Bandits with Side Observations: Bounded vs. Logarithmic Regret

Rémy Degenne
LPSM, CNRS, Sorbonne Université,
Université Paris Diderot, 75013 Paris, France;
CMLA, ENS Cachan, CNRS,
Université Paris-Saclay, 94235 Cachan, France

Evrard Garcelon
CMLA, ENS Cachan,
94235 Cachan, France

Vianney Perchet
CMLA, ENS Cachan, CNRS,
Université Paris-Saclay, 94235 Cachan, France
Criteo AI Lab
75009 Paris

Abstract

We consider the classical stochastic multi-armed bandit but where, from time to time and roughly with frequency ϵ, an extra observation is gathered by the agent for free. We prove that, no matter how small ϵ is the agent can ensure a regret uniformly bounded in time.

More precisely, we construct an algorithm with a regret smaller than $\sum_i \frac{\log(1/\epsilon)}{\Delta_i}$, up to multiplicative constant and $\log \log$ terms. We also prove a matching lower-bound, stating that no reasonable algorithm can outperform this quantity.

1 INTRODUCTION

We consider the celebrated multi-armed bandit framework (sometimes also called online learning), a repeated decision problem where an agent (or an algorithm, a machine, a player, etc.) takes sequentially decisions from a finite set. Each decision gives a stochastic reward to the agent of fixed expectation. The main objective is to derive an algorithm maximizing the cumulative reward or minimizing its normalized version, the so-called "regret". The latter is simply the difference between the cumulative expected reward of an agent knowing in hindsight the optimal decision, and the cumulative reward of the algorithm.

Online learning can be traced back to the 30's, when Thompson analysed random clinical trial using an analogy with finding the best slot-machine in a casino by pulling sequentially their arms in order to minimize the total loss. During the 20th century, many improvements have been made, at least on the asymptotic version of the problem. The quantity of theoretical studies and practical applications of bandits have exploded since the early 2000. There are several reasons for that. First of all, a simple yet almost optimal algorithm called UCB has been developed. Its simple structure allows to adapt it to many different settings. As a consequence, many possible applications of online learning have been developed. Amongst them, we can mention the routing problem: given a network with congested edges, one must find the quickest way from some origin to a destination (this setting incorporates a combinatorial structure); this can be used to send packets in a network, as well as finding the quickest itinerary from a point A to a point B. Online advertising is another application: given a possible set of ads, one must find the ad with the highest probability of click. The last application we mention is concerned with wireless network and/or cognitive radio, where either a radio can change from an available channel to other channels to improve its reception or emission quality, or alternatively a wireless source, in a relay selection problem where multiple relays are available, can explore those nodes to achieve better transmissions rates. One of the typical and crucial assumption of all these models is that the agent only observes the outcome of his decisions, but not what the other decisions would have given him. For instance, using a slot machine only gives you a feedback on the performance of that very machine, displaying an ad only gives information of the probability of clicks on that specific ad, etc. This assumption is actually called "bandit feedback". At the other end of the spectrum, the dual assumption (mostly used in non-stationary environment that we are not concerned with in that paper) is the "full information feedback", where all the outcomes of all decisions are observed at all stages. However, none of our motivating examples satisfies this strong assumption.

However, we argue that the bandit feedback is also too strong and that in many cases more informations are available to the agent. Typically, the agent will always observe the outcome of his own decision, but with some small probability he might also get one (or several, but that is irrelevant to our setting) extra "free" information. For instance, consider the original multi-armed bandit problem. A gambler is in a casino and wants to find out which slot machine is the best one. From time to time, he might observe other gamblers playing nearby machines. Even if this does not cost him anything, he gets feedback on the other machines. This effect also appears in other settings. In wireless network, a source with an allocated transmission capacity (because of a power-saving allocation protocol for instance) sends data through a relay and may have the opportunity to send another custom packet (so that the energy needed to send this packet is less than the available energy) through another relay in order to estimate transmissions rates. In online advertisement (and actually many other industrial markets), companies are willing to spend a small fraction of their data, say with probability ε as in the celebrated ε-greedy algorithm, just to acquire new information. An algorithm is only evaluated on the remaining (of proportion $1-\varepsilon$) fraction of the data treated. In a multi-armed bandit setting, this means that with probability ε, the next decision is "free". Finally, we can also think that in the congested network problem, an algorithm can from time to time send "fake", but free, packets to test the congestion; conversely, an app trying to minimize the congestion time of its users might be able to use free information if it notices that a bucket of users (for instance, those that are registered) might explore new road willingly, i.e., without uninstalling the app.

We therefore focus on the classical multi-armed bandits but where some extra and free information is available from time to time. Clearly, if the probability ε that it happens is arbitrarily close to 0, the improvement will be negligible. But we aim at constructing "optimal" algorithm, i.e., whose regret is small and in a multiplicative constant of the best regret achievable regret by "meaningful" algorithms. All these concepts are explained in details in the remaining of the paper that is organized as follows.

The model is introduced in Section 2, where we provide a very naïve algorithm achieving bounded regret (uniformly in time). We exhibit in Section 3 non-trivial lower bounds (we emphasize here that traditional bandit lower bounds are void in our setting). Algorithms are described and analysed in Section 4. Finally, Section 5 is dedicated to experiments illustrating the different guarantees and dependencies in the parameters of the models.

1.1 RELATED WORKS

This paper is not the first one to consider additional, free informations, available to the agents while optimizing. There are many different ways of modelling this idea, but our paper is the first one (to our knowledge) that also focus on strategical aspects of obtaining these free informations to reduce regret, especially in the stochastic case.

