bound on the variance.

Proof of the first bound of Theorem 7 We deal with the contribution of the initial exploration by using the inequality $\max(u+v) \leqslant \max u + \max v$, together with the fact that $y_{t,a} - y_{t,A_T} \leqslant M - m$ for any $a \in [K]$:

$$R_T(y_{1:T}) \leqslant \underbrace{\max_{a\in[K]} \sum_{t=1}^{K} y_{t,a} - \mathbb{E}\left[\sum_{t=1}^{K} y_{t,A_t}\right]}_{\leqslant K(M-m)} + \max_{a\in[K]} \sum_{t=K+1}^{T} y_{t,a} - \mathbb{E}\left[\sum_{t=K+1}^{T} y_{t,A_t}\right]. \qquad (14)$$

We now transform the pre-regret bound of Reminder 2, which is stated with the distributions q_t, into a pre-regret bound with the distributions p_t; we do so while substituting the bounds $B = C + KM/\gamma_T$ and $b = C + Km/\gamma_T$ implied by (10) and the fact that (γ_t) is non-increasing, and by using the definition $q_{t,a} = p_{t,a} - \gamma_t(1/K - q_{t,a})$ for all $a \in [K]$:

$$\begin{aligned} &\max_{k\in[K]} \sum_{t=K+1}^{T} \widehat{y}_{t,k} - \sum_{t=K+1}^{T}\sum_{a=1}^{K} p_{t,a}\,\widehat{y}_{t,a} + \sum_{t=K+1}^{T} \gamma_t \sum_{a=1}^{K} (1/K - q_{t,a})\,\widehat{y}_{t,a} \leqslant 2 \sum_{t=K+1}^{T} \delta_t \\ &\text{where} \qquad \sum_{t=K+1}^{T} \delta_t \leqslant \sqrt{\sum_{t=K+1}^{T}\sum_{a=1}^{K} q_{t,a}(\widehat{y}_{t,a} - C)^2 \ln K} + \frac{(M-m)K}{\gamma_T}\left(1 + \frac{2}{3}\ln K\right). \end{aligned} \tag{15}$$

As noted by Auer et al. (2002b), by the very definition (8) of the estimates,

$$\sum_{a=1}^{K} p_{t,a}\,\widehat{y}_{t,a} = y_{t,A_t}\,.$$

By (9), the tower rule and the fact that q_t is H_{t-1}–measurable, on the one hand, and the fact that the expectation of a maximum is larger than the maximum of expectations, on the other hand, the left-hand side of the first inequality in (15) thus satisfies

$$\begin{aligned} &\mathbb{E}\left[\max_{k\in[K]} \sum_{t=K+1}^{T} \widehat{y}_{t,k} - \sum_{t=K+1}^{T}\sum_{a=1}^{K} p_{t,a}\,\widehat{y}_{t,a} + \sum_{t=K+1}^{T} \gamma_t \sum_{a=1}^{K} (1/K - q_{t,a})\,\widehat{y}_{t,a}\right] \\ &\geqslant \max_{k\in[K]} \sum_{t=K+1}^{T} y_{t,k} - \mathbb{E}\left[\sum_{t=K+1}^{T} y_{t,A_t}\right] + \sum_{t=K+1}^{T} \gamma_t \left(\underbrace{\sum_{a=1}^{K} y_{t,a}/K}_{\in[m,M]} - \underbrace{\sum_{a=1}^{K} \mathbb{E}\big[q_{t,a}\big] y_{t,a}}_{\in[m,M]}\right) \\ &\geqslant \max_{k\in[K]} \sum_{t=K+1}^{T} y_{t,k} - \mathbb{E}\left[\sum_{t=K+1}^{T} y_{t,A_t}\right] - (M-m)\sum_{t=1}^{T} \gamma_t\,. \end{aligned}$$

