Multi-armed Bandit with Additional Observations

DONGGYU YUN, Naver Corporation
ALEXANDRE PROUTIERE, KTH
SUMYEONG AHN, JINWOO SHIN, and YUNG YI, KAIST

We study multi-armed bandit (MAB) problems with additional observations, where in each round, the decision maker selects an arm to play and can also observe rewards of additional arms (within a given budget) by paying certain costs. In the case of stochastic rewards, we develop a new algorithm KL-UCB-AO which is asymptotically optimal when the time horizon grows large, by smartly identifying the optimal set of the arms to be explored using the given budget of additional observations. In the case of adversarial rewards, we propose H-INF, an algorithm with order-optimal regret. H-INF exploits a two-layered structure where in each layer, we run a known optimal MAB algorithm. Such a hierarchical structure facilitates the regret analysis of the algorithm, and in turn, yields order-optimal regret. We apply the framework of MAB with additional observations to the design of rate adaptation schemes in 802.11-like wireless systems, and to that of online advertisement systems. In both cases, we demonstrate that our algorithms leverage additional observations to significantly improve the system performance. We believe the techniques developed in this paper are of independent interest for other MAB problems, e.g., contextual or graph-structured MAB.

CCS Concepts: • **Mathematics of computing → Probability and statistics**; • **Computing methodologies → Machine learning**; **Sequential decision making**;

ACM Reference Format:
Donggyu Yun, Alexandre Proutiere, Sumyeong Ahn, Jinwoo Shin, and Yung Yi. 2018. Multi-armed Bandit with Additional Observations. *Proc. ACM Meas. Anal. Comput. Syst.* 2, 1, Article 13 (March 2018), 22 pages. https://doi.org/10.1145/3179416

1 INTRODUCTION

Since the seminal work by Robbins [30], the multi-armed bandit (MAB) problem (or simply bandit problem) has received much attention due to a wide range of applications including medical trials, online advertising [12, 34], recommendation systems [9], search engines [33] and wireless networks [13]. In the classical MAB problem, the decision maker pulls in each round a single arm among many arms, and observes the corresponding *reward* at the end of the round. The goal of the decision maker is to maximize the cumulative reward (over a fixed number of rounds), or equivalently to minimize *regret*, defined as the difference between the cumulative reward and that achieved by always playing the best arm. The *expert* problem is a variation of the classical MAB problem and has also been extensively studied [4, 10, 11, 16, 19]: in this problem, at the end of round, the rewards of all arms are revealed to the decision maker[1]. In the MAB literature, two popular models on how rewards are generated have been considered: (a) stochastic – it assumes that rewards are i.i.d.

[1] Throughout this paper, the *classical* MAB (or bandit) problem refers to the problem where only the reward of the played arm is revealed.

Authors' addresses: Donggyu Yun, gue821@kaist.ac.kr, Naver Corporation; Alexandre Proutiere, alepro@kth.se, KTH; Sumyeong Ahn, sumyeongahn@kaist.ac.kr; Jinwoo Shin, jinwoos@kaist.ac.kr; Yung Yi, yiyung@kaist.edu, KAIST.

2476-1249/2018/3-ART13 $$15.00
https://doi.org/10.1145/3179416

across the various rounds and drawn from some unknown distributions, and (b) adversarial – it allows an arbitrary (and adversarial) sequence of rewards.

This paper studies the MAB problem with *additional observations*. In each round, the decision maker selects an arm to play, but also additional arms to observe the samples of their rewards. Observing additional arms possibly incurs a cost, and also might be constrained by a given budget (in terms of a maximum number of additional observations up to a given round). Similar versions of this problem were first suggested by [32] and studied by [3, 20, 31], but only for the adversarial rewards. They generalize both the classical MAB and expert problems: the former problem corresponds to having no extra observations, whereas the latter corresponds to observing all arms in each round. In this paper, we consider adversarial rewards, and provide, for the first time in this setting, algorithms with tight regret upper bounds. We also consider stochastic rewards, and develop asymptotically optimal algorithms. To our knowledge, MAB problems with additional observations with stochastic rewards have not been investigated in the literature.

Additional observations come with a cost, which naturally arises in practical MAB applications due to monetary or computational reasons. For example, in medical trials, a patient may consult various doctors to get treatment advices which has to be paid. In wireless communication systems, where packets have to be sent at different rates to probe the radio conditions of the channel, one has to allocate some time for such a probing procedure [13]; this time cannot be used to transmit actual data. For online advertisement in web services, one may wish to maintain 'test' websites, simultaneously with 'real' ones, for additional feedback on advertisements, news articles, etc.

1.1 Contributions

For stochastic rewards, we first derive an asymptotic lower bound on regret satisfied by any algorithm. We then design KL-UCB-AO, an algorithm whose regret matches our lower bound. In particular, when there exists no cost of additional observations, our results imply that a budget of additional observations growing logarithmically with the time horizon T is necessary and sufficient for obtaining a *constant regret*. The key idea of our algorithm is as follows. We first note that similarly to [25], under any reasonable[2] algorithm, every suboptimal arm, say a, should be observed at least $f_a(T) = C_a \log T$ times by round T, for some well identified constant C_a. Then, the idea is to sort the arms in the ascending order of their expected rewards and use the observation budget for each suboptimal arm a in that order to get $f_a(T)$ observations, provided that the budget is enough and observing arm a is more valuable than paying cost. Here the role of sorting is to play the arms with larger expected rewards towards less regret.

For adversarial rewards, we restrict our attention to observation budgets growing linearly with the time horizon, since sub-linear budgets can only have marginal impacts on regret in view of the best known regret lower bound [31]. Due to this, unlike the case of stochastic rewards, we consider no cost of additional observations: otherwise the regret should grow linearly. Note that we still have a fixed budget for additional observations: a player can observe the rewards of a constant number M of arms in each round for some $M \geq 1$ and there is no extra cost for observing the additional arms. This problem is a special case of known settings in the literature [3, 20, 31]. For this problem, we design H-INF, an algorithm with regret $O\left(\max\left\{\sqrt{\frac{N}{M}T}, \sqrt{T\log N}\right\}\right)$ for N arms, M observations, and T rounds. The regret of H-INF improves those of the known algorithms [3, 20, 31] and matches the lower bound derived in [31]. The design of H-INF is based on a novel two-layered structure, where each layer runs the optimal INF algorithm [4]. In H-INF, one can replace the INF algorithm by any other algorithm, as long as it is order-optimal for both the classical MAB (i.e., $M = 1$) and

[2] i.e., uniformly good.

the expert problem (i.e., $M = N$). This simple, yet powerful hierarchical structure, allows us to decouple the algorithmic components into layers, which significantly facilitates the regret analysis.

1.2 Related Work

We first review the literature for the classical settings of MAB and expert problems. For the classical MAB problem under stochastic rewards, Lai and Robbins [25] derived an asymptotic regret lower bound satisfied by *any* algorithm. They also designed an algorithm based on the so-called *upper confidence bound* (UCB), whose variants including UCB1 [5], Bayes-UCB [21] and KL-UCB [17] have been proposed to improve the regret upper bound. Note that the optimal algorithm for the expert problem (i.e., $M = N$) under stochastic rewards is somewhat trivial, i.e., following the decision of the best arm in its empirical mean. For the classical MAB problem under adversarial rewards, the EXP3 algorithm achieves the regret $O\left(\sqrt{TN\log N}\right)$ [6], and the order-optimal regret $O\left(\sqrt{TN}\right)$ is achieved by the INF algorithm [4]. The well-known regret lower bound for the expert problem under adversarial rewards is $O\left(\sqrt{T\log N}\right)$, which can be achieved by several algorithms [4, 11, 16].

The *additional observations or experts* (i.e., $1 \le M \le N$) studied in this paper generalizes the above classical settings. A version of this problem for the adversarial rewards was first suggested by [32] as an open question and resolved by [20]. The proposed algorithm in [20] have regret $O\left(\sqrt{TN}\right)$ in our setting and it is clearly sub-optimal as it does not even depend on M.[3] This is primarily because the author studies a more general assumption on experts than ours: each expert provides a probability distribution over all arms for selecting which arm to play, whereas we force every expert i to play the arm i. Similar problems with the same assumption on experts as ours were also studied [3, 31], where both algorithms have regret $O\left(\sqrt{\frac{N}{M}T\log N}\right)$ in our setting, which is sub-optimal by a logarithmic factor $O\left(\sqrt{\log N}\right)$ for $1 \le M \le N/\log N$. They consider more general budget constraints than ours for additional observations: time-varying budget [31] and cost-related budget [3], which makes our H-INF algorithm not directly applicable to their settings. We again emphasize that the above prior work only considered the adversarial rewards. For stochastic rewards, we are unaware of any attempt to design asymptotically optimal algorithms considering additional observations. Table 1 summarizes the related work as well as our contributions. In summary, we study a special version of problems studied in [3, 20, 31] for the adversarial rewards, and its stochastic version is first studied in this paper. For both types of rewards, we provide the first optimal regret upper bounds.

We also remark that the MAB problem with *graph-structured feedbacks* [2, 8, 9, 23, 26] has a similar flavor to our problem. However, the difference is that observations in prior work are given by an external graph (i.e., playing an arm reveals the rewards of all neighboring arms under the graph), while we can choose them adaptively over time without graphical restrictions, but with cost and/or budget constraints. Nevertheless, we believe that the design principle of our algorithm might also apply to the MAB problems with graph-structured feedback as well as other related ones [3, 20, 31]. As other related work, paying for exploration is also studied in the principal-agent framework [15, 27]. The authors assume that there is a principal or an organizer facing a MAB problem, which, however, does not pull an arm directly: a myopic agent who arrives at each round plays an arm instead of the principal. Since each myopic agent typically selects the arm with the

[3] The regret upper bound of [20] is $4\sqrt{\frac{\min\{K,M\}N\log\frac{8M}{\min\{K,M\}}}{M}T}$ for K arms and N experts. In our setting, $K = N$, hence it becomes $O\left(\sqrt{TN}\right)$.

