# Anytime optimal algorithms in stochastic multi-armed bandits

**Rémy Degenne** REMY.DEGENNE@MATH.UNIV-PARIS-DIDEROT.FR
LPMA, Université Paris Diderot

**Vianney Perchet** VIANNEY.PERCHET@NORMALESUP.ORG
CREST, ENSAE

## Abstract

We introduce an anytime algorithm for stochastic multi-armed bandit with optimal distribution free and distribution dependent bounds (for a specific family of parameters). The performances of this algorithm (as well as another one motivated by the conjectured optimal bound) are evaluated empirically. A similar analysis is provided with full information, to serve as a benchmark.

## 1. INTRODUCTION

The classical sequential decision problem known as the "multi-armed bandit problem" (Thompson, 1933; Robbins, 1985) has been widely used in operations research, computer science and economics, in sequential clinical trials or to construct policies maximizing the "click-throughrate" (i.e., finding the ad with the highest probability of click). At each stage $t \in \mathbb{N}$, an agent takes a decision $\pi_t \in \{1, \dots, K\}$ (or "he pulls arm $\pi_t$") and gets a random reward, the objective being to maximize the cumulative reward. A popular class of algorithms designed for this problem use the principle of "optimism under uncertainty" (Auer et al., 2002; Audibert & Bubeck, 2010b).

There are several different frameworks that have been studied, depending on the feedbacks available to the algorithm (*bandit* vs. *full information*), on the rewards generating processes (*stochastic* vs. *adversarial*) and on the duration of the problem (*anytime* vs. *fixed horizon*).

**Feedbacks:** In the bandit setting, the feedback is only the reward of the pulled arm while with full information, it is the rewards of all the arms.

**Processes:** In the stochastic case we consider, the successive rewards of an arm $k \in \{1, \dots, K\}$ are i.i.d. samples from a distribution $\nu^{(k)}$ with mean $\mu^{(k)}$, such that

$$\forall y \in \mathbb{R}, \log \mathbb{E}_{\nu^{(k)}} \left[ e^{y(X-\mu^{(k)})} \right] \leq y^2/8 \, ,$$

that is, they are $\frac{1}{2}$ sub-Gaussian, up to the centering. The main example of such rewards are distributions with support in $[0, 1]$. In fact, while most papers on the subject restrict themselves to such bounded rewards, often only the more general sub-Gaussian assumption is used. The results can be extended to other sub-Gaussian variables by normalizing the constants in the bounds accordingly.

**Duration:** With a fixed known horizon $T$, the performance of an algorithm is only evaluated after $T$ stages. On the other hand, the performance of an anytime algorithm is evaluated at all stages $t \in \mathbb{N}$, up the appropriate time renormalization.

The objectif is to define an anytime policy $(\pi_t)_{t\geq 1}$ that maximizes the expected sum of rewards $\mathbb{E}[\sum_{s=1}^t X^{(\pi_s)}]$ with $X^{(\pi_s)} \sim \nu^{(\pi_s)}$, or equivalently that minimizes the expected regret $\mathbb{E}R_t$, where

$$R_t = \sum_{s=1}^{t} (\mu^* - \mu^{(\pi_s)}) \, .$$

Let us denote by $\mu^*$ the mean of the best arm, $\Delta_k = \mu^* - \mu^{(k)}$ the gap between arm $k$ and the optimal arm and $T^{(k)}(t)$ the number of times arm $k$ was pulled until time $t$. The regret is then equivalently written as

$$R_t = \sum_{k=1}^{K} \Delta_k T^{(k)}(t) \, .$$

In that setup, we can exhibit three main types of regret bounds at some stage $t$:

**Distribution free bound:** This bound is independent of the set of parameters $\Delta_k$ and only depends on $K$ .

*Proceedings of the 33^rd International Conference on Machine Learning*, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

**Distribution dependent:** This bound depends explicitly on the whole set of parameters $\{\Delta_1, \ldots, \Delta_k\}$ .

**Single parameter dependent:** Intermediate concept, this bounds depends solely on $\min_{k:\Delta_k>0} \Delta_k$ and $K$ .

We first prove that with full information, the simple follow-the-leader algorithm (FTL), that pulls at each stage the arm with the highest empirical mean, has explicit bounds of the three types. Namely, $\mathbb{E}R_t$ is simultaneously smaller, at all stage $t \in \mathbb{N}$ and up to universal multiplicative constant, than $\sqrt{\log(K)t}$, $\sum_k 1/\Delta_k$ and $\log(K)/\Delta_{\min}$ .

In the bandit case, the algorithm MOSS (Audibert & Bubeck, 2010b) have explicit optimal distribution free and single parameter dependent bounds. Successive Elimination (SE) (Perchet & Rigollet, 2013; Perchet et al., 2015) and Improved-UCB (Auer & Ortner, 2010) have explicit distribution free and distribution dependent bounds. Unfortunately, all those algorithms are not anytime (they require prior knowledge of the horizon), and using the classic doubling trick would only improve one or the other bounds, but not all of them at all stages. The algorithm UCB2 (Auer et al., 2002) is anytime with (slightly suboptimal) distribution free and distribution dependent bounds, but it is less natural than the other algorithms: it uses a rigid epoch system requiring that an arm be pulled multiple times in a row.

We introduce a new algorithm inspired from MOSS that fulfils the goal of both optimal distribution free and single parameter dependent bounds at all stages and that has the major practical advantage of being anytime. We show experimentally that the new algorithm presents a clear improvement upon MOSS used with a doubling trick. We also introduce a new anytime algorithm motivated by the conjectured optimal distribution dependent bound, and we compare empirically its behaviour to the aforementioned other algorithms.

## 2. FULL INFORMATION

The goal of this section is to present the full information case. It will be used as a benchmark to compare the different types of bounds obtained in the bandit setting with the ones of the FTL algorithm.