As for the right-hand side of the second inequality in (15), we first note that by definition (see line 4 in Algorithm 1), $p_{t,a} \geqslant (1-\gamma_t)q_{t,a}$ with $\gamma_t \leqslant 1/2$ by assumption on the extra-exploration rate, so that $q_{t,a} \leqslant 2p_{t,a}$; therefore, by substituting first this inequality and then by using Jensen's inequality,

$$\begin{aligned} \mathbb{E}\left[\sqrt{\sum_{t=K+1}^{T}\sum_{a=1}^{K} q_{t,a}(\widehat{y}_{t,a} - C)^2 \ln K}\right] &\leqslant \sqrt{2}\,\mathbb{E}\left[\sqrt{\sum_{t=K+1}^{T}\sum_{a=1}^{K} p_{t,a}(\widehat{y}_{t,a} - C)^2 \ln K}\right] \\ &\leqslant \sqrt{2}\sqrt{\sum_{t=K+1}^{T}\sum_{a=1}^{K} \mathbb{E}\Big[p_{t,a}(\widehat{y}_{t,a} - C)^2\Big] \ln K}\,. \end{aligned} \tag{16}$$

Standard calculations (see Auer et al., 2002b again) show, similarly to (9), that for all $a \in [K]$,

$$\mathbb{E}\Big[p_{t,a}(\widehat{y}_{t,a} - C)^2 \,\Big|\, H_{t-1}\Big] = \mathbb{E}\left[\frac{(y_{t,A_t} - C)^2}{p_{t,a}} \mathbb{1}_{\{A_t = a\}}\right] = (y_{t,a} - C)^2 \leqslant (M-m)^2\,,$$

where the last inequality comes from (10). By the tower rule, the same upper bound holds for the (unconditional) expectation. Therefore, taking the expectation of both sides of (15) and collecting all bounds together, we proved so far

$$R_T(y_{1:T}) \leqslant \underbrace{2\sqrt{2}}_{\leqslant 3}(M-m)\sqrt{KT\ln K} + (M-m)\frac{K\ln K}{\gamma_T}\underbrace{\left(\frac{2+\gamma_T}{\ln K}+\frac{4}{3}\right)}_{\leqslant 5} + (M-m)\sum_{t=K+1}^{T}\gamma_t\,,$$

where we used $\gamma_T \leqslant 1/2$ and $\ln K \geqslant \ln 2$ as $K \geqslant 2$. ■

Appendix C. Proof of Theorem 9

Given the decomposition (1) of the regret, it is necessary and sufficient to upper bound the expected number of times $\mathbb{E}[N_a(t)]$ any suboptimal arm a is drawn, where by definition of Algorithm 1,

$$\mathbb{E}[N_a(t)] = 1 + \mathbb{E}\left[\sum_{t=K+1}^{T}\left((1-\gamma_t)q_{t,a} + \frac{\gamma_t}{K}\right)\right] \leqslant 1 + \sum_{t=K+1}^{T}\mathbb{E}[q_{t,a}] + \frac{1}{K}\sum_{t=K+1}^{T}\gamma_t\,.$$

We show below (and this is the main part of the proof) that

$$\sum_{t=K+1}^{T}\mathbb{E}[q_{t,a}] = \mathcal{O}(\ln T)\,. \tag{17}$$

The straightforward calculations (13) already showed that

$$\frac{1}{K}\sum_{t=K+1}^{T}\gamma_t \leqslant \sqrt{\frac{5\ln K}{(1-\alpha)K}}\,T^{1-\alpha}\,.$$

Substituting the value (11) of $\Phi^{\mathrm{AHB}}_{\mathrm{free}}(T) = \Phi_{\mathrm{adv}}(T)$ and using the decomposition (1) of $R_T(\underline{\nu})$ into $\sum \Delta_a\,\mathbb{E}[N_a(t)]$ then yield

$$\frac{R_T(\underline{\nu})}{T/\Phi^{\mathrm{AHB}}_{\mathrm{free}(T)}} \leqslant \sum_{a\in[K]}\Delta_a\sqrt{\frac{5\ln K}{(1-\alpha)K}}\left(3+\frac{5}{\sqrt{1-\alpha}}\right)\sqrt{K\ln K}\big(1+o(1)\big) + \mathcal{O}\!\left(\frac{\ln T}{T^{1-\alpha}}\right),$$

from which the stated bound follows, via the crude inequality $3\sqrt{5}\sqrt{1-\alpha}+5 \leqslant 12$.