There exists models where when a specific decision is taken, automatically (resp. with some probability), the performance of some other decision are observed [Alon et al., 2015, Chen et al., 2016, Caron et al., 2012]. Those models assume that there exists a directed (resp. weighted) graph whose set of nodes is the set of decisions. When the agent takes a decision, he also observes the outcome of any node linked (resp. with a probability proportional to the weight of the edge) to the current decision node. Our passive model could be recast as a specific case of that setting, but our results are much finer than the ones available for the general case.

In [Yu and Mannor, 2009] the rewards are stochastic but their means change at unknown time points. Free additional informations are queried by the algorithm in order to detect these change points. They however are not used to decrease the regret of the base bandit algorithm.

Another trend of literature of additional free information in multi-armed bandit studies the "adversarial" case, where no stationary assumption is made on the sequence of rewards (namely, there are not i.i.d.)[Audibert and Bubeck, 2010, Cesa-Bianchi et al., 2006, Mannor and Shamir, 2011]. However the rate of convergence in the two extreme cases (bandit and full information) have the same dependency in T, the total number of stages. To be precise, the regret is either of the order of $\sqrt{KT}$ (in the bandit case) or $\sqrt{\log(K)T}$ (in the full information case), where K is the number of decisions. Intermediate settings (where $1+M$ observations are available at each stage) interpolate between those two cases.

In the stochastic case though, regret is uniformly bounded with full information and grows logarithmically in the bandit case. As a consequence, even the rate of convergence will depend on the size of free informations.

2 MULTI-ARMED BANDITS, REGRET MINIMIZATION AND FEEDBACKS

In that section, we describe precisely the stochastic multi-armed bandit problem and its objective, the minimization of regret.

2.1 STOCHASTIC MULTI-ARMED BANDITS

2.1.1 Bandit vs Full-Information

At each successive stage $t \in \mathbb{N}^*$, an agent takes a decision (or *pulls an arm* using the multi-armed bandit lingo)

i_t in the finite set $[K] := \{1, \ldots, K\}$. After pulling this arm, the agent receives the reward $X_t^{(i_t)} \in \mathbb{R}$, which is sampled from a real reward distribution $\nu^{(i_t)}$ of expectation $\mu^{(i_t)}$. As a consequence, the stochastic bandit problem is parametrised by the vector of distribution, $(\nu^{(1)}, \ldots, \nu^{(K)})$, or alternatively in the non-parametric case, by the vector of expected rewards $(\mu^{(1)}, \ldots, \mu^{(K)})$. Throughout the paper, the results are stated using the arbitrary ordering $\mu^{(1)} > \mu^{(2)} \geq \ldots \geq \mu^{(K)}$. Obviously, those vectors are unknown to the agent, who is aiming at optimizing her cumulative expected reward $\sum_{t=1}^{T} \mu^{(i_t)}$. Actually, instead of this cumulative reward, the objective is normalized into *cumulative regret* minimization.

The cumulative regret (or simply regret) of an algorithm at stage T is defined as

$$R_T = T \max_{i \in [K]} \mu^{(i)} - \sum_{t=1}^{T} \mu^{(i_t)} \ ,$$

i.e., it is the difference between the maximal possible cumulative reward up to stage T and the expectation of the reward gained by the successive choices of arms $i_1, \ldots, i_T$. Following the classical notations, we define $\mu^\star = \max_{i \in [K]} \mu^{(i)}$ and the gaps $\Delta_i = \mu^\star - \mu^{(i)}$. In the non-parametric case, these gaps are the relevant quantities characterising the complexity of a bandit problem.

There are different standard assumption on the feedbacks available to the agent before taking a new decision. In the *bandit* setting, she observes only her reward $X_t^{(i_t)}$ (and, specifically, not the other $X_t^{(k)}$) at the end of stage $t \in \mathbb{N}^*$. In the *full information* setting, she observes the full vector of rewards $(X_t^{(1)}, \ldots, X_t^{(K)}) \in \mathbb{R}^K$. With full information, the *Follow The Leader* (FTL) algorithm that selects the arg max of the empirical average $\overline{X}_t^{(i)} := \frac{1}{t} \sum_{s=1}^{t} X_s^{(i)}$ attains a uniformly bounded regret (with respect to T). In the bandit setting, FTL gets a linear regret, yet the logarithmic optimal dependency in T is achieved by many algorithms. One of the most popular, called *Upper Confidence Bound* (UCB), selects the argmax of the empirical average augmented of an error term $\hat{\mu}_t^{(i)} + \sqrt{6 \frac{\log(t)}{N_i(t)}}$ where $N_i(t)$ is the number of pulls of arm i up to stage t, while $\hat{\mu}_t^{(i)} := \frac{1}{N_i(t)} \sum_{s: i_s = i} X_s^{(i_s)}$.

Many other algorithms are variants of UCB, by modifying the error term, changing some parameters, specifying it for a given class of parametric distributions, etc.

2.1.2 Additional Informations

As specified and motivated in the Introduction, we aim at analysing intermediate settings between bandit and full information, in which a subset of the reward vector might be observed. More precisely, at some stages, the agent not only observes an arm by pulling it but might also observe a second arm for free, i.e., without getting a reward (and without incurring any regret). We consider several ways in which these free observations can be obtained: they can be deterministically available periodically (for instance every $1/\varepsilon$ rounds) or arrive randomly (at each stage with probability ε); the agent can also be a *passive observer* if she can not choose from which arm she gets an extra information (the environment chooses it for her, in a manner to be specified latter on), or she can be an *active observer* if she can choose the arm to observe freely.

We end this section with some notations. In the random time arrival of free information, we assume that at each stage $t \in \mathbb{N}^*$ a Bernoulli random variable Z_t with expectation ϵ_t (whose law is denoted by $\mathrm{Ber}(\epsilon_t)$) is sampled and a free observation is available if $Z_t = 1$. The particular setting in which ϵ_t is constant will be called static random. We will denote by i_t the arm pulled and by f_t the arm chosen to be observed using the free information (if available). The total number of pulls of arm i up to stage t is $N_i(t)$, the number of free observations $F_i(t)$ and the total number of observation of arm i is $O_i(t) = N_i(t) + F_i(t)$.