In the full information setting the rewards from all arms are observed after each stage. An algorithm can then attain finite regret, as we will prove for FTL, with a logarithmic dependency on the number of arms $K$. We will first consider the simpler case where all gaps are equal, then with general gaps. This bound also leads to a $\sqrt{t\log K}$ distribution-free upper bound. To the best of our knowledge, these distribution-dependent results do not appear in the literature yet they are quite interesting to establish a baseline against which we can compare the bandit settings.

As mentioned before, we start by a bound in the case where all suboptimal arms have the same gap $\Delta$. This upper bound will be used as a proof step for a bound for general gaps. We mention that in the case of two arms, it reduces to $4/\Delta$ which is indeed optimal since the expected regret is lower bounded by $1/4\Delta$, see e.g. (Bubeck et al., 2013): .

**Lemma 1.** *The expected regret of FTL in the full information setting with $K$ arms with equal gaps verifies,*

$$\mathbb{E}R_t \leq \frac{2}{\Delta}(2+\log(K-1)),\ \forall t \in \mathbb{N}\,.$$

*Proof.* For simplicity we prove the lemma in the case of an unique optimal arm, that we call arm 1. The adaptation to several optimal arms is straightforward. The expected regret incurred by following the policy $(\pi_s)_{s\geq 1}$ can be bounded as

$$\begin{aligned}\mathbb{E}R_t &\leq \Delta\mathbb{E}[\sum_{s=1}^{\infty}\mathbb{I}_{\{\pi_s\neq 1\}}] \\ &= \Delta\sum_{s=1}^{\infty}\mathbb{P}\{\exists k\in[2,K], \overline{X}_s^{(k)} > \overline{X}_s^{(1)}\}\,.\end{aligned}$$

If the algorithm pulls a suboptimal arm then either the optimal arm was underestimated or one of the suboptimal arms was overestimated. Let $\delta \in (0,\Delta)$, whose value is to be chosen later. We get

$$\begin{aligned}&\mathbb{P}\{\exists k\in[2,K], \overline{X}_s^{(k)} > \overline{X}_s^{(1)}\} \\ &\leq \mathbb{P}\{\overline{X}_s^{(1)}{\leq}\mu^{(1)}{-}\delta\} + \mathbb{P}\{\exists k\in[2,K], \overline{X}_s^{(k)}{>}\mu^{(1)}{-}\delta\}\,.\end{aligned}$$

By Hoeffding's inequality,

$$\Delta\sum_{s=1}^{\infty}\mathbb{P}\{\overline{X}_s^{(1)} \leq \mu^{(1)}-\delta\} \leq \Delta\sum_{s=1}^{\infty}e^{-2s\delta^2} \leq \frac{\Delta}{2\delta^2}\,,$$

and

$$\begin{aligned}&\mathbb{P}\{\exists k\in[2,K], \overline{X}_s^{(k)} > \mu^{(1)}-\delta\} \\ &\quad = 1-\prod_{k=2}^{K}(1-\mathbb{P}\{\overline{X}_s^{(k)}-\mu^{(k)} > \Delta-\delta\}) \\ &\quad \leq 1-(1-\exp(-2s(\Delta-\delta)^2))^{K-1}\,.\end{aligned}$$

Finally, with standard computations, we obtain

$$\begin{aligned}\mathbb{E}R_t &\leq \frac{\Delta}{2\delta^2} + \Delta\sum_{s=1}^{\infty}(1-(1-\exp(-2s(\Delta-\delta)^2))^{K-1}) \\ &\leq \frac{\Delta}{2\delta^2} + \frac{\Delta}{2(\Delta-\delta)^2}\sum_{k=1}^{K-1}\frac{1}{k}\,.\end{aligned}$$

With $\delta = \frac{\Delta}{2}$, we get

$$\mathbb{E}R_t \le \frac{2}{\Delta}(1 + \sum_{k=1}^{K-1} \frac{1}{k}) \le \frac{2}{\Delta}(2 + \log(K-1)) \,,$$

hence the result. □

The next theorem presents a similar bound for general gaps. For notational convenience, we reorder the arms by decreasing means, $0 = \Delta_1 \le \Delta_2 \le \ldots \le \Delta_K$. Of course, algorithms cannot use that information.

**Theorem 1.** *FTL in the full information setting verifies for any $k_0 \in \{1, \ldots, K\}$, for $t \ge 1$,*

$$\mathbb{E}R_t \le t\Delta_{k_0} + \frac{8}{\Delta_{k_0+1}}(2 + \log(K-1)).$$

*In particular,*

$$\mathbb{E}R_t \le \frac{8}{\Delta_{\min}}(2 + \log(K-1)) \,.$$

*Proof of Theorem 1.* The regret is bounded as

$$\mathbb{E}R_t \le t\Delta_{k_0} + \sum_{s=1}^{+\infty} \sum_{k>k_0} \Delta_k \mathbb{P}\{\pi_s = k\} \,.$$

Trying to immediately generalize the proof of Lemma 1 with different gaps leads to cumbersome, heavy computations. We therefore use instead a peeling idea.

We put the arms with gaps greater than $\Delta_{k_0}$ in $M$ groups $G_1, \ldots, G_M$ with increasing gaps and bound the regret for each group as in the case of equal arms. Let $K_m$ be the number of arms in group $G_m$ and $\Delta_{G_m,\min}, \Delta_{G_m,\max}$ be the smallest and biggest gaps in group $G_m$. We denote by $\tilde{\log}$ the function with value 0 when $\log$ is not defined and is equal to $\log$ otherwise.

$$\begin{aligned}\mathbb{E}R_t &\le t\Delta_{k_0} + \sum_{m=1}^{M} \frac{2\Delta_{G_m,\max}}{\Delta^2_{G_m,\min}}(2 + \tilde{\log}(K_m - 1)) \\ &\le t\Delta_{k_0} + 2(2 + \log(K-1)) \sum_{m=1}^{M} \frac{\Delta_{G_m,\max}}{\Delta^2_{G_m,\min}} \,.\end{aligned}$$