Table 1. Current best results for our problem of MAB with additional observations, where N and M denote the number of arms and the upper bound on the number of additional observations per round, respectively. The cases $M = 1$ and $M = N$ correspond to the classical MAB and expert problems, respectively. We note that the case of $1 < M < N$ for stochastic rewards is first studied in this paper.

regret	reward	$M = 1$	$M = N$	$1 < M < N$
lower bound	stochastic	[25]	trivial	this paper
	adversarial	[6]	[10, 11], etc	[3, 20, 31]
upper bound	stochastic	[17, 21]	trivial	this paper
	adversarial	[4]	[11, 16], etc	this paper

current largest expected reward, there may not be enough exploration for all arms. The principal should induce the agents to play other arms through incentives such as payment. The problem studied in this paper is fundamentally different, since the payment in the prior work is not for *additional* observations, where the principal can observe the reward of the played arm only and she pays for incentivizing exploration in selecting the arm to play.

1.3 Organization

Section 2 provides the detailed description of the MAB problem with additional observations. Sections 3 and 4 describe our main results for stochastic and adversarial rewards, respectively. We report our numerical results with applications in Section 5, and conclude in Section 6.

2 PROBLEM FORMULATION

We introduce the multi-armed bandit (MAB) problem with additional observations. In each round $t = 1, 2, \dots,$ a decision maker chooses an arm A_t to *play* in the set $\mathcal{N} = \{1, 2, \dots, N\}$ of arms, and schedules additional arms from which she observes the reward at the end of the round. There exists a *budget* for additional observations so that the decision maker spends a unit budget if she wants to observe an arm once. The budget is given in the form of a cumulative budget function $b(t)$ which is the total allowable additional observations up to the t-th round. Namely, at the t-th round, the decision maker receives a budget $b(t) - b(t-1)$ of additional observations, but she is allowed to save the budget to use it later. In addition, we consider the scenario that each additional observation comes with a cost $c \geq 0$. For example, when the budget function $b(t) = (N-1)t$ and the cost $c = 0$, she can always observe the rewards of all arms in every round, which corresponds to the expert problem. The classical MAB problem corresponds to the case $b(t) = 0$ or $c = \infty$. We assume that the budget function $b(t)$ is at most $(N-1)t$ for all $t \in \mathbb{N}$. The decision maker is said to "observe" an arm when the arm is played or observed. Let $\mathcal{O}_t$ be the set of observed arms in round t, e.g., $A_t \in \mathcal{O}_t$. We denote by $X_a(t) \in [0, 1]$ the reward of arm a in round t. Denote by $A_a(t)$ and $O_a(t)$ the numbers of times the arm a has been played and observed, respectively, up to the t-th round. We often use $X_{a,s}$ to denote the reward of arm a at the s-th observation of a, i.e., filtering out the rounds when arm a is unobserved, i.e., $X_{A_t}(t) = X_{A_t, O_{A_t}(t)}$. We consider the two following models specifying how the rewards are generated.

- **Stochastic rewards:** For any $a \in \mathcal{N}$, the sequence of rewards $(X_a(t))_{t \geq 1}$ over rounds is i.i.d drawn from the distribution $\nu(\theta_a)$ which is parameterized by θ_a in some set Θ. In particular, we focus on Bernoulli reward distributions but the results can be readily extended to one-parameter exponential family of distributions. We use μ_a to denote the expected reward of arm a and/or the parameter of the Bernoulli reward distribution of arm a. We assume that there exists a unique *optimal arm* $a^\star$ which has the highest mean reward, where for simplicity we write $\mu^\star = \mu_{a^\star}$.

Without loss of generality, we order arms with respect to their expected rewards, i.e., $a^\star = N$ and $\mu_1 \leq \cdots < \mu_N = \mu^\star$. Let Ω be the set $\{\boldsymbol{\mu} \in (0,1)^N : \mu_1 \leq \mu_2 \leq \ldots < \mu_N\}$.

- **Adversarial rewards:** A *non-oblivious* adversary arbitrarily chooses rewards of arms, where she may take into account the player's past decisions on assigning rewards. The adversary has only to decide the sequence of rewards $(X_a(t) : a \in \mathcal{N})$ before the decision maker selects the observed arms O_t in each t-th round. It is the most general scenario in the adversarial reward setting.

In each round, a decision rule or algorithm selects an arm to play and additional arms to observe, depending on the arms observed in earlier rounds, their observed rewards and the budget function $b(t)$. The goal is to minimize the (expected) regret up to the round T, defined as:

$$R(T) = \max_{a \in \mathcal{N}} \mathbb{E}\left[\sum_{t=1}^{T} X_a(t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} X_{A_t}(t) - c\left(|O_t| - 1\right)\right]$$

where the expectation $\mathbb{E}$ is taken over the possibly random rewards and the possible randomness in the selected arms. Note that in the stochastic setting, the regret can be written as:

$$\begin{aligned} R(T) &= \mu^\star T - \sum_{a \in \mathcal{N}} \mu_a \mathbb{E}\left[A_a(T)\right] + c \sum_{a \in \mathcal{N}} \mathbb{E}\left[O_a(T) - A_a(T)\right] \\ &= \sum_{a \neq a^\star} \Delta_a \mathbb{E}\left[A_a(T)\right] + c \sum_{a \in \mathcal{N}} \mathbb{E}\left[O_a(T) - A_a(T)\right], \end{aligned}$$

where $\Delta_a = \mu^\star - \mu_a$.

3 STOCHASTIC REWARDS

In this section, we study the MAB problem with additional observations in the stochastic setting. For the classical MAB problem, it is known that any algorithm has $\Omega(\log(T))$ regret [25]. One can expect that additional observations would naturally reduce the regret. Given budget function $b(t)$ on additional observations up to the t-th round, we first derive an asymptotic lower bound on the regret satisfied by any algorithm. Then, we devise KL-UCB-AO (KL-UCB with Additional Observations), an algorithm whose regret matches the lower bound.

3.1 Regret Lower Bound

Recall that we assume that arms are ordered with respect to their expected rewards, i.e., $a^\star = N$. Given the budget $b(t)$ and the parameters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N) \in \Omega$, we define $a^\dagger \neq a^\star$ as the arm with minimum index such that

$$\sum_{i=1}^{a^\dagger} \frac{1}{KL(\mu_i \,\|\, \mu_N)} \geq \lim_{T \to \infty} \frac{b(T)}{\log T},$$

where $KL(p \,\|\, q) := p \log(\frac{p}{q}) + (1-p)\log(\frac{1-p}{1-q})$ denotes the Kullback-Leibler (KL) divergence between two Bernoulli distributions with means p and q. If there is no such arm, we set $a^\dagger = a^\star$. We also define the arm $a^\ddagger$ by $a^\ddagger := \min\{i \in \{1, \ldots, N-1\} \mid \Delta_i \leq c\}$. Similar to $a^\dagger$, if there is no such arm, we let $a^\ddagger = a^\star$.

We say that an algorithm is *uniformly good* if its regret satisfies $R(T) = o(T^\alpha)$ for all $\alpha > 0$ regardless of $\boldsymbol{\mu} \in \Omega$. The following theorem states an asymptotic lower bound on regret satisfied by any uniformly good algorithm. Note that uniformly good algorithms exist, e.g., KL-UCB without spending any observation budget is uniformly good.

Theorem 3.1. *For all $\boldsymbol{\mu} = (\mu_1, \dots \mu_N) \in \Omega$, any uniformly good algorithm satisfies: if $a^\dagger < a^\ddagger$,*

$$\liminf_{T\to\infty} \frac{R(T)}{\log T} \geq \sum_{i=a^\dagger+1}^{N-1} \frac{\Delta_i}{KL(\mu_i \,\|\, \mu_N)} + c \lim_{T\to\infty} \frac{b(T)}{\log T} + \Delta_{a^\dagger} \left(\sum_{i=1}^{a^\dagger} \frac{1}{KL(\mu_i \,\|\, \mu_N)} - \lim_{T\to\infty} \frac{b(T)}{\log T} \right).$$

otherwise,

$$\liminf_{T\to\infty} \frac{R(T)}{\log T} \geq \sum_{i=a^\ddagger}^{N-1} \frac{\Delta_i}{KL(\mu_i \,\|\, \mu_N)} + \sum_{i=1}^{a^\ddagger-1} \frac{c}{KL(\mu_i \,\|\, \mu_N)}$$

The proof of Theorem 3.1 is given in Section 3.3. It is based on the following observations: (i) in order to minimize the regret, we should not make additional observations for arms i having smaller Δ_i than c, (ii) since playing an arm with a smaller expected reward incurs larger regret, we should use the additional budget to observe such arms. These immediately induce a virtual policy and no algorithm can have a better regret than this. Here, the two arms $a^\dagger$ and $a^\ddagger$ become the *threshold* arms. The observations of arms having lower expected reward than $a^\dagger$ are done by only using the budget. On the contrary, the observations of arms having higher expected reward than $a^\ddagger$ and $a^\ddagger$ are done by only playing these arms, i.e., the budget is never used for arms $\{a^\ddagger, \dots, N\}$

When $b(t) = 0$, i.e., no additional observation, the regret lower bound in Theorem 3.1 matches that of the classical MAB problem [25]. Moreover, Theorem 3.1 implies that a small budget $b(t) = o(\log t)$ does not have any effect on the asymptotic regret lower bound. However, when the budget is enough so that $\lim_{T\to\infty} \frac{b(T)}{\log T} > \sum_{i=1}^{N-1} \frac{1}{KL(\mu_i \,\|\, \mu_N)}$, and we can use the budget with no cost (i.e., $c = 0$), the regret might be sub-logarithmic with respect to the number of rounds, i.e., $R(T)/\log T \longrightarrow 0$. This is confirmed by the algorithm we propose in the following section.

3.2 KL-UCB-AO Algorithm

Inspired by the regret lower bound derived in the previous section, we design the following algorithm, called KL-UCB-AO, which is based on KL-UCB [17] designed for the classical MAB problem.

The following theorem states the asymptotic optimality of KL-UCB-AO.