2.2 A FINITE REGRET SETTING

It is not really difficult to devise a naïve algorithm with a (uniformly) bounded regret at least in the deterministic case, when a free observation is obtained every $1/\epsilon$ round. We consider for simplicity the case of $K = 2$ arms in this section as it gives all the intuitions. Consider the following (heavily sub-optimal) strategy, which we denote by FTL-robin: pull the *leading arm* (the one with the highest empirical average $\hat{\mu}_t^{(i)}$) and when a free sample is available, observe arms in a round-robin fashion.

After a period of $1/\epsilon$ stages, both arms have their observation counters increased by at least one. As a consequence, this simple algorithm FTL-robin can be seen as a full-information algorithm which would take $1/\epsilon$ stages to get the observations. To simplify intuitions

Lemma 1. *The regret of the FTL-robin algorithm on the deterministic setting with $K = 2$ satisfies*

$$\mathbb{E}\, R_T \leq \frac{\overline{c}}{\epsilon} \frac{1}{\Delta} \ , \ \textit{where} \ \Delta = |\mu^{(1)} - \mu^{(2)}|,$$

and there exist distributions $(\nu^{(1)}, \nu^{(2)})$ such that

$$\frac{\underline{c}}{\epsilon} \frac{1}{\Delta} \leq \mathbb{E}\, R_T,$$

where $\underline{c}, \overline{c} > 0$ are universal constants that do not involve any parameter of the problem.

This lemma shows that even the simplest algorithm gets a finite regret in this setting. The proof is almost trivial and omitted. To provide some insights, just assume that $\nu^{(1)} = \mathcal{N}(\Delta, 1)$ and $\nu^{(2)} = \delta_0$. Then the regret of FTL-robin is equal to the Δ/ϵ times the number of times that $\overline{X}_t^{(1)}$ is smaller than 0. Basic computations show that this number is of order $\frac{1}{\Delta^2}$.

The relevant question is then not the asymptotic regime, but what is the precise optimal dependency on ϵ. Indeed, when $\epsilon < \frac{1}{\log T}$, this bound gets larger than the $O(\log T)$ regret of another naive approach, which is to use an algorithm for bandits and discard the additional information.

This free information problem is characterized by a transition from "small" ϵ, where the amount of additional information is not enough to improve the performance of bandit algorithm, to "big" ϵ, where the regret is finite and the setting is closer to full-information.

We answer the question of what "small" and "big" mean in this context and where the transition occurs and we display algorithms enjoying both logarithmic regret when ϵ is small and finite regret when it is big.

3 LOWER BOUNDS

We first consider the definition of *optimality* of an algorithm, that is, what is the minimal regret achievable by any "reasonable" algorithm, in a sense we will make precise. Our lower bounds will highlight a transition from logarithmic (with respect to the horizon T) to finite regimes when ϵ gets big enough.

There are now quite standard techniques to devise lower bounds for stochastic bandits problems, but surprisingly these techniques are inadequate in our case, due to the finiteness of the optimal regret. As a finite regret is possible, a traditional, asymptotic lower bound for $\frac{\mathbb{E}\, R_T}{\log T}$ [Lai and Robbins, 1985] could only be 0 and hence would not be informative. We can obtain a finite time version of this type of bound as in [Garivier et al., 2016] by imposing that our algorithm should perform better than a reference algorithm.

Definition 1. An algorithm is said to be sub-logarithmic with constants C, C_0 if on all bandit problems it verifies for all stages $T \in \mathbb{N}^*$,

$$\mathbb{E}\, R_T \leq C \sum_{i=1}^{K} \frac{\log T}{\Delta_i} + C_0 \sum_{i=2}^{K} \Delta_i \,.$$

There exists sub-logarithmic algorithms (UCB for example, with constants $C = 8$, $C_0 = (1 + \pi^2/3)$ [Auer et al., 2002]). A sub-logarithmic algorithm is performing at least as good as the UCB baseline. This finite time constraint on the performance of the algorithm translates into a lower bound: to perform relatively well on all bandit problems, an algorithm cannot outperform the lower bound guarantee on any of them.

3.1 PASSIVE OBSERVER

When the observer is passive (i.e., she does not choose the arm f_t to observe freely), we assume that f_t is equal to $i \in [K]$ with probability $p_t^{(i)}$ chosen by the environment. Consider the static setting in which for all t, $Z_t \sim \mathrm{Ber}(\epsilon)$ and the probabilities $p_t^{(i)}$ do not depend on the stage t (we will thereafter omit the subscript t).

Standard lower bound techniques proceed as follows: at stage T, the expected number of pulls of an arm is linked to the Kullback-Leibler divergence between the bandit problem studied and a related alternative, in which this arm would be the best one (roughly speaking, in order to be able to “test” that the problem is not the alternative one, a minimum number of samples of that arm must be gathered in the original problem).

A bound on this divergence gives a constraint of the form $\mathbb{E}\, O_i(T) \geq h_i(t)/\Delta_i^2$ for some function $h_i(T) = O(\log T)$. Then a lower bound for the regret is the minimal value of $\sum_{i=2}^{K} \Delta_i \,\mathbb{E}\, N_i(t)$ respecting all these constraints, that can be computed through some linear program. With this proof technique, we obtain lemma 2 .

Lemma 2. *The regret of a sub-logarithmic algorithm with constants C, C_0 must verify*

$$\mathbb{E}_1\, R_T \geq \sum_{i=2}^{K} \max\left\{0, \frac{h_i(T)}{2\Delta_i} - \epsilon p^{(i)} T \Delta_i)\right\} \,.$$

where $h_i(T) = O(\log T)$ (see appendix for a detailed definition).