Let the groups be such that $i \in G_m \Leftrightarrow \Delta_i \in [\eta^{m-1}\Delta_{k_0+1}, \eta^m \Delta_{k_0+1})$ for some $\eta > 1$ that will be chosen later. A group can be empty. $M$, number of groups, is such that $\eta^{M-1}\Delta_{k_0+1} \le \Delta_{\max} < \eta^M \Delta_{k_0+1}$.

$$\begin{aligned}\mathbb{E}R_t &\le t\Delta_{k_0} + 2(2 + \log(K-1)) \sum_{m=1}^{M} \frac{\Delta_{G_m,max}}{\Delta^2_{G_m,min}} \\ &\le t\Delta_{k_0} + \frac{2}{\Delta_{k_0+1}}(2 + \log(K-1)) \sum_{m=1}^{M} \frac{1}{\eta^{m-2}} \\ &= t\Delta_{k_0} + \frac{2}{\Delta_{k_0+1}}(2 + \log(K-1))\eta \frac{\eta - (\frac{1}{\eta})^{M-1}}{\eta - 1} \,.\end{aligned}$$

We use $\eta^{M-1}\Delta_{k_0+1} \le \Delta_{\max}$ to get $(\frac{1}{\eta})^{M-1} \ge \frac{\Delta_{k_0+1}}{\Delta_{\max}}$,

$$\mathbb{E}R_t \le t\Delta_{k_0} + \frac{2}{\Delta_{k_0+1}}(2 + \log(K-1))\eta \frac{\eta - \frac{\Delta_{k_0+1}}{\Delta_{\max}}}{\eta - 1} \,.$$

We now choose $\eta$ to minimize this last expression, taking $\eta = 1 + \sqrt{1 - \frac{\Delta_{k_0+1}}{\Delta_{\max}}}$, and get

$$\begin{aligned}\mathbb{E}R_t &\le t\Delta_{k_0} + \frac{2}{\Delta_{k_0+1}}(2 + \log(K-1)) \left(1 + \sqrt{1 - \frac{\Delta_{k_0+1}}{\Delta_{\max}}}\right)^2 \\ &\le t\Delta_{k_0} + \frac{8}{\Delta_{k_0+1}}(2 + \log(K-1)) \,,\end{aligned}$$

which entails the result. □

The proof of Lemma 1 for $K = 2$ yields a regret smaller than $4/\Delta$. Using the same arguments for all the arms independently, gives that the regret of FTL is also upper bounded as $4\sum_{k,\Delta_k>0} 1/\Delta_k$. This can be a better bound than the one of Theorem 1 if $\Delta_{\min}$ is very small compared to the other gaps.

Another main purpose of Theorem 1 is that it entails a distribution-free upper bound by choosing $k_0$ such that $\Delta_{k_0} \le 2\sqrt{2t(2 + \log(K-1))} < \Delta_{k_0+1}$. We recall here that the arms have been reordered, to ease notations, with respect to the size of the gaps.

**Theorem 2.** *FTL in the full information satisfies,*

$$\sup_{\textit{distributions}} \mathbb{E}R_t \le 2\sqrt{2t(2 + \log(K-1))}, \ \forall t \in \mathbb{N} \,.$$

This distribution free bound in $O(\sqrt{t \log K})$ is optimal (Cesa-Bianchi & Lugosi, 2006).

We have therefore obtained regret bounds of the three different types for the algorithm FTL. In the next section, we introduce in the bandit setting an anytime algorithm with a single parameter dependent bound in $\frac{K}{\Delta_{\min}} \log(\frac{t\Delta^2_{\min}}{K})$ and an optimal distribution free bound $\sqrt{tK}$. This is the first anytime algorithm with both properties.

## 3. ANYTIME MINIMAX OPTIMAL ALGORITHM FOR BANDITS

In the bandit setting, algorithms based on optimism under uncertainty, like the seminal example UCB (Auer et al., 2002), were introduced to get an upper bound on the regret matching the lower bound for the stochastic bandit setting that was derived in (Lai & Robbins, 1985). We report a weaker version of this bound in Lemma 2.

**Lemma 2** ((Lai & Robbins, 1985))**.** *For all reward distributions, all strongly consistent policies (policies with*

$\mathbb{E}R_t = o(t^a)$ *for all* $a > 0$) *satisfy, for all suboptimal arms* $k \in \{1, \dots, K\}$,

$$\liminf_{t \to +\infty} \frac{\mathbb{E}T^{(k)}(t)}{\log t} \geq \frac{1}{KL(\nu^{(k)}, \nu^*)} ,$$

*where* $KL(\nu^{(k)}, \nu^*)$ *is the Kullback-Leibler divergence between the distribution* $\nu^{(k)}$ *and the distribution* $\nu^*$ *of the rewards of the optimal arm.*

Algorithms designed to get matching upper bounds for certain classes of distributions include Thompson Sampling (Kaufmann et al., 2012), KL-UCB (Cappé et al., 2013) and DMED (Honda & Takemura, 2010). The algorithms we study however have bounds expressed as functions of the gaps $\Delta_k$ and place emphasis on being applicable without prior knowledge of the reward distributions. Indeed the only hypothesis on these distributions is the subgaussian property.

Kullback's inequality on the Kullback-Leibler divergence implies that $KL(\nu^{(k)}, \nu^*) \geq 2\Delta_k^2$ where the constant 2 is the best possible. Optimal upper bounds for $T^{(k)}(t)$ in the sense of Lemma 2 expressed as functions of the gaps are of order $\log(t)/\Delta_k^2$. While the goal of the algorithm introduced here is not primarily to match closely this first lower bound, we want to maintain the correct dependencies, notably the logarithmic dependency in $t$.

In the bandit setting, a minimax optimal algorithm (optimal in the distribution free sense) has a regret at most proportional to $\sqrt{Kt}$, matching the following lower bound.