Theorem 3.2. *For all $\boldsymbol{\mu} = (\mu_1, \dots \mu_N) \in (0,1)^N$, the regret of KL-UCB-AO satisfies:*

i) if $a^\dagger < a^\ddagger$, then

$$\limsup_{T\to\infty} \frac{R(T)}{\log T} \leq \sum_{i=a^\dagger+1}^{N-1} \frac{\Delta_i}{KL(\mu_i \,\|\, \mu_N)} + c \lim_{T\to\infty} \frac{b(T)}{\log T} + \Delta_{a^\dagger} \left(\sum_{i=1}^{a^\dagger} \frac{1}{KL(\mu_i \,\|\, \mu_N)} - \lim_{T\to\infty} \frac{b(T)}{\log T} \right).$$

ii) if $a^\dagger \geq a^\ddagger$, $a^\ddagger \neq N$ or $c \neq 0$, $a^\dagger = a^\ddagger = N$, then

$$\limsup_{T\to\infty} \frac{R(T)}{\log T} \leq \sum_{i=a^\ddagger}^{N-1} \frac{\Delta_i}{KL(\mu_i \,\|\, \mu_N)} + \sum_{i=1}^{a^\ddagger-1} \frac{c}{KL(\mu_i \,\|\, \mu_N)}.$$

iii) otherwise (i.e., $c = 0$, $a^\dagger = a^\ddagger = N$),

$$\limsup_{T\to\infty} R(T) < \infty.$$

KL-UCB-AO Algorithm for the MAB problem with additional observations under stochastic rewards

```
for $t = 1$ to $T$ do
    Step 1: Sort arms.
    Sort arms in the ascending order of their empirical mean rewards, i.e., $\hat{\mu}_a(t) = \frac{\sum_{u=1}^{O_a(t)} X_{a,u}}{O_a(t)}$ and
    $\hat{\mu}_{\sigma(1)}(t) \leq \cdots \leq \hat{\mu}_{\sigma(N)}(t)$. Further define $J_1(t)$ and $J_2(t)$ as follows:
        $J_1(t) := \min \left\{ j \leq N-1 : \sum_{i=1}^{j} \frac{1}{KL(\hat{\mu}_{\sigma(i)} \,\|\, \hat{\mu}_{\sigma(N)})} \geq \frac{b(t)}{\log t} \right\}$
    and $J_1(t) := N$ if $\sum_{i=1}^{N-1} \frac{1}{KL(\hat{\mu}_{\sigma(i)} \,\|\, \hat{\mu}_{\sigma(N)})} < \frac{b(t)}{\log t}$,
    $J_2(t) := \min \left\{ j \leq N-1 : \hat{\mu}_{\sigma(N)} - \hat{\mu}_{\sigma(j)} \leq c \right\}$ and $J_2(t) := N$ if $\hat{\mu}_{\sigma(N)} - \hat{\mu}_{\sigma(N-1)} > c$.
    Then, the arms are divided into two sets $L(t)$ and $U(t)$ depending on $J_1(t)$ and $J_2(t)$, where
    if $J_1(t) < J_2(t)$, $L(t) := \{\sigma(1), \ldots, \sigma(J_1(t))\}$ and $U(t) := \{\sigma(J_1(t)), \ldots, \sigma(N)\}$,
    otherwise, $L(t) := \{\sigma(1), \ldots, \sigma(J_2(t) - 1)\}$ and $U(t) := \{\sigma(J_2(t)), \ldots, \sigma(N)\}$.
    Step 2: Select arms to additionally observe.
    Among the arms in $L(t)$, observe arms according to the order defined in Step 1, as long as
    their KL-UCB indices ($I_a(t)$ for arm $a$ defined below) are higher than $\hat{\mu}_{\sigma(N)}(t)$ and the budget
    is not exhausted.
    Step 3: Play an arm.
    Play arm $A_t \in U(t)$ having the highest KL-UCB index, i.e., $A_t \in \arg\max_{a \in U(t)} I_a(t)$ where
        $I_a(t) := \max\{q \in (0,1) :$
            $O_a(t) KL(\hat{\mu}_a(t) \,\|\, q) \leq \log t + 3 \log \log t\}$.
end for
```

The proof of the above theorem is provided in Section 3.4. In essence, by law of large numbers, the sample mean rewards become close to the true mean rewards after enough rounds have passed. Thus, $\sigma(J_1(t))$ and $\sigma(J_2(t))$ converge to $a^\dagger$ and $a^\ddagger$, respectively. After convergence, $L(t)$ becomes the set of arms that require additional observations and $U(t)$ becomes the set of arms, one of which should be played. In other words, the algorithm utilizes the observation budget only for arms in $L(t)$ and it chooses the arm to play only within $U(t)$. This coincides with our intuition used for deriving the regret lower bound in Theorem 3.1.

Theorem 3.2 shows the asymptotic optimality of KL-UCB-AO for Bernoulli distributed rewards. Furthermore, it implies that if we use the budget for free (i.e., $c = 0$) and the budget is enough so that

$$\lim_{T \to \infty} \frac{b(T)}{\log T} \geq \sum_{i=1}^{N-1} \frac{1}{KL(\mu_i \,\|\, \mu_N)},$$

then the regret does not grow with respect to the number of rounds, i.e., it is finite.

3.3 Proof of Theorem 3.1

Let $\mathbb{P}_\mu$ (resp. $\mathbb{E}_\mu$) denote the probability measure (resp. expectation) under which the average reward vector is $\boldsymbol{\mu}$. We establish that under any uniformly good algorithm, the following holds for any $\boldsymbol{\mu} \in \Omega$ and any suboptimal arm $a \neq N$,

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[O_a(T)]}{\log T} \geq \frac{1}{KL(\mu_a \,\|\, \mu_N)}. \tag{1}$$

This inequality can be proved as in the classical bandit literature [25] using a so-called change-of-measure argument. Recently the authors of [22] simplified this argument (see Theorem 21 and its proof), and we can directly use it to show (1). More precisely, for a suboptimal arm $a \neq N$, consider a new reward vector $\tilde{\boldsymbol{\mu}} = (\mu_1, \ldots, \mu_{a-1}, \tilde{\mu}, \mu_{a+1}, \ldots, \mu_N)$ in which only μ_a is replaced by $\tilde{\mu}$, where $\mu_N < \tilde{\mu} < 1$. Hence, the arm a becomes the unique optimal arm under $\tilde{\boldsymbol{\mu}}$. We can then show, using the same reasoning as in [22] (refer to the appendix for a detailed proof), that $\liminf_{T\to\infty} \frac{\mathbb{E}[O_a(T)]}{\log T} \geq \frac{1}{KL(\mu_a \,\|\, \tilde{\mu})}$. Since we can take arbitrarily $\mu_N < \tilde{\mu} < 1$ and $KL(\mu_a \,\|\, \tilde{\mu})$ is continuous w.r.t. $\tilde{\mu}$, (1) can be derived. Note that several uniformly good algorithms (e.g., KL-UCB) show the tightness of the inequality (1) (see Theorem 1 in [17]), which readily implies that every suboptimal arm a should be observed asymptotically $\log T / KL(\mu_a \,\|\, \mu_N)$ times by round T to have $R(T) = o(T^\alpha)$ for any $\alpha > 0$.

Now, we just minimize the regret subject to the above constraints. To do so, notice the following: (i) the regret increment by playing an arm in $\{a^\ddagger, \ldots, N-1\}$ is less than or equal to c, (ii) playing an arm with smaller μ incurs a larger regret. From (i) and (ii), the solution is as follows: Sort the arms in the ascending order of the expected rewards and spend the observation budget on each arm $a \in \{1, \ldots, a^\ddagger - 1\}$ in that order just as much as (asymptotically) $\log T / KL(\mu_a \,\|\, \mu_N)$ until the budget is exhausted (in that case, the budget becomes empty in arm $a^\dagger$'s turn). Then, we have

if $a^\dagger < a^\ddagger$,

$$\liminf_{T\to\infty} \frac{\mathbb{E}[A_a(T)]}{\log T} \geq \begin{cases} 0 & \text{if } a \in \{1, \ldots, a^\dagger - 1\}, \\ \displaystyle\sum_{i=1}^{a^\dagger} \frac{1}{KL(\mu_a \,\|\, \mu_N)} - \lim_{T\to\infty} \frac{b(T)}{\log T} & \text{if } a = a^\dagger, \\ \dfrac{1}{KL(\mu_a \,\|\, \mu_N)} & \text{if } a \in \{a^\dagger + 1, \ldots, N-1\}, \end{cases}$$

and

$$\liminf_{T\to\infty} \frac{\mathbb{E}[O_a(T) - A_a(T)]}{\log T} \geq \begin{cases} \dfrac{1}{KL(\mu_a \,\|\, \mu_N)} & \text{if } a \in \{1, \ldots, a^\dagger - 1\}, \\ \displaystyle\lim_{T\to\infty} \frac{b(T)}{\log T} - \sum_{i=1}^{a^\dagger - 1} \frac{1}{KL(\mu_a \,\|\, \mu_N)} & \text{if } a = a^\dagger, \\ 0 & \text{if } a \in \{a^\dagger + 1, \ldots, N-1\}, \end{cases}$$

otherwise, i.e., $a^\dagger \geq a^\ddagger$,

$$\liminf_{T\to\infty} \frac{\mathbb{E}[A_a(T)]}{\log T} \geq \begin{cases} 0 & \text{if } a \in \{1, \ldots, a^\ddagger - 1\}, \\ \dfrac{1}{KL(\mu_a \,\|\, \mu_N)} & \text{if } a \in \{a^\ddagger, \ldots, N-1\}, \end{cases}$$

and

$$\liminf_{T\to\infty} \frac{\mathbb{E}[O_a(T) - A_a(T)]}{\log T} \geq \begin{cases} \dfrac{1}{KL(\mu_a \,\|\, \mu_N)} & \text{if } a \in \{1, \ldots, a^\ddagger - 1\}, \\ 0 & \text{if } a \in \{a^\ddagger, \ldots, N-1\}, \end{cases}$$

which concludes the proof.

3.4 Proof of Theorem 3.2

Since KL-UCB-AO mimics the KL-UCB algorithm, one can use the proof arguments in [17] and easily check that KL-UCB-AO is also a uniformly good algorithm satisfying

$$\liminf_{T\to\infty} \frac{\mathbb{E}[O_a(T)]}{\log(T)} \geq \frac{1}{KL(\mu_a \,\|\, \mu_N)} \tag{2}$$

$$\exists D > 0 : \ \forall a \neq N, \quad \limsup_{T\to\infty} \frac{\mathbb{E}[O_a(T)]}{\log(T)} \leq D. \tag{3}$$

In the proof of Theorem 3.2, we distinguish three cases (as stated in the theorem).