As mentioned above, this lower bound is void as it reaches 0 as soon as T is big enough, bigger than $\frac{1}{\epsilon} \max_{j \geq 2} \frac{h_j(T)}{2p^{(j)}\Delta_j^2}$.

We want to explain why this lower bound fails to provide relevant informations as our algorithm (see Section 4) are somehow inspired by this. Recall that the lower bound only states that any reasonable algorithm must have gathered, for each sub-optimal arm, a given number of observations, namely $\frac{h_i(T)}{2\Delta_i^2}$. However, $h_i(T)$ grows sublinearly, while the number of free observations grows linearly. So if T is large enough, there will be in total enough free observations to allocate $\frac{h_i(T)}{2\Delta_i^2}$ of them to arm i and an optimal algorithm should somehow have used only free information to explore.

However, this is only possible if the εT free observations were gathered *at the beginning* of the problem and not scarcely with time! Indeed, in the traditional lower bounds techniques, the fact that arm i is observed at the beginning or at the end of time is irrelevant (since the cost of one pull is constant throughout time). They totally discard the fact that the quantities $\mathbb{E}\, N_i(t)$ and $\mathbb{E}\, R_t$ must be non-decreasing. Tighter, relevant lower bounds can be recovered using this monotonicity.

Theorem 1. *The regret of a sub-logarithmic algorithm with constants C, C_0 must verify*

$$\mathbb{E}\, R_T \geq \sum_{i=2}^{K} \frac{1}{2\Delta_i} r_T^{(i)}$$

where

$$r_T^{(i)} = \log(\frac{T\Delta_i^2}{2C\log T \sum_{j\neq i}\frac{\Delta_i}{\Delta_i+\Delta_j}}) + \eta_i(T) - 2\epsilon p^{(i)}\Delta_i^2 T$$

if $T \leq 1/(2\epsilon p^{(i)}\Delta_i^2)$ and otherwise

$$r_T^{(i)} = \Bigg[\log\left(\frac{1}{\epsilon}\frac{1}{4Cp^{(i)}\sum_{j\neq i}\frac{\Delta_i}{\Delta_i+\Delta_j}}\right) - \log\log(\frac{1}{2\epsilon p^{(i)}\Delta_i^2}) + \eta_i(\frac{1}{2\epsilon p^{(i)}\Delta_i^2}) - 1\Bigg] .$$

The function $\eta_i(T)$ goes to zero in $O(1/\log T)$. See appendix for details.

Theorem 1 correctly reports a lower bound increasing with the horizon. It shows a transition from a $O(\log T)$ optimal regret for $T \ll 1/(2\epsilon p^{(i)}\Delta_i^2)$ to a finite regret function of ϵ when T gets bigger. According to Theorem 1, the correct dependency in ϵ in the regret should be in $O(\log(1/\epsilon))$, not $O(1/\epsilon)$ as seen for the naive FTL-robin algorithm.

We can also wonder what is the most favorable passive setting. Simple computations show that free observations should be drawn according to the probability vector $(p_\star^{(1)}, \ldots, p_\star^{(K)})$ where $p_\star^{(i)}$ is proportional to $\frac{1}{\Delta_i}$ (here, we actually ignore the $\log\log$ and η terms of Theorem 1), leading to a lowest lower bound

$$\begin{aligned}\mathbb{E}_1\, R_T &\geq \sum_{i=2}^{K}\frac{1}{2\Delta_i}\log\left(\frac{1}{\epsilon}\frac{\sum_{j=2}^{K}\frac{1}{\Delta_j}}{4C\sum_{j\neq i}\frac{1}{\Delta_i+\Delta_j}}\right) + \alpha \\ &\geq \sum_{i=2}^{K}\frac{1}{2\Delta_i}\log\left(\frac{1}{4C\epsilon}\right) + \alpha\ ,\end{aligned}$$

where α regroups the $\log\log$ and η terms in theorem 1. This lower bound shows in particular that when all suboptimal arms have the same gap, the optimal sample distribution is uniform and the lower bound is of order $\frac{K}{\Delta}\log(\frac{1}{\epsilon})$.

3.2 ACTIVE OBSERVER

An active observer has the possibility to chose the weights $p_t^{(i)}$ at each stage $t \leq T$, potentially achieving a much better distribution of the free observations up to stage T than any static distribution. As before, standard techniques give the following lower bound.

Lemma 3. *The regret of a sub-logarithmic algorithm with constants C, C_0 verifies*

$$\mathbb{E} R_T \geq \sum_{i=2}^{k}\frac{h_i(T)}{2\Delta_i} - \Delta_k(\epsilon T - \sum_{j>k}\frac{h_j(T)}{2\Delta_j^2})\ ,$$

where $k = \min\{i \in \{2, \ldots, K\}\ :\ \sum_{j>i}\frac{h_j(T)}{2\Delta_j^2} \leq \epsilon T\}$.

The structure of the solution to the optimization problem in this case is again educational: an optimal algorithm presented with a given amount of free observations would spend them at the beginning, before costly pulls, and will spend them on the worst arms. This intuition drove the construction of algorithms for active observer in section 4:

First gather free observations, ideally accordingly to the proportion $(p_\star^{(1)}, \ldots, p_\star^{(K)})$ then discards arms for which enough information were gathered, and use a standard optimal bandit algorithm on the remaining ones.

As in the passive observer case, although this lower bound can be meaningful for small horizon T, it becomes void for larger horizons. A better lower bound using the monotony of the number of pulls and of the regret is provided in the next theorem.