**Lemma 3** ((Auer et al., 1995))**.** *For the multi-armed stochastic bandit setting with* $K$ *arms,*

$$\inf_{policies} \sup_{distributions} \mathbb{E}R_t \geq \sqrt{Kt}/20 \, .$$

If a horizon $T$ is known in advance, the MOSS algorithm (Audibert & Bubeck, 2010b) enjoys this optimal $\sqrt{KT}$ upper bound while also having a single parameter dependent upper bound with a $\log T$ dependency. Contrary to the bound we proved for FTL, this is not valid for $t \neq T$. MOSS can be converted to the anytime setting with a doubling trick and keep an optimal $\sqrt{Kt}$ bound without the $\log t$ dependency optimal with respect to Lemma 2.

There are other algorithms with bounds close to $\sqrt{TK}$, such as Successive Elimination (SE) (Perchet & Rigollet, 2013; Perchet et al., 2015) and Improved UCB (Auer & Ortner, 2010), that both enjoy a bound with a $\sqrt{TK \log K}$ dependency. These algorithms are however not anytime. UCB2 and Thompson sampling with Gaussian priors (Agrawal & Goyal, 2013) are anytime algorithm with the same $\sqrt{TK \log K}$ bound. UCB2 also enjoy a distribution-dependent bound in $\sum_{k, \Delta_k > 0} \log(t\Delta_k^2)/\Delta_k$.

We introduce an algorithm inspired from MOSS, named MOSS-anytime. It is anytime minimax optimal and has an optimal single parameter dependent $O\left(\frac{K}{\Delta_{\min}} \log(\frac{t\Delta_{\min}^2}{K})\right)$ bound.

**Algorithm 1** MOSS-anytime.

```
Input: $\alpha > 0$.
Pull each arm once.
For $1 \leq k \leq K$, set $s_k = 1$.
**for** $t \geq 1$ **do**
    Pull arm $k$ that maximizes
        $\overline{X}^{(k)}_{s_k} + \sqrt{\frac{(1+\alpha)}{2} \frac{\max(0, \log(\frac{t}{K s_k}))}{s_k}}$ .
    Update the number of pulls: $s_k \leftarrow s_k + 1$.
**end for**
```

### 3.1. Anytime optimal upper bound

The best upper bound previously attained by an anytime algorithm with the $\log t$ distribution-dependent behaviour was $\sqrt{tK \log K}$, as obtained by UCB2 (Auer et al., 2002). While nearly anytime minimax optimal, UCB2 however used a block structure for the pulls and was thus not as convenient to use as UCB1 or MOSS. We were able to remove this structure and prove with a refined analysis that a single-pull variant of UCB2 enjoys the same bounds (see the supplementary material) while keeping the simplicity of UCB.

MOSS-anytime improves over UCB2 with respect to the distribution free bound by removing the $\log K$ gap. It improves over MOSS with a doubling trick by having both single parameter dependent and distribution free optimal bounds.

In the following, we define $\overline{\log}(x) := \max\{1, \log(x)\}$.

**Theorem 3** (Upper bounds for MOSS-anytime)**.** *In the* $K$ *arms bandit setting, for* $\alpha = 1.35$*, the expected regret of MOSS-anytime satisfies for all* $t \geq 1$

$$\mathbb{E}R_t \leq 75 \frac{K}{\Delta_{\min}} \left( \overline{\log}(\frac{2t\Delta_{\min}^2}{K}) + 1 \right) + \Delta_{\max}$$

*and*

$$\mathbb{E}R_t \leq 113\sqrt{Kt} + \Delta_{\max} \, .$$

Theorem 3 states a result only for $\alpha = 1.35$ but there exists similar bounds for all $\alpha \in [0, 1.35]$ (and for bigger $\alpha$ with modifications of the proof), with bigger constants. The constants go to infinity when $\alpha$ goes to zero. However, in contradiction with this analysis, experimental examination shows a better performance of the algorithm for $\alpha$ closer to zero.

The first upper bound matches a known single parameter dependent (and $K$ dependent) lower bound in the case of equal gaps that refines the dependency on $\Delta$ over the lower bound of Lemma 2.

**Lemma 4** ((Mannor & Tsitsiklis, 2004), (Bubeck et al., 2013) for $K = 2$)**.** *There exists a positive constant $C$ such that for all policies, for all $K \geq 2$ and all $\Delta > 0$, there exists a problem with gaps all equal to $\Delta$ such that for all $t \geq 1$,*

$$\mathbb{E}R_t \geq C\frac{K}{\Delta}\log\left(\frac{t\Delta^2}{K}\right) .$$

MOSS-anytime is the first anytime algorithm matching this lower bound while being minimax optimal at all stages.

*Sketch of the proof of Theorem 3.* The beginning of this proof uses a decoupling of the arms inspired from the proof of the upper bounds of MOSS (Audibert & Bubeck, 2010b) but then departs from it to control the probabilities of the suboptimal pulls in an anytime fashion. In this second part, the critical arguments are well chosen relative weights for the different sources of regret, the use of Hoeffding's maximal inequality and a peeling technique.

Let $k_0$ be an integer in $[1, K]$ that will be chosen later. Let $\epsilon_{t,s} = \sqrt{\frac{(1+\alpha)}{2}\frac{\max(0,\log(\frac{t}{Ks}))}{s}}$ be the exploration term of the algorithm and $\delta > 0$ a constant to be chosen later. For $k \in \{k_0+1,\dots,K\}$, we define $z_k = \mu^* - \delta\frac{\Delta_k}{2}$, $z_{k_0} = +\infty$ and $z_{K+1} = 0$. We will consider the smallest value possibly taken by the index of the optimal arm after time $t$,

$$A^*_t = \min_{s\geq 1}\min_{u\geq t}\overline{X}^*_s + \epsilon_{u,s} ,$$

and after $r$ pulls of suboptimal arms,

$$B^*_r = \min_{s\geq 1}\min_{u\geq r+s}\overline{X}^*_s + \epsilon_{u,s} .$$

**Step 1: separating the events that the optimal arm is underestimated or that a suboptimal arm is overestimated.** We allow a regret of $\Delta_{k_0}$ at each stage,

$$\mathbb{E}R_t \leq t\Delta_{k_0} + \mathbb{E}[\sum_{k=k_0+1}^{K}(\Delta_k - \Delta_{k_0})T^{(k)}(t)] .$$