Structure of the proof of (17). Let $a^\star$ denote an optimal arm. By definition of $q_{t,a}$ and by lower bounding a sum of exponential terms by any of the summands, we get

$$q_{t,a} = \frac{\exp\left(\eta_t \displaystyle\sum_{s=K+1}^{t-1}\widehat{y}_{t,a}\right)}{\displaystyle\sum_{k=1}^{K}\exp\left(\eta_t\sum_{s=K+1}^{t-1}\widehat{y}_{t,k}\right)} \leqslant \exp\left(\eta_t\sum_{t=K+1}^{t-1}(\widehat{y}_{t,a}-\widehat{y}_{t,a^\star})\right).$$

Then, by separating cases, depending on whether $\sum_{t=K+1}^{t-1}(\widehat{y}_{t,a} - \widehat{y}_{t,a^\star})$ is smaller or larger than the threshold $-(t-1-K)\Delta_a/2$, and by remembering that the probability $q_{t,a}$ is always smaller than 1, we get

$$\begin{aligned}\sum_{t=K+1}^{T} \mathbb{E}[q_{t,a}] \leqslant \quad & \sum_{t=K+1}^{T} \mathbb{E}\biggl[\exp\biggl(-\eta_t \frac{(t-1-K)\Delta_a}{2}\biggr)\biggr] \\ & + \sum_{t=K+1}^{T} \mathbb{P}\Biggl[\sum_{s=K+1}^{t-1} (\widehat{y}_{s,a} - \widehat{y}_{s,a^\star}) \geqslant -\frac{(t-1-K)\Delta_a}{2}\Biggr] .\end{aligned} \tag{18}$$

We show that the sums in the right-hand side of (18) are respectively $\mathcal{O}(1)$ and $\mathcal{O}(\ln T)$.

First sum in the right-hand side of (18). Given the definition of the learning rates (see the statement of Algorithm 1), namely,

$$\eta_t = \ln K \Bigg/ \sum_{s=K+1}^{t-1} \delta_s \,, \tag{19}$$

we are interested in upper bounds on the sum of the δ_s. Such upper bounds were already derived in the proof of Theorem 7; the second inequality in (15) together with the bound $q_{t,a} \leqslant 2p_{t,a}$ stated in the middle of the proof immediately yield

$$\begin{aligned}\sum_{s=K+1}^{t-1} \delta_s &\leqslant \sqrt{\sum_{s=K+1}^{t} \sum_{a=1}^{K} q_{s,a}\bigl(\widehat{y}_{s,a} - C\bigr)^2 \ln K} + \frac{(M-m)K}{\gamma_t}\biggl(1 + \frac{2}{3}\ln K\biggr) \\ &\leqslant \sqrt{2}\sqrt{\sum_{s=K+1}^{t} \sum_{a=1}^{K} p_{s,a}\bigl(\widehat{y}_{s,a} - C\bigr)^2 \ln K} + \frac{(M-m)K}{\gamma_t}\biggl(1 + \frac{2}{3}\ln K\biggr) .\end{aligned}$$

Unlike what we did to complete the proof of Theorem 7, we do not take expectations and rather proceed with deterministic bounds. By the definition (8) of the estimated payoffs for the equality below, by (10) for the first inequality below, and by the fact that the exploration rates are non-increasing for the second inequality below, we have, for all $s \geqslant K+1$,

$$\sum_{a=1}^{K} p_{s,a}\bigl(\widehat{y}_{s,a} - C\bigr)^2 = \frac{\bigl(y_{s,A_s} - C\bigr)^2}{p_{s,A_s}} \leqslant \frac{(M-m)^2}{\gamma_s/K} \leqslant \frac{(M-m)^2}{\gamma_t/K} \,. \tag{20}$$

Therefore,

$$\sum_{s=K+1}^{t-1} \delta_s \leqslant \sqrt{2}(M-m)\sqrt{\frac{t\,K\ln K}{\gamma_t}} + \frac{(M-m)K}{\gamma_t}\biggl(1 + \frac{2}{3}\ln K\biggr) \stackrel{\text{def}}{=} D_t = \Theta\Bigl(\sqrt{t/\gamma_t} + 1/\gamma_t\Bigr) .$$

For the sake of concision, we denoted by D_t the obtained bound. Via the definition (19) of η_t, the sum of interest is in turn bounded by

$$\sum_{t=K+1}^{T} \exp\biggl(-\eta_t\bigl(t-1-K\bigr)\frac{\Delta_a}{2}\biggr) \leqslant \sum_{t=K+1}^{T} \exp\biggl(-\frac{\Delta_a \ln K}{2}\,\frac{t-1-K}{D_t}\biggr) = \mathcal{O}(1) \,,$$

where the equality to $\mathcal{O}(1)$, i.e., the fact that the considered series is bounded, follows from the fact that

$$-(t-1-K)/D_t = \Theta\Big(\sqrt{t\gamma_t} + t\gamma_t\Big) = \Theta\big(t^{(1-\alpha)/2} + t^{1-\alpha}\big)\,.$$

Second sum in the right-hand side of (18). We will use Bernstein's inequality for martingales, and more specifically, the formulation of the inequality by Freedman (1975, Theorem 1.6)—see also Massart (2007, Section 2.2)—, as stated next.