Theorem 2. *For $k \in \{2, K-1\}$ let $t_k = \max\{t \geq 1\ :\ \sum_{j=k+1}^{K}\frac{h_j(t)}{2\Delta_j^2} > \epsilon t\}$. The regret of any active sub-logarithmic algorithm with constants C, C_0 verifies*

$$\mathbb{E}\, R_T \geq \max_{k:t_k\leq T}\sum_{i=2}^{k}\frac{1}{\Delta_i}\Bigg[\log(\frac{1}{\epsilon}\frac{\sum_{j=k+1}^{K}\frac{\Delta_i^2}{\Delta_j^2}}{4C\sum_{j\neq i}\frac{\Delta_i}{\Delta_i+\Delta_j}}) - \log\log(\frac{1}{\epsilon}\sum_{j=k+1}^{K}\frac{1}{2\Delta_j^2}) + \eta(\frac{1}{\epsilon}\sum_{j=k+1}^{K}\frac{1}{2\Delta_j^2})\Bigg] .$$

When all gaps are equal to the same value $\Delta > 0$, the leading term of this lower bound is of the form

$$\max_{k:t_k\leq T}\frac{k-1}{\Delta}\log(\frac{1}{\epsilon}\frac{K-k}{K})\ .$$

In particular, this result states that as T goes to infinity, the regret is asymptotically lower bounded by $\frac{K-1}{\Delta}\Big[\log(\frac{1}{\epsilon}) - \log\log(\frac{e}{\epsilon})\Big]$.

4 ALGORITHMS AND UPPER-BOUNDS

In this section, we exhibit algorithms matching the lower bounds derived in the previous section, up to $\log\log(\cdot)$ terms, showing that they indeed represent accurately the problem complexity.

4.1 PASSIVE OBSERVER

A passive observer does not get to choose the arms on which free information is gained. As in the classical stochastic multi-armed bandit, the only decision is therefore which arm to pull. It is then natural to extend known algorithms by taking into account all observations from both provenances.

As UCB pulls the arm with maximal index $\overline{X}_t^{(i)} + \sqrt{\frac{6\log t}{N_i(t)}}$, we extend it by using all available observations both in the empirical mean and exploration term. Algorithm 1 pulls $i_t = \arg\max_i \overline{X}_t^{(i)} + \sqrt{\frac{6\log t}{O_i(t)}}$.

Algorithm 1 UCB with passive observations.

```
Pull each arm once.
loop: at stage $t$,
    $i_t = \arg\max_i \overline{X}_t^{(i)} + \sqrt{\frac{6\log t}{O_i(t)}}$
    Pull arm $i_t$, observe $X_t^{(i)}$.
    If $Z_t = 1$, sample $f_t$ and observe $X_t^{(f_t)}$.
    Update $\overline{X}_t$, $N_i(t)$, $F_i(t)$, $O_i(t) = N_i(t)+F_i(t)$.
end loop
```

Theorem 3. *Consider the static passive observer case, where f_t follows the categorical distribution with parameters $(p^{(1)}, \dots, p^{(K)})$ and the probability of getting a free observation is $\epsilon \in (0, 1]$ for all stages $t \geq 1$.*

Then the regret of ucb verifies both

$$\mathbb{E}\, R_T \leq \sum_{i=2}^{K} \frac{24}{\Delta_i} \log T\,,$$

and

$$\mathbb{E}\, R_T \leq \sum_{i=2}^{K} \frac{24}{\Delta_i} \log \frac{50}{\epsilon p^{(i)}} + \sum_{i=2}^{K} \frac{24}{\Delta_i} \max\left\{\log\frac{1}{e\Delta_i^2}, \log\log\frac{20}{\epsilon p^{(i)}}\right\}\,.$$

Hence UCB with passive observations recovers the $\log(\frac{1}{\epsilon})$ dependency in ϵ, up to a doubly logarithmic term when $\epsilon p^{(i)}$ is small compared to the squared gaps. When the dominant term in this maximum is $\log \frac{1}{e\Delta_i^2}$, the regret due to arm i has the form $\frac{1}{\Delta_i}\log\frac{1}{\epsilon p^{(i)}\Delta_i^2}$, which is sub-optimal with respect to Δ_i (see Theorem 1). This is due to the sub-optimality of UCB itself: while the regret of UCB on a bandit problem is $O(\sum_{i=2}^{K}\frac{\log T}{\Delta_i})$, other algorithms of the same family like UCB2 [Auer et al., 2002], Improved-UCB, [Auer and Ortner, 2010] or MOSS [Audibert and Bubeck, 2009, Degenne and Perchet, 2016] get an improved regret of order $O(\sum_{i=2}^{K}\frac{\log(T\Delta_i^2)}{\Delta_i})$.

The dependency in $\log(\frac{1}{\epsilon})$ means that ϵ as small as $\frac{1}{T}$ gives useful information to a learner. Obviously there is no gain to be had if $\epsilon < \frac{1}{T}$, as there is in average less than one additional observation before T, but few more free observations are enough to improve the regret.

4.2 ACTIVE OBSERVER

While a uniform allocation of the free observations over the arms gets the right $\log(\frac{1}{\epsilon})$ dependency in ϵ, having the choice of the arm which will be observed allows an algorithm to get the right dependency in the parameters of the bandit problem. In the active setting, the algorithm can choose freely which of the $[K]$ arms will get an additional observation, when such an observation is available.

To devise an algorithm taking advantage of this possibility, we try to mimic the lower bound for fixed stage, as in Lemma 3. A good algorithm should use the available free observations first to discard the worse arms, before using costly pulls only on the remaining arms.