Proof for $a^\dagger < a^\ddagger$. Our strategy is to take a simple sample path analysis. Note that when $a^\dagger < a^\ddagger$, $a^\dagger$ should not be N and thus $B := \lim_{T\to\infty} \frac{b(T)}{\log(T)}$ exists. We denote by $F_a(t)$ the number of observations gathered on arm a using the extra budget up to the t-th round, so that $O_a(t) = A_a(t) + F_a(t)$.

We now fix a sample path. From (2), we know that for every arm a, $\liminf_{t\to\infty} O_a(t) = \infty$, and hence $\lim_{t\to\infty} \hat{\mu}_a(t) = \mu_a$. Thus for all $\delta > 0$, there exists t_0 such that for all $t \geq t_0$:

$$\hat{\mu}_1(t) \leq \ldots < \hat{\mu}_N(t), \tag{4}$$

$$|\hat{\mu}_a(t) - \mu_a| < \delta, \quad \forall a, \tag{5}$$

$$O_a(t) \geq (1-\delta)\frac{\log(t)}{KL(\mu_a \,\|\, \mu_N)}, \tag{6}$$

$$J_1(t) = a^\dagger, \tag{7}$$

$$J_2(t) = a^\ddagger. \tag{8}$$

Now consider a suboptimal arm $a \neq N$ and $t \geq t_0$, and denote by t'_a the last round before t where arm a was observed. If we take sufficiently larger t than t_0, we can find $t_1 \geq t_0$ such that for all $a \neq N$, $t'_a \geq t_1$. Since a is observed in the t'_a-th round, we deduce that its KL-UCB index is larger than $\hat{\mu}_N(t)$, and thus:

$$\begin{aligned} O_a(t'_a) &\leq \frac{f(t'_a)}{KL(\hat{\mu}_a(t'_a) \,\|\, \hat{\mu}_N(t'_a))} \\ &\leq \frac{f(t)}{KL(\mu_a + \delta \,\|\, \mu_N - \delta)}, \end{aligned} \tag{9}$$

where $f(t) = \log(t) + 3\log\log(t)$.

Combining (6) and (9), we have shown that for all $\delta > 0$, and for all $t \geq t_1$,

$$(1-\delta)\frac{\log(t)}{KL(\mu_a \,\|\, \mu_N)} \leq O_a(t) \leq \frac{f(t)}{KL(\mu_a + \delta \,\|\, \mu_N - \delta)}.$$

Hence, from the continuity of the function $KL(\cdot \,\|\, \cdot)$, we conclude that:

$$\lim_{T\to\infty} \frac{O_a(T)}{\log(T)} = \frac{1}{KL(\mu_a \,\|\, \mu_N)}. \tag{10}$$

Note that for all $t \geq t_0$, in view of (7) and (8), if $a < a^\dagger$, by design of the algorithm, a is not played at the t-th round, and hence we also have

$$\lim_{T\to\infty} \frac{F_a(T)}{\log(T)} = \frac{1}{KL(\mu_a \,\|\, \mu_N)}.$$

Again by design of the KL-UCB-AO algorithm, the remaining budget can only be used to observe arm $a^\dagger$ in every t-th round for $t \geq t_0$, which implies:

$$\lim_{T\to\infty} \frac{F_{a^\dagger}(T)}{\log(T)} = B - \sum_{a=1}^{a^\dagger-1} \frac{1}{KL(\mu_a \,\|\, \mu_N)}. \tag{11}$$

From (10), where we apply $a = a^\dagger$, the number of rounds when $a^\dagger$ is played satisfies:

$$\lim_{T\to\infty} \frac{A_{a^\dagger}(T)}{\log(T)} = \sum_{a=1}^{a^\dagger} \frac{1}{KL(\mu_a \,\|\, \mu_N)} - B. \tag{12}$$

Finally for every t-th round for $t \geq t_0$, by design of the algorithm, the budget of observations is not used to sample an arm $a > a^\dagger$. Using (10) applied to arm $a > a^\dagger$, we get: for all $a > a^\dagger$,

$$\lim_{T\to\infty} \frac{A_a(T)}{\log(T)} = \frac{1}{KL(\mu_a \,\|\, \mu_N)}. \tag{13}$$

To conclude the proof of the case $a^\dagger < a^\ddagger$, using the dominated convergence theorem, we deduce from (3), (12) and (13) that:

$$\lim_{T\to\infty} \frac{R(T)}{\log(T)} = \sum_{i=a^\dagger+1}^{N-1} \frac{\Delta_i}{KL(\mu_i \,\|\, \mu_N)} + c \lim_{T\to\infty} \frac{b(T)}{\log T} + \Delta_{a^\dagger} \left(\sum_{i=1}^{a^\dagger} \frac{1}{KL(\mu_i \,\|\, \mu_N)} - \lim_{T\to\infty} \frac{b(T)}{\log T} \right).$$

Proof for $a^\dagger \geq a^\ddagger$, $a^\ddagger \neq N$ or $c \neq 0$, $a^\dagger = a^\ddagger = N$. The technique used in the proof of the previous case can be applied to this case in a straightforward way. Then, for all $t \geq t_0$, we obtain $L(t) = \{1, \ldots, a^\ddagger - 1\}$ for additional observations and $U(t) = \{a^\ddagger, \ldots, N\}$ for playing. From (10), the above statement implies that for all $a < a^\ddagger$,

$$\lim_{T\to\infty} \frac{F_a(T)}{\log(T)} = \frac{1}{KL(\mu_a \,\|\, \mu_N)} \tag{14}$$

and for all $a \geq a^\ddagger$,

$$\lim_{T\to\infty} \frac{A_a(T)}{\log(T)} = \frac{1}{KL(\mu_a \,\|\, \mu_N)}. \tag{15}$$

Combining the dominated convergence theorem with (3), (14) and (15), we conclude

$$\lim_{T\to\infty} \frac{R(T)}{\log(T)} = \sum_{i=a^\ddagger}^{N-1} \frac{\Delta_i}{KL(\mu_i \,\|\, \mu_N)} + \sum_{i=1}^{a^\ddagger-1} \frac{c}{KL(\mu_i \,\|\, \mu_N)}.$$

Proof for $c = 0$, $a^\dagger = a^\ddagger = N$. Similar to the above cases, there exists t_0 such that for all $t \geq t_0$, $\hat{\mu}_1(t) \leq \ldots < \hat{\mu}_N(t)$ and $J_1(t) = J_2(t) = N$. Then, $U(t)$ turns out to contain only the optimal arm N for all $t \geq t_0$. However, unlike the previous case, no charge is made for observing arms in $L(t)$. Therefore, the regret $R(t)$ does not grow any more after t_0, which completes the proof of the last case.

H-INF Algorithm for the MAB problem with additional observations under adversarial rewards

```
Parameters:
• Continuously differentiable potential functions $\Psi^b, \Psi^e : \mathbb{R}^*_- \longrightarrow \mathbb{R}^*_+$, for INF bandit/expert algorithm [4], respectively, where $\Psi' > 0$, $\lim_{x \to 0} \Psi(x) \geq 1$, $\lim_{x \to -\infty} \Psi^b(x) < 1/\lceil N/M \rceil$, $\lim_{x \to -\infty} \Psi^e(x) < 1/M$.
• Estimates $v_a^b(t)$ and $v_a^e(t)$ of $X_a(t)$ based on the observed rewards at (and before) round $t$ for INF bandit/expert algorithm, respectively.

Initialization:
• Partition arms $\mathcal{N}$ into $M$ groups $\{\mathcal{G}_1, \ldots, \mathcal{G}_M\}$ with $\bigcup_i \mathcal{G}_i = \mathcal{N}$ and $|\mathcal{G}_i| \in \{\lceil N/M \rceil, \lfloor N/M \rfloor\}$. The set of arms in $\mathcal{G}_i$ is denoted by $\{i_1, \ldots, i_{|\mathcal{G}_i|}\}$.
• At the initial round $t = 1$, randomly select a candidate arm $c_i \in \mathcal{G}_i$ for every $\mathcal{G}_i$ and then randomly select $A_1$ among $\{c_1, \ldots, c_M\}$. Play $A_1$ and observe the rewards of all $\{c_1, \ldots, c_M\}$.

for $t = 2$ to $T$ do
    Layer 1: Find the per-group local best arm.
        For every group $\mathcal{G}_i$,
        Step 1: Build the estimate $v_i^b(t-1) = (v_{i_1}^b(t-1), \ldots, v_{i_{|\mathcal{G}_i|}}^b(t-1))$ of $(X_{i_1}(t-1), \ldots, X_{i_{|\mathcal{G}_i|}}(t-1))$ and let $V_i^b(t-1) = \sum_{s=1}^{t-1} v_i^b(s) = (V_{i_1}^b(t-1), \ldots, V_{i_{|\mathcal{G}_i|}}^b(t-1))$.
        Step 2: Compute the normalized constant $C_{i,t-1} = C(V_i^b(t-1))$.
        Step 3: Compute the probability distribution $\mathbf{p_{i,t}} = (p_{i_1,t}, \ldots, p_{i_{|\mathcal{G}_i|},t})$, where $p_{i_a,t} = \Psi^b(V_{i_a}(t-1) - C_{i,t-1})$.
        Step 4: Draw a candidate arm $c_{i,t} \in \mathcal{G}_i$ from the probability distribution $\mathbf{p_{i,t}}$.
    Layer 2: Find the best group.
        Step 5: Build the estimate $v^e(t-1) = (v_{\mathcal{G}_1}^e(t-1), \ldots, v_{\mathcal{G}_M}^e(t-1))$ of $(X_{\mathcal{G}_1}(t-1), \ldots, X_{\mathcal{G}_M}(t-1))$ and let $V^e(t-1) = \sum_{s=1}^{t-1} v^e(s) = (V_{\mathcal{G}_1}^e(t-1), \ldots, V_{\mathcal{G}_M}^e(t-1))$, where each group is considered as a virtual arm so that the reward of group $\mathcal{G}_i$, $X_{\mathcal{G}_i}(t)$ is the reward of the corresponding candidate arm at that round $X_{c_{i,t}}(t)$.
        Step 6: Compute the normalized constant $C_{t-1} = C(V^e(t-1))$.
        Step 7: Compute the probability distribution $\mathbf{p_t} = (p_{\mathcal{G}_1,t}, \ldots, p_{\mathcal{G}_M,t})$, where $p_{\mathcal{G}_i,t} = \Psi^e(V_{\mathcal{G}_i}(t-1) - C_{t-1})$.
        Step 8: Draw a group $\mathcal{G}_t$ from the probability distribution $\mathbf{p_t}$. Play the candidate arm of $\mathcal{G}_t$ and observe the rewards of all $\{c_{1,t}, \ldots, c_{M,t}\}$.
end for
```

4 ADVERSARIAL REWARDS

In the adversarial setting, the reward sequences can be arbitrary. We first describe a known lower bound on regret for our problem and then present H-INF (Hierarchical INF), an algorithm whose regret matches the lower bound order-wise.