We will bound the regret incurred for $k > k_0$. We note $\pi_s$ the arm pulled at time $s$.

$$\begin{aligned}&\mathbb{E}R_t - t\Delta_{k_0}\\ &\quad\leq \mathbb{E}[\sum_{k=k_0+1}^{K}(\Delta_k - \Delta_{k_0})\sum_{s\geq 0}\mathbb{I}_{\{k \text{ pulled at time } s\}}]\\ &\quad\leq \mathbb{E}[\sum_{k=k_0+1}^{K}\sum_{j=k_0}^{K}(\Delta_k - \Delta_{k_0})\sum_{s\geq 0}\mathbb{I}_{\{\pi_s=k, A^*_s\in[z_{j+1},z_j)\}}] .\end{aligned}$$

We now cut this quantity into two sums: one quantifying the event that the optimal arm is underestimated (against values depending on the arms) and a second one quantifying the event that one of the suboptimal arms is pulled even if the optimal arm is not underestimated.

$$\begin{aligned}&\mathbb{E}R_t - t\Delta_{k_0}\\ &\quad\leq \sum_{s\geq 0}\mathbb{E}[\sum_{j=k_0}^{K}\sum_{k=k_0+1}^{j}(\Delta_k - \Delta_{k_0})\mathbb{I}_{\{\pi_s=k, A^*_s\in[z_{j+1},z_j)\}}]\\ &\quad+ \sum_{s\geq 0}\mathbb{E}[\sum_{j=k_0}^{K}\sum_{k=j+1}^{K}(\Delta_k - \Delta_{k_0})\mathbb{I}_{\{\pi_s=k, A^*_s\in[z_{j+1},z_j)\}}]\end{aligned}$$

**Step 2: bounding the probability that the optimal arm is underestimated.** Using some standard computations (see the supplementary material for more details), we can reorder the sum and get

$$\begin{aligned}&\sum_{j=k_0}^{K}\sum_{k=k_0+1}^{j}(\Delta_k - \Delta_{k_0})\mathbb{I}_{\{\pi_s=k, A^*_s\in[z_{j+1},z_j)\}}\\ &\quad\leq \mathbb{I}_{\{\pi_s\in[k_0+1,K]\}}\sum_{j=k_0+1}^{K}(\Delta_j - \Delta_{j-1})\mathbb{I}_{\{A^*_s<z_j\}}\end{aligned}$$

We will rewrite the sum over $s$ of such terms as a sum over $r$, number of times that an arm in $[k_0+1, K]$ has been pulled. Note that if we know that suboptimal arms were pulled at least $r$ times before a time $s$, we get $A^*_s \geq B^*_r$.

$$\begin{aligned}&\sum_{s\geq 0}\mathbb{I}_{\{\pi_s\in[k_0+1,K]\}}\sum_{k=k_0+1}^{K}(\Delta_k - \Delta_{k-1})\mathbb{I}_{\{A^*_s<z_k\}}\\ &\quad\leq \sum_{r\geq 0}\sum_{k=k_0+1}^{K}(\Delta_k - \Delta_{k-1})\mathbb{I}_{\{B^*_r<z_k\}} .\end{aligned}$$

Intuitively, for each $r$, this is of the form $(\Delta_{k_r} - \Delta_{k_0})$ for some $k_r \geq k_0$ and is thus of the order of one $\Delta_{k_r}$.

The probability of the events in the sum is bounded as

$$\mathbb{P}\{B^*_r < z_k\} \leq \mathbb{P}\{\exists s\geq 1, \exists t' \geq r+s, \overline{X}^*_s + \epsilon_{t',s} < z_k\} .$$

We use the monotonicity of $u \mapsto \epsilon_{u,s}$ to simplify the event,

$$\begin{aligned}&\mathbb{P}\{\exists s\geq 1, \exists t'\geq r+s, \overline{X}^*_s+\epsilon_{t',s}<z_k\}\\ &\quad\leq \mathbb{P}\{\exists s\geq 1, \overline{X}^*_s + \epsilon^*_{r+s,s} < z_k\} .\end{aligned}$$

**Step 3: bounding the probability that a suboptimal arm is overestimated.** Similarly to the previous step, we reorder the sum describing this event,

$$\sum_{j=k_0}^{K}\sum_{k=j+1}^{K}(\Delta_k-\Delta_{k_0})\mathbb{I}_{\{\pi_s=k,A_s^*\in[z_{j+1},z_j)\}}$$
$$\leq \sum_{k=k_0+1}^{K}\Delta_k\mathbb{I}_{\{\pi_s=k,A_s^*\geq z_k\}}$$

As we did above, we replace the sum over the time of such terms by sums over the number of pulls of arms. Let $P_r^{(k)}$ be the event ”arm $k$ was pulled for the $r^{th}$ time”.

$$\sum_{s\geq 0}\sum_{k=k_0+1}^{K}\Delta_k\mathbb{I}_{\{\pi_s=k,A_s^*\geq z_k\}}$$
$$\leq \sum_{r\geq 0}\sum_{k=k_0+1}^{K}\Delta_k\mathbb{I}_{\{P_r^{(k)},\text{ at time } t_r\text{, and } \overline{X}_r^{(k)}+\epsilon_{t_r,r}\geq z_k\}}$$
$$\leq \sum_{r\geq 0}\sum_{k=k_0+1}^{K}\Delta_k\mathbb{I}_{\{\exists t'\geq r,\overline{X}_r^{(k)}+\epsilon_{t',r}\geq z_k\}}\ .$$

For this sum, for each $r$ we get a sum that intuitively can be of order $\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k_0})$, that is roughly $K$ times larger than the sum depending on the optimal arm.