We introduce an algorithm combining two subroutines: an Explore-Then-Commit (ETC) [Even-Dar et al., 2006, Perchet and Rigollet, 2013] algorithm on the free observations is used to narrow the set of arms which need to be pulled and an algorithm of the UCB family is used on this set. As we seek for optimality with respect to the problem parameters we use OCUCB-n [Lattimore, 2016], which is the UCB-type algorithm closest to it. ETC is described in Algorithm 3. OCUCB-n with parameters $\eta > 1$ and $\rho \in [1/2, 1]$ pulls at stage $t \in \mathbb{N}^*$ the arm with maximal index

$$\overline{X}_t^{(i)} + \sqrt{\frac{2\eta \log B_{t-1}^{(i)}}{N_i(t)}}$$

where

$$B_{t-1}^{(i)} = \max\left\{e, \log(t), \frac{t\log t}{\sum_{i=1}^{K}\min\{N_i, N_j^{\rho}N_i^{1-\rho}\}}\right\}$$

where N_i is a shorthand notation for $N_i(t)$.

The main algorithm use a succession of epochs. In epoch number $m \in \mathbb{N}$, the ETC subroutine collects (free) information on all the arms in $[K]$, while OCUCB-n pulls arms in an available subset of the arms S_m. At the end of epoch m, the free observations gathered are used to discard arms from $[K]$ which are not optimal with high enough confidence, forming S_{m+1}. There is a finite $m_i \in \mathbb{N}$ depending on ϵ and the gaps such that with high probability, $i \notin S_m$ for $m > m_i$, hence arm i contributes to the regret only up to epoch m_i and the regret is finite.

Algorithm 2 Active Algorithm.

```
Require: parameters $\rho \in [1/2, 1]$, $\alpha \geq 1, \eta > 1$.
    Initialize $S_0 = [K]$.
    loop: at epoch $m$, with duration $d_m = 2^{2^m}$,
        Pull arms according to OCUCB-n with parameters $\eta$ and $\rho$ on $S_m$,
        Use free observations according to ETC with parameter $\alpha$ and horizon $T = d_{m+1}^{3/2} \log d_{m+1}$.
        Set $S_{m+1}$ to the set returned by ETC.
    end loop
```

Algorithm 3 Explore-Then-Commit

```
Require: parameter $\alpha \geq 1$, horizon $T \in \mathbb{N}^*$.
    Initialize $s = 0$, $S = [K]$.
    loop
        Observe all arms in $S$.
        Discard from $S$ any arm $i$ such that
    $\hat{\mu}_s^{(i)} + \sqrt{\frac{2\alpha}{s}\log(\frac{T}{s})} < \max_{j \in S} \hat{\mu}_s^{(j)} - \sqrt{\frac{2\alpha}{s}\log(\frac{T}{s})}$.
        $s \leftarrow s + 1$.
    end loop
    return $S$.
```

In order to write a regret upper bound for our active algorithm, we introduce quantities $H_{i,\rho}$ for $i \in \{2, \dots, K\}$ and $\rho \in [1/2, 1]$,

$$H_{i,\rho} = \frac{i}{\Delta_i^2} + \sum_{j=i+1}^{K} \frac{1}{\Delta_i^{2(1-\rho)} \Delta_j^{2\rho}} \,.$$

These constants transcribe the difficulty of the problem. A number of observations of order $\frac{1}{\epsilon} H_{i,1}$ will be necessary for ETC to eliminate arm i with high confidence.

Theorem 4. *The regret of the active algorithm 2 with parameters $\rho \in [1/2, 1]$ and $\alpha = 1$ on problems with rewards in $[0,1]$ is*

$$\mathbb{E}\, R_T \leq C_\eta \sum_{i=2}^{K} \frac{4}{\Delta_i} \max\left\{ \log(\frac{1}{\epsilon}), \log\sqrt{H_{i,\rho}} \right\} + 51K + O\left(\sum_{i=2}^{K} \frac{1}{\Delta_i} (\log\log \frac{H_{i,1}}{\epsilon})^2 \right)$$

with C_η a constant that depends only on η (see [Lattimore, 2016] for details on C_η).

Our analysis of Explore-Then-Commit relies on a new maximal concentration inequality which can be of independent interest.

Lemma 4. *Let Z_t be a σ^2-sub-Gaussian martingale difference sequence then, for every $\delta \in (0, 0.2]$ and every integers $T \in \mathbb{N}^*$,*

$$\mathbb{P}\left\{ \exists t \leq T, \overline{Z}_t \geq \sqrt{\frac{2\sigma^2}{t} \log(\frac{T}{\delta t})} \right\} \leq 6\delta \sqrt{\log(\frac{1}{\delta})} \,.$$

Asymptotically, we obtain

$$\limsup_{\delta \to 0} \frac{\mathbb{P}\left\{ \exists t \leq T, \overline{Z}_t \geq \sqrt{\frac{2\sigma^2}{t} \log(\frac{T}{\delta t})} \right\}}{\delta \sqrt{\log(\frac{1}{\delta})}} \leq \sqrt{e/8}.$$

This value is $\sqrt{e/8} \approx 0.6$.

4.2.1 Heuristics and Influence of ϵ

Besides the algorithm already discussed, we also experimented on the following heuristic: choose a bandit algorithm of the UCB family, which pulls the arm with a maximal index; use it to pull the arm with maximal index and if an observation is available, observe the second maximal arm. We provide no regret analysis for this heuristic but study its performance in the experimental section.

Concerning the dependency in ε, we can make the following interesting remark. To simplify notations, we will assume that all arms have the same gap Δ and we remove constants for this analysis. With these simplifications, we proved that regret at stage T is of the order of $R_T \simeq \frac{K}{\Delta} \log(\frac{1}{\epsilon})$. Obviously, if ϵ is almost equal to 0, this upper-bound is void and the algorithm should not depend on the free observations. One might ask what is the threshold at which free informations become relevant at stage T.