Reminder 3 *Let $(X_n)_{n\geqslant 1}$ be a martingale difference sequence with respect to a filtration $(\mathcal{F}_n)_{n\geqslant 0}$, and let $N \geqslant 1$ be a summation horizon. Assume that there exist real numbers b and v_N such that, almost surely,*

$$\forall n \leqslant N, \quad X_n \leqslant b \qquad \textit{and} \qquad \sum_{n=1}^{N} \mathbb{E}\big[X_n^2 \,\big|\, \mathcal{F}_{n-1}\big] \leqslant v_N\,.$$

Then for all $\delta \in (0,1)$,

$$\mathbb{P}\left[\sum_{n=1}^{N} X_n \geqslant \sqrt{2v_N \ln\frac{1}{\delta}} + \frac{b}{3}\ln\frac{1}{\delta}\right] \leqslant \delta\,.$$

For $s \geqslant K+1$, we consider the increments $X_s = \Delta_a - \widehat{y}_{s,a^\star} + \widehat{y}_{s,a}$, which are adapted to the filtration $\mathcal{F}_s = \sigma(A_1, Z_1, \ldots, A_s, Z_s)$, where we recall that $Z_1, \ldots, Z_s$ denote the payoffs obtained in rounds $1, \ldots, s$. Also, as p_s is measurable with respect to past information $\mathcal{F}_{s-1}$ and since payoffs are drawn independently from everything else (see Section 2), we have, by the definition (8) of the estimated payoffs (where we rather denote by $Y_{s,a}$ the payoffs drawn at random according to ν_a, to be in line with the notation of Section 2 for stochastic bandits): for all $a \in [K]$,

$$\mathbb{E}\big[\,\widehat{y}_{s,a} \,\big|\, \mathcal{F}_{s-1}\big] = \frac{\mathbb{E}[Y_{s,a} \,|\, \mathcal{F}_{s-1}] - C}{p_{s,a}} \mathbb{1}_{\{A_s = a\}} + C = \frac{\mu_a - C}{p_{s,a}} \mathbb{1}_{\{A_s = a\}} + C = \mu_a\,.$$

As a consequence, $\mathbb{E}[X_s \,|\, \mathcal{F}_{s-1}] = \mathbb{E}\big[\Delta_a - \widehat{y}_{s,a^\star} + \widehat{y}_{s,a} \,|\, \mathcal{F}_{s-1}\big] = 0$. Put differently, $(X_s)_{s\geqslant K+1}$ is indeed a martingale difference sequence with respect to the filtration $(\mathcal{F}_s)_{s\geqslant K}$.

We now check that the additional assumptions of Reminder 3 are satisfied. Manipulations and arguments similar to the ones used in (10) and (20) show that for all $s \geqslant K+1$,

$$\begin{aligned}\Delta_a - \widehat{y}_{s,a^\star} + \widehat{y}_{s,a} &\leqslant \Delta_a - \frac{Y_{s,a^\star} - C}{p_{s,a}} \mathbb{1}_{\{A_s = a^\star\}} + \frac{Y_{s,a} - C}{p_{s,a}} \mathbb{1}_{\{A_s = a\}} \\ &\leqslant (M-m)(1 + K/\gamma_s) \leqslant b \overset{\text{def}}{=} (M-m)(1+K/\gamma_t)\,.\end{aligned}$$