4.1 Regret Lower Bound

It is known that for the adversarial classical MAB problem (i.e., $b(t) = 0$), the regret scales at least as $\Omega(\sqrt{NT})$. On the other hand, for the expert problem (i.e., $b(t) = (M-1)t$), the regret under any

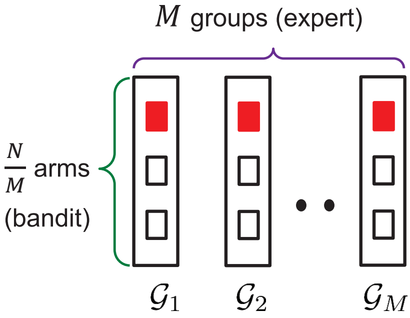


Fig. 1. Colored ones represent arms to be observed, and their rewards regarded as those of groups at the corresponding round. Under H-INF, the bandit algorithm is operated inside a group and the expert algorithm is executed over groups.

algorithm scales as $\Omega(\sqrt{T \log N})$. These regret lower bounds are achieved by the INF algorithm [4]. It implies that the regret inevitably grows at the rate of $\sqrt{T}$ regardless of the amount of additional observations. Hence, we restrict our attention to linear budget functions, i.e., $b(t) = (M-1)t$ for some $M = 1, \ldots, N$. Due to this, we consider only the case where additional observations have no cost (i.e., $c = 0$), otherwise, the regret would grow linearly.[4] Such linear budget is needed to hope to significantly reduce the regret, and is indeed considered in [3, 20, 31]. Note that none of these papers proposes an order-optimal algorithm. For $b(t) = (M-1)t$, the authors of [31] provides the regret lower bound

$$\inf_{\text{algorithm}} \sup_{\text{adversary}} R(T) \geq 0.03\sqrt{\frac{N}{M}T}, \text{ for } T \geq \frac{3N}{16M}, \tag{16}$$

where the infimum is over any algorithm and the supremum is over all oblivious adversaries[5]. Since we assume a *non-oblivious* adversary who is the most general one, this lower bound is valid for our problem as well. The regret lower bound on the expert problem in [10] also provides another lower bound of the problem, which is

$$\inf_{\text{algorithm}} \sup_{\text{adversary}} \limsup_{T\to\infty} \frac{R(T)}{\sqrt{T \log N/2}} \geq 1. \tag{17}$$

Therefore, from (16) and (17), the regret of any algorithm should satisfy

$$\sup_{\text{adversary}} \limsup_{T\to\infty} \frac{R(T)}{\max\left\{\sqrt{\frac{N}{M}T}, \sqrt{T \log N}\right\}} \geq 0.03. \tag{18}$$

4.2 H-INF Algorithm

The proposed H-INF algorithm for any $1 \leq M \leq N$ has a hierarchical structure consisting of two layers. We explain the key idea of H-INF using an illustration in Figure 1. The set of arms is

[4] Note that [3, 31] studied adversarial MAB problems considering costs for additional observations, but they considered settings that differ from ours. First, [31] assumes that the arms played are not necessarily observed ones, which makes the regret lower bound different from ours. Second, [3] considers regrets independent of costs, while we assume that each additional observation increases the regret.

[5] An oblivious adversary generates rewards independent of the past decisions of the player.

first divided into M groups $\{\mathcal{G}_1, \ldots, \mathcal{G}_M\}$ of almost equal sizes, thus a single group has $\approx N/M$ arms. Then, in Layer 1, we run the INF bandit algorithm inside each group and select per-group locally-best arms ($c_i : i = 1, 2, \cdots, M$), which correspond to the arms to be played and/or observed. In Layer 2, we run the INF expert algorithm that regards each group as a *virtual arm,*where each c_i's reward is provided as a reward of the group i (or the virtual arm i) to the INF expert algorithm.

To explain how the INF algorithm works in more detail, it calculates the estimated cumulative reward of each arm (Steps 1 and 5). It assigns $p_{a,t}$ to arm a, which is the potential function value of the normalized estimated cumulative reward of arm a (Steps 3 and 7). The algorithm selects an arm to play/observe according to the distribution p_t (Steps 4 and 8). Note that the more rewarded arm a has been received, the more likely it will be selected at the next round because $\Psi' > 0$. The normalization in Step 2 (or Step 6) ensures that p_{t+1} in Step 3 (or Step 7) forms a probability distribution. The existence of function C in Steps 2 and 6 is guaranteed in the work of Audibert and Bubeck [4] (see Lemma 1 in their paper). The authors provide some potential functions and reward estimators resulting in the $O(\sqrt{NT})$ regret for the classical MAB setting, and $O(\sqrt{T \log N})$ regret for the expert setting, where the following functions are the examples:

$$\Psi^b(x) = \frac{5t}{x^2} + \min\left(\frac{1}{2\lceil N/M \rceil}, \sqrt{\frac{3}{t\lceil N/M \rceil}}\right), \quad \Psi^e(x) = \left(\frac{1.8\sqrt{t \log M}}{-x}\right)^{3 \log M}$$

$$v_a^b(t) = \frac{X_a(t)}{p_{a,t}} \mathbb{1}_{A_t = a}, \quad v_a^e(t) = X_a(t)$$

In H-INF, we utilize this INF algorithm in a hierarchical manner such that additional observation opportunities are smartly used. This simple, yet intelligent "recycling" of the known optimal algorithms enables us to decouple the algorithmic components, which significantly facilitates the regret analysis, as formally stated in the following theorem whose proof is given in Section 4.3.

Theorem 4.1. *For all $T \geq 1$, the regret of H-INF satisfies that*

$$\begin{aligned} R(T) &= O\left(\max\left\{\sqrt{\frac{N}{M}T}, \sqrt{T \log N}\right\}\right) \\ &= \begin{cases} O\left(\sqrt{\frac{N}{M}T}\right) & \textit{if} \quad M \leq N/\log N \\ O\left(\sqrt{T \log N}\right) & \textit{otherwise} \end{cases} . \end{aligned} \tag{19}$$

The above theorem implies that the regret upper bound (19) of the H-INF algorithm is order-optimal matching the regret lower bound (18). As we mentioned in Section 1.2, the best known regret bound is $O\left(\sqrt{\frac{N}{M}T \log N}\right)$ by [3, 31] which is larger than (19) by a logarithmic factor $O\left(\sqrt{\log N}\right)$ for $1 \leq M \leq N/\log N$. The fundamental reason why the algorithms in [3, 31] are sub-optimal is because they are based on the EXP3 algorithm [6] that is optimal only in the expert problem. The core strength of H-INF lies in constructing a *hierarchical* structure utilizing the optimal INF expert/bandit algorithms as basic building blocks. We remark that it is not mandatory to use the INF algorithm to achieve the order-optimality, and any order-optimal bandit and/or expert algorithm can be used at each layer. One can easily see that our proof similarly works as long as the algorithm used at each layer is order-optimal in terms of its regret bound.

4.3 Proof of Theorem 4.1

At the T-th round, we denote the best arm by

$$a^\star = \arg\max_{a \in \mathcal{N}} \mathbb{E}\left[\sum_{t=1}^{T} X_a(t)\right]$$

and the best group by $\mathcal{G}^\star$ which contains $a^\star$. Then, it follows that

$$\begin{aligned} R(T) &= \max_{a \in \mathcal{N}} \mathbb{E}\left[\sum_{t=1}^{T} X_a(t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} X_{A_t}(t)\right] \\ &= \max_{a \in \mathcal{G}^\star} \mathbb{E}\left[\sum_{t=1}^{T} X_a(t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} X_{A_t}(t)\right] \\ &\leq \max_{a \in \mathcal{G}^\star} \mathbb{E}\left[\sum_{t=1}^{T} X_a(t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} X_{A_t}(t)\right] \\ &\quad + \max_{\mathcal{G}_i \in \{\mathcal{G}_1, \ldots, \mathcal{G}_M\}} \mathbb{E}\left[\sum_{t=1}^{T} X_{\mathcal{G}_i}(t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} X_{\mathcal{G}^\star}(t)\right], \end{aligned} \tag{20}$$

where $X_{\mathcal{G}_i}(t)$ denotes the reward of $\mathcal{G}_i$ at the t-th round, defined as the reward of the candidate arm of $\mathcal{G}_i$ selected at Layer 1 of H-INF. We now focus on the group $\mathcal{G}^\star$. One can observe that inside the group $\mathcal{G}^\star$, the algorithm runs the INF bandit algorithm over the arms in $\mathcal{G}^\star$ only. Since there are at most $\lceil N/M \rceil$ arms in $\mathcal{G}^\star$, we obtain

$$\max_{a \in \mathcal{G}^\star} \mathbb{E}\left[\sum_{t=1}^{T} X_a(t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} X_{\mathcal{G}^\star}(t)\right] = O\left(\sqrt{\lceil N/M \rceil T}\right) = O\left(\sqrt{\frac{N}{M} T}\right) \tag{21}$$

from the known regret bound of the INF bandit algorithm [4].[6]

Similarly, for Layer 2 where the INF expert algorithm is applied on M groups, we have