We use the monotonicity of $u\mapsto\epsilon_{u,s}$ to simplify the event,

$$\mathbb{P}\{\exists t'\geq r,\overline{X}_r^{(k)}+\epsilon_{t',r}\geq z_k\}\leq\mathbb{P}\{\overline{X}_r^{(k)}+\epsilon_{t,r}\geq z_k\}\ .$$

**Step 4: Controlling the probabilities.** Putting the two previous steps together we get the inequality

$$\begin{aligned}\mathbb{E}R_t &\leq t\Delta_{k_0}\\ &+\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k-1})\sum_{r\geq 0}\mathbb{P}\{\exists s\geq 1,\overline{X}_s^*+\epsilon_{r+s,s}^*<z_k\}\\ &+\sum_{k=k_0+1}^{K}\Delta_k\sum_{r\geq 0}\mathbb{P}\{\overline{X}_r^{(k)}+\epsilon_{t,r}\geq z_k\}\ .\end{aligned}$$

The next step is to control the sums of probabilities, which are small for $r$ big enough. To this effect we cut the sums in two, a first part for small $r$ for which the probability is upper bounded by 1 and a second part for big $r$. As noted previously, intuitively the first sum tend to be $K$ times smaller than the second one. Thus we cut the sums at indices that differ by a factor $K$ (up to a $(1+80\alpha)\overline{\log}(\frac{2t\Delta_k^2}{K})$ term).

Let $\tilde{r}_k$ be the largest integer such that $\tilde{r}_k\leq\frac{K}{2\Delta_k^2}+1$ and $\tilde{r}'_k$ the largest integer such that $\tilde{r}'_k\leq\frac{(1+80\alpha)\overline{\log}(\frac{2t\Delta_k^2}{K})}{2\Delta_k^2}$.

$$\begin{aligned}\mathbb{E}R_t &\leq t\Delta_{k_0}+\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k-1})\tilde{r}_k\\ &+\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k-1})\sum_{r>\tilde{r}_k}\mathbb{P}\{\exists s\geq 1,\overline{X}_s^*+\epsilon_{r+s,s}^*<z_k\}\\ &+\sum_{k=k_0+1}^{K}\Delta_k\tilde{r}'_k+\sum_{k=k_0+1}^{K}\Delta_k\sum_{r>\tilde{r}'_k}\mathbb{P}\{\overline{X}_r^{(k)}+\epsilon_{t,r}\geq z_k\}\\ &=t\Delta_{k_0}+A+B+C+D\ ,\end{aligned}$$

where $A,B,C,D$ are the four sums of the previous equation.

*Bounding term A.* Since $\tilde{r}_k\leq\frac{K}{2\Delta_k^2}+1$ ,

$$\begin{aligned}A&\leq\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k-1})(\frac{K}{2\Delta_k^2}+1)\\ &\leq\frac{K}{\Delta_{k_0+1}}+\Delta_K\ .\end{aligned}$$

*Bounding term B.* With Lemma 4 of the supplementary material,

$$B\leq\frac{4(1+\alpha)^{3/2}}{\alpha^2\log(1+\alpha)}\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k-1})\frac{K}{\Delta_k^2}\log(\frac{2et\Delta_k^2}{K})$$

We compute the sum with a sum-integral comparison and get

$$B\leq\frac{8(1+\alpha)^{3/2}}{\alpha^2\log(1+\alpha)}\left(\overline{\log}(\frac{2t\Delta_{k_0+1}^2}{K})\frac{K}{\Delta_{k_0+1}}+\frac{2K}{\Delta_{k_0+1}}\right)$$

*Bounding term C.* $\tilde{r}'_k\leq\frac{(1+80\alpha)\overline{\log}(\frac{2t\Delta_k^2}{K})}{2\Delta_k^2}$ and then

$$\begin{aligned}C&\leq\sum_{k=k_0+1}^{K}\Delta_k\frac{(1+80\alpha)\overline{\log}(\frac{2t\Delta_k^2}{K})}{2\Delta_k^2}\\ &\leq(1/2+40\alpha)\frac{K}{\Delta_{k_0+1}}\overline{\log}(\frac{2t\Delta_{k_0+1}^2}{K})\ .\end{aligned}$$

*Bounding term D.* Since for $r>\tilde{r}'_k$, $r>\frac{(1+80\alpha)\overline{\log}(\frac{2t\Delta_k^2}{K})}{2\Delta_k^2}\geq\frac{1}{2\Delta_k^2}$, it is not difficult to see that $\epsilon_{t,r}\leq\Delta_k\sqrt{\frac{1+\alpha}{1+80\alpha}}$ and thus that Lemma 2 of the supplementary material applies.

$$D\leq\sum_{k=k_0+1}^{K}\frac{32}{\alpha^2\Delta_k}\leq\frac{32K}{\alpha^2\Delta_{k_0+1}}\ .$$

**Step 5: Putting things together.** Define $C_\alpha$ and $C'_\alpha$ as

$$C_\alpha = \frac{8(1+\alpha)^{3/2}}{\alpha^2 \log(1+\alpha)} + 40\alpha + 1/2$$
$$C'_\alpha = \frac{16(1+\alpha)^{3/2}}{\alpha^2 \log(1+\alpha)} + 1 + \frac{32}{\alpha^2} .$$

We get the following bound for the regret, for any $k_0$,

$$\mathbb{E}R_t \le t\Delta_{k_0} + C_\alpha \frac{K}{\Delta_{k_0+1}} \overline{\log}(\frac{2t\Delta_{k_0+1}^2}{K}) + C'_\alpha \frac{K}{\Delta_{k_0+1}} + \Delta_K .$$

Using the same argument as for Theorem 1, we can then get two particular upper-bounds. The first one is obtained with $k_0$ the number of the last optimal arm,

$$\mathbb{E}R_t \le C_\alpha \frac{K}{\Delta_{\min}} \overline{\log}(\frac{2t\Delta_{\min}^2}{K}) + C'_\alpha \frac{K}{\Delta_{\min}} + \Delta_K ,$$

then one upper-bound independent of the distributions, by taking $k_0$ such that $\Delta_{k_0} \le \sqrt{\frac{K}{t}} < \Delta_{k_0+1}$ ,

$$\mathbb{E}R_t \le \sqrt{Kt}(1 + C_\alpha \log(2) + C'_\alpha) + \Delta_K .$$

For $\alpha = 1.35$, the maximum value allowed by Lemma 2 of the supplementary material, $C_\alpha \le 75, C'_\alpha \le 60$ and $(1 + C_\alpha \log(2) + C'_\alpha) \le 113$. Hence the result.