For the variance bound, we first note that for all $s \leqslant t-1$, we have $(\widehat{y}_{s,a} - C)(\widehat{y}_{s,a^\star} - C) = 0$ because of the indicator functions, and therefore,

$$\begin{aligned}\mathbb{E}\Big[\big(\Delta_a - \widehat{y}_{s,a^\star} + \widehat{y}_{s,a}\big)^2 \,\Big|\, \mathcal{F}_{s-1}\Big] &\leqslant \mathbb{E}\Big[\big(\widehat{y}_{s,a^\star} + \widehat{y}_{s,a}\big)^2 \,\Big|\, \mathcal{F}_{s-1}\Big] \\ &\leqslant \mathbb{E}\Big[\big(\widehat{y}_{s,a^\star} - C\big)^2 \,\Big|\, \mathcal{F}_{s-1}\Big] + \mathbb{E}\Big[\big(\widehat{y}_{s,a} - C\big)^2 \,\Big|\, \mathcal{F}_{s-1}\Big]\,;\end{aligned}$$

in addition, for all $a \in [K]$, including $a^\star$,

$$\mathbb{E}\Big[\big(\widehat{y}_{s,a} - C\big)^2 \,\Big|\, \mathcal{F}_{s-1}\Big] = \mathbb{E}\bigg[\frac{(Y_{s,A_s} - C)^2}{p_{s,a}^2}\mathbb{1}_{\{A_s=a\}} \,\bigg|\, \mathcal{F}_{s-1}\bigg] \leqslant \frac{(M-m)^2}{p_{s,a}} \leqslant \frac{(M-m)^2 K}{\gamma_t}\,.$$

Therefore

$$\sum_{s=K+1}^{t-1} \mathbb{E}\Big[\big(\Delta_a - \widehat{y}_{s,a^\star} + \widehat{y}_{s,a}\big)^2 \,\Big|\, \mathcal{F}_{s-1}\Big] \leqslant \frac{2K(M-m)^2(t-1-K)}{\gamma_t} \leqslant v_t \overset{\text{def}}{=} \frac{2(M-m)^2 tK}{\gamma_t}\,.$$

Bernstein's inequality (Reminder 3) may thus be applied; the choice $\delta = 1/t$ therein leads to

$$\mathbb{P}\Bigg[\sum_{s=K+1}^{t-1} \big(\Delta_a - (\widehat{y}_{s,a^\star} - \widehat{y}_{s,a})\big) \geqslant \underbrace{2(M-m)\sqrt{\frac{tK}{\gamma_t}\ln t} + \frac{M-m}{3}\left(1 + \frac{K}{\gamma_t}\right)\ln t}_{\overset{\text{def}}{=} D'_t}\Bigg] \leqslant \frac{1}{t}\,.$$

As $\sqrt{t/\gamma_t} = \mathcal{O}(t^{(1+\alpha)/2})$ and $1/\gamma_t = \mathcal{O}(t^\alpha)$ as $t \to \infty$, where $\alpha < 1$, and as $\Delta_a > 0$ given that we are considering a suboptimal arm a, there exists $t_0 \in \mathbb{N}$ such that for all $t \geqslant t_0$,

$$D'_t \leqslant \frac{(t-1-K)\Delta_a}{2}$$

thus

$$\begin{aligned}
&\mathbb{P}\Bigg[\sum_{s=K+1}^{t-1} (\widehat{y}_{s,a} - \widehat{y}_{s,a^\star}) \geqslant -\frac{(t-1-K)\Delta_a}{2}\Bigg] \\
&= \mathbb{P}\Bigg[\sum_{s=K+1}^{t-1} \big(\Delta_a - (\widehat{y}_{s,a^\star} - \widehat{y}_{s,a})\big) \geqslant \frac{(t-1-K)\Delta_a}{2}\Bigg] \\
&\leqslant \mathbb{P}\Bigg[\sum_{s=K+1}^{t-1} \big(\Delta_a - (\widehat{y}_{s,a^\star} - \widehat{y}_{s,a})\big) \geqslant D'_t\Bigg] \leqslant \frac{1}{t}\,.
\end{aligned}$$

Therefore, as $T \to \infty$

$$\sum_{t=1}^{T} \mathbb{P}\Bigg[\sum_{t=K+1}^{t-1} (\widehat{y}_{t,a} - \widehat{y}_{t,a^\star}) \geqslant -\frac{(t-1-K)\Delta_a}{2}\Bigg] = \mathcal{O}(\ln T)\,,$$

as claimed. This concludes the proof.

References

C. Allenberg, P. Auer, L. Györfi, and G. Ottucsák. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In *Proccedings of the 17th International Conference on Algorithmic Learning Theory (ALT'06)*, pages 229–243. Springer, 2006.