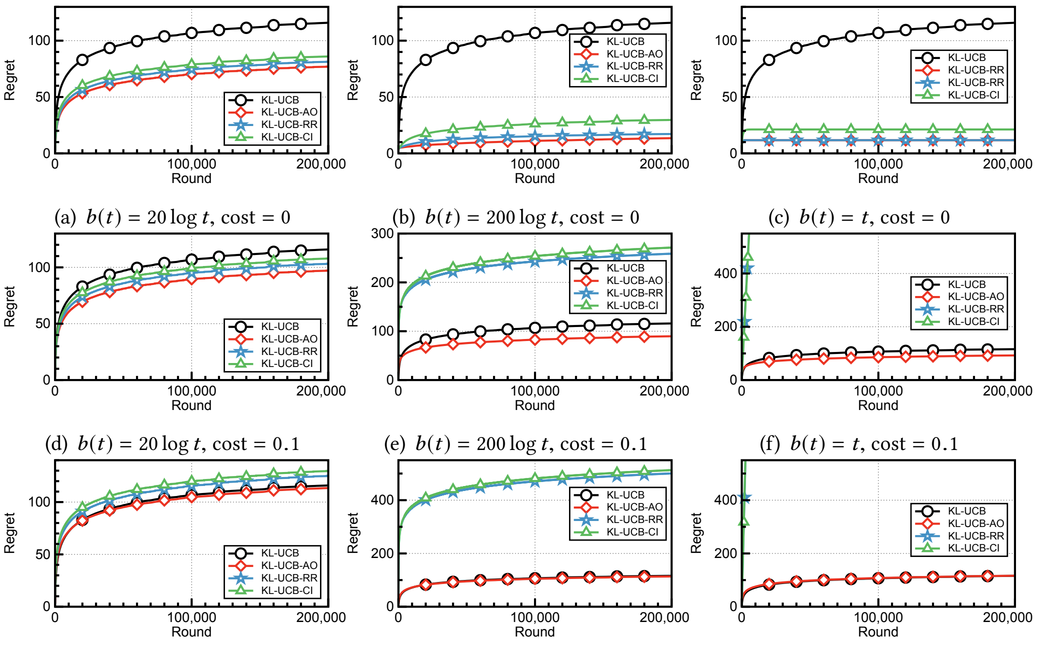
$$\begin{aligned} \max_{\mathcal{G}_i \in \{\mathcal{G}_1, \ldots, \mathcal{G}_M\}} \mathbb{E}\left[\sum_{t=1}^{T} X_{\mathcal{G}_i}(t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} X_{A_t}(t)\right] \\ = O\left(\sqrt{T \log M}\right) \end{aligned} \tag{22}$$

from the known regret bound of the INF expert algorithm [4]. Therefore, combining (20), (21) and (22) leads to

$$\begin{aligned} R(T) &= O\left(\sqrt{\frac{N}{M} T} + \sqrt{T \log M}\right) \\ &= O\left(\max\left\{\sqrt{\frac{N}{M} T}, \sqrt{T \log N}\right\}\right). \end{aligned}$$

This completes the proof of Theorem 4.1.

[6] Theorems in [4] assume that T is already known, but using the "doubling trick" in [11] can lead to the same-order regret bound under unknown T.



(a) $b(t) = 20 \log t$, cost = 0

(b) $b(t) = 200 \log t$, cost = 0

(c) $b(t) = t$, cost = 0

(d) $b(t) = 20 \log t$, cost = 0.1

(e) $b(t) = 200 \log t$, cost = 0.1

(f) $b(t) = t$, cost = 0.1

(g) $b(t) = 20 \log t$, cost = 0.2

(h) $b(t) = 200 \log t$, cost = 0.2

(i) $b(t) = t$, cost = 0.2

Fig. 2. Regret comparisons under various costs and observation budget functions.

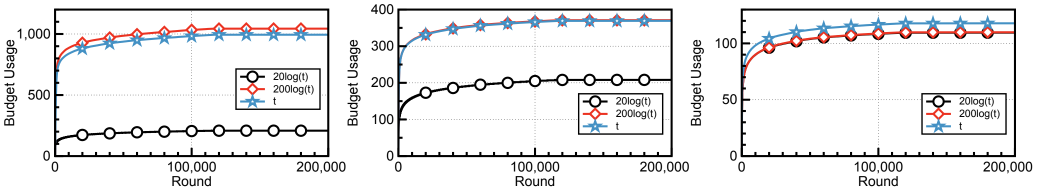
5 EVALUATION AND APPLICATIONS

In this section, we illustrate the performance of our algorithms on synthetic and real-world scenarios. We only consider KL-UCB-AO, and do not run H-INF in our experiments since (a) KL-UCB-AO is better suited to practical scenarios with limited budget of additional observations (b) the adversarial setting is more difficult to simulate and (c) stochastic rewards provide a more appropriate model in many real-world applications.

We first consider synthetic setups where we investigate how KL-UCB-AO performs compared to other heuristic based algorithms for various cost values and budget functions. Next, we consider two interesting real-world applications: (i) link rate adaptation in wireless networks and (ii) online web advertisement. For both applications, we use real traces to extract the results of our simulations (the average rewards of the various arms). We derive the results in the synthetic setting when both the budget and the cost are given. For the rate adaptation scenario, we vary the cost but consider an infinite budget, since in this case, there is no reason for us to restrict our analysis to algorithms exploiting a few additional observations only (see Section 5.2 for details). In the case of online advertisement, we assume that the budget is limited and the cost is equal to 0 (see Section 5.3 for detailed explanations). The above three scenarios illustrate the fact that KL-UCB-AO can deal with different cost and budget settings.

5.1 Synthetic evaluation

Setup. We consider 6 arms with Bernoulli distributed rewards of expectations $\boldsymbol{\mu} = (0.59, 0.64, 0.67, 0.68, 0.78, 0.85)$, drawn from the uniform distribution in $[0.5, 0.9]$. Considering the fact that the



(a) Budget usage, cost = 0 (b) Budget usage, cost = 0.1 (c) Budget usage, cost = 0.2

Fig. 3. Budget usage comparison under various costs and observation budget functions.

mean reward of each arm is randomly chosen over the interval $[0.5, 0.9]$, our cost ranges from 0 to about 30% of the average mean reward across arms. Specifically, we set the cost values to 0, 0.1 and 0.2. This is because the maximum difference between the average reward of two arms is $0.85 - 0.59 = 0.26$, and hence if the cost is larger than 0.26, then KL-UCB-AO does not use any additional observation. We compare KL-UCB-AO with the following two baseline algorithms:

- *KL-UCB-RR:* it plays the arm with the highest KL-UCB index and spends the budget to observe all other arms in a round-robin manner so that the budget is fairly distributed over all arms.
- *KL-UCB-CI:* it also plays the arm with the highest KL-UCB index. However, it spends the budget on the arms with wider *confidence interval*, defined as the difference between the KL-UCB index and the empirical average reward. This algorithm is natural: since the confidence interval quantifies the uncertainty of the empirical reward, it may be desirable to additionally observe arms with wide confidence intervals.

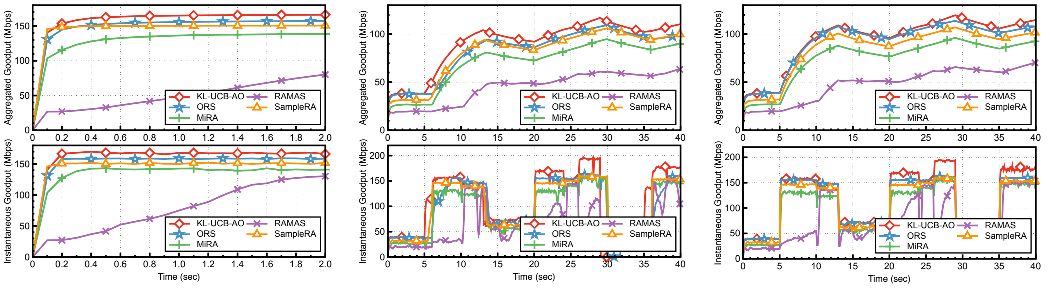
All the curves in each figure correspond to the regret averaged over 3, 000 random runs.

Results. Figure 2 contains 9 plots, each of which corresponds to a specific configuration of cost value and budget function. We note that from our analysis in Section 3, under the zero cost, the minimum budget needed to achieve a constant regret in KL-UCB-AO is:

$$\sum_{a=1}^{N-1} \frac{\log t}{KL(B(\mu_a) \,||\, B(\mu_N))} \simeq 93 \log t.$$

Using this criterion, we classify the cases $b(t) = 20 \log t$ as "small budgets", $b(t) = 200 \log t$ as "large budget". We also test the "linear budget" $b(t) = t$, which represents the $\omega(\log(t))$ order budget functions. Note that KL-UCB's plot (this corresponds to having no budget for additional observations) is identical in all figures. As expected, compared to KL-UCB without additional observations, one can confirm that additional observations are indeed effective for reducing the regret as long as the observation cost is not too high. In particular, as in Figure 2(b), having a budget function $b(t) = 200 \log t$ and no cost leads to a constant regret. As seen in Figure 2, the regret of KL-UCB-AO always decreases with increasing budget and is the lowest among those of tested algorithms, while regrets of KL-UCB-RR and KL-UCB-CI often increase. This is because additional observations might incur higher regret if they are not used appropriately.

We also observe that the performance of KL-UCB-AO in the large budget ($b(t) = 200 \log(t)$) case is similar to that in the linear ($b(t) = t$) case as shown in Figures 2(b),2(e),2(h),2(c),2(f), and 2(i). Especially, if the cost increases, the budget is used less and proceeds in the direction of gain in exploration. To explain why, as reported in Figure 3, KL-UCB-AO always uses budget as log order, no matter which budget function is given. This is because $b(t) = O(\log t)$ is enough for achieving constant regret due to our analysis in Section 3.



(a) Stationary Channel (b) Non-stationary Channel, Periodical Reset (c) Non-stationary Channel, Ideal Reset

Fig. 4. Rate adaptation. Throughput comparison under various algorithms in 802.11 systems

5.2 Rate Adaptation in 802.11 Systems

Rate adaptation (RA) is one of the critical modules in 802.11-based wireless systems. Its objective is to identify the optimal rate at which packets should be sent from the transmitter to the receiver, where the optimal rate is the one that maximizes the throughput, i.e., the product of the rate and the (unknown) probability of successful transmission at that rate. For each packet transmission, the transmitter selects a rate, and receives an acknowledgement of transmission success or failure from the receiver. This acknowledgement constitutes the unique feedback, based on which the transmitter learns the optimal rate. Regarding each rate as an arm, and transmitting a packet as pulling an arm, the design of RA schemes can be seen as a stochastic MAB problem, where the mean reward of an arm with X Mbps and the success probability p is the $X \times p$.