□

### 3.2. Comparison with other algorithms

While MOSS-anytime is optimal with respect to the distribution-free bound and the single parameter dependent bound of Lemma 4 with equal gaps, other algorithms can have a better behaviour when the gaps are not all equal. In particular, MOSS-anytime does not have a distribution dependent upper bound of the form of Lemma 2, in $\sum_{k,\Delta_k>0} \frac{\log t}{\Delta_k}$.

UCB, UCB2 and its single-pull variant, Improved UCB (Auer & Ortner, 2010) and SE are algorithms that all get upper bounds expressed as sums of terms depending on each gap $\Delta_k$, contrary to MOSS and MOSS-anytime that only gets a bound expressed as a function of the smallest gap $\Delta_{\min}$. UCB2, Improved UCB and SE get bounds of the form $\mathbb{E}R_T \le C\sum_{k,\Delta_k>0} \frac{\log(T\Delta_k^2)}{\Delta_k}$. When $\Delta_{\min}$ is very small compared to the other gaps, the problem can be significantly easier than expressed by an upper bound written as function of $\Delta_{\min}$ and the bounds for these other algorithms can be more advantageous.

Finding an algorithm that is optimal with respect to the lower bounds of the 3 precedent Lemmas is still an open problem. We conjecture that there exists an algorithm that enjoys a bound of the form $\left(C\sum_{k,\Delta_k>0} \frac{\overline{\log}(tH)}{\Delta_k} + C_K\right)$, where $C$ and $C_K$ does not depend on the gaps or $t$, $C$ does nor depend on $K$ and the constant $H = \left(\sum_{k,\Delta_k>0} \frac{1}{\Delta_k^2}\right)^{-1}$ is believed to describe the difficulty of the problem (Audibert & Bubeck, 2010a). This bound has the same form as the one of MOSS-anytime in the case of equal gaps and gets smaller when some gaps are bigger.

It is shown in (Lattimore, 2015) that such a bound with $C_K = 0$ cannot be attained for a horizon $T = K^2$ but this does not exclude that the bound be valid for larger horizons or for $C_K$ a non-zero function of $K$ that would be the dominant term for small $T$.

## 4. EXPERIMENTS

This section is dedicated to the experimental comparison of the algorithms introduced, to which we add a new algorithm for which we do not provide theoretical analysis but that seems promising in the experiments. This new algorithm, based on the constant $H$ presented in the previous sections, is a variant of UCB with the index

$$I_t^{(k)} = \overline{X}_t^{(k)} + \sqrt{\frac{1}{2s_k} \max\left\{0, \log\left(t(\sum_{j,s_j<\sqrt{t}} s_j)^{-1}\right)\right\}}$$

where the $s_j$ are the number of pulls of the arms until time $t$. The idea is that dividing by one $s_j$ will lead to a $1/\Delta_j^2$ in the logarithm appearing in the bound, similarly to what happens in UCB2 and MOSS. Summing the number of pulls that are smaller than $\sqrt{t}$ is a way to sum only the pulls of the suboptimal arms and obtain an approximation of $H$. Indeed an optimal arm would have been pulled a linear number of times and will not be included in the sum, whereas a suboptimal arm should have $s_j \propto \log t$ and will be summed. The hope is that this algorithm will enjoy a bound of the form $\sum_{k,\Delta_k>0} \overline{\log}(tH)/\Delta_k$, at least for large $t$, and we include it here to show as a first step that it has a good experimental performance.

We compare experimentally the different anytime algorithms studied here on synthetic data. The reward variables used are all Gaussian with variance $\sigma^2 = 1/2$. While the unique best arm will always have mean 0, the gaps between this arm and the 9 suboptimal arms are the main parameters influencing the behavior of the algorithms and depend on the experiment.

UCB will be omitted in the figures because its high regret otherwise distorts the plots and masks the differences between the other algorithms.

Note that MOSS, although included in the figures for comparison, is not an anytime algorithm.

In the first case, reported in Figure 1, all 9 suboptimal arms

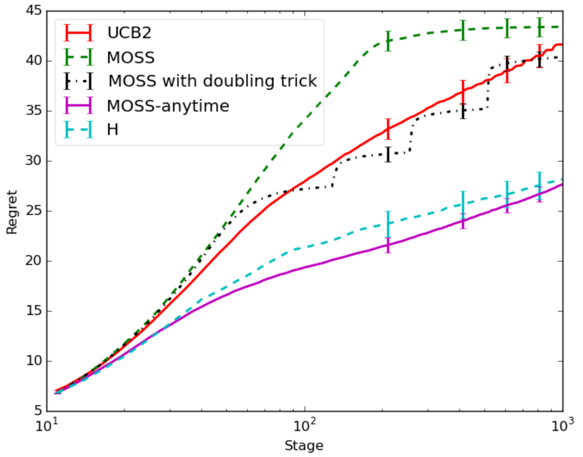

*Figure 1.* Regrets of the algorithms in the equal gaps case, averaged over 100 runs.