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. Omnipress, 2009.

J.-Y. Audibert, R. Munos, and C. Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):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.

D. Baudry, P. Saux, and O.-A. Maillard. From optimality to robustness: Adaptive resampling strategies in stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 34, pages 14029–14041, 2021.

S. Bubeck, M.B. Cohen, and Y. Li. Sparsity, variance and curvature in multi-armed bandits. In *Proceedings of the 29th International Conference on Algorithmic Learning Theory (ALT'18)*, volume 83 of PMLR, pages 111–127, 2018.

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.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games.* Cambridge University Press, 2006.

N. Cesa-Bianchi and O. Shamir. Bandit regret scaling with the effective loss range. In *Proceedings of the 29th International Conference on Algorithmic Learning Theory (ALT'18)*, volume 83 of PMLR, pages 128–151, 2018.

N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2-3):321–352, 2007.

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

W. Cowan and M.N. Katehakis. An asymptotically optimal policy for uniform bandits of unknown support, 2015. Preprint, arXiv:1505.01918.

W. Cowan, J. Honda, and M.N. Katehakis. Normal bandits of unknown means and variances. *Journal of Machine Learning Research*, 18(154):1–28, 2018.

S. De Rooij, T. van Erven, P.D. Grünwald, and W.M. Koolen. Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15(37):1281–1316, 2014.

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

D.A Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1): 100–118, 1975.

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, 2019a. Preprint, arXiv:1805.05071.

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, 2019b.

S. Gerchinovitz and T. Lattimore. Refined lower bounds for adversarial bandits. In *Advances in Neural Information Processing Systems*, volume 29, pages 1198–1206, 2016.

H. Hadiji. Polynomial cost of adaptation for $\mathcal{X}$-armed bandits. In *Advances in Neural Information Processing Systems*, volume 32, pages 1029–1038, 2019.

J. Honda and A. Takemura. An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Machine Learning*, 85:361–391, 2011.

J. Honda and A. Takemura. Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards. *Journal of Machine Learning Research*, 16(113):3721–3756, 2015.

J. Huang, Y. Dai, and L. Huang. Scale-free adversarial multi-armed bandit with arbitrary feedback delays, 2021. Preprint, arXiv:2110.13400.

R. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems*, volume 17, pages 697–704, 2004.

J. Kwon and V. Perchet. Gains and losses are fundamentally different in regret minimization: The sparse case. *Journal of Machine Learning Research*, 17(227):1–32, 2016.

T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

T. Lattimore. A scale free algorithm for stochastic bandits with bounded kurtosis, 2017. Preprint arXiv:1703.08937, later published, with the omission of some remarks, in *Advances in Neural Information Processing Systems*, volume 30, pages 1584–1593, 2017.

P. Massart. *Concentration Inequalities and Model Selection*, volume XXXIII of *Ecole d'Eté de Probabilités de Saint-Flour*. Springer, 2007. Lectures given in 2003, published in 2007.

S.R. Putta and S. Agrawal. Scale-free adversarial multi armed bandits. In *Proceedings of the 33rd International Conference on Algorithmic Learning Theory (ALT'22)*, volume 167 of PMLR, pages 910–930, 2022.

Y. Seldin and G. Lugosi. An improved parametrization and analysis of the EXP3++ algorithm for stochastic and adversarial bandits. In *Proceedings of the 30th Annual Conference on Learning Theory (COLT'17)*, volume 65 of PMLR, pages 1743–1759, 2017.

G. Stoltz. *Incomplete Information and Internal Regret in Prediction of Individual Sequences*. PhD thesis, Université Paris-Sud, 2005. URL `https://tel.archives-ouvertes.fr/tel-00009759/document`.

T.S. Thune and Y. Seldin. Adaptation to easy data in prediction with limited advice. In *Advances in Neural Information Processing Systems*, volume 31, pages 2909–2918, 2018.

C.-Y. Wei and H. Luo. More adaptive algorithms for adversarial bandits. In *Proceedings of the 31st Conference On Learning Theory (COLT'18)*, volume 75 of PMLR, pages 1263–1291, 2018.

J. Zimmert and T. Lattimore. Connections between mirror descent, Thompson sampling and the information ratio. In *Advances in Neural Information Processing Systems*, volume 32, pages 11973–11982, 2019.

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'20)*, volume 89 of PMLR, pages 467–475, 2019.