Setup. In our evaluation, we consider a 802.11n MIMO system with 16 rate configurations in Mbps, (single stream: 13.5, 27, 40.5, 54, 81, 108, 121.5, 135, and double stream: 27, 54, 81, 108, 162, 216, 243, 270). We assume a stationary channel condition, and to realistically assign the success probabilities to each rate, we use the real traces used in [14], which record the mapping between channel measurements, the SNR and diffSNR (i.e., the maximal gap between the SNRs measured at various antennas), and the packet transmission success probabilities. The size of a MAC packet is set to be 7969 bytes (header size 34 bytes and maximum payload size 7935 bytes), where we assume popular frame aggregation in 802.11n.

KL-UCB-AO in rate adapation. We first start by presenting our implementation of KL-UCB-AO in the context of the rate adaptation application. In KL-UCB-AO, with each data packet transmission is associated a sequence of probe packets. Probe packets are those used for exploration with additional observations (**Step 2** of KL-UCB-AO), whereas data packets are for both exploration and exploitation (**Step 3** of KL-UCB-AO). Note that the rates used for the data packets and each associated probe packets are determined by KL-UCB-AO (see Section 3.2). Thus, using probe packets naturally incur a certain cost, because their transmissions waste bandwidth, and the amount of cost due to probe packets depends on the size of probe packets. We use data MAC (Medium Access Control) packets of size 7969 bytes (this corresponds to the largest packet size with frame aggregation in 802.11n), and probe packets of size 398 bytes, which corresponds to a cost of 5% compared to data packet, i.e., $7969 \times 0.05 = 398$. Note that in the rate adaptation application, we do not impose any budget constraint (i.e., infinite budget), because the efficiency of the rate adaptation schemes is measured through the long-term net-throughput, for which it is natural that the transmitter is assigned a full degree of freedom in how aggressively it utilizes bandwidth-wasting probe packets. Thus, without the budget constraint, KL-UCB-AO behaves in such a way that it smartly controls the amount of the probe packets over time to maximize the net-throughput computed for data packets only.

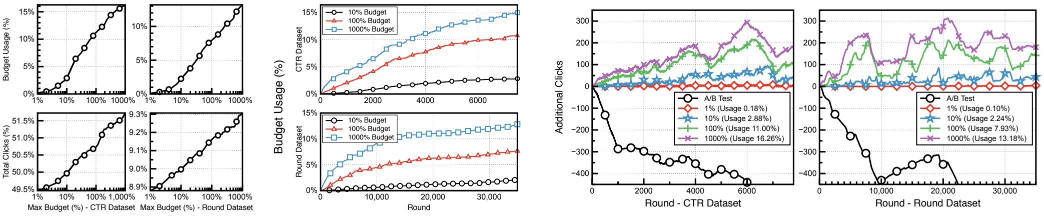
Tested algorithms for comparison.

We test four baseline algorithms SampleRA [7], MiRA [29], RAMAS [28], and ORS [13] for comparison to KL-UCB-AO. Briefly speaking, all these algorithms except for ORS heuristically use the idea of trading exploration and exploitation in a different manner. ORS is the rate adaptation algorithm that comes from the explicit formulation of the problem as a MAB problem, but does not perform additional observations, relying on the pulled arm for exploration and exploitation, as in the conventional MAB. MiRA performs additional observations via probe packets similarly to KL-UCB-AO, where over some intervals probe packets are transmitted, with the intervals being determined in an adaptive manner. The performance gap between MiRA and KL-UCB-AO allows us to quantify the gain of a theory-driven optimal algorithm with additional observations. SampleRA is the earliest version of rate adaptation in 802.11 systems, where exploration is executed every 10 packets. RMAMS is a threshold-based algorithm, where some statistics (e.g., success rate and retry count) are checked and the selected rate is changed if those statistics are above or below some threshold.

Results. We performed trace-based simulations for stationary and non-stationary channel scenarios. We use the real traces from [14]. In the stationary scenario, the channel condition between the transmitter and the receiver is chosen so that the maximum transmission rate is 162 Mbps. We observe similar trends for different channel conditions. We generate non-stationary scenarios by choosing 8 channel conditions from the traces. These conditions change at times 5, 10, 13, 20, 24, 26, 30, 36 secs. As summarized earlier, ORS corresponds to KL-UCB without additional observations, whereas other algorithms implicitly mix exploration and exploitation in a heuristic manner. Figure 4(a) shows the results of instantaneous (measured every 0.1 sec intervals) and aggregate goodputs under the stationary scenario. We observe that KL-UCB-AO finds the optimal rate slightly faster than ORS, and that other algorithms perform sub-optimally. This result illustrates the fact that additional observations enable to dynamically utilize exploration chances so that the optimal arm is quickly found. The advantage of additional observations becomes even clearer in the non-stationary scenario. Figures 4(b) and 4(c) show the gootputs when the learning algorithms are periodically and ideally reset, respectively. By ideally, we mean that the transmitter magically knows when the channel condition changes and resets its parameters in the algorihtms. These two ways of 'periodic and ideal reset' are just the baseline mechanisms that respond to non-stationary condition changes, and developing a better mechanism is beyond the scope of this paper. However, we are able to quantify how beneficial the idea of using additional observations like KL-UCB-AO is, compared to those without it. In Figures 4(b) and 4(c), we observe that KL-UCB-AO achieves the goodputs of upto 50 Mbps higher than other algorithms, and overall it quickly finds the optimal rate and thus achieves larger goodputs over time.

5.3 Online Advertisement

Online advertisement is one of the major income sources for the Internet industry today [18]. An advertiser who wants to promote new products would buy an ad placement on a popular website. Whenever a user visits the website, the web administrator should choose a few ads to display among many candidate ones. For example, in case of sponsored advertising [18] in a search engine such as Google, an advertiser bids for keywords related to her product or service to display her ad. For example, car manufacturers might want to advertise their products when the keyword *'car'* is queried. In this case, the search engine should display an ad among those of car manufacturers that actually bid for the keyword. It should select the ad with the highest click-through-rate



(a) Budget usage, click (b) Budget usage per round (c) Additional clicks beyond KL-UCB (CTR, Round dataset)

Fig. 5. Online advertisement. Comparison under various datasets and budget sizes

(CTR). Therefore, the search engine should learn which ad has the highest CTR. This is a popular application of MAB problems, where each candidate ad corresponds to an arm.

Setup. In order to evaluate the proposed KL-UCB-AO algorithm compared to other algorithms, we use a real-world online advertising dataset from KDD CUP 2012 [1], which is obtained from a Tencent proprietary search engine, soso.com. Each data instance corresponds to an impression that tells which ad is displayed and whether the ad is clicked or not. In other words, whether an advertisement is displayed or not corresponds to whether the value of the real trace is visible or not.

In this dataset, there exist a huge number of ads (641707 ads), from which we extract two subsets of top 10 ads based on the number of times they are displayed (we refer to the corresponding dataset as the Round dataset) and on their CTRs (we refer to the corresponding dataset as CTR dataset). The Round and CTR datasets contain 700,000 and 100,000 data, respectively, and the CTRs (i.e., mean rewards in the MAB problem) of the 10 top ads are equal to [0.05065, 0.00001, 0.00007, 0.03681, 0.03468, 0.000189, 0.02023, 0.03158, 0.0101, 0.08369] and [0.45514, 0.31748, 0.29142, 0.19736, 0.15904, 0.14182, 0.14768, 0.12895, 0.10791, 0.10273], respectively. For the Round (or CTR) dataset, the 700,000 (or 100,000) data are divided into 14 (or 10) sets and we report the averaged performance among them.

Algorithm. In practice, additional observations required for the KL-UCB-AO algorithm can be obtained in several ways. For example, whenever a user visits the web-site, the web administrator can ask the user to click on 'fake' ads whenever she finds one interesting. In other words, the KL-UCB-AO based algorithm selects one ad and method (real or fake) to display at each round. Such additional information obtained from users might be considered as being free (i.e., zero cost), but it degrades the reputation of the web-site. Therefore, ideally one should not use too many of these additional observations. Each web-site operator can manage their own preferences by selecting an appropriate budget of additional observations to increase the overall CTR performance while not sacrificing its reputation. In order to evaluate the impact of additional observations for KL-UCB-AO, we experiment with various linear budget functions of $0.01t, 0.1t, t, 10t$, i.e., do not show a pop-up querying 0.01, 0.1, 1, 10 times more than the user actually has a chance to click 'real' ads. We remark that irrespective of the linear maximum budget allowed, KL-UCB-AO only uses its log portion as we explained in Section 5.1. We also compare KL-UCB-AO with the online A/B Test algorithm [24] which is popular for the online advertisement scenario and selects the ad with the highest empirical mean value among candidates.

Result. First, Figure 5(a) reports the total number of clicks and budget usage according to the budget given for the two test datasets. As expected, the number of clicks increases as the budget usage increases. However, as mentioned earlier, KL-UCB-AO only uses a small portion of the maximum budget allowed for additional observations. Second, Figure 5(b) shows the pattern of actual budget usage per round. Even if the budget is linearly increasing with time, additional observations are used a sub-linear (i.e., log) number of times. Third, 5(c) represents the number of additional clicks achieved by KL-UCB-AO and A/B Test compared to that by KL-UCB, which quantifies the advertisement gain achieved thanks to additional observations. As reported in 5(c), the total number of clicks under KL-UCB-AO achieves the best performance. Overall, we observe that additional observations lead to up to $4 \sim 5\%$ improvement in terms of the total number of clicks.

6 CONCLUSION

In this paper, we studied the multi-armed bandit with additional observations which provides a natural extension between the bandit problem and the expert problem. For stochastic rewards, we derive an asymptotic lower bound on regret, satisfied by any uniformly good algorithm. Motivated by the lower bound, we developed KL-UCB-AO, an asymptotically optimal algorithm. For adversarial rewards, we propose a hierarchical algorithm whose regret is order-optimal. We present two applications in rate adaptation over wireless networks and online advertisement, where we showed the proposed algorithms' value in practice. We believe that our ideas in designing bandit algorithms are of boarder interest to study similar problems, e.g., the contextual or graph-structured bandit problems.