Notice that standard information theory arguments yield that if $\epsilon T \leq \frac{K}{2\Delta^2}$, and even if the free observations were gathered at the begining of the problem, only $\frac{K}{2}$ arms could be removed (with high probability) from the set of possible optimal arms. Hence these free information are not useful for at least $K/2$ arms and regret will have to scale as $\frac{K}{2\Delta} \log(\frac{2T\Delta^2}{K})$, the optimal rate for the bandit problem with $K/2$ arms with equal gaps Δ.

On the other hand, if $\epsilon T \geq \frac{K}{2\Delta^2}$, then (up to multiplicative constant), $\frac{K}{\Delta} \log(\frac{1}{\epsilon})$ dominates $\frac{K}{\Delta} \log(\frac{T\Delta^2}{K})$. As a consequence, the relevant threshold for the probability of free observations after T stages is

$$\varepsilon^* = \frac{1}{T} \frac{K}{\Delta^2} \,.$$

5 EXPERIMENTS

All experiments are performed with Gaussian rewards with unit variance.

Influence of ϵ. The goal of this first experiment is to confirm the scaling of the regret with ϵ. That is to say, the regret scales with $\sum_{i:\Delta_i>0} \frac{1}{\Delta_i} \log(\frac{1}{\epsilon \Delta_i^2})$. The experiment is performed with a passive observer with either a uniform distribution or the optimal one, as defined in Section 3.1. To do so, the experiment is performed in the passive setting associated with a uniform distribution and the optimal one, as defined in Section 4.1. Also, when free observations are scarce, $\epsilon \sim \frac{1}{T}$, the average number of those is approximately 1 during the experience. Therefore, the regret is similar to the one suffered by an UCB algorithm in a classic multi-armed bandit setting, a behaviour captured by the function f. On Figure 1 and 2, experiments are run on four Gaussian arms with expectations 2, 1.8, 0.5, 0.2, the error bars are quantile at 10% and 90%.

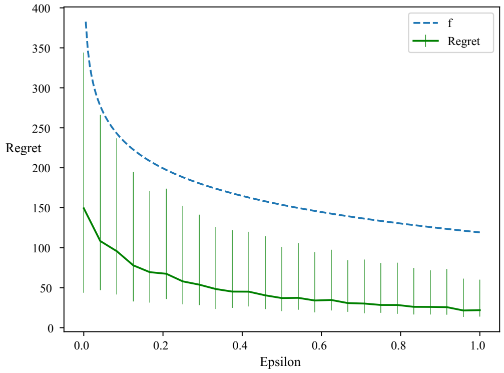

Figure 1: Dependence on ϵ of the regret of UCB as passive observer, with a uniform distribution of the free observations, averaged over 300 runs.

Passive Observer: optimal sampling distribution. This second experiment illustrates the induced regret in the passive setting with a probability distribution $p^{(i)} = \frac{1}{\Delta_i}$. This distribution is considered to be optimal because, as mentioned in Section 3.1, it achieves the lowest lower bound. It also suggests a paradigm for algorithms in the active setting i.e sampling freely as much as possible the arm with the lowest Δ_i. A way to do so is to run an UCB type algorithm to choose which arm to pull, and use another UCB type algorithm on other arms to determine which will be observed if a free observation is available. The results of this type of policy is presented in the next paragraph.

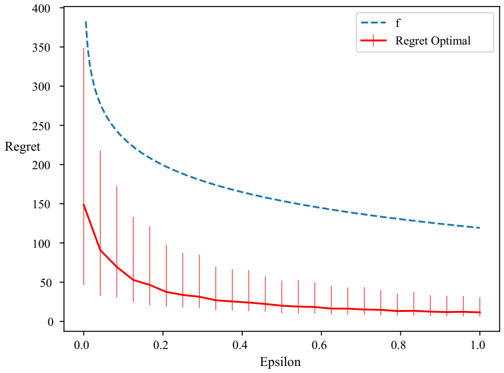

Figure 2: Dependence on ϵ of the regret of UCB as passive observer, with the optimal distribution of the free observations, averaged over 300 runs.

The experiment is run on the same set of arms as previously with a uniform distribution, the optimal distribution and a suboptimal one such that $p^{(i)} = \frac{1}{\Delta_i^2}$, referred as SubOptimal in Figure 3. Color filled regions are 25% and 75% quantiles.

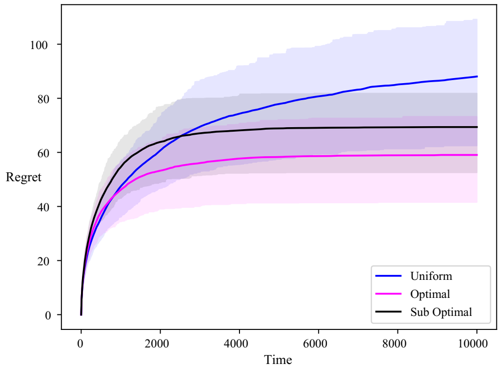

Figure 3: Regret averaged over 300 runs

Active Observer: comparison of algorithms. This subsection is dedicated to the comparaison of algorithms introduced earlier : UCB1-Double, ETC-OCUCB and ETC-OCUCB-2.
UCB1-Double uses a UCB algorithm and select the free observation as the second index maximising arm. The optimal allocation in the passive setting samples better arms more often, therefore we use the free observation to sample the arm next to optimal (according to its UCB index). The second algorithm, referred to as ETC-OCUCB, is the algorithm studied in the above section. In particular, its ETC subroutine checks for potentially

removable arms every $C|S|$ pulls, with C a fixed parameter and S the set of currently active arm. Finally, the algorithm referred to as ETC-OCUCB-2 is a variant of ETC-OCUCB where elimination checks are made every 2^k stages, thus behaving less aggressively than ETC-OCUCB. In addition, we introduced in this experiment a parameter p so that the epoch length is $d_m = p^{p^m}$ in ETC-OCUCB. This enables us to adapt the growth of epochs to the horizon, here $T = 10^4$. Other parameters are : $\alpha = 1$, $\rho = \frac{1}{2}$, $\eta = 2$ and $C = 10$.
The experiments is run on five Gaussian arms with expectations 2, 1.8, 1.5, 1 and 0.5. Color filled regions are 25% and 75% quantiles.