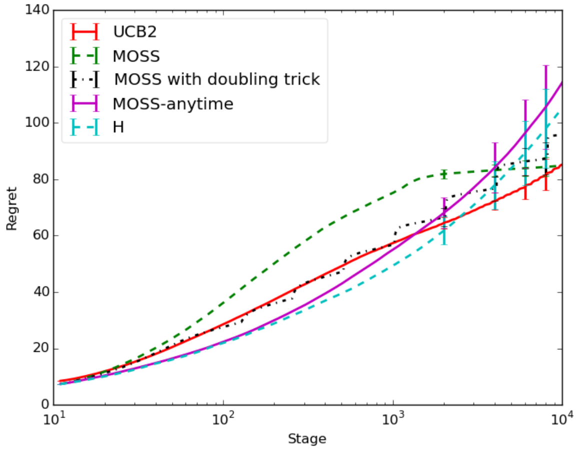

*Figure 2.* Regrets of the algorithms in the increasing gaps case, averaged over 800 runs.

have the same gap $\Delta = \sigma$. This is the best scenario for the bound of MOSS and MOSS-anytime. In practice, while MOSS and MOSS with a doubling trick behave similarly to UCB2, MOSS-anytime significantly outperforms the other algorithms.

Figure 2 shows the results of an experiment with increasing gaps: the 9 suboptimal arms have gaps increasing linearly between $\sigma$ and $3\sigma$. This scenario with different gaps theoretically favours UCB2 over MOSS and its variants. We note that MOSS-anytime has a regret slightly higher than the ones of the other algorithms and has a higher variance. More detailed examination of the data showed that this variance is due to rare runs with very high regret.

In both experiments, the $H$-inspired algorithm shows promising experimental behavior by performing similarly to MOSS-anytime on those ranges of number of steps. This said, we expect that the bigger the number of steps, the better the $H$-inspired algorithm behaves.

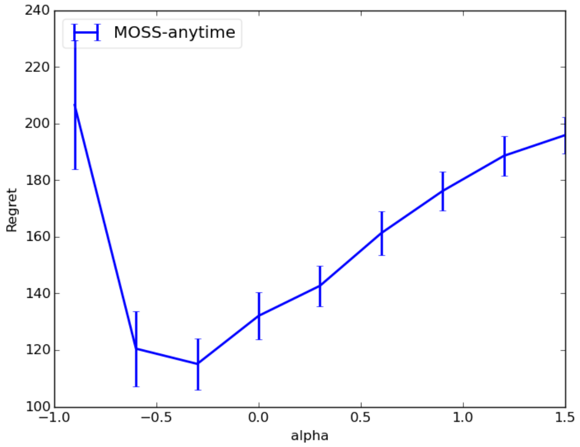

*Figure 3.* Regret of MOSS-Anytime in the increasing gaps case for different values of $\alpha$, averaged over 400 runs.

In both experiments, $\alpha$ was taken equal to 0.1. This is theoretically worse than the 1.35 value used in Theorem 3 but an experimental study indicated than smaller values of $\alpha$ were better (see Figure 3). Negative values of $\alpha$ are not covered by our theoretical analysis and lead to a high variance ($\alpha = -1$ gives the FTL algorithm) so we took $\alpha$ small and positive.

As a last note, further experiments show that the single-pull UCB2 algorithm mentioned in section 3.1 is experimentally very close to UCB2 in all cases (see the supplementary material): the complicated structure of pulls of UCB2 can be safely removed from both theoretical and practical points of view.

## 5. CONCLUSION

In the full information and in the bandit stochastic settings, we investigated anytime algorithms with regret bounded optimally both from a distribution dependent (or single parameter dependent) and a distribution free points of view. We proved that this is realized by the classic follow-the-leader algorithm in full information. In the bandit setting, we introduced the MOSS-anytime algorithm and proved that it is minimax optimal and has an optimal single parameter dependent bound function of the minimum gap $\Delta_{\min}$.

An algorithm that gets the same properties as MOSS-anytime and additionally a bound using the gaps of all the arms is still an open problem. A candidate algorithm for this task shows promising empirical performance and would require involved theoretical analysis.

## Acknowledgements

The second author would like to thank the fundings from the ANR under grant number ANR-13-JS01-0004 as well as from the PGMO Program under the project "Nougat" and the CNRS under the project "Parasol".

## References

Agrawal, Shipra and Goyal, Navin. Further Optimal Regret Bounds for Thompson Sampling. *Aistats*, 31:99–107, 2013.

Audibert, Jean-Yves and Bubeck, Sébastien. Best arm identification in multi-armed bandits. In *Proceedings of the 23th Conference on Learning Theory (COLT)*, pp. 13–p, 2010a.

Audibert, Jean-yves and Bubeck, Sbastien. Regret Bounds and Minimax Policies under Partial Monitoring. *Journal of Machine Learning Research*, 11:2785–2836, 2010b.

Auer, Peter and Ortner, Ronald. Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.

Auer, Peter, Cesa-Bianchi, Nicolo, Freund, Yoav, and Schapire, Robert E. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings ot the 36th Annual Symposium on Foundations of Computer Science*, pp. 322–331. IEEE, 1995.

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

Bubeck, Sébastien, Perchet, Vianney, and Rigollet, Philippe. Bounded regret in stochastic multi-armed bandits. *Journal of Machine Learning Research*, 2013.

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

Cesa-Bianchi, Nicolo and Lugosi, Gábor. *Prediction, learning, and games*. Cambridge University Press, 2006.

Honda, Junya and Takemura, Akimichi. An asymptotically optimal bandit algorithm for bounded support models. In *Proceedings of the 23rd Conference on Learning Theory (COLT)*, pp. 67–79. Citeseer, 2010.

Kaufmann, Emilie, Korda, Nathaniel, and Munos, Rémi. Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory*, pp. 199–213. Springer, 2012.

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

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

Mannor, Shie and Tsitsiklis, John N. The sample complexity of exploration in the multi-armed bandit problem. *The Journal of Machine Learning Research*, 5:623–648, 2004.

Perchet, Vianney and Rigollet, Philippe. The multi-armed bandit problem with covariates. *The Annals of Statistics*, 41(2):693–721, 2013.

Perchet, Vianney, Rigollet, Philippe, Chassang, Sylvain, and Snowberg, Erik. Batched bandit problems. *To appear in The Annals of Statistics*, 2015.

Robbins, Herbert. Some aspects of the sequential design of experiments. In *Herbert Robbins Selected Papers*, pp. 169–177. Springer, 1985.

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