ACKNOWLEDGMENTS

The work of Y. Yi is supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science and ICT(No. 2016R1A2A2A05921755). The work of A. Proutiere is supported by the ERC consolidator grant 308267 - Fluid Spectrum Access, by VR, and SRA TNG framework. The work of J.W. Shin is supported by the ICT R&D program of MSIP/IITP. [2016-0-00563, Research on Adaptive Machine Learning Technology Development for Intelligent Autonomous Digital Companion]

REFERENCES

[1] KDD cup 2012 Track 2. http://www.kddcup2012.org/c/kddcup2012-track2.

[2] Noga Alon, Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. 2013. From bandits to experts: A tale of domination and independence. In *Proceedings of NIPS.*

[3] Kareem Amin, Satyen Kale, and Gerald Tesauro Deepak Turaga. 2015. Budgeted Prediction With Expert Advice. In *Proceedings of AAAI.*

[4] Jean-Yves Audibert and Sébastien Bubeck. 2010. Regret bounds and minimax policies under partial monitoring. *The Journal of Machine Learning Research* 11 (2010), 2785–2836.

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

[6] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. 2002. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.* 32, 1 (2002), 48–77.

[7] J. Bicket. 2005. Bit-rate selection in wireless networks. In *PhD thesis, Massachusetts Institute of Technology.*

[8] Swapna Buccapatnam, Atilla Eryilmaz, and Ness B Shroff. 2014. Stochastic bandits with side observations on networks. In *Proceedings of ACM SIGMETRICS.*

[9] Stéphane Caron, Branislav Kveton, Marc Lelarge, and Smriti Bhagat. 2012. Leveraging Side Observations in Stochastic Bandits. In *Proceedings of UAI.*

[10] Nicolo Cesa-Bianchi, Yoav Freund, David Haussler, David P Helmbold, Robert E Schapire, and Manfred K Warmuth. 1997. How to use expert advice. *Journal of the ACM (JACM)* 44, 3 (1997), 427–485.

[11] Nicolò Cesa-Bianchi and Gábor Lugosi. 2006. *Prediction, Learning, and Games.* Cambridge University Press.

[12] Richard Combes, Chong Jiang, and R Srikant. 2015. Bandits with Budgets: Regret Lower Bounds and Optimal Algorithms. In *Proceedings of ACM SIGMETRICS.*

[13] Richard Combes, Alexandre Proutiere, Donggyu Yun, Jungseul Ok, and Yung Yi. 2014. Optimal rate sampling in 802.11 systems. In *Proceedings of IEEE INFOCOM*.

[14] L. Deek, E. Garcia-Villegas, E. Belding, S.-J. Lee, and K. Almeroth. 2013. Joint rate and channel width adaptation in 802.11 MIMO wireless networks. In *Proceedings of IEEE SECON*.

[15] P. Frazier, D. Kempe, J. Kleinberg, and R. Kleinberg. 2014. Incentivizing exploration. In *Proceedings of the fifteenth ACM conference on Economics and computation*. 5–22.

[16] Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences* 55, 1 (1997), 119–139.

[17] A. Garivier and O. Cappé. 2011. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of COLT*.

[18] Thore Graepel, Joaquin Q Candela, Thomas Borchert, and Ralf Herbrich. 2010. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*.

[19] Adam Kalai and Santosh Vempala. 2005. Efficient algorithms for online decision problems. *J. Comput. System Sci.* 71, 3 (2005), 291–307.

[20] Satyen Kale. 2014. Multiarmed bandits with limited expert advice. In *Proceedings of COLT*.

[21] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. 2012. On Bayesian upper confidence bounds for bandit problems. In *Proceedings of AISTATS*.

[22] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. 2016. On the Complexity of Best Arm Identification in Multi-Armed Bandit Models. *The Journal of Machine Learning Research* 17 (2016), 1–42.

[23] Tomáš Kocák, Gergely Neu, Michal Valko, and Remi Munos. 2014. Efficient learning by implicit exploration in bandit problems with side observations. In *Proceedings of NIPS*.

[24] Ron kohavi. 2015. Online Controlled Experiments: Lessons from Running A/B/N Tests for 12 Years. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[25] T.L. Lai and H. Robbins. 1985. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* 6, 1 (1985), 4–2.

[26] Shie Mannor and Ohad Shamir. 2011. From bandits to experts: On the value of side-observations. In *Proceedings of NIPS*.

[27] Y. Mansour, A. Slivkins, and V. Syrgkanis. 2015. Bayesian incentive-compatible bandit exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*. 565–582.

[28] J. Garcia N. Duy. 2011. A practical approach to rate adaptation for multi-antenna systems. In *Proceedings of 19th IEEE International Conference on Network Protocols*. 331–340.

[29] Ioannis Pefkianakis, Yun Hu, Starsky HY Wong, Hao Yang, and Songwu Lu. MIMO rate adaptation in 802.11n wireless networks. In *proceedings of ACM Mobicom*.

[30] Herbert Robbins. 1952. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* 58, 5 (1952), 527–535.

[31] Yevgeny Seldin, Peter Bartlett, Koby Crammer, and Yasin Abbasi-Yadkori. 2014. Prediction with Limited Advice and Multiarmed Bandits with Paid Observations. In *Proceedings of ICML*.

[32] Yevgeny Seldin, Koby Crammer, and Peter Bartlett. 2013. Open Problem: Adversarial Multiarmed Bandits with Limited Advice. In *Proceedings of COLT*.

[33] Aleksandrs Slivkins, Filip Radlinski, and Sreenivas Gollapudi. 2013. Ranked bandits in metric spaces: learning diverse rankings over large document collections. *The Journal of Machine Learning Research* 14, 1 (2013), 399–436.

[34] Min Xu, Tao Qin, and Tie-Yan Liu. 2013. Estimation Bias in Multi-Armed Bandit Algorithms for Search Advertising. In *Proceedings of NIPS*.

APPENDIX A: REGRET LOWER BOUND

For the sake of completeness, we provide a more detailed proof of (1), following the arguments in [22] to prove Theorem 21. Let $a \neq N$ be a suboptimal arm. Let us do the following change-of-measure: consider a new reward vector $\tilde{\boldsymbol{\mu}} = (\mu_1, \ldots, \mu_{a-1}, \tilde{\mu}, \mu_{a+1}, \ldots, \mu_N)$ in which only μ_a is replaced by $\tilde{\mu}$, where $\mu_N < \tilde{\mu} < 1$. Hence, the arm a becomes the unique optimal arm under $\tilde{\boldsymbol{\mu}}$. Denote by Z_T the set of observations up to time T (the arms selected and their observed rewards up to time T), and by $\mathcal{F}_T = \sigma(Z_T)$ the corresponding σ-algebra. Further denote by $L = \log(\frac{\mathbb{P}_{\boldsymbol{\mu}}[Z_T]}{\mathbb{P}_{\tilde{\boldsymbol{\mu}}}[Z_T]})$ the log-likelihood ratio of Z_T under $\mathbb{P}_{\boldsymbol{\mu}}$ and $\mathbb{P}_{\tilde{\boldsymbol{\mu}}}$. Then, by changing the measure from $\mathbb{P}_{\boldsymbol{\mu}}$ to $\mathbb{P}_{\tilde{\boldsymbol{\mu}}}$, we get (see Lemma 18 in [22]): for all $E \in \mathcal{F}_T$,

$$\mathbb{P}_{\tilde{\boldsymbol{\mu}}}[E] = \mathbb{E}_{\boldsymbol{\mu}}[1_E \exp(-L)].$$

Now applying Wald lemma and the data processing inequality, we obtain (see Lemma 19 in [22]):

$$\mathbb{E}_{\boldsymbol{\mu}}[L] = \mathbb{E}_{\boldsymbol{\mu}}[O_a(T)]KL(\mu_a \,\|\, \tilde{\mu}) \geq KL(\mathbb{P}_{\boldsymbol{\mu}}[E] \,\|\, \mathbb{P}_{\tilde{\boldsymbol{\mu}}}[E]).$$

Now select $E = \{A_N(T) \leq T - \sqrt{T}\}$. Markov inequality yields:

$$\mathbb{P}_{\boldsymbol{\mu}}[E] = \mathbb{P}_{\boldsymbol{\mu}}[T - A_N(T) \geq \sqrt{T}] \leq \frac{\sum_{a \neq N} \mathbb{E}_{\boldsymbol{\mu}}[A_a(T)]}{\sqrt{T}}$$

$$\mathbb{P}_{\tilde{\boldsymbol{\mu}}}[E^c] \leq \frac{\mathbb{E}_{\tilde{\boldsymbol{\mu}}}[A_N(T)]}{T - \sqrt{T}} \leq \frac{\sum_{j \neq a} \mathbb{E}_{\tilde{\boldsymbol{\mu}}}[A_j(T)]}{T - \sqrt{T}}$$

Since the algorithm is uniformly good, we must have for all $\alpha > 0$, $\sum_{a \neq N} \mathbb{E}_{\boldsymbol{\mu}}[A_a(T)] = o(T^\alpha)$ and $\sum_{j \neq a} \mathbb{E}_{\tilde{\boldsymbol{\mu}}}[A_j(T)] = o(T^\alpha)$. Hence $\mathbb{P}_{\boldsymbol{\mu}}[E] \to 0$ and $\mathbb{P}_{\tilde{\boldsymbol{\mu}}}[E] \to 1$ as $T \to \infty$. We conclude that:

$$\begin{aligned}\frac{KL(\mathbb{P}_{\boldsymbol{\mu}}[E] \,\|\, \mathbb{P}_{\tilde{\boldsymbol{\mu}}}[E])}{\log(T)} &\overset{T \to \infty}{\sim} \frac{1}{\log(T)} \log(\frac{1}{\mathbb{P}_{\tilde{\boldsymbol{\mu}}}[E^c]}) \\ &\geq \frac{1}{\log(T)} \log(\frac{T - \sqrt{T}}{\sum_{j \neq a} \mathbb{E}_{\tilde{\boldsymbol{\mu}}}[A_j(T)]}).\end{aligned}$$

The r.h.s. of the latter inequality is $1 + o(1)$ since again the algorithm is uniformly good. We have proved that:

$$\liminf_{T \to \infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[O_a(T)]}{\log(T)} \geq \frac{1}{KL(\mu_a \,\|\, \tilde{\mu})}.$$

Inequality (1) is obtained by letting $\tilde{\mu}$ tend to $\mu^\star$.