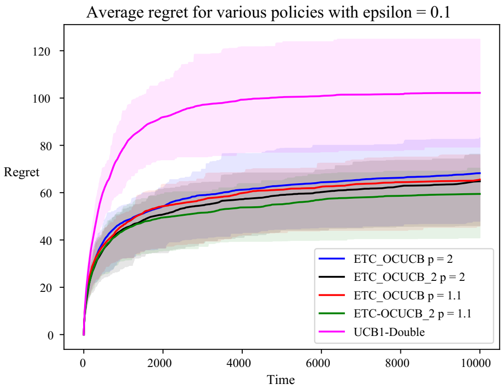

Figure 4: Regret for $\epsilon = 0.1$ averaged over 100 runs

Figure 4 illustrates that:

- UCB1-Double reaches rapidly its final regret value after a logarithmic exploration phase where informations are gathered so that the policy doesn't pull an other suboptimal arm after this phase.
- ETC-OCUCB and ETC-OCUCB-2 algorithms have similar performances and the parameter p offers a control how often the set of active arms is updated which offers a slight performance increase for lower p.

ETC-OCUCB and ETC-OCUCB-2 maintain two distinct tracks of rewards, one for rewards obtained after pulling an arm and the other for rewards after sampling freely an arm. Therefore, it may be possible to increase their performance by using both sources of information in both subroutines. In the Figure below, these variants are referred as ETC-OCUCB-all-info and ETC-OCUCB-all-info-2.

This simple modification provides a clear improvement whether for the final regret or the speed at which this value is reached.

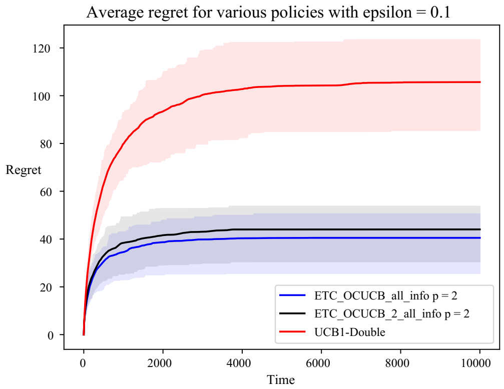

Figure 5: Regret for $\epsilon = 0.1$ averaged over 300 runs for $p = 2$

6 CONCLUSION

We analysed the multi-armed bandit problem with just a few extra free information. Interestingly, as the regret is uniformly bounded in time, standard lower bounds are void. However, a careful analysis allowed us to exhibit non-trivial guarantee that no reasonable algorithm can out-perform and we finally provided an optimal algorithm, whose regret matches the lower bound up to doubly logarithmic terms.

We would like to finally emphasize that our algorithm can be used even if the εT observations are not free. Since we used ETC on these observations, we get that our algorithm has a regret smaller (discarding multiplicative constants and $\log\log$ terms) than

$$\sum_{i=2}^{K} \frac{\log(\varepsilon T \Delta_i^2)}{\Delta_i} + \sum_{i=2}^{K} \frac{\log(1/\varepsilon)}{\Delta_i}$$

where the first term is the guarantee of ETC on εT samples, and the second one is the guarantee of our algorithm with "free" observations. As a consequence, no matter the value of ε (as long as the $\log\log$ terms do not become dominant), its dependency vanishes, and we recover the expected performance of ETC.

Acknowledgements

V. Perchet has benefited from the support of the ANR (grant n.ANR-13- JS01-0004-01), of the FMJH Program Gaspard Monge in optimization and operations research (supported in part by EDF), from the Labex LMH and from the CNRS, PEPS project Lacreme.

References

N. Alon, N. Cesa-Bianchi, O. Dekel, and T. Koren. Online learning with feedback graphs: Beyond bandits. In *Conference on Learning Theory*, pages 23–35, 2015.

J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, pages 217–226, 2009.

J.-Y. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11(Oct):2785–2836, 2010.

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

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

S. Caron, B. Kveton, M. Lelarge, and S. Bhagat. Leveraging side observations in stochastic bandits. *arXiv preprint arXiv:1210.4839*, 2012.

N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Regret minimization under partial monitoring. *Mathematics of Operations Research*, 31(3):562–580, 2006.

I. Chatzigeorgiou. Bounds on the lambert function and their application to the outage analysis of user cooperation. *IEEE Communications Letters*, 17(8):1505–1508, 2013.

W. Chen, Y. Wang, Y. Yuan, and Q. Wang. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *The Journal of Machine Learning Research*, 17(1):1746–1778, 2016.

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

E. Even-Dar, S. Mannor, and Y. Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(Jun):1079–1105, 2006.

A. Garivier, P. Ménard, and G. Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *arXiv preprint arXiv:1602.07182*, 2016.

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

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

S. Mannor and O. Shamir. From bandits to experts: On the value of side-observations. In *Advances in Neural Information Processing Systems*, pages 684–692, 2011.

M. Okamoto. Some inequalities relating to the partial sum of binomial probabilities. *Annals of the institute of Statistical Mathematics*, 10(1):29–35, 1959.

V. Perchet and P. Rigollet. The multi-armed bandit problem with covariates. *The Annals of Statistics*, pages 693–721, 2013.

J. Y. Yu and S. Mannor. Piecewise-stationary bandit problems with side observations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1177–1184. ACM, 2009.